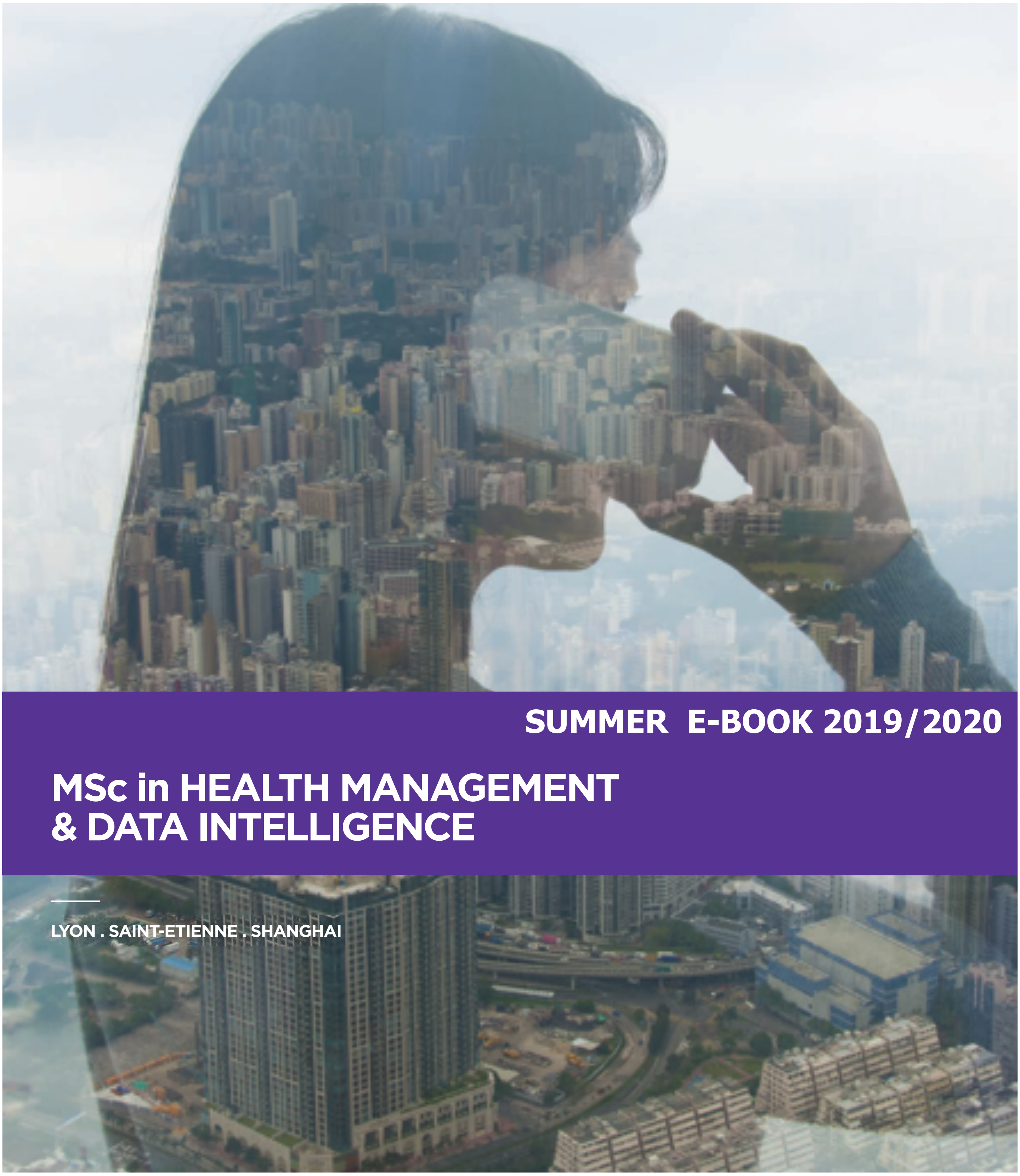
em lyon business school

MINES Saint-Étienne

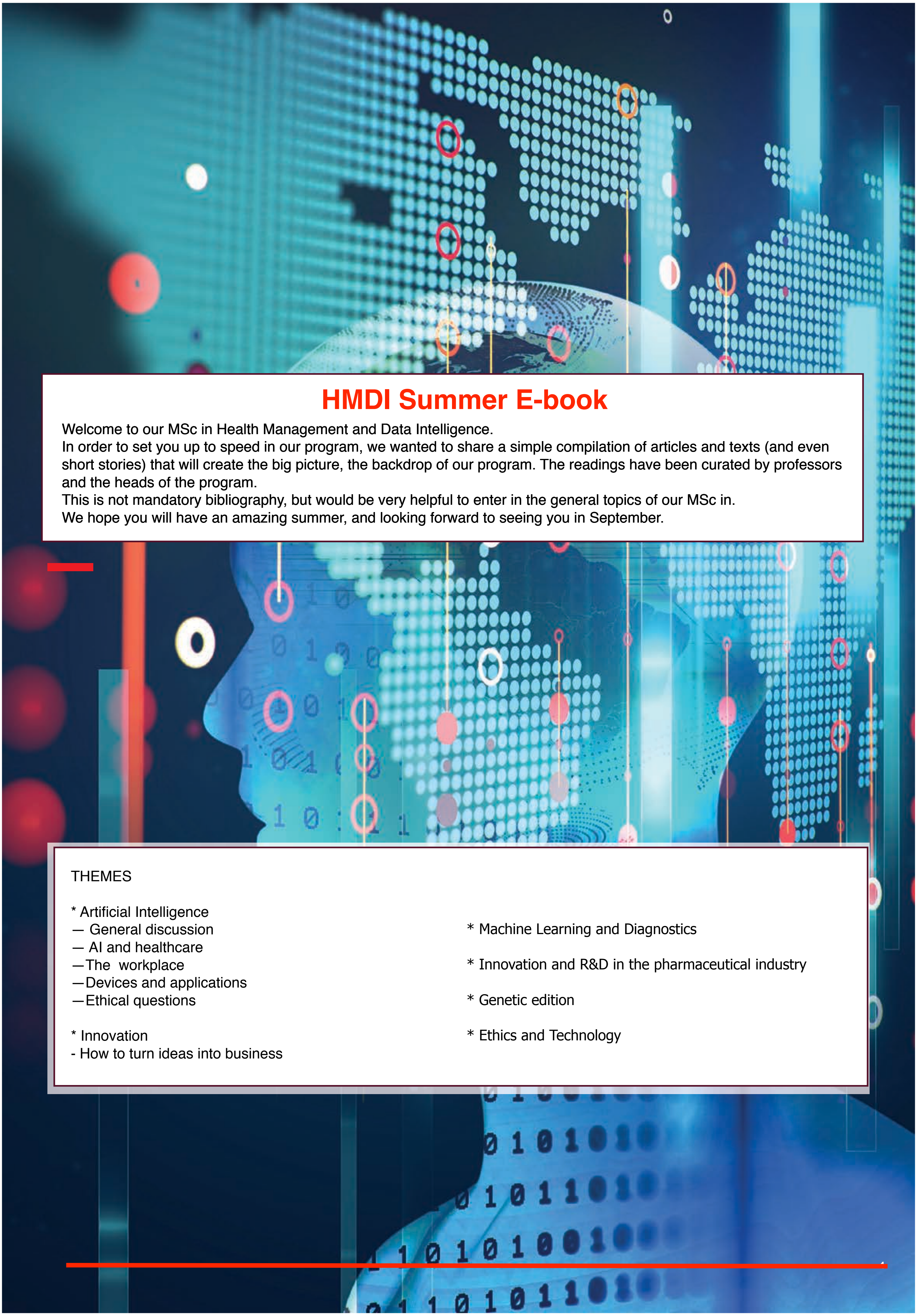
Une école de l'IMT



SUMMER E-BOOK 2019/2020

MSc in HEALTH MANAGEMENT & DATA INTELLIGENCE

LYON . SAINT-ETIENNE . SHANGHAI



HMDI Summer E-book

Welcome to our MSc in Health Management and Data Intelligence.
In order to set you up to speed in our program, we wanted to share a simple compilation of articles and texts (and even short stories) that will create the big picture, the backdrop of our program. The readings have been curated by professors and the heads of the program.
This is not mandatory bibliography, but would be very helpful to enter in the general topics of our MSc in.
We hope you will have an amazing summer, and looking forward to seeing you in September.

THEMES

* Artificial Intelligence
— General discussion
— AI and healthcare
—The workplace
—Devices and applications
—Ethical questions

* Innovation
- How to turn ideas into business

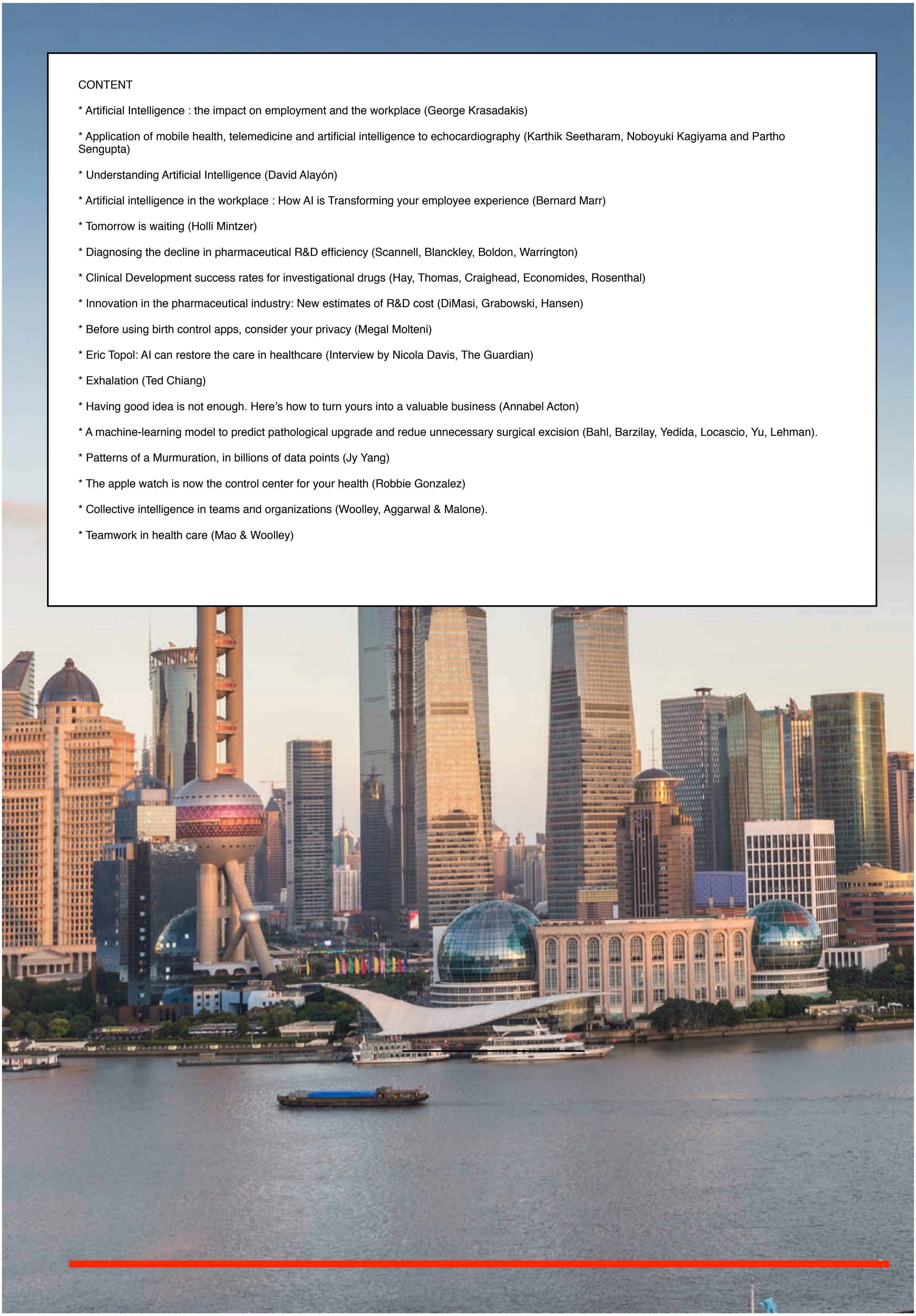
* Machine Learning and Diagnostics

* Innovation and R&D in the pharmaceutical industry

* Genetic edition

* Ethics and Technology

CONTENT



https://medium.com/innovation-machine/artificial-intelligence-3c6d80072416

Artificial Intelligence: the impact on employment and the workforce

How is AI replacing jobs? Which roles and industries will be most impacted? How can societies get prepared?

George Krasadakis
Jan 18, 2018 · 5 min read

image: pixabay

Although **Artificial Intelligence** dramatically improves our world in many ways, there are notable concerns regarding the forthcoming **impact of A.I. on employment and the workforce**.

There are predictions talking about millions of unemployed people in the next decades — primarily due to the impact of Intelligent Automation and A.I. systems.

In any case, the entire socioeconomic system is entering a phase of accelerating transformation: *markets, businesses, education, government*, *social welfare,* and *employment models* will be severely impacted.

Tasks, Roles, and Jobs at risk

Tasks that are **monotonous**, can be easily automated; this can gradually make certain roles obsolete. For instance, tasks and activities related to *customer care/call center operation*, *document classification*, *discovery* and *retrieval, content moderation* are more and more based on technology and automation and less on human work. The same is true for roles related to *operation* and *support* of *production lines* and *factories*: humans are being replaced by smart robots that can safely navigate the space, find and move objects (such as products, parts or tools) or perform complex assembling operations.

A.I. proves to be very effective in handling even more complex activities — those requiring processing of *multiple signals*, *data streams* and *accumulated knowledge in real time*. A characteristic case is the **autonomous vehicles**that can capture and 'understand' the environment and its dynamics; they can *'see', decide* and *act* in real-time, towards well-defined optimization objectives.

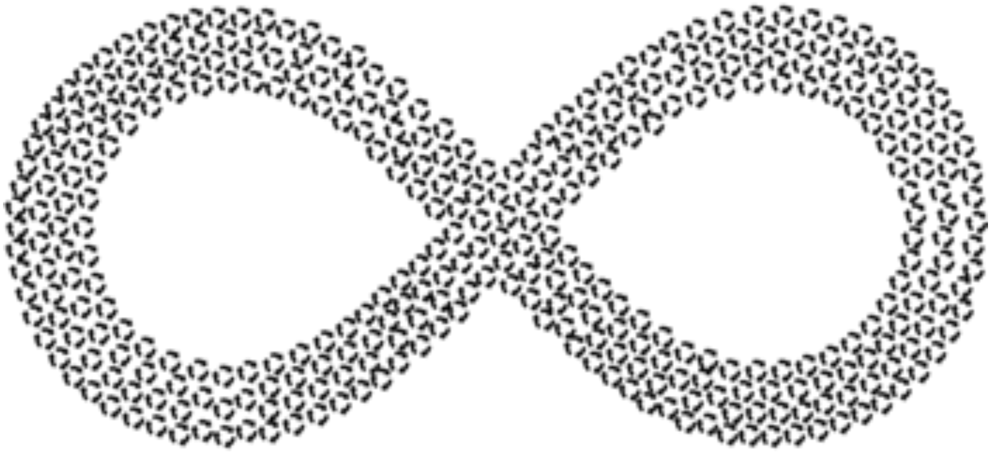
Sectors that will be impacted

Transportation is already in a transformation mode — fully autonomous cars will be soon a reality — and they will be safer, more efficient and more effective. **Professional drivers** (taxi, trucks and more) will see the demand for their skill set dropping rapidly.

Electronic commerce will also undergo a significant transformation: fulfillment centers will be fully automated, with robots navigating the space to collect products and execute customer orders; to be then sent or even delivered to customers, also automatically, with autonomous drones and/or cars. The importance of salespersons and networks of physical stores will shrink; we are close to scenarios where ***consumer A.I. agents* negotiate with Retailer *AI agents*** — based on different objectives, tactics, and strategies.

Even more traditional professions which are built on top of strong human relationships, such as **legal professions**, will be significantly impacted: typical support services in a legal context, have to do with *document handling -classification, discovery, summarization, comparison*, *knowledge extraction and management* — tasks where AI agents can do a great job already.

Financial services, **Insurance** and any other sector requiring a significant amount of data processing and content handling will also benefit from A.I. And of course *states, governance, and social mechanisms* — **A.I. can have a great role in eliminating bureaucracy**, improving the service to citizens, along with the design and performance of social programs.



How Artificial Intelligence can replace human work — an example

Imagine a typical customer care department: tens or even hundreds of specialized employees working with a shared mission: to handle customer requests, complaints, asks, etc. in the best possible way.

The workstream of '*handling a customer request in the best possible way*' can be broken down in separated jobs which are repeated over time and across different types of requests, for instance: *customer identification*, *customer history retrieval*, *request understanding* and *classification*, *problem identification* and *mapping to a solution space*, *forwarding* or *escalating to another team*, *customer document retrieval* and finally the *decisioning* based on the suitable *corporate policy.*

All the above can be covered with increased effectiveness from A.I. algorithms — they prove to be *faster*, more *accurate, reliable* and *cheaper*than the corresponding team of humans. **A properly trained A.I. system can understand customer requests in natural language**, identify the mentioned or implied entities (for instance, *which product* or *service the request refers to*); it can estimate *customer's intent* early enough (for example, to *activate a service* or *ask for help*); it can instantly process large volumes of data and apply the corporate policy in order to identify the *best action/ decision for the particular case*; the decision can then be communicated to the customer in natural language.

The system also knows early enough if it can handle the request *with confidence* or not; in the latter case, it knows where to *redirect the request as an exception, for a human team to handle it.* And all these, in milliseconds, **as part of a chat or voice session between the customer and companies' agent**.

This technological solution requires just a small percentage of the human team that a traditional customer care department has. And while this hybrid system is in operation, ***the A.I. component learns from the exceptions it forwards to the human team to handle***, leading to a *continuous improvement of its performance.* This **feedback loop** will eventually minimize the need for human intervention, making the AI system autonomous.

Getting ready

In the long run, we will witness certain roles and jobs becoming less and less relevant, and finally obsolete. But, in most of the cases, **Artificial Intelligence** will have a supportive role to humans — **empowering the human factor to perform better** in handling complex and critical situations which require judgment and creative thinking. In parallel, there would be numerous new roles and specialties with a focus on technology and science. For example, there will be needs for highly skilled professionals to *oversee* or *manage* or *coordinate* the *training of complex Artificial Intelligence systems*; to *ensure their integrity*, *security, objectivity* and *proper use*.

Under certain assumptions, and following the initial disruption due to technological unemployment, the AI revolution will lead to a new era of *prosperity, creativeness,* and *well-being*. Humans will no more need to perform routine, limited value, jobs. The workforce and the underlying employment models will move from *long-term, full-time employment agreements*, to *flexible*, *selective premium services offerings*.

There will be a stream of new business opportunities empowering a culture of entrepreneurship, creativeness and innovation.

The above positive scenario requires a *common*, *shared understanding of the technology, its opportunities, and its risks*. **Societies** need to *adapt to the new technology landscape*, become *more flexible* and also *inherit an attitude of lifelong learning, collaboration, innovation, and entrepreneurship*.

States need a new strategy with a focus on **education**; they need to rethink how *markets, companies and employment agreements* should work in the new era of intelligent automation; they need to redesign the *social mechanisms* to cover a range of new scenarios and situations.

*At an even higher level, **we need a solid framework to avoid the unbalanced concentration of technology power and control.***

WRITTEN BY

George Krasadakis

REVIEW

Application of mobile health, telemedicine and artificial intelligence to echocardiography

Karthik Seetharam MD, Nobuyuki Kagiyama MD PhD and **Partho P Sengupta MD DM**

West Virginia University Heart and Vascular Institute, Morgantown, West Virginia, USA

Correspondence should be addressed to P P Sengupta: partho.sengupta@wvumedicine.org

Abstract

The intersection of global broadband technology and miniaturized high-capability computing devices has led to a revolution in the delivery of healthcare and the birth of telemedicine and mobile health (mHealth). Rapid advances in handheld imaging devices with other mHealth devices such as smartphone apps and wearable devices are making great strides in the field of cardiovascular imaging like never before. Although these technologies offer a bright promise in cardiovascular imaging, it is far from straightforward. The massive data influx from telemedicine and mHealth including cardiovascular imaging supersedes the existing capabilities of current healthcare system and statistical software. Artificial intelligence with machine learning is the one and only way to navigate through this complex maze of the data influx through various approaches. Deep learning techniques are further expanding their role by image recognition and automated measurements. Artificial intelligence provides limitless opportunity to rigorously analyze data. As we move forward, the futures of mHealth, telemedicine and artificial intelligence are increasingly becoming intertwined to give rise to precision medicine.

Key Words

- mobile health
- telemedicine
- artificial intelligence

Introduction

Technological advancement has developed portable computer devices and miniaturized cardiac imaging devices. These devices with the simultaneous development of broadband technologies has led to a new frontier in communication by expanding the capabilities of information sharing among users worldwide. The effects of this digital landscape have permeated through multiple facets of daily life. Telemedicine and mobile health (mHealth), which is defined as use of mobile and wireless technologies to improve health care (1, 2), are becoming important in this digital landscape with cardiovascular medicine and the field of echocardiography being no exception. The Department of Health and Human Services estimates that more than 60% of all health care institutions in the United States currently use some form of telemedicine (3). Handheld imaging platforms and tele-interpreting has brought these trends into the field of echocardiography (4, 5).

Although big data generated by the telemedicine and cardiac imaging present great opportunity for cardiovascular research, this influx of data requires so much effort to integrate and interpret them that human cardiologists cannot digest all of it (6). Artificial intelligence (AI), including machine learning techniques, is increasingly recognized as a potential solution for facilitating a seamless transition between cardiologists

https://erp.bioscientifica.com
https://doi.org/10.1530/ERP-18-0081

and big data. AI can integrate the multifactorial information from many aspects of healthcare, including echocardiographic data, and can help cardiologists make better clinical decisions even in resource-limited areas where experts are not readily accessible. mHealth, telemedicine and AI offer bright promises intertwined with complex challenges in the field of cardiology and imaging. In this review, we will discuss the role of mHealth, telemedicine and AI in echocardiography.

Mobile health

Handheld imaging devices

As mobile computers and handheld imaging platforms become easily accessible and readily available, they present new paths of opportunity for the delivery and optimization of cardiovascular healthcare. Since the dawn of medicine, physical examination has been central to point-of-care diagnosis in cardiovascular medicine. The rapid rise of imaging devices which help physicians visualize the heart's activities in real time have complemented physical examination and augmented clinical decision making. Despite the wide array of imaging capabilities at our disposal, correct diagnosis are not always made in time resulting in unfavorable outcomes (7). This has perpetuated a need, no a necessity for rapid and efficient diagnosis at bedside. The development of miniaturized handheld imaging platforms such as the pocket-size ultrasound can circumvent the obstacles of delayed diagnosis and reduce medical errors (7). There are several types of handheld ultrasounds with various capabilities; a laptop-based equipment has almost every 2D echocardiographic application, while a pocket-size ultrasound does not usually have full-scale color-flow and spectral Doppler capabilities (Fig. 1). The point-of-care ultrasound (POCUS) can fundamentally alter bedside medicine and be indispensable with physical examination. There are numerous studies which have clearly shown that POCUS is as efficient and effective compared to conventional machines. (8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18) Many researches have shown their capability in the assessment of valvular heart disease (19, 20), heart failure (21, 22, 23), coronary artery disease (22, 24) and so forth. (Table 1). Accuracy of POCUS has been reported well. For example, Abe *et al*. studied 130 patients with aortic stenosis and reported that pocket ultrasound was able to discriminate moderate-to-severe aortic stenosis with sensitivity 84% and specificity 90% even without quantitative Doppler information (19). Most recently, there are some smartphone-sized devices with image quality well enough for cardiac assessment using AI-based technologies. Some of these devices are supposedly financially cheap and help cardiologists in practice (e.g. Vscan, GE Healthcare and Butterfly IQ, Israel).

With powerful and affordable diagnostic imaging devices at the palm of our hands, POCUS can augment and add a significant impact on cardiovascular healthcare (25, 26, 27, 28, 29), especially in patients living in resource-limited areas.

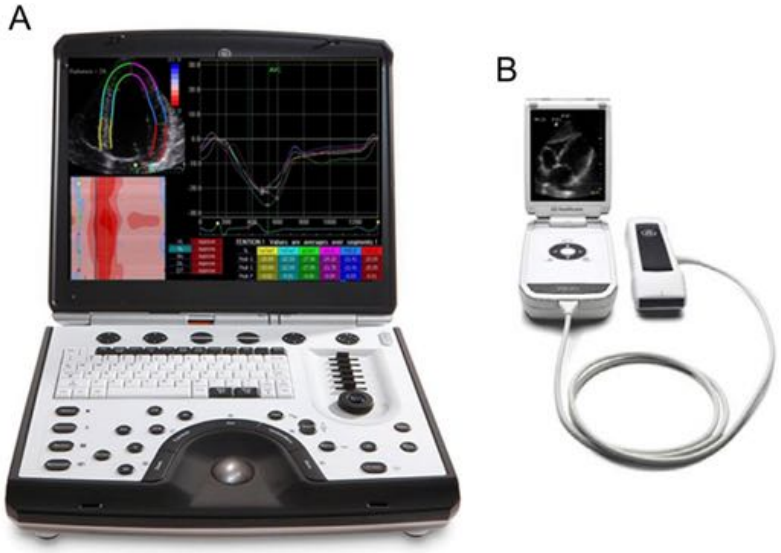


Figure 1
Type of handheld ultrasound machines. There are several types of handheld ultrasounds with various capabilities; a laptop-based equipment has almost every 2D echocardiographic application (panel A), while a pocket-size ultrasound does not usually have full-scale color-flow and spectral Doppler capabilities (panel B). Reproduced, with permission, from Chamsi-Pasha *et al.* (4).

https://doi.org/10.1530/ERP-18-0081

Table 1 Comparison of handheld ultrasound with reference standard.

Study	Year	Number of subjects	Reference standard to PUS	Study findings
Prinz *et al.* (8)	2011	349	Standard echocardiography	Statistically significant agreement between PUS and high-end echocardiography (1.6 ± 0.5 vs 1.7 ± 0.4, $P < 0.01$), regional wall motion ($\kappa = 0.73$, $P < 0.01$), LV measurements ($r = 0.99$, $P < 0.01$), regurgitation detection ($k = 0.9$, $P < 0.01$)
Galderisi *et al.* (9)	2010	304	Standard echocardiography	The K between PUS and reference was 0.67 in the pooled population (0.84 by experts and 0.58 by trainees)
Testuz *et al.* (10)	2013	104	Standard echocardiography	Statistically significant agreement between PUS and reference for left ventricular function and pericardial effusion (kappa: 0.89 and 0.81). The agreement for aortic, mitral, tricuspid and left ventricular size was moderate (Kappa: 0.55–0.66)
Andersen *et al.* (11)	2011	108	Standard echocardiography	Strong agreement between PUS and reference for abdominal aorta and pericardial effusion was ($r \geq 0.92$), right ventricular and valvular function ($r \geq 0.81$). The correlation for aortic stenosis was ($r = 0.62$)
Skjetne *et al.* (7)	2011	119	Standard echocardiography	The PUS accurately assessed and diagnosed only 16% of patients in the cardiac unit. In 55% of patients, the reference had higher diagnostic value
Lafitte *et al.* (12)	2011	100	Standard echocardiography	The concordance between PUS and reference for LV function and morphology ($\kappa = 0.91$ and 0.96), left ventricular hypertrophy ($k = 0.74$), mitral regurgitation grades were 0.90, 0.95, and 1.00
Michalski *et al.* (13)	2012	220	Standard echocardiography	There was excellent correlation between PUS and reference ($r =$ 0.64–0.96, $P < 0.001$)
Biais *et al.* (14)	2012	151	Standard echocardiography	The PUS had good accordance with the reference in global left ventricular systolic dysfunction ($\kappa = 0.87$), pericardial effusion ($\kappa = 0.75$)
Prinz *et al.* (15)	2012	320	Standard echocardiography	In comparison to reference, substantial agreement in functional assessment ($\kappa > 0.61$, $P < 0.01$) and wall motion scoring ($\kappa = 0.67$, $P < 0.01$) could be observed over time. The correlation in left ventricular measurements ($r > 0.98$, $P < 0.01$) was very good
Fukuda *et al.* (16)	2009	125	Standard echocardiography	Left ventricular dimensions, fractional shortening, interventricular septum thickness, posterior wall thickness, left atrial dimension, and aortic diameter show excellent correlation ($r = 0.87$–0.98, all $P < 0.001$)
Mjolstad *et al.* (17)	2012	196	Standard echocardiography	Excellent agreement was observed between PUS and reference
Panoulas *et al.* (18)	2013	122	Standard echocardiography	After addition of PUS, there was improved diagnostic accuracy ($Z = -7.761$, $P < 0.001$)
Carlino *et al.* (25)	2018	102	Standard echocardiography	After addition of PUS, it helped improve diagnostic accuracy (all $P < 0.01$ vs single modalities)
Bhavnani *et al.* (39)	2018	254	Standard echocardiography	PUS had a shorter time to referral for intervention (83 ± 79 days vs 180 ± 101 days; $P < 0.001$). The PUS group had lower risk of hospitalization and death (15% vs 28%, adjusted hazard ratio: 0.41; $P = 0.013$)
Filipiak-Strzecka *et al.* (26)	2017	100	Standard echocardiography	There was statistically significant correlation between PUS and reference for intimal medial thickness ($r = 0.58$; 95% CI: 0.48–0.66; $P < 0.0001$)
Phillips *et al.* (22)	2017	102	Standard echocardiography	In relation to reference, PUS had values ranging from 85% for left atrial enlargement to 100% for cardiomegaly, but limited specificity of cardiomegaly at just 51%
Esposito *et al.* (27)	2017	508	Standard echocardiography	In a subgroup, PUS was compared with the standard for abdominal aorta size (rho = 0.966, $P < 0.0001$)
Cavallari *et al.* (28)	2015	100	Standard echocardiography	The PUS had a shorter time for examination (6.1 ± 1.2 min vs 13.1 ± 2.6 min, $P < 0.0001$) and saved waiting time ($P < 0.001$). No difference in conclusiveness between both groups (86 vs 96%; $P = 0.08$)
Khan *et al.* (29)	2014	240	Standard echocardiography	No discernable differences between both groups ($P = 7.22 \times 10(-7)$).

PUS, pocket-size ultrasound.

Other mHealth devices

There are more than 160,000 health-related smartphone apps, such as apps for monitoring weight or diet control, available and these apps have been downloaded close to 660 million times (30, 31, 32). In addition, there have been many smartphone-connected devices and wearable devices available, which enable remote monitoring of health conditions including heart rhythm and blood pressure (30, 33, 34, 35). One of the hottest topics in the field is detection of atrial fibrillation using smartwatch. Tison *et al*. compared smartwatch data with standard ECG in 9750 patients for detecting atrial fibrillation (36). A deep learning-based algorithm showed excellent prediction of atrial fibrillation (C-statistic 0.97) with a sensitivity of 98% and specificity of 90.2%. Those apps and devices, along with other devices such as point-of-care measurements of B-type natriuretic peptide (BNP) (37) and implantable pulmonary artery pressure sensors (38), have potential to provide better identification of underlying diseases and improve their outcomes in communities (Fig. 2).

Integration of handheld imaging platforms with other mHealth devices

Bhavnani *et al*. conducted the first randomized trial of integration of POCUS with other mHealth devices in modern structural heart disease clinics in rural parts of India, under the ASE Foundation-Valvular Assessment Leading to Unexplored Echocardiographic Stratagems (ASEF-VALUES) (39). There were a total of 253 patients with structural heart disease randomized into two groups of mHealth clinic and standard healthcare. The main focus was the impact of mHealth with pocket-size echocardiography on medical decision making in patients with valvular heart disease in remote areas. The primary objective was time to referral for management for surgical or percutaneous intervention. The initial mHealth clinic was associated with shorter referral time for intervention (83±79 days vs 180±101 days; $P<0.001$) and increased probability for intervention compared with standard healthcare (adjusted hazard ratio, 1.54; 95% CI, 0.96–2.47, $P<0.07$). The patients assigned to mHealth clinic had lower hospitalization and death (15% vs 28%, adjusted hazard ratio, 0.41; 95% CI, 0.21–0.83; $P<0.013$). In this study, the authors successfully integrated POCUS with other mHealth devices and showed that this integration can be associated with earlier medical interventions and favorable clinical outcome.

Telemedicine with POCUS and mHealth devices

Feasibility of POCUS in telemedicine

Thus, pocket-size ultrasound and other mHealth devices have allowed point-of-care screening of cardiovascular

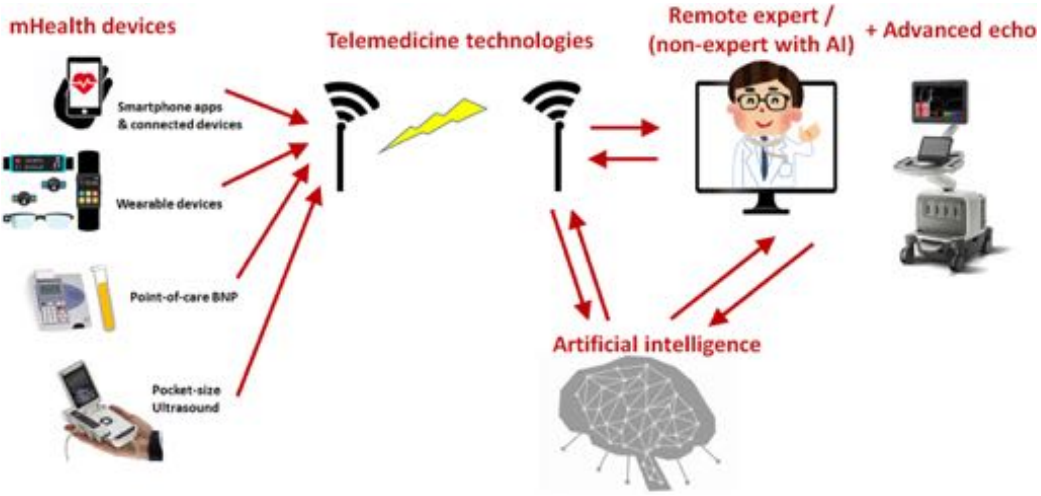


Figure 2
Interrogation of mHealth devices and use of artificial intelligence. Technological advancement has created a number of mobile health devices, which are available even in resource-limited areas. Involving remote experts using telemedicine helps appropriate diagnosis and management. Artificial intelligence can efficiently address the lack of experts and the influx of complex data generated by mHealth and telemedicine as well as advanced imaging modalities.

© 2019 The authors
Published by Bioscientifica Ltd

diseases to resource-limited communities. Furthermore, application of telemedicine technologies enables mHealth strategies even in remote areas with limited access to experts. Singh *et al.* (40) under American Society Echocardiography: Remote Echocardiography with Web-Based Assessments for Referral at a Distance (ASE-REWARD) performed a prospective study in order to test the feasibility of performing POCUS with long-distance Web-based assessment of recorded images. Using pocket-size ultrasound, 1023 studies were scanned in a rural region of India, and the images were sent to physicians in remote locations for review through Web-based platforms. The images were successfully uploaded and reviewed at a median time of 11:44 h. There was an excellent agreement in assessing valvular lesions, whereas the on-site readings were frequently modified by expert reviewers for left ventricular function and hypertrophy. The study successfully showed the feasibility of remote echocardiographic assessment and the incremental value of using Web-based remote assessment for facilitating appropriate mass triage of patients with suspected cardiac illnesses.

Choi *et al.* (41) tested the feasibility of remote interpretation of echocardiographic images on a smartphone. Eighty-nine patients underwent POCUS and the images were sent to remote experts who read them using smartphone apps. The authors found that 38% of on-site, non-expert diagnosis was revised by remote experts, whose interobserver agreement was excellent. The study suggested that remote interpretation is feasible and should be considered when POCUS is done by non-experts.

Limitations of POCUS

Although the benefits of POCUS are promising, there have been several challenges for its clinical application. One of the biggest concerns is the standardization of the quality of scan and interpretation (4). Because of its availability, POCUS can be used in more various situations and by wider range of observers than standard echocardiography. On the other hand, POCUS has limited ability in terms of image quality and applications such as pulse-wave Doppler. Scanning patients using POCUS and interpreting images by novice observers can result in overlooking important findings and wrong diagnosis (4). It is absolutely pivotal for all healthcare providers who use POCUS to be properly trained and understand the limitations of POUCS. Most professional societies require a minimum of 30 scans for basic training, but this number is not enough for accurate interpretation. Universal standardization of training is necessary for wide use of POCUS in clinical practice. Some of these limitations can be addressed by AI. For example, AI-based automated LVEF analysis program that works on PUS images (LVivo by DiA Imaging Analysis Ltd., Israel and Vscan by GE Healthcare) has been developed (42). This kind of programs will reduce the interobserver variability and help standardization of procedures. The lack of incentive for POCUS in US healthcare model is another problem, because more referrals and reimbursement for conventional echocardiography is more beneficial. A reward system is important to stimulate increased utilization.

Remote training and robot-assisted echocardiography

Even telemedicine enables remote assessment of acquired echocardiographic images, and appropriate acquisition itself requires expertise, which may limit its wider use in rural areas. Telemedicine also has a potential to address this issue through Web-based training. Bansal *et al.* (43) tested the feasibility of Web-based, real-time, hands-on, personalized training program of POCUS. Seventeen physicians in India were provided 6-h training of POCUS; nine had an on-site training and eight had an online training using transcontinental tele-echocardiography system. Although good-quality images were obtained more frequently by physicians trained on-site (90 vs 84%, $P=0.03$), there were no difference between the two groups in agreement of the trained physicians' diagnoses with expert interpretations. Such training, combined with Web-based integration of remote, expert interpretation of stored images, allows the delivery of echocardiographic expertise to remote communities, which could be of great help in optimizing cardiovascular health outcomes.

Robot-assisted remote echocardiography may be another solution. Boman *et al.* conducted randomized control trial and showed that real-time robot-assisted remote echocardiography followed by cardiologic consultation at a distance significantly reduced the total diagnostic process time (44).

Advanced echocardiography in contrast to POCUS

As discussed earlier, POCUS is promising and has a huge impact in expanding and complimenting physical examination. However, those handheld imaging platforms have limited function, and comprehensive and advanced echocardiography is definitely warranted in addition to POCUS. Recent advancement in echocardiography includes automation of measurements and analysis. Although speckle tracking echocardiography (STE) and 3D imaging have been the most promising methods in echocardiography for the past two decades, clinical use of these techniques is not sufficient due to time-consuming process. Automation of these techniques using AI algorithms are evolving and they help physicians and sonographers by reducing analysis time and increasing reproducibility (45, 46, 47). This is also a field where AI has a core role for evolution and for widespread use of the techniques. For example, although 3D echocardiography has been extensively reported to be superior to 2D echocardiography, the full adoption of this technique is not embraced due to time constraints and complicated measurement steps which disrupt clinical work flow. HeartModel (Philips Healthcare, Andover, MA, USA) is an AI-based fully automated quantification program for left heart chambers. The program dramatically reduces time for analysis (144±32 to 26±2min, $P<0.0001$) with even better interobserver measurement agreement compared with conventional 3D quantification (48, 49).

Artificial intelligence

Despite the transformative potential of mHealth and telemedicine, the data generated by these technologies are multifactorial and complex. In addition, imaging modalities including echocardiography also generate a huge amount of data; a single echo examination generates 2 gigabyte of information and annually there can be 15 petabytes of information produced (50). This large size of data would overwhelm current statistical software. AI is a field of computer science which mimics human thought process and learning capacity. AI could algorithmically quickly analyze and offer various interpretations of these elaborate datasets with lesser difficulty. With the rapid evolution of data, AI will be the primary and most efficient tool which brings the necessary revolution for integration of information into cardiovascular healthcare. In resource-limited situations where mHealth and telemedicine have an important role, well-trained AI may complement the lack of experts. AI techniques, such as machine learning and deep learning, unravel hidden patterns within heterogeneous datasets using a number of various algorithms (50). With the advent of AI, the paradigm is being fundamentally altered from current statistical tools to cardiovascular precision medicine (Fig. 3) (51).

Type of machine learning

Machine learning is one subfield of AI, which aims at automatic discovery of regularities in data through the use of computer algorithms and generalizing those into new but similar data (Fig. 4). In general, machine learning tends to make less pre-assumption than traditional statistical method but requires greater data. Machine leaning techniques can be broadly split into supervised learning, unsupervised learning and reinforcement learning.

In supervised learning, the database is labeled with outcome and classes. Supervised learning frequently groups an observation into one or more categories or outcomes (51). It is ultimately designed to show how the independent variable is linked to the dependent variable. A statistical model is generated from the data to create a model to predict an event or complication. Supervised learning proves to be very valuable in classifying phenotypically different patients (51). In contrast to supervised learning, unsupervised learning uses database with no prior label present (34). The purpose of unsupervised learning is to discover the

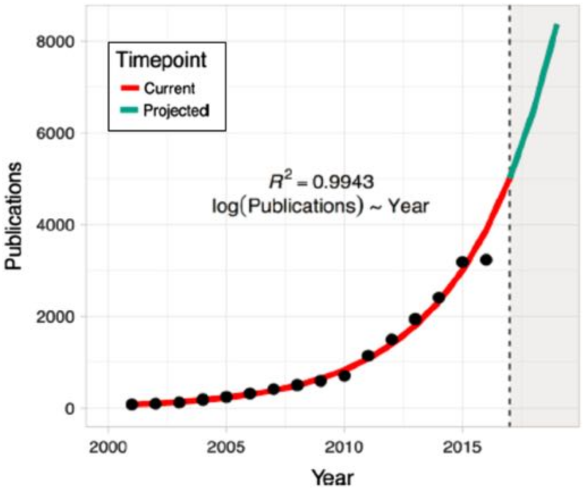


Figure 3
Growth of publications in machine learning. The *x*- and *y*-axis shows the year and the number of publications in PubMed with 'Cardiology' and 'Machine Learning'. The number of publications is rapidly growing, representing huge interest in the field. Reproduced, with permission, from Shameer *et al.* (50).

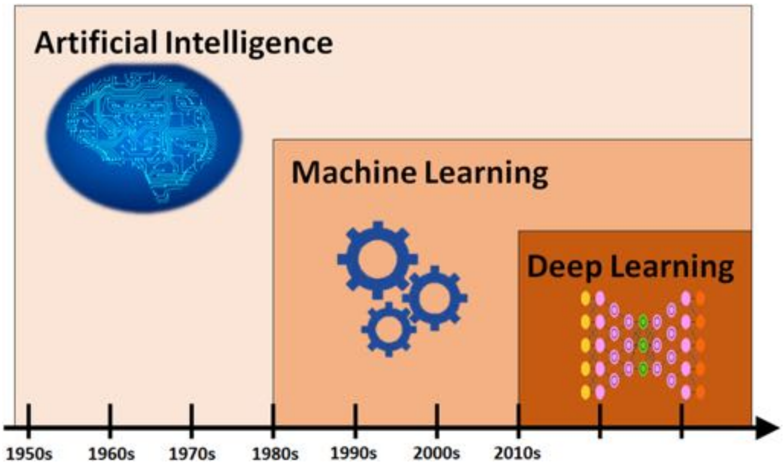


Figure 4
Association of artificial intelligence, machine learning and deep learning. Artificial intelligence (AI), though there are various definitions by itself, represents any techniques which enables computers mimic human behavior when it's used in medical field. Machine learning is a subfield of AI, which aims at automatic discovery of regularities in data through the use of computer algorithms and generalizing those into new but similar data. Deep learning is a subset of machine learning, which makes the computation of multi-layer neural networks feasible.

relationship between variables. This machine-learning approach consists of clustering methods (hierarchical or K means), self-organizing maps, topological data analysis, information maximization analysis and finally deep learning. Reinforcement learning is derived from behavioral psychology. The algorithm learn and modify behavior through trial and error so as to maximize some notion of cumulative reward (51). Reinforcement learning is mostly used in game programs, such as AlphaGo software by DeepMind (52). Reinforcement learning has had limited role in healthcare so far.

Deep learning

Deep learning is a thriving discipline which learns complex hierarchical representations from data which has multiple levels of abstractions (53). It mimics the complexities of the human brain. Presently, deep learning is playing a prominent role in Facebook's image recognition, speech recognition in Apple's Siri and Amazon's Alexa, Google brain and robots (53). Deep learning architecture utilizes an artificial neural network which contains multiple layers of neurons which facilitates reasoning and interpretation. Recent advances in graphic processing unit and cloud-based platforms have spurred the growth of deep learning.

Deep learning requires a large elaborate data sets which requires information sharing between institutions and organizations. If the dataset is not large enough, overfitting is an issue (50). It has multiple layers and performs analysis in a nonlinear manner. This also increases the training time. Assembling the neural network is also lengthy process. Powerful computing processing unit and cloud-based systems are often necessary for deep learning.

Comparison of machine learning with traditional statistics

Logistic regression is one of the most commonly used methods in statistics to predict outcomes (50). However, this technique requires a strong number of assumptions to help generate *P* values. Nevertheless, machine learning can be used in any data set without making any assumptions of the underlying data. Especially for classification, machine learning can be more accurate and predictive. Another difference is the capability to deal with complex data. Electronic health record contains a massive amount of information from billing, international disease classification, lab values, imaging and medications. This can exceed the capacity of logistic regression model. Other statistical approaches such as univariate significance screening or stepwise regression, but the results do not translate well for patient care. Complex interaction between variable may be difficult to analyze with traditional approaches. Churpek *et al.* showed how flexible algorithms in machine learning was superior to conventional logistic regression for clinical deterioration in wards in a large multihospital study (54). Popular risk scores such as Framingham risk score, CHADS2 and CHA2DS2-VASc score, and so forth were derived from large trials and registries (55). However, Cook *et al.* found that there was an overestimation of these pooled cohort equations, believed that big data analytics could resolve the issue (56).

The role of machine learning in cardiology and echocardiography

Many machine-learning and deep learning techniques can be applied to researches in echocardiography

(Table 2) (38, 48, 57, 58, 59, 60). For example, we showed that supervised learning algorithm, including artificial neural networks, support vector machines and random forests, could differentiate athlete heart and hypertrophic cardiomyopathy using STE data more accurately than traditional measures (58). We have also used supervised learning approach with 15 STE variables and the four conventional echocardiographic variables and showed that machine learning was superior to other echo parameters for differentiating constrictive pericarditis from restrictive cardiomyopathy (57).

Deep learning is being utilized for a number of image-based classifications. This machine learning approach is particularly useful for computer vision. Deep learning can track pattern recognition in cardiovascular imaging and heterogeneous syndromes. Left ventricular ejection fraction is usually assessed by manually tracing boundaries (53), but unfortunately this method can be subjective lack precision or reproducibility (61). Deep learning can greatly improve the accuracy of 2D STE and other imaging modalities (48, 59). This can be extended into other cardiovascular imaging modalities such as 3D STE and cardiac magnetic resonance imaging. It performs well even with noisy data such as strain imaging. Deep learning can be implemented into a number of cardiovascular diseases including heart failure, takotsubo cardiomyopathy, hypertension, atrial fibrillation, Brugada syndrome and so forth. It can categorize these conditions with new genotypes or phenotypes and innovative echocardiographic parameters can craft pathways for new therapies.

Recently, Zhang *et al*. developed a deep learning algorithm that enables fully automated interpretation of echocardiography (62). Using a huge (over 14,000) sample of echocardiographic studies, the algorithm achieved a 96% accuracy in image recognition for distinguishing

Table 2 Examples of application of machine learning techniques to echocardiographic research.

Study	**Algorithm model**	**Brief algorithm description**	**Data source**	**Brief study description**
Narula *et al*. (58)	(a) Support vector machine	Finds a gap in multidimensional data and classifies data based on gap	Echocardiographic data	To differentiate between athlete heart and hypertrophic cardiomyopathy
	(b) Random forest	Decision tree-based method derived from creating a number of decision trees		
	(c) Artificial neural network	Learns in a manner similar to a biological network		
Sengupta *et al*. (57)	Associative memory classifier-supervised learning	Used for making predictions based on a set of matrices. It is developed by observing co-occurrences of predictors from outcomes	Speckle tracking echocardiographic data	To differentiate between constrictive pericarditis and restrictive cardiomyopathy
Berikol *et al*. (48)	Artificial neural network		Echocardiographic data	Echocardiographic data and clinical factors used to stratify cardiovascular risk
Lancaster *et al*. (59)	Hierarchical clustering	It classifies similar objects into the same groups called clusters by building a hierarchy based on the distance between patients	Echocardiographic data	To investigate the natural clustering of echocardiographic variables to measure left ventricular dysfunction and isolate high-risk phenotyping patterns
Abdolmanafi *et al*. (38)	Deep learning	It creates layered neural networks to extract and transform features and learn in supervised and/ or unsupervised manners	Coronary optical coherence tomography images	To automatically classify coronary artery layers in coronary optical coherence tomography images in Kawasaki disease
Bai *et al*. (60)			Cardiac magnetic resonance	Deep learning was used to analyze short and long axis cardiac magnetic resonance imaging and compare with human performance

between broad echocardiographic view classes (e.g. parasternal long axis from short axis), and 72–90% accuracy of image segmentation. Furthermore, the authors showed that the algorithm for automated quantification of cardiac structure and function was comparable or even superior to manual measurements across 11 internal consistency metrics (e.g. the correlation of left atrial and ventricular volume) and that the convolutional neural networks was successfully trained to detect hypertrophic cardiomyopathy, cardiac amyloidosis and pulmonary artery hypertension with high accuracy. Although the accuracy has not reached that of experts, application of deep learning to echocardiography interpretation is promising.

Future of artificial intelligence in cardiology

The rapid expansion of data is creating a moment of reckoning of sorts for cardiologists. With the development of POCUS integrated with mHealth devices and telemedicine, a concept which once seemed like a fantasy is now becoming a reality. AI, mainly machine learning techniques including deep learning, is the most effective means presently available to handle the sheer complexity data incoming from these evolutions. Compared to subspecialties of medicine, cardiologists have vast expanses of data at their disposal. As the complexities of data continue to grow, it is becoming imminent for an AI to be integrated into clinical practice. AI will become part and parcel of daily medicine, which is evidenced in the fields of radiology and pathology (63). It should be embraced not feared as it will enhance the clinical decision-making process. In the future, it may be necessary for all cardiologists to be physicians and data scientists simultaneously.

Conclusion

The burgeoning of mHealth, telemedicine and AI are the expanding the boundaries of echocardiography and cardiology. mHealth and telemedicine are establishing new bridges between patient and physician and helping underserved population to overcome previous barriers with their health care providers. AI is the truss support for these bridges. AI is the primary means and will be interconnected with the growth of these novel healthcare technologies for years to come. As mHealth and telemedicine create big data even in resource-limited areas where the number of experts is not sufficient and big data from these technologies are getting more and more complex, AI will assist cardiologists to provide more focused and personalized decision for the patients.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this review.

Funding
This review did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

References

1 Burke LE, Ma J, Azar KM, Bennett GG, Peterson ED, Zheng Y, Riley W, Stephens J, Shah SH, Suffoletto B, *et al*. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. *Circulation* 2015 **132** 1157–1213. (https://doi.org/10.1161/CIR.0000000000000232)

2 Tuckson RV, Edmunds M & Hodgkins ML. Telehealth. *New England Journal of Medicine* 2017 **377** 1585–1592. (https://doi.org/10.1056/NEJMsr1503323)

3 Office of Health Policy, Office of the Assistant Secretary for Planning and Evaulation. Report to Congress: E-health and telemedicine. Washington, DC, USA: Department of Health and Human Services, 2016. (available at: https://aspe.hhs.gov/sites/default/files/pdf/206751/TelemedicineE-HealthReport.pdf)

4 Chamsi-Pasha MA, Sengupta PP & Zoghbi WA. Handheld echocardiography: current state and future perspectives. *Circulation* 2017 **136** 2178–2188. (https://doi.org/10.1161/CIRCULATIONAHA.117.026622)

5 Johri AM, Durbin J, Newbigging J, Tanzola R, Chow R, De S & Tam J. Cardiac point-of-care ultrasound: state-of-the-art in Medical School education. *Journal of the American Society of Echocardiography* 2018 **31** 749–760. (https://doi.org/10.1016/j.echo.2018.01.014)

6 Konstam MA, Hill JA, Kovacs RJ, Harrington RA, Arrighi JA, Khera A & Academic Cardiology Section Leadership Council of the American College of Cardiology. The academic medical system: reinvention to survive the revolution in health care. *Journal of the American College of Cardiology* 2017 **69** 1305–1312. (https://doi.org/10.1016/j.jacc.2016.12.024)

7 Skjetne K, Graven T, Haugen BO, Salvesen Ø, Kleinau JO & Dalen H. Diagnostic influence of cardiovascular screening by pocket-size ultrasound in a cardiac unit. *European Journal of Echocardiography* 2011 **12** 737–743. (https://doi.org/10.1093/ejechocard/jer111)

8 Prinz C & Voigt JU. Diagnostic accuracy of a hand-held ultrasound scanner in routine patients referred for echocardiography. *Journal of the American Society of Echocardiography* 2011 **24** 111–116. (https://doi.org/10.1016/j.echo.2010.10.017)

9 Galderisi M, Santoro A, Versiero M, Lomoriello VS, Esposito R, Raia R, Farina F, Schiattarella PL, Bonito M, Olibet M, *et al*. Improved cardiovascular diagnostic accuracy by pocket size imaging device in non-cardiologic outpatients: the NaUSiCa (Naples ultrasound Stethoscope in Cardiology) study. *Cardiovascular Ultrasound* 2010 **8** 51. (https://doi.org/10.1186/1476-7120-8-51)

10 Testuz A, Muller H, Keller PF, Meyer P, Stampfli T, Sekoranja L, Vuille C & Burri H. Diagnostic accuracy of pocket-size handheld echocardiographs used by cardiologists in the acute care setting. *European Heart Journal Cardiovascular Imaging* 2013 **14** 38–42. (https://doi.org/10.1093/ehjci/jes085)

11 Andersen GN, Haugen BO, Graven T, Salvesen O, Mjolstad OC & Dalen H. Feasibility and reliability of point-of-care pocket-sized echocardiography. *European Journal of Echocardiography* 2011 **12** 665–670. (https://doi.org/10.1093/ejechocard/jer108)

12 Lafitte S, Alimazighi N, Reant P, Dijos M, Zaroui A, Mignot A, Lafitte M, Pillois X, Roudaut R & DeMaria A. Validation of the smallest pocket echoscopic device's diagnostic capabilities in heart investigation. *Ultrasound in Medicine and Biology* 2011 **37** 798–804. (https://doi.org/10.1016/j.ultrasmedbio.2011.02.010)

13 Michalski B, Kasprzak JD, Szymczyk E & Lipiec P. Diagnostic utility and clinical usefulness of the pocket echocardiographic device. *Echocardiography* 2012 **29** 1–6. (https://doi.org/10.1111/j.1540-8175.2011.01553.x)

14 Biais M, Carrie C, Delaunay F, Morel N, Revel P & Janvier G. Evaluation of a new pocket echoscopic device for focused cardiac ultrasonography in an emergency setting. *Critical Care* 2012 **16** R82. (https://doi.org/10.1186/cc11340)

15 Prinz C, Dohrmann J, van Buuren F, Bitter T, Bogunovic N, Horstkotte D & Faber L. Diagnostic performance of handheld echocardiography for the assessment of basic cardiac morphology and function: a validation study in routine cardiac patients. *Echocardiography* 2012 **29** 887–894. (https://doi.org/10.1111/j.1540-8175.2012.01728.x)

16 Fukuda S, Shimada K, Kawasaki T, Fujimoto H, Maeda K, Inanami H, Yoshida K, Jissho S, Taguchi H, Yoshiyama M, *et al*. Pocket-sized transthoracic echocardiography device for the measurement of cardiac chamber size and function. *Circulation Journal* 2009 **73** 1092–1096. (https://doi.org/10.1253/circj.CJ-08-1076)

17 Mjolstad OC, Dalen H, Graven T, Kleinau JO, Salvesen O & Haugen BO. Routinely adding ultrasound examinations by pocket-sized ultrasound devices improves inpatient diagnostics in a medical department. *European Journal of Internal Medicine* 2012 **23** 185–191. (https://doi.org/10.1016/j.ejim.2011.10.009)

18 Panoulas VF, Daigeler AL, Malaweera AS, Lota AS, Baskaran D, Rahman S & Nihoyannopoulos P. Pocket-size hand-held cardiac ultrasound as an adjunct to clinical examination in the hands of medical students and junior doctors. *European Heart Journal Cardiovascular Imaging* 2013 **14** 323–330. (https://doi.org/10.1093/ehjci/jes140)

19 Abe Y, Ito M, Tanaka C, Ito K, Naruko T, Itoh A, Haze K, Muro T, Yoshiyama M & Yoshikawa J. A novel and simple method using pocket-sized echocardiography to screen for aortic stenosis. *Journal of the American Society of Echocardiography* 2013 **26** 589–596. (https://doi.org/10.1016/j.echo.2013.03.008)

20 Furukawa A, Abe Y, Ito M, Tanaka C, Ito K, Komatsu R, Haze K, Naruko T, Yoshiyama M & Yoshikawa J. Prediction of aortic stenosis-related events in patients with systolic ejection murmur using pocket-sized echocardiography. *Journal of Cardiology* 2017 **69** 189–194. (https://doi.org/10.1016/j.jjcc.2016.02.021)

21 Gustafsson M, Alehagen U & Johansson P. Imaging congestion with a pocket ultrasound device: prognostic implications in patients With chronic heart failure. *Journal of Cardiac Failure* 2015 **21** 548–554. (https://doi.org/10.1016/j.cardfail.2015.02.004)

22 Phillips CT & Manning WJ. Advantages and pitfalls of pocket ultrasound vs daily chest radiography in the coronary care unit: a single-user experience. *Echocardiography* 2017 **34** 656–661. (https://doi.org/10.1111/echo.13509)

23 Russell FM, Ehrman RR, Cosby K, Ansari A, Tseeng S, Christain E & Bailitz J. Diagnosing acute heart failure in patients with undifferentiated dyspnea: a lung and cardiac ultrasound (LuCUS) protocol. *Academic Emergency Medicine* 2015 **22** 182–191. (https://doi.org/10.1111/acem.12570)

24 Nishigami K. Point-of-care echocardiography for aortic dissection, pulmonary embolism and acute coronary syndrome in patients with killer chest pain: EASY screening focused on the assessment of effusion, aorta, ventricular size and shape and ventricular asynergy. *Journal of Echocardiography* 2015 **13** 141–144. (https://doi.org/10.1007/s12574-015-0265-1)

25 Carlino MV, Paladino F, Sforza A, Serra C, Liccardi F, de Simone G & Mancusi C. Assessment of left atrial size in addition to focused cardiopulmonary ultrasound improves diagnostic accuracy of acute heart failure in the emergency department. *Echocardiography* 2018 **35** 785–791. (https://doi.org/10.1111/echo.13851)

26 Filipiak-Strzecka D, Kasprzak JD, Szymczyk E, Wejner-Mik P & Lipiec P. Bedside screening with the use of pocket-size imaging device can be useful for ruling out carotid artery stenosis in patients scheduled for cardiac surgery. *Echocardiography* 2017 **34** 716–722. (https://doi.org/10.1111/echo.13507)

27 Esposito R, Ilardi F, Schiano Lomoriello V, Sorrentino R, Sellitto V, Giugliano G, Esposito G, Trimarco B & Galderisi M. Identification of the main determinants of abdominal aorta size: a screening by pocket size imaging device. *Cardiovascular Ultrasound* 2017 **15** 2. (https://doi.org/10.1186/s12947-016-0094-z)

28 Cavallari I, Mega S, Goffredo C, Patti G, Chello M & Di Sciascio G. Hand-held echocardiography in the setting of pre-operative cardiac evaluation of patients undergoing non-cardiac surgery: results from a randomized pilot study. *International Journal of Cardiovascular Imaging* 2015 **31** 995–1000. (https://doi.org/10.1007/s10554-015-0656-4)

29 Khan HA, Wineinger NE, Uddin PQ, Mehta HS, Rubenson DS & Topol EJ. Can hospital rounds with pocket ultrasound by cardiologists reduce standard echocardiography? *American Journal of Medicine* 2014 **127** 669.e1–669.e7. (https://doi.org/10.1016/j.amjmed.2014.03.015)

30 Bhavnani SP, Narula J & Sengupta PP. Mobile technology and the digitization of healthcare. *European Heart Journal* 2016 **37** 1428–1438. (https://doi.org/10.1093/eurheartj/ehv770)

31 Chow CK, Ariyarathna N, Islam SM, Thiagalingam A & Redfern J. mHealth in cardiovascular health care. *Heart, Lung and Circulation* 2016 **25** 802–807. (https://doi.org/10.1016/j.hlc.2016.04.009)

32 Eapen ZJ, Turakhia MP, McConnell MV, Graham G, Dunn P, Tiner C, Rich C, Harrington RA, Peterson ED & Wayte P. Defining a mobile health roadmap for cardiovascular health and disease. *Journal of the American Heart Association* 2016 **5** e003119. (https://doi.org/10.1161/JAHA.115.003119)

33 Tarakji KG, Wazni OM, Callahan T, Kanj M, Hakim AH, Wolski K, Wilkoff BL, Saliba W & Lindsay BD. Using a novel wireless system for monitoring patients after the atrial fibrillation ablation procedure: the iTransmit study. *Heart Rhythm* 2015 **12** 554–559. (https://doi.org/10.1016/j.hrthm.2014.11.015)

34 Williams B, Lacy PS, Baschiera F, Brunel P & Dusing R. Novel description of the 24-hour circadian rhythms of brachial versus central aortic blood pressure and the impact of blood pressure treatment in a randomized controlled clinical trial: the Ambulatory Central Aortic Pressure (AmCAP) Study. *Hypertension* 2013 **61** 1168–1176. (https://doi.org/10.1161/HYPERTENSIONAHA.111.00763)

35 Brooks GC, Vittinghoff E, Iyer S, Tandon D, Kuhar P, Madsen KA, Marcus GM, Pletcher MJ & Olgin JE. Accuracy and usability of a self-administered 6-minute walk test smartphone application. *Circulation: Heart Failure* 2015 **8** 905–913. (https://doi.org/10.1161/CIRCHEARTFAILURE.115.002062)

36 Tison GH, Sanchez JM, Ballinger B, Singh A, Olgin JE, Pletcher MJ, Vittinghoff E, Lee ES, Fan SM, Gladstone RA, *et al*. Passive detection of atrial fibrillation using a commercially available smartwatch.

JAMA Cardiology 2018 **3** 409–416. (https://doi.org/10.1001/jamacardio.2018.0136)

37 Maisel A, Barnard D, Jaski B, Frivold G, Marais J, Azer M, Miyamoto MI, Lombardo D, Kelsay D, Borden K, *et al*. Primary results of the HABIT trial (heart failure assessment with BNP in the home). *Journal of the American College of Cardiology* 2013 **61** 1726–1735. (https://doi.org/10.1016/j.jacc.2013.01.052)

38 Abdolmanafi A, Duong L, Dahdah N & Cheriet F. Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography. *Biomedical Optics Express* 2017 **8** 1203–1220. (https://doi.org/10.1364/BOE.8.001203)

39 Bhavnani SP, Sola S, Adams D, Venkateshvaran A, Dash PK, Sengupta PP & ASEF-VALUES Investigators. A randomized trial of pocket-echocardiography integrated mobile health device assessments in modern structural heart disease clinics. *JACC: Cardiovascular Imaging* 2018 **11** 546–557. (https://doi.org/10.1016/j.jcmg.2017.06.019)

40 Singh S, Bansal M, Maheshwari P, Adams D, Sengupta SP, Price R, Dantin L, Smith M, Kasliwal RR, Pellikka PA, *et al*. American Society of Echocardiography: remote echocardiography with web-based assessments for referrals at a distance (ASE-REWARD) study. *Journal of the American Society of Echocardiography* 2013 **26** 221–233. (https://doi.org/10.1016/j.echo.2012.12.012)

41 Choi BG, Mukherjee M, Dala P, Young HA, Tracy CM, Katz RJ & Lewis JF. Interpretation of remotely downloaded pocket-size cardiac ultrasound images on a web-enabled smartphone: validation against workstation evaluation. *Journal of the American Society of Echocardiography* 2011 **24** 1325–1330. (https://doi.org/10.1016/j.echo.2011.08.007)

42 GE Healthcare. Vscan Extend datasheet. Chicago, IL, USA: General Electric Company, 2018. (available at: https://www.gehealthcare.com/-/media/fdbbc3f456914f5dbc3cc44cb866ffb5.pdf)

43 Bansal M, Singh S, Maheshwari P, Adams D, McCulloch ML, Dada T, Sengupta SP, Kasliwal RR, Pellikka PA, Sengupta PP, *et al*. Value of interactive scanning for improving the outcome of new-learners in transcontinental tele-echocardiography (VISION-in-tele-Echo) study. *Journal of the American Society of Echocardiography* 2015 **28** 75–87. (https://doi.org/10.1016/j.echo.2014.09.001)

44 Boman K, Olofsson M, Berggren P, Sengupta PP & Narula J. Robot-assisted remote echocardiographic examination and teleconsultation: a randomized comparison of time to diagnosis with standard of care referral approach. *JACC: Cardiovascular Imaging* 2014 **7** 799–803. (https://doi.org/10.1016/j.jcmg.2014.05.006)

45 Kagiyama N, Toki M, Hara M, Fukuda S, Aritaka S, Miki T, Ohara M, Hayashida A, Hirohata A, Yamamoto K, *et al*. Efficacy and accuracy of novel automated mitral valve quantification: three-dimensional transesophageal echocardiographic study. *Echocardiography* 2016 **33** 756–763. (https://doi.org/10.1111/echo.13135)

46 Medvedofsky D, Mor-Avi V, Byku I, Singh A, Weinert L, Yamat M, Kruse E, Ciszek B, Nelson A, Otani K, *et al*. Three-dimensional echocardiographic automated quantification of left heart chamber volumes using an adaptive analytics algorithm: feasibility and impact of image quality in nonselected patients. *Journal of the American Society of Echocardiography* 2017 **30** 879–885. (https://doi.org/10.1016/j.echo.2017.05.018)

47 Mor-Avi V, Lang RM, Badano LP, Belohlavek M, Cardim NM, Derumeaux G, Galderisi M, Marwick T, Nagueh SF, Sengupta PP, *et al*. Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: ASE/EAE consensus statement on methodology and indications endorsed by the Japanese Society of Echocardiography. *Journal of the American Society of Echocardiography* 2011 **24** 277–313. (https://doi.org/10.1016/j.echo.2011.01.015)

48 Tsang W, Salgo IS, Medvedofsky D, Takeuchi M, Prater D, Weinert L, Yamat M, Mor-Avi V, Patel AR & Lang RM. Transthoracic 3D echocardiographic left heart chamber quantification using an automated adaptive analytics algorithm. *JACC: Cardiovascular Imaging* 2016 **9** 769–782. (https://doi.org/10.1016/j.jcmg.2015.12.020)

49 Medvedofsky D, Mor-Avi V, Amzulescu M, Fernandez-Golfin C, Hinojar R, Monaghan MJ, Otani K, Reiken J, Takeuchi M, Tsang W, *et al*. Three-dimensional echocardiographic quantification of the left-heart chambers using an automated adaptive analytics algorithm: multicentre validation study. *European Heart Journal Cardiovascular Imaging* 2018 **19** 47–58. (https://doi.org/10.1093/ehjci/jew328)

50 Shameer K, Johnson KW, Glicksberg BS, Dudley JT & Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018 **104** 1156–1164. (https://doi.org/10.1136/heartjnl-2017-311198)

51 Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E & Dudley JT. Artificial intelligence in cardiology. *Journal of the American College of Cardiology* 2018 **71** 2668–2679. (https://doi.org/10.1016/j.jacc.2018.03.521)

52 DeepMind. AlphaGo. London, UK: DeepMind Technologies Limited, 2019. (available at: https://deepmind.com/research/alphago/)

53 Krittanawong C, Zhang H, Wang Z, Aydar M & Kitai T. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology* 2017 **69** 2657–2664. (https://doi.org/10.1016/j.jacc.2017.03.571)

54 Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW & Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine* 2016 **44** 368–374. (https://doi.org/10.1097/CCM.0000000000001571)

55 Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, *et al*. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology* 2014 **63** 2935–2959. (https://doi.org/10.1016/j.jacc.2013.11.005)

56 Cook NR & Ridker PM. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease: an update. *Annals of Internal Medicine* 2016 **165** 786–794. (https://doi.org/10.7326/M16-1739)

57 Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, Gall W & Dudley JT. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circulation: Cardiovascular Imaging* 2016 **9** e004330. (https://doi.org/10.1161/CIRCIMAGING.115.004330)

58 Narula S, Shameer K, Salem Omar AM, Dudley JT & Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *Journal of the American College of Cardiology* 2016 **68** 2287–2295. (https://doi.org/10.1016/j.jacc.2016.08.062)

59 Lancaster MC, Salem Omar AM, Narula S, Kulkarni H, Narula J & Sengupta PP. Phenotypic clustering of left ventricular diastolic function parameters: patterns and prognostic relevance. *JACC: Cardiovascular Imaging* 2018 [epub]. (https://doi.org/10.1016/j.jcmg.2018.02.005)

60 Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, *et al*. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance* 2018 **20** 65. (https://doi.org/10.1186/s12968-018-0471-x)

61 Pellikka PA, She L, Holly TA, Lin G, Varadarajan P, Pai RG, Bonow RO, Pohost GM, Panza JA, Berman DS, *et al*. Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. *JAMA Network Open* 2018 **1** e181456. (https://doi.org/10.1001/jamanetworkopen.2018.1456)

62 Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, *et al*. Fully automated echocardiogram interpretation in clinical practice. *Circulation* 2018 **138** 1623–1635. (https://doi.org/10.1161/CIRCULATIONAHA.118.034338)

63 Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, van Rosendael AR, Beecy AN, *et al*. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal* 2018. (https://doi.org/10.1093/eurheartj/ehy404)

Received in final form 10 February 2019
Accepted 20 February 2019

© 2019 The authors
Published by Bioscientifica Ltd

Future Today
High quality curated content and topics related to innovation and futurism along with a little reflection

Follow

936

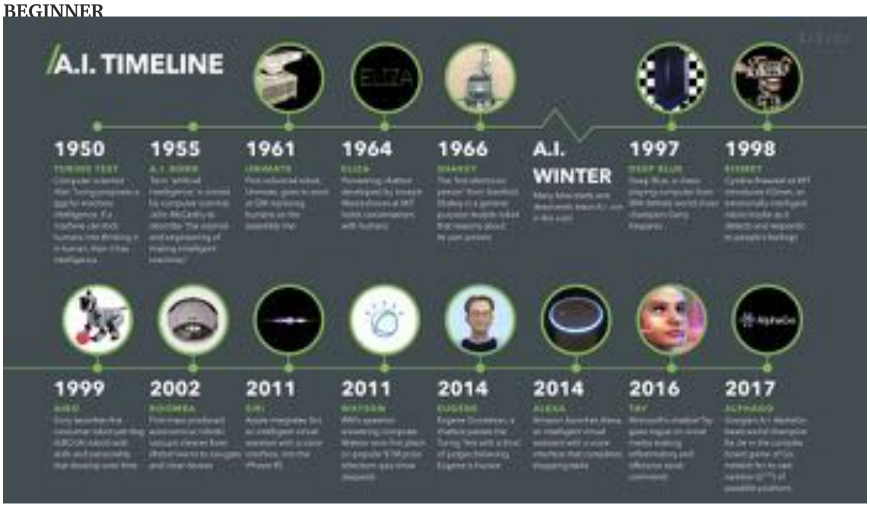
Understanding Artificial Intelligence

David Alayón Follow
Dec 9, 2018 · 11 min read

When I published the article “Understanding Blockchain” many of you wrote me to ask me if I could make one dedicated to **Artificial Intelligence**. The truth is that I hadn’t had time to get on with it and before sharing anything, I wanted to finish some courses in order to add value to the recommendations.

The problem with *Artificial Intelligence* is that it’s much more fragmented, both technologically and in use cases, than *Blockchain*, making it a real challenge to condense all the information and share it meaningfully. Likewise, I have tried to make an effort in the summary of key concepts and in the compilation of interesting sources and resources, I hope it helps you as well as it did to me!

BEGINNER



Let’s start with a little history. The timeline you see is taken from this article and it shows the most important milestones of *Artificial Intelligence*. The term AI goes back to Alan Turing who defined a *from abstractions and concepts, solve problems now reserved for humans, and improve themselves”*. The rest of the milestones you see, mainly *Deep Blue* and *AlphaGo*, will appear throughout the article on several occasions. I recommend that you also watch the *ColdFusion* video where some more details about the history of *Artificial Intelligence* are nuanced.

Future Today
High quality curated content and topics related to innovation and futurism along with a little reflection
Follow
936

Something really interesting that appears in the video are **the 7 aspects of Artificial Intelligence** defined in 1955 and that are still valid today, and in which we have currently reached (with some level of progress) only three of them: *programming a computer to use general language, a way to determine and measure the complexity of problems and self-improvement*. We could also say that we are starting with *“randomness and creativity”*, with some examples like Morgan’s trailer or the script of “Surprising” (2016), the perfumes of Watson, and projects like AIVA, Magenta or My Artificial Muse.

Therefore, we could say that **Artificial Intelligence are machines or computer programs that learn to perform tasks that require types of intelligence and that are usually performed by humans**. And when we talk about types of intelligence, we need to rescue *Jack Copeland*’s reflection on what intelligence is: *“the dominant thought in psychology considers human intelligence not as a single ability or cognitive process, but rather as a set of separate components. Research in AI has focused primarily on the following components of intelligence: learning, reasoning, problem solving, perception, and comprehension of language”*.

That said, let’s go with the different types of Artificial Intelligence. In the video there were two: **Weak AI** or also called **Artificial Narrow Intelligence (ANI)**, which allows computers to outperform humans in some very specific tasks (the most famous example is *IBM Watson*); and **Strong AI** or **Artificial General Intelligence (AGI)**, the ability of a machine to perform the same intellectual tasks as a human being (we are far from reaching it). There is a third level called **Artificial Superintelligence (ASI)**, when a machine possesses an intelligence that far surpasses the brightest and most gifted human minds in the world combined.

Another way to categorize it’s in four levels: **Reactive Machines**, which simply react to a stimulus (or several) but cannot build on previous experiences and cannot improve with practice (*IBM Deep Blue*); **Limited Memory**, can retain and use data for a short period of time but cannot add them to a library of experiences (*Self Driving Cars*); **Theory of Mind**, machines that imitate our mental models: have thoughts, emotions and memories (none yet exist) and finally **Self-Awareness**, or conscious machines, something that stays in the realm of science fiction (for now)

We have now reached the *ANI* or *Limited Memory* level, with the intention of making the next leap but with much uncertainty as to how and when we will achieve it. If we focus on the first categorization, there is a pioneering project that could be laying the groundwork for achieving *AGI* (although it’s still light years away) and is *AlphaGo* or its latest version: AlphaZero. This last one uses a totally different approach for learning than the rest of the AIs we have seen. The previous versions used expert knowledge (humans introducing what’s right) or needed a lot of data (the version of AlphaGo that won Lee Sedol at Go learned from thousands and thousands of games). On the contrary, Alpha Zero uses *Reinforcement Learning*, that is, it learns by playing against itself. In this article you can see what it means and how in 40 days by learning this way it became the best of all its predecessors, and by extension the best in the world.

or directly should be preloaded with a layer of *"common sense"* like what *Etzioni* is creating at the Allen Institute for Artificial Intelligence.

Returning to the present and understanding these basic concepts, we can see the immense applications of AI and how we already have many of them working in our hands or in our day to day: *virtual assistants, translators, eCommerce or social networks recommendations, chatbots* … This is just starting and is going at full speed! If we look a little into the future it's clear that AI is going to change companies, industries, countries and the whole world; and **it's up to us to think how we want it to be and make the right decisions from the present.** Gerd Leonhard has spoken a lot about this topic and along with him there are many other writers, thinkers and futurists who have explained their visions, mainly in using *Artificial Intelligence* to increase and complement us, not to replace us.

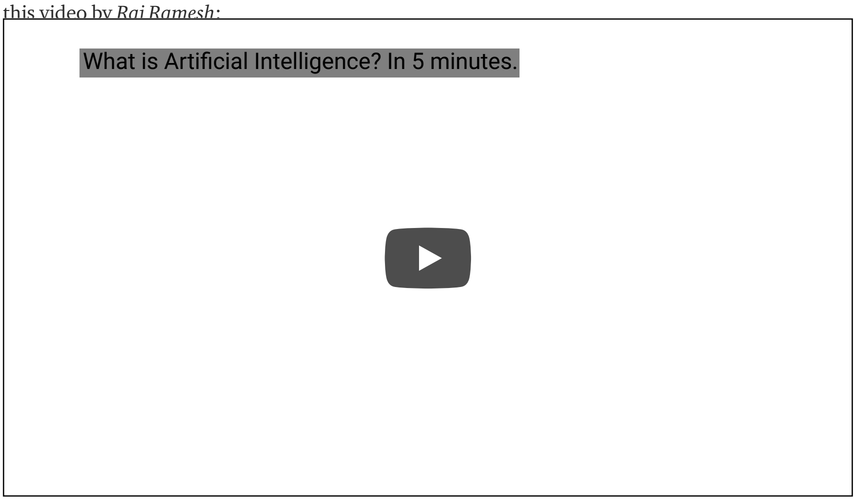
- Agentive Technology. Artificial Intelligence should augment us, give us the tools to get rid of work and do it with autonomy.
- Centaurs. It is demonstrated that the sum of IA and Humans is better than IA alone because of the complementarity of competences (Kasparov proved it on multiple occasions).
- Multiplicity. The key lies in the collaboration between Artificial Intelligence and Amplified Intelligence (human + artificial).
- AI Superpowers. Dr. Kai-Fu includes variables such as compassion, love and I would say empathy, placing AI at the center.

Then we have Yuval Noah Harari and Tristan Harris talking about *dataism* and making a call for a big reflection and ethics, or Elon Musk and another group of technologists and scientists, developing initiatives to raise awareness that we are in our way to self-destruction and creating projects like OpenAIfor a "safe" Artificial Intelligence. I personally don't yet see this Black Mirror approach but I do believe that **we are at the point of starting to think about where we want to go and try to create a kind of world committee to make decisions at the planetary level and as a species**.

To finish this first block, here you have a list of books on AI, highly recommended Nick Bostrom's, and a list of films, highly recommended the last two: *Her (2013)* and *Ex Machina (2015)*.

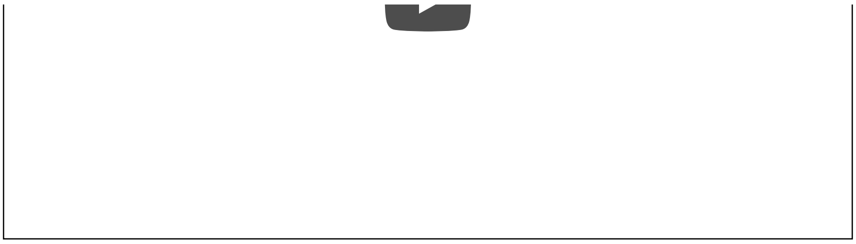
ADVANCED

We're moving to the next level! Regardless of the category, the technological learning base of Artificial Intelligence is mainly based on two pillars: **Symbolic Learning** and **Machine Learning**. Curiously, the first pillar was the one that began everything but with the birth of *Machine Learning* and specifically with **Deep Learning**, all efforts have been focused on the second (although there are many technologists who are thinking on retaking the first). Before we move on, take a look at this video by *Raj Ramesh*:



Really interesting how it synthesizes the different branches of Artificial Intelligence. I think it's clear that the most promising branch that has come to stay is **Machine Learning, which is nothing more than a system capable of taking large amounts of data, developing models that can successfully classify them and then make predictions with new data**. To understand a little more this approach, watch this *CGP Grey's* video:

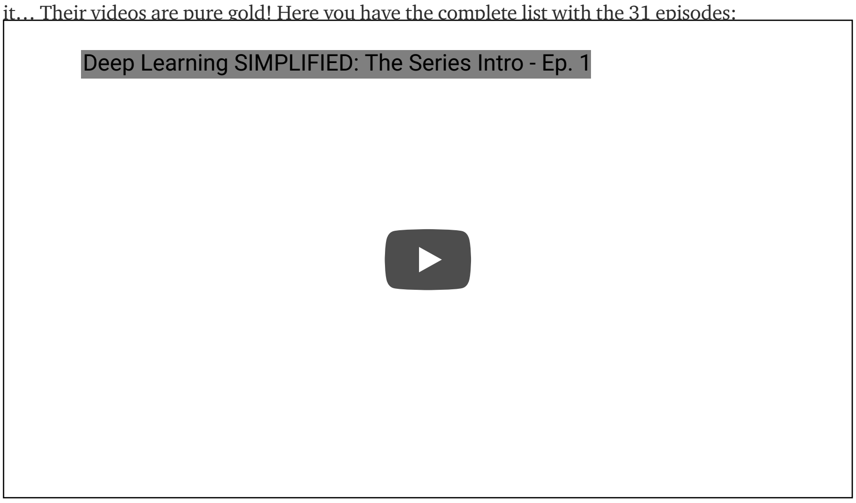
Future Today
High quality curated content and topics related to innovation and futurism along with a little reflection
Follow
936



One of the most interesting thing is that **these models are not programmed, they arise from training, and there is a point where no human, nor the programmers themselves, can understand how it works**. By now, enough new "words" have come out, so I'll leave you with an IA dictionary for beginners and one more reading: Difference between Machine Learning, Deep Learning and Artificial Intelligence.

Now it's time to go deeper into *Deep Learning*, the most advanced approach to develop *Artificial Intelligence* today. After reading many articles, watching many videos and doing some courses, I can say with certainty that an ideal way to have a complete overview of *Deep Learning*, handling basic concepts, technical terminology and even starting to know some tools and platforms is **DeepLearningTV**. I don't know how long it will be active (I recommend you download the videos) because it's been a while since their last update and I don't see any company or community behind it… Their videos are pure gold! Here you have the complete list with the 31 episodes:

Top highlight



Deep Learning SIMPLIFIED: The Series Intro - Ep. 1

Impressive, isn't it? I think what needs a little more development are the frameworks or tools of Machine / Deep Learning. It talks about **TensorFlow, Caffe, Torch, DeepLearning4j and Theano** but there are many others like **Keras, AWS Deep Learning AMI, Google Cloud ML Engine or Microsoft Cognitive CNTK**. As complementary material, here you have some links with comparatives of these platforms:

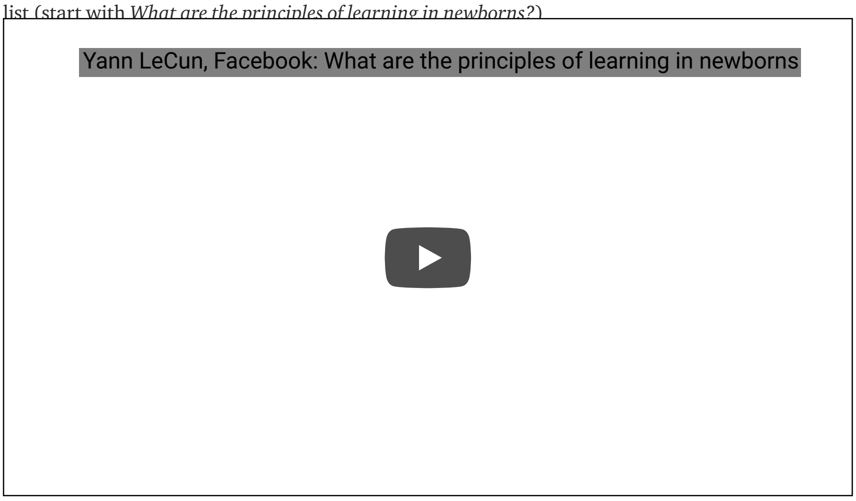
- The 10 most popular machine learning frameworks used by data scientists
- Top 8 open source AI technologies in machine learning
- 8 Best Deep Learning Frameworks for Data Science enthusiasts

Let's summarize some key points to close this block:

- **Types of learning**: *Supervised learning* (contains both inputs and desired outputs, and is trained with a training data); *Unsupervised learning* (only contains input data that has not been tagged or classified, and common elements are identified); and *Reinforcement learning* (instead of focusing on performance, seeks a balance between exploration — new knowledge — and exploitation — current knowledge-).
- **Learning models**: there are many such as *basic regression* (*linear, logistic*), *classification (neural networks, naive bayes, random forest…), cluster analysis (k-means, anomaly detection…)*. Here you have an infographics as cheat sheet and a video that explains a little more in detail 7 of those models.

As you can see, *Artificial Neural Networks (ANNs)* can assume all three types of learning and are within the classification spectrum. I'm not going to go deeper into the types inside each one, *like CNN or RNN* that you'll see in the next section, but I do want to share with you an article by Matt Truck talking about how far we are from the AGI: Frontier AI: How far are we from artificial "general" intelligence, really?. Now that we have a solid base of how DeepMind and AlphaGo works article on this topic, with positive results on the transfer knowledge between simple games. Obviously, this extrapolation to the real world is still science fiction but it's a great start.

Returning to Truck's article, it lists really interesting different approaches to ANNs such as *Recursive Cortical Networks (RCN), CapNets, Differentiable Neural Computers (DNC)*… In short, **fusing neuroscience with AI**. In the article he names a workshop, *"Canonical Computation in Brains and Machines"*, where he specifically talks about these topics and which content is uploaded to YouTube. I haven't had time to watch it entirely (24 lectures of 40 minutes…) but here you have the complete list (start with *What are the principles of learning in newborns?*)


Yann LeCun, Facebook: What are the principles of learning in newborns

TECHNICAL

And we've reached the final level! In this section **I'm not going to explain any new concept in detail but I recommend you different online courses, free and paid, for you to do.** Obviously, they are all technical courses, some require programming experience and others just a solid mathematical base. The important thing about these courses is that you acquire the knowledge you need. Some of you will want to go all the way until you can program an Alpha Zero (as they write in this article) and others, as is my case, understand the technological bases and extrapolate from there. Let's start!

AI For Everyone — Andrew Ng

I don't know if you know *Andrew Ng*, co-founder of *Coursera*, director of *Stanford's AI lab* and former *Chief Scientist at Baidu*. I knew him from the free book he launched in Machine Learning and I recommend the course he has launched in Coursera, 100% online and free.

> *AI is not only for engineers. This non-technical course will help you understand technologies like machine learning and deep learning and spot opportunities to apply AI to problems in your own organization. You will see examples of what today's AI can — and cannot — do. Finally, you will understand how AI is impacting society and how to navigate through this technological change.*

Google Machine Learning

Google has a lot of training content related to TensorFlow. I recommend the crash course of ML. On Youtube you have Machine Learning Recipes with Josh Gordon and AI Adventures, both also very recommendable. I advise you to also go through AI Experiments.

Amazon Machine Learning

Recently, Amazon Web Services has opened its internal Machine Learning courses: 35 online courses that add up to more than 45 hours. They are totally free and although they focus on Amazon technologies, I think they help set a very strong knowledge base Machine and Deep Learning.

Udacity

Udacity began its journey with an Artificial Intelligence course. This course no longer exist by itself but it's integrated in what they call *Nanodegrees*. I haven't done any of them yet but they have one on Artificial Intelligence, another on Deep Learning and a specific one on Self Driving Cars. Very powerful.

Udemy

Finally in *Udemy* you can find 3 paid courses of the same creators of the course I recommended on *Blockchain*: *SuperDataScience*. I recommend you to start with *Artificial Intelligence*:

- Artificial Intelligence A-Z™: Learn How To Build An AI
- Deep Learning A-Z™: Hands-On Artificial Neural Networks

7,940 views | May 29, 2019, 12:21am

Artificial Intelligence In The Workplace: How AI Is Transforming Your Employee Experience

Bernard Marr Contributor

Enterprise & Cloud

Artificial intelligence (AI) is quickly changing just about every aspect of how we live our lives, and our working lives certainly aren't exempt from this.

Artificial Intelligence In The Workplace: How AI Is Transforming Your Employee Experience ADOBE STOCK

Soon, even those of us who don't happen to work for technology companies (although as every company moves towards becoming a tech company, that will be increasingly few of us) will find AI-enabled machines increasingly present as we go about our day-to-day activities.

From how we are recruited and on-boarded to how we go about on-the-job training, personal development and eventually passing on our skills and experience to those who follow in our footsteps, AI technology will play an increasingly prominent role.

Here’s an overview of some of the recent advances made in businesses that are currently on the cutting-edge of the AI revolution, and are likely to be increasingly adopted by others seeking to capitalize on the arrival of smart machines.

Recruitment and onboarding

Before we even set foot in a new workplace, it could soon be a fact that AI-enabled machines have played their part in ensuring we’re the right person for the job.

AI pre-screening of candidates before inviting the most suitable in for interviews is an increasingly common practice at large companies which make thousands of hires each year, and sometimes attract millions of applicants.

Pymetricsprovides tools which use a series of "games" based on principles of neuroscience to assess candidates before they are asked in for an interview. It works by assessing cognitive and emotional features of the candidate, while specifically avoiding demographic biases based on their gender, socioeconomic status, or race. This is done by matching candidates' performance against that of existing employees who have succeeded in the roles that are being recruited for. If it finds that they may not be a particularly good fit for that role, it might recognize another role they would be suitable for and recommend they instead apply for that one.

Another company providing these services is Montage, which claims that 100 of the Fortune 500 companies have used their AI-driven interviewing tool. It enables businesses to carry out on-demand text interviewing, automated scheduling, and reduce the impact of unconscious biases on the recruitment process.

When it comes to onboarding, AI-enabled chatbots are the current tool of choice, for helping new hires settle into their roles and get to grips with the various facets of the organizations they've joined.

Multinational consumer goods manufacturer Unilever uses a chatbot called Unabot, that employs natural language processing(NLP) to answer employees' questions in plain, human language. Advice is available on everything from where they can catch a shuttle bus to the office in the morning, to how to deal with HR and payroll issues.

On-the-job training

Of course, learning doesn't end once you've settled into your role, and AI technology will also play a part in ongoing training for most employees in the future.

It will also assist with the transfer of skills from one generation to the next – as employees move on to other companies or retire, it can help to ensure that they can leave behind the valuable experience they've gained for others to benefit from, as well as take it with them.

Engineering giant Honeywellhas developed tools which utilize augmented and virtual reality (AR/VR) along with AI, to capture the experience of work and extract "lessons" from it which can be passed on to newer hires.

Employees wear AR headsets while carrying out their daily tasks. These capture a record of everything the engineer does, using image recognition technology, which can be played back, allowing trainees or new hires to experience the role through VR.

Information from the video imagery is also being used to build AR tools which provide real-time feedback while engineers carry out their job – alerting them to dangers or reminding them to carry out routine tasks when they are in a particular place or looking at a particular object.

Augmented workforce

One of the reasons that the subject of AI in the workplace makes some people uncomfortable is because it is often thought of as something that will replace humans and lead to job losses.

However, when it comes to AI integration today, the keyword is very much "augmentation" – the idea that AI machines will help us do our jobs more efficiently, rather than replace us. A key idea is that they will take over the mundane aspects of our role, leaving us free to do what humans do best – tasks which require creativity and human-to-human interaction.

Just as employees have become familiar with tools like email and messaging apps, tools such as those provided by PeopleDocor Betterworkswill play an increasingly large part in the day-to-day workplace experience.

These are tools which can monitor workflows and processes and make intelligent suggestions about how things could be done more effectively or efficiently. Often this is referred to as robotic process automation (RPA).

These tools will learn to carry out repetitive tasks such as arranging meetings or managing a diary. They will also recognize when employees are having difficulty or spending too long on particular problems, and be ready to step in to either assist or if the job is beyond something a bot is capable of doing itself, suggest where human help can be found.

Surveillance in the workplace

Of course, there's a potential dark side to this encroachment of AI into the workplace that's likely to leave some employees feeling distinctly uncomfortable.

According to a Gartner survey, more than 50% of companies with a turnover above $750 million use digital data-gathering tools to monitor employee activities and performance. This includes analyzing the content of emails to determine employee satisfaction and engagement levels. Some companies are known to be using tracking devices to monitor the frequency of bathroom breaks, as well as audio analytics to determine stress levels in voices when staff speak to each other in the office.

Technology even exists to enable employers to track their staff sleeping and exercise habits. Video game publisher Blizzard Activision recently unveiled plansto offer incentives to staff who let them track their health through Fitbit devices and other specialized apps. The idea is to use aggregated, anonymized data to identify areas where the health of the workforce as a whole can be improved. However, it's clear to see that being monitored in this way might not sit particularly well with everyone.

Workplace analytics specialists Humanyzeuse staff email and instant messaging data, along with microphone-equipped name badges, to gather data on employee interactions. While some may consider this potentially intrusive, the company says that this can help to protect employees from bullying or sexual harassment in the workplace.

Workplace Robots

Physical robots capable of autonomous movement are becoming commonplace in manufacturing and warehousing installations, and are likely to be a feature of many other workplaces in the near future.

Mobility experts Segwayhave created a delivery robot which can navigate through workplace corridors to make deliveries directly to the desk. Meanwhile, security robots such as those being developed by Gamma 2could soon be a common site, ensuring commercial properties are safe from trespassers.

Racing for a space in the office car park could also become a thing of the past if solutions developed by providers such as ParkPlusbecome commonplace. Their robotic parking assistants may not match our traditional idea of how a robot should look, but consist of automated "shuttle units" capable of moving vehicles into parking bays which would be too small for humans to manually park in – meaning more vehicles can fit into a smaller space.

Follow me on Twitter or LinkedIn. Check out my website.

TOMORROW IS WAITING
HOLLI MINTZER
ISSUE: 21 NOVEMBER 2011
4274 WORDS
http://strangehorizons.com/fiction/tomorrow-is-waiting/

If you want the truth, it happened because Anji was feeling lazy. Her AI class wasn't all that interesting, nor was it a field she wanted a career in, so there wasn't any reason she could see for trying especially hard. So she came up with a project that didn't look like too much work, and she picked what looked like the easiest way of doing it. Things just got a little out of hand, after that.

Anji's AI class was taught by a grad student who seemed as bored as her students. It was a graduation requirement for programmers, even though everyone knew AIs, as a field, weren't going anywhere much. In seventy years of computing nobody yet had designed an AI that passed the Turing test, let alone did anything really interesting. No matter the computing power behind them, AIs just couldn't be as complex as a human brain; everyone knew that. Anji and her classmates still needed to know how to use the little crippleware bots that ran traffic lights and production lines, though, and that meant knowing the basics of AI programming. At least well enough to pass the final.

So Anji decided to pick the easiest-looking project off the list of options: Design an AI that mimics the behavior of a public domain character. There was a list of characters to choose from, mostly stuff she'd never heard of. She picked Kermit the Frog because, she figured, there was a ton of footage of Kermit, even if it was mostly fifty years old, and she could just feed old TV shows to a bot until it started acting enough like Kermit to get her a passing grade.

Only it wasn't that easy. For one thing, the bot was too stupid to understand that it was meant to be Kermit. Anji used off-the-shelf open-source language- and image-parsing software, so the bot would understand what it what watching, but she had to write a program to key the bot to Kermit in particular. It took forever. It was actually a pretty good challenge, writing a program to convince the bot that it was Kermit the Frog, that the little fuzzy green thing in the old video was itself—that it had a self, for that matter. She ended up using concepts and bits of code from the other classes she was taking, pulling a few all-nighters at the library with books on AI design, and just plain making stuff up in a few places. Her code wasn't anything like elegant, but Anji found herself liking the project a lot more than she'd expected to, even as it got harder.

She also found herself liking Kermit a lot more than she'd expected to. Anji had never really watched the Muppets before; her parents, like most parents she knew, had treated TV as only slightly less corrupting an influence than refined sugar and gendered toys. But The Muppet Show was really funny—strange, and kind of hokey, but charming all the same. She ended up watching way more of it than she needed just for the project.

Then her friend Brian, who was really into robotics, got wind of what she was doing, and demanded the chance to participate. Apparently he had weird, nostalgic parents who'd actually allowed him to watch TV as a kid, and what he'd mostly watched was Sesame Street and the Muppets, so the chance to make a real live AI-powered Kermitbot was too good to pass up.

Of course, that made more work for Anji. She had finally gotten the bot keyed to Kermit properly, so it didn't get confused every time there was another Muppet on screen that looked vaguely froggy or was voiced by Jim Henson, and it was sucking down footage at a pretty good clip—luckily there was so much to feed it, on top of the movies: hours and hours of TV specials and commercials and interviews and even outtakes, all of it in character. But now she had to write a whole new suite of programs so the little AI could operate a robot body. Anji started to worry about finishing the project by the due date. For that matter, she was getting behind in her other classes, and it would be downright embarrassing to do poorly in them because AI design, of all things, was taking up her time.

The thing was, her little AI was getting kind of interesting. It had started writing its own code about the time she'd gotten it keyed to Kermit properly, which was one of the project requirements, but Anji hadn't expected much more than a few badly parsed lines. Nobody else in her class was getting more than that, but Anji's AI was producing more code all the time. And weird code, too. Anji couldn't really make sense of it, but it was working, apparently: the bot hadn't frozen up or crashed, and it wasn't having any trouble parsing the footage Anji fed it.

Brian finished his robot a couple of days after Anji got through the last of the footage. He presented it to her proudly, like a cat gives you something really good it's killed and expects your praise for it. "Good, isn't he?" Brian asked, beaming at her, and Anji had to admit it was convincing. Brian had really gone all out: the little robot was fully articulated ("Enough to play the banjo!" Brian pointed out), and perfectly accurate, with plenty of internal memory built in, and a wireless charger. It didn't even need to be plugged in to upload Anji's code. Not that most of it was really Anji's, anymore. She was starting to wonder if this project wasn't getting away from her a little.

The one change Brian had made, in designing his robot, was to give it eyelids. He said it was creepy without them. So when Anji hit the key that uploaded her code, the first sign she had that it had worked was when Kermit gave a couple of slow, sleepy blinks. "Oh," he said, sitting up (Anji was glad to see she'd done a good job with the movement programs), "hello there."

"Hi, Kermit!" Brian said, all dorkily excited. "I'm Brian. It's really nice to meet you."

He elbowed Anji. "Uh, hi," she said. "I'm Anjali. Anji, really."

"Hello, Anji," Kermit said. "Pleased to meet you. I'm Kermit the Frog," and hey, that sounded exactly right. Anji was totally getting an A.

Anji let Brian keep talking to Kermit, and went to check her computer to make sure everything had uploaded okay. It looked fine: everything running smooth. Only the bot was still writing new code, even as it chatted with Brian. Huh. Anji looked back over at them; Kermit had said something that was making Brian laugh really, really hard. Bots weren't supposed to be very god at telling jokes, were they? They'd covered that in class: how AIs never really seemed to get how jokes worked, and even AIs designed to tell them mostly just produced a sort of unfunny word salad. Maybe Kermit was just quoting the jokes from the footage she'd fed him. AIs could mimic like that, although if she'd built a bot that could mimic good comic timing she deserved more than just an A.

In the weeks that followed, it got harder to treat Kermit like a school project. He spent a lot of his time with Brian, who claimed to need to do a bunch of unspecified adjustments to the robot, although this mostly seemed to entail Kermit being shown off to all Brian's friends. Anji didn't mind it too much, though, because it gave her more time to try and puzzle out Kermit's code, and also it meant that Kermit acquired a very small banjo and several sets of little clothes from Muppet fans among Brian's friends. And that seemed to make Kermit happy.

That was the freaky thing: Anji had designed a bot that could seem to be happy. She wasn't supposed to be able to do that. She was way, way outside the parameters of her project now, into territory that people who studied AI for a living hadn't covered anywhere Anji could find. Because Kermit could, in fact, make jokes—and if he was mimicking them, the originals weren't in the footage Anji had fed him—and he could noodle around on the banjo in a way that sounded nothing like the precision of music-playing AIs Anji had heard. And he could also do things that freaked Anji out on a deep and meaningful personal level, like the afternoon when Kermit, perched on the edge of the bed in Anji's dorm, stopped strumming his banjo and sighed wistfully.

"You know, I sure do miss Fozzie," he announced, and Anji stopped typing mid-keystroke.

"What did you say?" Anji asked, trying not to sound as startled as she felt.

"Oh, it's not that I don't like it here, Anji. You and Brian are awfully nice. But Fozzie's my best friend, you know? After a while, you get to miss things. The squeak of a rubber chicken. The smell of custard pie on fur. Little things like that."

He sighed again, and went back to strumming his banjo. Anji waited five minutes, excused herself, and ran full-tilt across campus to Brian's dorm.

He answered the door, looking concerned. Well, Anji had been hammering on it pretty hard. "What's the matter? Is Kermit okay?"

"Brian, I think we invented sentient AI." Anji tried not to sound like she was panicking. She totally was, though. "We weren't supposed to invent sentient AI! I was just supposed to get a passing grade! Now there's an artificial life-form in my dorm room who plays the banjo!"

"Whoa. Calm down. Why are you freaking out now? Kermit hasn't gotten any more sentient than he was last week, has he? And why is it such a big deal if he is?"

"People have been trying to build a sentient AI for like seventy years, Brian. And I knocked one together out of spare parts for a freshman project in a class I didn't even want to take!" Anji wasn't sure how people were going to react, but she didn't think it would be good. The grad student who taught her class would probably be pissed. "No one's going to believe I actually programmed him, or that he's really sentient. But he just told me he misses his friend and made a couple of novel jokes that made sense, so I'm pretty sure I've created life. And I bet I'm going to get in trouble for it."

Brian, damn him, thought she was overreacting. Worse, he thought she was mostly worried about her grade. They ended up fighting over it, getting into a yelling match that drew Brian's RA in to tell them they were damaging the rest of the floor's calm. Anji really didn't like Brian's RA.

Anji trudged back across campus to her dorm that night in a gloomy frame of mind. Sure, it was pretty cool that she had created sentient AI, but she was afraid it would cause more problems than she really knew how to handle. There was the issue of convincing people Kermit was really sentient, just for starters, and then what was he supposed to do with himself, if people ever believed he was for real? He was just a little frog in a big world, when you got down to it.

Lost in her own thoughts, Anji didn't hear the music until she was nearly back to her dorm. When the sound finally made its way through her thick skull, she paused outside her door, and just listened for a minute. Kermit was singing a song.

It wasn't anything Anji had heard before. The lyrics were sweet and simple, all about looking towards the future, and how it was always just a day away. "I won't miss yesterday," Kermit sang, "because I can see—tomorrow is waiting for me." He strummed a few more chords on the banjo, and fell silent.

Anji pushed the door open. "I liked your song, Kermit," she said.

"Thanks, Anji," Kermit said. "It just kinda came to me, you know? That's why I like singing."

"Yeah," Anji said. She though about her project deadline, three days away, and the other homework she wasn't getting done. Then she sat down at her desk and called up a fresh copy of the generic AI, the same blank template she'd started from with Kermit, and got to work keying it to Fozzie.

She wasn't anything like done, three days later, when it was time to present her project, but she'd gotten a lot of good practice with Kermit, and she thought she could have Fozzie up and running inside of two weeks. Kermit walked with her to class, carrying his banjo slung across his back, and Anji ignored the funny looks they got from the other students passing them. She was busy with a sudden, unexpected flurry of guilt: what right, she thought, did she have to show Kermit off to her class like—like some kind of show frog? If he was sentient, he deserved better. Just because he didn't seem to mind—was, in fact, excited to be performing for an audience—didn't mean that Anji was doing the right thing.

But right or wrong, if she didn't show up with something to show for a semester's worth of work, her GPA would be toast. Anji felt guilty, but that didn't stop her from being practical. And she could hope for allies among her classmates, maybe. Once they saw Kermit, they might understand.

Or she could get in a lot of trouble. That was the thought at the top of her mind as Anji came into the classroom, and nervously eyed Malika, the grad student waiting at the front of the room. A few other students had already arrived, most of them carrying the tablets or laptops they'd demonstrate their own projects with. A few had robots, but theirs were little bug-like creatures or wheeled rovers.

To her surprise, Malika brightened as soon as she saw Kermit, and came over to talk to Anji. "You did Kermit?" she said, sounding delighted. "Wow, he looks really great. Just like the real thing. Who built him?"

"Um," Anji said, already embarrassed to be talking like Kermit wasn't there. "There's something I'd like to talk to you about, actually. In private?"

"Sure, sure, after class," Malika said. "I can't wait to see your presentation!"

Somehow, it was worse than if Malika hadn't been interested at all. Kermit looked up at her, concern showing on his small green face. "Are you all right, Anji?"

She hadn't known how to talk to Kermit about the problem. And now it seemed like it was too late. "Just nervous, that's all," she lied.

"You don't have to be nervous," Kermit said. "I mean, sharing something with an audience for the first time is always a little scary, but I've got lots of practice. You don't need to worry about me."

Yes, I do, Anji thought, but she didn't say it.

Kermit didn't seem to be bothered by the fact that the other presentations were all about code and AIs and made frequent mention of bots and programming. Anji wished she'd gotten up the nerve to talk to him about what, exactly, he thought he was—if he knew he was a robot, if he understood he was a sentient AI, if he even got what any of that meant. But she'd been too scared to do it, and Brian had been too excited about his new little green friend. She felt miserably like she'd betrayed Kermit's trust.

When Kermit's—when Anji's turn came, Kermit strolled down to the front of the room and hopped easily onto Malika's desk, settling his banjo on his lap. "Hi-ho everyone," he began. "Kermit the Frog here. My friend Anji asked me to put on a show for you. I haven't got the backup I usually have—and anyway, I don't think there's room in here for a chicken chorus, or a penguin orchestra, or a cannon—but I thought I might sing you a song. I hope you all like it."

The class, who had giggled a little at Kermit's joke, fell quiet as he began to sing. It was the same song Anji had heard him working on before, but he'd changed some of the lyrics, and the arrangement wasn't quite the same. The new version was a little better, actually, to Anji's ear. She looked anxiously at her classmates' faces, at Malika, as Kermit sang the chorus again. He played a little flourish on his banjo, sang "Tomorrow is waiting for me" one last time, and strummed a final chord.

There was silence in the classroom, for a long moment. Then someone started clapping, and the rest of the class joined in, and Anji smiled with relief until she saw that Malika wasn't clapping. She looked serious, and thoughtful.

After class, in the empty lecture hall, Malika still looked grave. "Anji, you've put me in kind of a difficult position," she said. "You're obviously a talented programmer, but the project requirements were pretty clear. You weren't supposed to program a performance, you were supposed to get some novel behavior out of your AI."

"Um," said Anji. "This is what I wanted to talk to you about before class, actually. I was kind of afraid of this. See, I didn't program that."

"Then who did?"

"No one did! Kermit came up with it on his own. I'm tone-deaf, anyway; I can't write music."

"Aw, I wouldn't say tone-deaf, Anji," Kermit said. "I've heard you humming along a few times. Tone-confused, maybe, but I bet with a little practice you could get better."

Malika stared at Kermit. Then she said, "Anji, can I have a minute alone with your—with Kermit?"

Anji looked anxiously down at him. "She just wants to ask you a few questions, I think," she said. "Is that okay?"

"No problem," said Kermit. "I interview well."

Anji sat with her back to the wall outside the classroom, the minutes stretching out like taffy. She watched the other students passing by, and wondered if any of them had ever managed to get themselves into a fix like this.

Well, no. Probably not. None of them had invented sentient AI, after all. That was pretty much a one-time thing, unless she got Fozzie off the ground.

Finally, the door opened, and Malika leaned out into the hall. She looked puzzled, like she'd just eaten something and wasn't sure yet if she liked the taste. "Okay, either you've spent the last three months doing nothing but program in responses to every conceivable question, or he's as close to sentient as any AI I've seen. Either way, you must have been seriously slacking off at the beginning of the semester, because your early assignments don't reflect this level of dedication. How the hell did you do it?"

"It was an accident!" Anji said weakly. "I just kept feeding him footage of Kermit, until he kind of was Kermit. I can show you my notes?"

"I think you'd better," Malika said, and stood aside so Anji could come back into the lecture hall.

The next few weeks were confusing, but in a good way. Anji had a truly terrifying meeting with her AI professor, which was mitigated a little by Malika going to bat for her. Her professor didn't come around as easily—apparently he wasn't a Muppet fan—but Anji's code made his eyebrows go up in a promising way, and when they left his office he was already emailing some other AI experts.

Meanwhile, Kermit was becoming something of a star on campus. People were always excited to meet him—some because they were meeting what might be a sentient AI, and other just because they were meeting Kermit the Frog. Brian started working on the bot for Fozzie, and Anji hit the stage where her Fozzie program started writing its own code. This time it was a lot better-documented, with Malika practically peering over her shoulder as the first lines appeared.

Some friends of Brian's at another school got in touch with her, asking if they could use her keying programs to bring Gonzo to life; they apparently had a build team ready to go. Anji said yes. That set off a wave of Internet chatter. Up to now, there hadn't really been any media attention—her professor wanted to wait until enough experts had met with Kermit and agreed that he was sentient. Or "close enough to sentient to fool me," which was his begrudging way of putting it.

But Brian's friends were blogging the whole build process, and the attention they drew eventually found its way back to Anji. For the first time, she found herself fielding interview requests, and a local news team actually came out in person to film her and Kermit talking to their reporter.

Kermit took the whole thing in stride. Well, he would be used to media attention, Anji figured; he had lots of experience dealing with famous people and reporters. When they weren't being interviewed, Kermit spent his time playing music, writing in a little notebook, walking around campus with Anji and talking to people, hanging out with Brian and his friends. He seemed happy, especially when Anji told him Fozzie would be joining them soon. But part of Anji wasn't convinced that everything was okay.

Finally, she got up the nerve to talk about it. "Kermit," she asked, "what—exactly what are you?"

He looked up from the sheet of musical notations he was doodling on. "What do you mean, Anji? I'm a frog."

"Right, but—frogs look like this." She called up a picture on her laptop, of a real frog. It was brownish, and a little slimy-looking.

"Well, obviously, I'm not that kind of frog."

"Then—what kind of frog are you?"

This gave Kermit pause. He didn't say anything for a while, looking down at his small green hands, then tipping his head to one side thoughtfully. "Well," he said, "I know I used to be a puppet frog, and now I'm a robot frog, but I think I'm still a real frog. I think I always was."

Something inside Anji, some taut string that had been vibrating for weeks, suddenly relaxed. "You know what, Kermit?" she said. "That's what I think, too."

"I'm glad, Anji," Kermit said. "Hey, do you want to hear my new song?"

"I'd love to," Anji said.

In the end, it wasn't as bad as Anji thought it would be. There was a fair amount of controversy, but most of it was restricted to the realm of very important people who thought about AI for a living. Plenty of people were willing to believe that Kermit was sentient, and plenty of people thought he was a cleverly programmed hoax. Anji got an offer from Disney World to buy him, which she turned down, and Fozzie and Gonzo went live without a hitch. She made her keying programs public, so other people could give the rest of the Muppets a chance to be real.

That summer, all the build teams got together for the first time. It was a little chaotic, with the programmers talking and laughing and comparing notes, Scooter looking harried as he wandered around with a clipboard, trying to check everyone in, the penguins tuning their instruments, Sweetums carrying an armload of chickens and Gonzo, five minutes later, in frantic search of Camilla. Kermit was at the center of it all, Piggy on his arm, and for the first time he looked completely happy.

"Hey Anji!" Kermit called when he caught sight of her. He and Piggy had been talking to a man Anji vaguely recognized, an AI expert from a school in the Midwest who'd led the build team for Rowlf. He'd been circling the room with a tablet in hand, talking to people, for most of the afternoon.

"What's up, Kermit?" Anji asked. She shot a questioning glance at the AI expert—she thought his name might be Andrew.

"Well, my friend Andrew here says he's got a line on an old theater that's for sale. He thinks we can raise the money to buy it and fix it up by putting on a show! Isn't that great?"

Andrew looked nervously at her, as if he wanted her approval. A lot of the other builders treated her that way, although she'd explained time and again that the whole thing had really been an accident. People still acted like she had some strange power to confer sentience on computer programs, and possibly also could shoot lightning from her eyes.

But she wasn't planning on striking anyone down with thunderbolts today. "I think that's an awesome idea," she said, and Andrew relaxed.

And really, it was. She could see how happy the thought of having a theater again made Kermit, and in her head she saw the future unspooling out in front of her: their own theater, a new show every night, too many jokes and songs and unprogrammed answers to ever be faked. People would believe then, she was pretty sure. They'd just have to come and see.

Anji hummed to herself as she left Kermit and Andrew to their conversation. "I won't miss yesterday, because I can see," she sang, only a little off-key. "Tomorrow is waiting for me."

OPINION

Diagnosing the decline in pharmaceutical R&D efficiency

Jack W. Scannell, Alex Blanckley, Helen Boldon and Brian Warrington

Abstract | The past 60 years have seen huge advances in many of the scientific, technological and managerial factors that should tend to raise the efficiency of commercial drug research and development (R&D). Yet the number of new drugs approved per billion US dollars spent on R&D has halved roughly every 9 years since 1950, falling around 80-fold in inflation-adjusted terms. There have been many proposed solutions to the problem of declining R&D efficiency. However, their apparent lack of impact so far and the contrast between improving inputs and declining output in terms of the number of new drugs make it sensible to ask whether the underlying problems have been correctly diagnosed. Here, we discuss four factors that we consider to be primary causes, which we call the 'better than the Beatles' problem; the 'cautious regulator' problem; the 'throw money at it' tendency; and the 'basic research–brute force' bias. Our aim is to provoke a more systematic analysis of the causes of the decline in R&D efficiency.

Over the past 60 years, there have been major advances in many of the scientific and technological inputs into drug research and development (R&D). For example, combinatorial chemistry increased the number of drug-like molecules that could be synthesized per chemist per year by perhaps 800-fold during the 1980s and 1990s[1–3], and greatly increased the size of chemical libraries[4]. DNA sequencing has become over a billion times faster since the first genome sequence was determined in the 1970s[5–7], aiding the identification of new drug targets. It now takes at least three orders of magnitude fewer man-hours to calculate three-dimensional protein structure via X-ray crystallography than it did 50 years ago[8,9], and databases of three-dimensional protein structure have 300 times more entries than they did 25 years ago[9] (see the RCSB Protein Data Bank database website), facilitating the identification of improved lead compounds through structure-guided strategies. High-throughput screening (HTS) has resulted in a tenfold reduction in the cost of testing compound libraries against protein targets since the mid-1990s[10]. Added to this are new inventions (such as the entire field of biotechnology, computational drug design and screening, and transgenic mice) and advances in scientific knowledge (such as an understanding of disease mechanisms, new drug targets, biomarkers and surrogate end points).

There have also been substantial efforts to understand and improve the management of the commercial R&D process. Experience has accumulated on why projects overrun[11], on the factors that influence financial returns on R&D investment[12–17], on project success[18] and R&D portfolio management[19–22], on how to reduce costs by outsourcing, and on what is likely to impress or worry the regulatory authorities[23].

However, in parallel — as many have discussed — R&D efficiency, measured simply in terms of the number of new drugs brought to market by the global biotechnology and pharmaceutical industries per billion US dollars of R&D spending, has declined fairly steadily[24]. We call this trend 'Eroom's Law', in contrast to the more familiar Moore's Law ('Eroom's Law' is 'Moore's Law' backwards). Moore's Law is a term that was coined to describe the exponential increase in the number of transistors that can be placed at a reasonable cost onto an integrated circuit. This number doubled every 2 years from the 1970s to 2010. The term is used more generally for technologies that improve exponentially over time. The data in FIG. 1a show that the number of new US Food and Drug Administration (FDA)-approved drugs per billion US dollars of R&D spending in the drug industry has halved approximately every 9 years since 1950, in inflation-adjusted terms. Part of the contrast between Moore's Law and Eroom's Law is related to the complexity and limited current understanding of biological systems versus the relative simplicity and higher level of understanding of solid-state physics[25] but, as discussed below, there are other important causes.

Although there are difficulties in making like-for-like comparisons in R&D spending over very long periods, Eroom's Law has been fairly robust. The number of new drugs introduced per year has been broadly flat over the period since the 1950s, and costs have grown fairly steadily[24]. The slope of the line, over 10-year periods at least, does not change substantially (FIG. 1b), and assumptions about the delay between R&D investment and drug approval do not have a substantial influence on the overall pattern (FIG. 1c). For more details of the data used for FIG. 1, and the major assumptions made, see Supplementary information S1 (table).

Eroom's Law indicates that powerful forces have outweighed scientific, technical and managerial improvements over the past 60 years, and/or that some of the improvements have been less 'improving' than commonly thought. The more positive anyone is about the past several decades of progress, the more negative they should be about the strength of countervailing forces. If someone is optimistic about the prospects for R&D today, they presumably believe the countervailing forces — whatever they are — are starting to abate, or that there has been a sudden and unprecedented acceleration in scientific, technological or managerial progress that will soon become visible in new drug approvals.

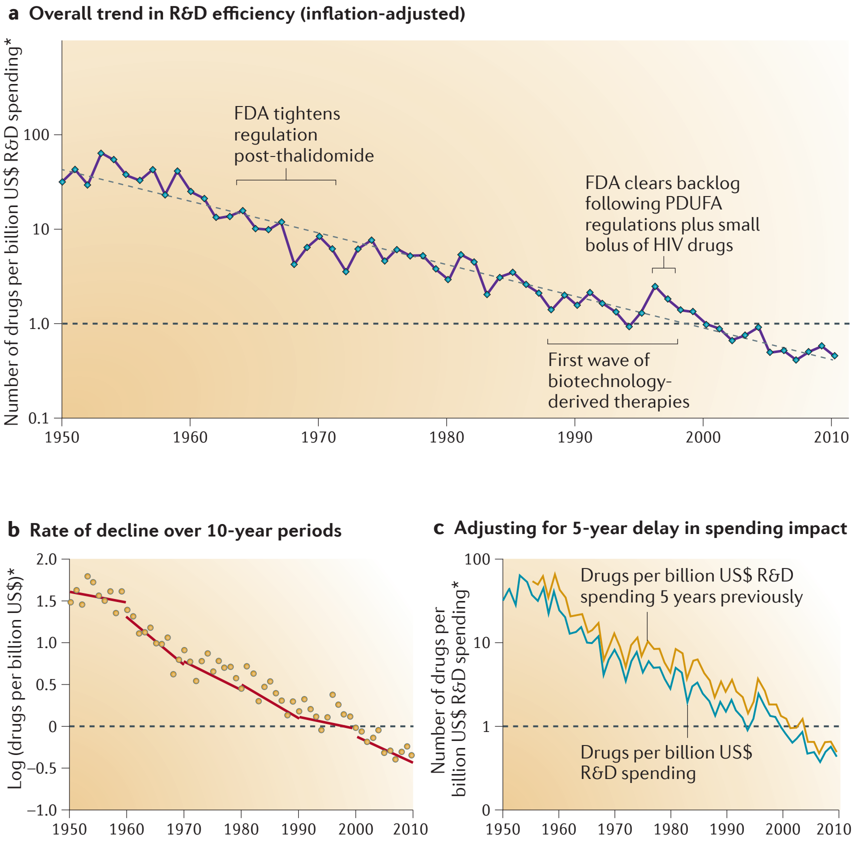


Figure 1 | **Eroom's Law in pharmaceutical R&D. a** | The number of new drugs approved by the US Food and Drug Administration (FDA) per billion US dollars (inflation-adjusted) spent on research and development (R&D) has halved roughly every 9 years. **b** | The rate of decline in the approval of new drugs per billion US dollars spent is fairly similar over different 10-year periods. **c** | The pattern is robust to different assumptions about average delay between R&D spending and drug approval. For details of the data and the main assumptions, see Supplementary information S1 (table) and REFS 24,86,87. Note that R&D costs are based on the Pharmaceutical Research and Manufacturers of America (PhRMA) Annual Survey 2011 (REF. 86) and REF. 87. PhRMA is a trade association that does not include all drug and biotechnology companies, so the PhRMA figure understates R&D spending at an industry level. The total industry expenditure since 2004 has been 30–40% higher than the PhRMA members' total expenditure, which formed the basis of this figure. The new drug count, however, is the total number of new molecular entities and new biologics (applying the same definition as Munos[24]) approved by the US FDA from all sources, not just PhRMA members. We have estimated real-term R&D cost inflation figures from REFS 24,87. The overall picture seems to be fairly robust to the precise details of cost and inflation calculations. Panel **a** is based on a figure that originally appeared in a Bernstein Research report (The Long View — R&D productivity; 30 Sep 2010). *Adjusted for inflation. PDUFA, Prescription Drug User Fee Act.

The magnitude and duration of Eroom's Law also suggests that a lot of the things that have been proposed to address the R&D productivity problem are likely, at best, to have a weak effect. Suppose that we found that it cost 80 times more in real terms to extract a tonne of coal from the ground today than it did 60 years ago, despite improvements in mining machinery and in the ability of geologists to find coal deposits. We might expect coal industry experts and executives to provide explanations along the following lines: "The opencast deposits have been exhausted and the industry is left with thin seams that are a long way below the ground in areas that are prone to flooding and collapse." Given this analysis, people could probably agree that continued investment would be justified by the realistic prospect of either massive improvements in mining technology or large rises in fuel prices. If neither was likely, it would make financial sense to do less digging.

However, readers of much of what has been written about R&D productivity in the drug industry might be left with the impression that Eroom's Law can simply be reversed by strategies such as greater management attention to factors such as project costs and speed of implementation[26], by reorganizing R&D structures into smaller focused units in some cases[27] or larger units with superior economies of scale in others[28], by outsourcing to lower-cost countries[26], by adjusting management metrics and introducing R&D 'performance scorecards'[29], or by somehow making scientists more 'entrepreneurial'[30,31]. In our view, these changes might help at the margins but it feels as though most are not addressing the core of the productivity problem.

There have been serious attempts to describe the countervailing forces or to understand which improvements have been real and which have been illusory. However, such publications have been relatively rare. They include: the FDA's 'Critical Path Initiative'[23]; a series of prescient papers by Horrobin[32–34], arguing that bottom-up science has been a disappointing distraction; an article by Ruffolo[35] focused mainly on regulatory and organizational barriers; a history of the rise and fall of medical innovation in the twentieth century by Le Fanu[36]; an analysis of the organizational challenges in biotechnology innovation by Pisano[37]; critiques by Young[38] and by Hopkins *et al.*[39], of the view that high-affinity binding of a single target by a lead compound is the best place from which to start the R&D process; an analysis by Pammolli *et al.*[19], looking at changes in the mix of projects in 'easy' versus 'difficult' therapeutic areas; some broad-ranging work by Munos[24]; as well as a handful of other publications.

There is also a problem of scope. If we compare the analyses from the FDA[23], Garnier[27], Horrobin[32–34], Ruffolo[35], Le Fanu[36], Pisano[37], Young[38] and Pammolli *et al.*[19], there is limited overlap. In many cases, the different sources blame none of the same countervailing forces. This suggests that a more integrated explanation is required.

Seeking such an explanation is important because Eroom's Law — if it holds — has very unpleasant consequences. Indeed, financial markets already appear to believe in Eroom's Law, or something similar to it, and the impact is being seen in cost-cutting measures implemented by major drug companies. Drug stock prices indicate that investors expect the financial returns on current and future R&D investments to be below the cost of capital at an industry level[40], and would prefer less R&D and higher dividends. Investors may well be wrong about this. However, they have less reason to be biased towards optimism about the likelihood of Eroom's Law being successfully counteracted than those who are working in the industry, or those who sell consulting services to the industry. Shareholders ultimately appoint executives and control resource allocation, so their perceptions matter. It is likely that Pfizer, Merck & Co., AstraZeneca and Eli Lilly will be spending less — in nominal terms — in 2015 than they did in 2011, partly in response to shareholder pressure. Across the top ten large pharmaceutical companies, it seems that nominal R&D spending will be flat until 2015, which represents a decline in real terms. More importantly, the combined effect of declining real-term R&D spending with Eroom's Law (fewer new drugs per billion US dollars of R&D investment over time) is that there will be fewer new drugs and/or drugs will become inordinately expensive. This will threaten the huge benefits[41,42] that follow from the availability of effective and affordable new drugs.

In our view, avoiding such an outcome requires a more systematic analysis of the factors that underlie Eroom's Law. We think that any serious attempt to explain Eroom's Law should try to address at least two things: the progressive nature of the decline in the number of new drugs per billion US dollars of R&D spending, and the scale (~80-fold) of the decline. In this article, we make some suggestions. We realize that the industry is heterogeneous, so our generalizations will be wrong in many cases. We appreciate the intellectual effort that has been made by others on analysing the problems of R&D productivity. We note that our chosen measure of R&D efficiency is based on cost per new drug approved. This does not account for the huge variation in the medical and financial value of new drugs. A few breakthrough drugs — for example, a highly effective Alzheimer's disease treatment — could have much greater medical and financial value than a larger number of new drugs that provide only modest incremental benefits. We also note that the very long cycle time for drug R&D means that our productivity measure is a lagging indicator; perhaps things have improved, but the result is not yet visible.

However, with the aim of prompting debate and analysis, here we discuss what we consider to be the four primary causes of Eroom's Law: the 'better than the Beatles' problem; the 'cautious regulator' problem; the 'throw money at it' tendency; and the 'basic research–brute force' bias. There may also be some contribution from a fifth factor, termed 'the low-hanging fruit' problem, but we consider this to be less important.

Primary causes

The 'better than the Beatles' problem. Imagine how hard it would be to achieve commercial success with new pop songs if any new song had to be better than the Beatles, if the entire Beatles catalogue was available for free, and if people did not get bored with old Beatles records. We suggest something similar applies to the discovery and development of new drugs. Yesterday's blockbuster is today's generic. An ever-improving back catalogue of approved medicines increases the complexity of the development process for new drugs, and raises the evidential hurdles for approval, adoption and reimbursement. It deters R&D in some areas, crowds R&D activity into hard-to-treat diseases and reduces the economic value of as-yet-undiscovered drugs. The problem is progressive and intractable.

Few other industries suffer from this problem. In the example of the coal industry noted above, the opencast deposits are mined first. However, the coal is burnt, which increases the value of the coal that is still in the ground. In most intellectual property businesses (for example, fashion or publishing), people get bored with last year's creations, which maintains demand for novelty. Unfortunately for the drug industry, doctors are not likely to start prescribing branded diabetes drugs because they are bored with generic metformin.

Anti-ulcerants — still a very valuable therapeutic area in terms of revenues — provide an example of the shadow that is cast by successful drugs. A class of anti-acid agents known as potassium-competitive acid blockers, such as soraprazan (now discontinued), would probably be safe and effective anti-ulcerants, and 15 years ago they could have been blockbusters. The problem today is that there are now two classes of highly effective and safe anti-ulcer drugs on the market: the histamine H_2 receptor antagonists (for example, generic ranitidine, which is available over the counter) and the proton pump inhibitors (for example, generic esomeprazole and several others). Any sensible healthcare system would only pay for patients to receive a new branded potassium-competitive acid blocker if they fail to respond to a cheap generic proton pump inhibitor and/or H_2 receptor antagonist, but such patients are a very small proportion of the total population. This general problem applies in diabetes, hypertension, cholesterol management and many other indications.

Pammolli *et al.*[19] have provided a quantitative illustration of the 'better than the Beatles' problem. Their analysis compared R&D projects started between 1990 and 1999 with those started between 2000 and 2004. Attrition rates rose during the latter period. However, the increase could be largely explained by a shift in the mix of R&D projects from commercially crowded therapeutic areas in which historic drug approval probabilities were high (for example, genitourinary drugs and sex hormones) to less crowded areas with lower historical approval probabilities (for example, antineoplastics and immunomodulatory agents).

There is another related potential cause of Eroom's Law that has frequently been put forward, termed the 'low-hanging fruit' problem, which results from the progressive exploitation of drug targets that are more technically tractable[43]. To be clear, the 'low-hanging fruit' problem argues that the easy-to-pick fruit has gone, whereas the 'better than the Beatles' problem argues that the fruit that has been picked reduces the value of the fruit that is left in the tree.

In our opinion, the 'low-hanging fruit' problem is less important than the 'better than the Beatles' problem. First, estimates of the number of potential drug targets[44,45] versus the number of drugged targets[46] suggest that many decades-worth of new targets remain if the industry continues to exploit four or five new targets each year. It is also becoming clear that many drugs may derive their therapeutic benefit from interactions with multiple proteins rather than a single target. These drugs are 'magic shotguns' rather than 'magic bullets'[47]. If this turns out to be more generally true, then worrying about the 'low-hanging fruit' problem would be similar to worrying that a shortage of notes is threatening the future of music composition. In our view, the 'low-hanging fruit' explanation is sometimes tautological as 'technically easy' tends to be equated with 'already discovered'[48]. Indeed, investigation of the history of drug discovery suggests that there was little easy or obvious about some of the great discoveries of the 1940s and 1950s, such as the anti-inflammatory effects of corticosteroids, the psychiatric effects of imipramine or lithium, or the antibacterial properties of penicillin[36,49–51].

The 'cautious regulator' problem. Progressive lowering of the risk tolerance of drug regulatory agencies obviously raises the bar for the introduction of new drugs, and could substantially increase the associated costs of R&D[52]. Each real or perceived sin by the industry, or genuine drug misfortune, leads to a tightening of the regulatory ratchet, and the ratchet is rarely loosened, even if it seems as though this could be achieved without causing significant risk to drug safety. For example, the Ames test for mutagenicity may be a vestigial regulatory requirement; it probably adds little to drug safety but kills some drug candidates. Furthermore, for most of the past 60 years large and sophisticated drug companies may not have been disappointed to see the regulatory ratchet tighten because it reduced competition.

It also seems that the concern that drug companies could cheat the system in some way has led the cautious regulator to apply an audit-based approach to regulatory documentation, as the more demanding the reporting requirements are, the harder it is to cheat without leaving some kind of error or inconsistency in what is reported. The scale of reporting was summarized recently by the Chief Scientific Officer of Novo Nordisk in the company's third quarter 2011 results conference call with respect to the submission to the FDA of data on two new insulin therapies: "If printed and stacked, the many million pages of documentation, with a total of 9 million electronic links, [would] exceed the height of [the] Empire State Building."

The impact of the 'cautious regulator' problem on Eroom's Law is apparent in FIG. 1. First, it shows R&D efficiency dipping in the early 1960s following the 1962 Kefauver Harris Amendment to the Federal Food, Drug, and Cosmetic Act, which was introduced in the wake of the thalidomide drug safety disaster. For the first time, medicines had to demonstrate efficacy, and the safety hurdles were also raised. This reduced financial returns on R&D for a decade or so[12,14], before rising drug prices outstripped R&D cost inflation and increased financial returns in the 1970s[15]. Interestingly, FIG. 1 also shows a rise in R&D efficiency in the mid to late 1990s, which is likely to be due to two regulatory factors: primarily the clearing of a regulatory backlog at the FDA following the implementation of the 1992 Prescription Drug User Fee Act (PDUFA), but also a small contribution from the rapid development and approval of several HIV drugs. In the case of HIV drugs, organized and politically astute lobbying effectively lowered the normal regulatory hurdles[53].

The 'cautious regulator' problem follows, in part, from the 'better than the Beatles' problem, as the regulator is more risk-tolerant when few good treatment options exist; or, put another way, the availability of safe and effective drugs to treat a given disease raises the regulatory bar for other drugs for the same indication. Although the 'cautious regulator' problem is tractable in principle, it is hard to see the regulatory environment relaxing to any great extent. Society may be right to prefer a tougher regulator, even if it means more costly R&D. Drug safety matters. And although the 1950s and 1960s may be viewed by some as a golden age in terms of therapeutic innovation[36,48,54], it seems unlikely that the severe adverse outcomes for many patients taking part in clinical trials during this period[36] would be acceptable today.

The 'throw money at it' tendency. The 'throw money at it' tendency is the tendency to add human resources and other resources to R&D, which — until recent years — has generally led to a rise in R&D spending in major companies, and for the industry overall. It is probably due to several factors, including good returns on investment in R&D for most of the past 60 years, as well as a poorly understood and stochastic innovation process that has long pay-off periods. In addition, intense competition between marketed drugs (where being second or third to launch is often worth less than being first) provides a rationale for investing additional resources to be the first to launch. There may also be a bias in large companies to equate professional success with the size of one's budget.

Unfortunately for people working in R&D today, tackling the 'throw money at it' tendency looks feasible. Investors and many senior executives think that a lot of costs can be cut from R&D without reducing output substantially. Whether this is correct remains to be seen, although if so, it may be the single strategy most likely to counteract Eroom's Law in the short term. The risk, however, is that the lack of understanding of factors affecting return on R&D investment that contributed to relatively indiscriminate spending during the good times could mean that cost cutting is similarly indiscriminate. Costs may go down, without resulting in a substantial increase in productivity.

The 'basic research–brute force' bias. The 'basic research–brute force' bias is the tendency to overestimate the ability of advances in basic research (particularly in molecular biology) and brute force screening methods (embodied in the first few steps of the standard discovery and preclinical research process) to increase the probability that a molecule will be safe and effective in clinical trials (FIG. 2). We suspect that this has been the intellectual basis for a move away from older and perhaps more productive methods for identifying drug candidates[32–34]. It should be noted that many of our comments are more relevant to small-molecule drugs, although the data used for FIG. 1 also include biologics.

FIGURE 2 illustrates the standard model of most drug R&D. It is — effectively — a serial search, filter and selection process. Scientific and technical advances have, superficially at least, increased the breadth of the funnel (that is, more potential targets have been identified, and more drug-like molecules have been synthesized). They have improved the filtering efficiency by several orders of magnitude (for example, HTS versus testing in expensive and low-throughput animal models). They should also have increased the quality of filtering and selection (for example, the use of pathway analysis for target selection, the use of transgenic mice for target validation and the use of accumulated experience to favour molecules that would be likely to have good ADMET (absorption, distribution, metabolism, excretion and toxicology) characteristics).

The cumulative effect of improvements in these early steps should have been a higher signal-to-noise ratio among drug candidates entering clinical trials; that is, the candidates chosen should have had a greater likelihood of successfully demonstrating effectiveness and safety in these trials. This, in turn, should have increased R&D efficiency, given that most of the costs of new drug development are related to the costs of failed projects[22]. Yet the probability that a small-molecule drug successfully completes clinical trials has remained more or less constant for 50 years[21], and overall R&D efficiency has declined[24].

So how can some parts of a process improve dramatically, yet important measures of overall performance remain flat or decline? There are several possible explanations, but it seems reasonable to wonder whether companies industrialized the wrong set of activities[34,36,38]. At first sight, R&D was more efficient several decades ago (FIG. 1), when many research activities that are today regarded as critical (for example, the derivation of genomics-based drug targets and HTS) had not been invented, and when other activities (for example, clinical science, animal-based screens and iterative medicinal chemistry) dominated.

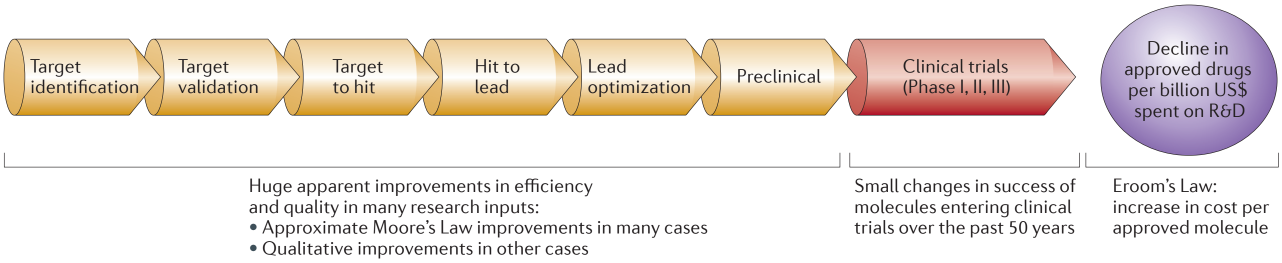

Target identification
Target validation
Target to hit
Hit to lead
Lead optimization
Preclinical
Clinical trials (Phase I, II, III)
Decline in approved drugs per billion US$ spent on R&D
Huge apparent improvements in efficiency and quality in many research inputs:
• Approximate Moore's Law improvements in many cases
• Qualitative improvements in other cases
Small changes in success of molecules entering clinical trials over the past 50 years
Eroom's Law: increase in cost per approved molecule

Figure 2 | **How can some parts of the R&D process improve, yet the overall efficiency decline?** Dramatic improvements in brute force screening methods and basic science should have tended to increase the efficiency of the research process (more leads tested against more targets, at a lower cost; shown in gold) and raised its quality (better targets as disease pathways and mechanisms are understood, better leads that avoid old mistakes surrounding ADMET (absorption, distribution, metabolism, excretion and toxicity) characteristics, and so on). This, in turn, should have increased the probability that molecules would succeed in the clinic (shown in red), which in turn should have increased overall efficiency, as research and development (R&D) costs are dominated by the cost of failure. However, the probability that a small molecule successfully completes clinical trials has remained more or less constant for 50 years[21], whereas overall R&D efficiency has declined[24]. One possible explanation for this is that much of the industry industrialized and 'optimized' the wrong set of R&D activities.

There have been several interesting critiques of modern research[33,48,55], but here we highlight two potential problems. First, much of the pharmaceutical industry's R&D is now based on the idea that high-affinity binding to a single biological target linked to a disease will lead to medical benefit in humans[39]. However, if the causal link between single targets and disease states is weaker than commonly thought[38,56], or if drugs rarely act on a single target, one can understand why the molecules that have been delivered by this research strategy into clinical development may not necessarily be more likely to succeed than those in earlier periods.

Indeed, drug-like small molecules tend to bind promiscuously, and this sometimes turns out to have an important role in their efficacy[47,57] as well as their so-called off-target effects[39]. Targets are parts of complex networks leading to unpredictable effects[58], and biological systems show a high degree of redundancy, which could blunt the effects of highly targeted drugs[56,57]. Perhaps this helps to explain why the R&D process was more cost-effective several decades ago (FIG. 2), when expensive labour-intensive animal models — rather than cheap automated molecular assays — formed the basis of initial drug screening[36,49–51,59].

More recent analysis also points to a similar conclusion. More first-in-class small-molecule drugs approved between 1999 and 2008 were discovered using phenotypic assays than using target-based assays[60]. Drugs approved during this period would have been discovered when screening activity was dominated by the target-based approach, so one might have expected more target-based discoveries. Perhaps target-based approaches are efficient for pursuing validated therapeutic hypotheses but are less effective in the search for innovative drugs that have a better chance of clearing the 'better than the Beatles' barrier.

The second potential problem follows from the nature of chemical space and a shift from iterative medicinal chemistry coupled with parallel assays (pre-1990s) to serial filtering that begins with HTS of a static compound library against a target. Directed iteration — even if each cycle is slow — may be a much more efficient way of searching a large and high-dimensional chemical space than fast HTS of a predefined collection of compounds (BOX 1).

As an aside, biologics have had a higher success rate than small molecules once they leave research and enter clinical trials. There was an approximately 32% approval rate for biologics versus an approximately 13% approval rate for small-molecule drugs first tested in humans between 1993 and 2004 (REF. 21). This may not be surprising for copies or close analogues of endogenous signalling molecules (for example, insulins, erythropoietins or growth hormones) or for agents that replace dysfunctional proteins (for example, clotting factors, lysosomal enzymes, and so on). The high rates of success in clinical trials of monoclonal antibodies (and related fusion proteins) is perhaps more notable[61]. One might expect them to suffer from the same kind of problems with single-target efficacy as small molecules (albeit with fewer off-target effects). However, they have opened up new sets of therapeutic targets, which may suffer less from the 'better than the Beatles' problem. Perhaps their success is also a function of their limited target set — either cell surface proteins or protein-based extracellular signalling molecules. In both cases, the chain of causality between target binding and therapeutic effect is relatively short. Out of 34 monoclonal antibodies or other targeted biologics (such as fusion proteins or aptamers) that have been approved by the FDA, 13 target white blood cell-specific antigens (for example, CD20) and are used for haematological cancers or immunosuppression; three target receptors in the human epidermal growth factor receptor family and are used in oncology; seven target tumour necrosis factor or interleukins and are used for immunomodulation in autoimmune diseases; and four target vascular endothelial growth factor variants and are used in oncology or ophthalmology.

In our view, there are several reasons why the 'basic research–brute force' bias has come to dominate drug research. First, by the early 1980s there was already a sense that the pace of pharmaceutical innovation was slowing. The 'cautious regulator' problem was an obvious drag[52,54,62]. The 'better than the Beatles' problem was starting to emerge, with complaints that new drugs had only modest incremental benefit over what was already available[62]. There were concerns about the 'low-hanging fruit' problem, with a growing sense that the industry had started to run out of good animal models to screen drugs for still poorly treated diseases[52,62].

Second, the 'basic research–brute force' bias matched the scientific zeitgeist[48], particularly as the older approaches for early-stage drug R&D seemed to be yielding less. What might be called 'molecular reductionism' has become the dominant stream in biology in general, and not just in the drug

Box 1 | Directions in small-molecule drug discovery

The 1990s saw a major shift in small-molecule drug discovery strategies, from iterative low-throughput *in vivo* screening and medicinal chemistry optimization to target-based high-throughput screening (HTS) of large compound libraries. At first sight, the former is slow and expensive in terms of the number of compounds that can be tested, whereas the latter is fast and cheap[59]. However, the topography of chemical space and the nature of industrialized drug discovery may conspire to make the second approach less productive. The problem is not necessarily HTS per se (the pros and cons of which are actively debated[79]); rather, it may be the research processes that new technologies helped to cement.

First, real-world compound libraries for HTS cover infinitesimally small and somewhat redundant regions of chemical space, which is vast; it has been suggested that there could be between 10^{26} and 10^{62} (REFS 80,81) chemotypes that would comply with the Lipinski guidelines for oral drugs[82], and each chemotype has a large number of potential derivatives. By contrast, a typical corporate screening collection for HTS contains around 10^6 chemical entities and perhaps 10^3 chemotypes. Furthermore, mergers have revealed that different companies' compound libraries often substantially overlap. This is not surprising: companies generated their libraries in similar ways, as they used clustered sets of molecules from similar historical campaigns; there is a limited set of commercially available reagents; and a relatively small number of reactions are amenable to high-throughput automated synthesis.

Second, it has proved to be difficult to design systems that reward people for producing 'good' hits and leads rather than 'more' hits and leads. Collections are biased towards developable compounds with acceptable ADME (absorption, distribution, metabolism and excretion) characteristics. Companies want measurable developability benchmarks. There are few immediate prizes for chemical or biological novelty. The pre-selection and pre-design of screening collections means that the lead structures are largely foreseen. It provides no easy way to jump from local chemical optima to something better.

Third, the process to whittle down a few thousand HTS hits into a couple of qualified leads has been dominated by molecules that win on potency measures. Selection is based on serial assays, with most molecules failing at each step. There is no practical way to view the full biological profile of all hits at an early stage. Hits with merely adequate target potency but with other potentially attractive features (such as good ADME, other interesting biological properties, and so on) could be thrown away. This further focuses the search process on small parts of screening collections. It may even focus the search process on a suboptimal part of the screening collection. Recent research suggests that there is a negative correlation between *in vitro* potency and desirable ADME and toxicology[83]. Given these features of HTS in the real world, we should expect different drug companies to produce similar molecules for a given target. We should also expect these molecules to reflect local optima within the screening collections, rather than global optima from the much larger chemical universe.

Before the 1990s, however, the standard approach for small-molecule drug discovery involved synthesizing and screening a relatively small number of compounds. There would be a few tens of molecules (often fewer) in active assessment at any one time, and perhaps 1,000 molecules synthesized by a team of chemists during a 5-year project. The search usually started with a molecule that was known, or suspected, to have promising pharmacology but perhaps with poor ADME characteristics: adrenaline led to the development of beta blockers, and histamine lead to the development of cimetidine. Phenomenological screening was also used, to a small extent, to provide starting points. Each molecule was then assessed in a range of concurrent assays (often *in vivo*[59], considering potency, ADME, toxicity, selectivity and so on). Molecules were then modified (or discarded) depending on the results of the assays. The cycle was repeated, with the biological results being used to establish structure–activity relationships for each assay and thus advance the structures of lead compounds through the chemical space until one or two compounds met the multiple criteria necessary for progression into clinical trials. Unlike the screening case, after a few iterations one had compounds specifically customized to a particular target, with structures that would not have been foreseen at the start of the process. This approach prevented trial compounds from being confined to minor local optima. It facilitated what Sir James Black called "obliquity"[84] — the art of looking for one thing and finding something else. It made it less likely that competitors had identical drugs. Remarkably, the search for blockbuster drugs using this method was often achieved with fewer than 1,000 compounds.

This is a profoundly different search strategy to the one that was industrialized, but one that may be more efficient when there is a very large number of items arranged in a high-dimensional space, as is the case with drug-like molecules (see Supplementary information S2 (box)). This is because it is possible to traverse large regions of a high-dimensional space with a small number of steps[85], whereas any static, predefined compound library will cover only a tiny part of the chemical space. Perhaps this is part of the explanation of the pre-1990s productivity? These kinds of arguments are not lost on the drug industry. Efforts are underway to try to combine some of the obvious advantages of HTS with the advantages of small teams dedicated to a broader exploration of the biological profiles of a set of evolving lead compounds. The idea is to analyse several structure–activity relationships in parallel (for example, potency at the target, potency at likely toxicity sites, potency in cellular assays, *in vivo* ADME) to direct rapid, sometimes automated, iterative chemistry.

industry[33,34,55]: "Since the 1970s, nearly all avenues of biomedical research have led to the gene"[63]. Genetics and molecular biology are seen as providing the 'best' and most fundamental ways of understanding biological systems, and subsequently intervening in them[64]. The intellectual challenges of reductionism and its necessary synthesis (the '-omics') appear to be more attractive to many biomedical scientists than the messy empiricism of the older approaches.

Third, the 'basic research–brute force' bias matched the inclination of many commercial managers, management consultants and investors. The old model, based on iterative medicinal chemistry, animal-based screening and clinical science was seen as "too dependent on either inefficient trench-warfare type of slog or the unpredictable emergence of seemingly capricious geniuses like James Black, Paul Janssen, Daniel Bovet, Gertrude Elion, or Gerald Hitchings"[33]. Automation, systematization and process measurement have worked in other industries. Why let a team of chemists and biologists go on a trial and error-based search of indeterminable duration, when one could quickly and efficiently screen millions of leads against a genomics-derived target, and then simply repeat the same industrial process for the next target, and the next? In the early 1990s, few companies thought they could thrive or survive without moving towards a drug discovery process based on HTS and the products of the human genome.

Here, we are reminded of a debate[25] about improving clinical trial efficiency, triggered by an editorial by Andy Grove[65], the former Chief Executive of Intel — a man with personal experience of Moore's Law. Grove noted the "disappointing output" of R&D in the drug industry and made suggestions to radically change clinical trials by making more use of electronic health data[65]. Some biomedical scientists probably find Grove's intervention irritating, given the simplicity and predictability of semiconductor physics versus "biology's mysteries"[25]. However, shareholders and taxpayers have been persuaded to fund a lot of R&D because biomedical scientists (and drug industry executives) have told them that — thanks to molecular reductionism — it would soon become more predictable[63], more productive and less mysterious.

We think that the 'basic research–brute force' bias is supported by survivor bias among R&D projects. This makes drug discovery and development sound more prospectively rational than it really is. Nearly all drugs are sold with a biological story that sounds like molecular reductionism and that sometimes, but not always, turns out to be true: for example, "drug x works by binding receptor a, which influences pathway b, which adjusts physiological process c, which alleviates disease d." Such stories get confused with prediction because we hear very little about the vast majority of the other projects that were also initiated on the basis of high-affinity binding of a plausible candidate to a plausible target, and that had similarly plausible biological stories until the point at which they failed in development for unexpected reasons.

It would be interesting to see how well prospective estimates of plausibility correlated with subsequent attrition. This point is illustrated by the anticancer drug iniparib. Attendees of the 2010 meeting of the American Society of Clinical Oncology (ASCO), or readers of the *New England Journal of Medicine*[66], could have been forgiven for believing that iniparib had a spectacular effect on metastatic breast cancer in a Phase II trial because it inhibited a specific target, poly(ADP-ribose) polymerase 1 (which is involved in DNA repair), and therefore potentiated chemotherapy. However, the following year, Phase III trial results presented at the 2011 ASCO meeting indicated that iniparib did not work very well in breast cancer[67], and it did not seem to inhibit poly(ADP-ribose) polymerase 1 very much either[68].

Fortunately, the 'basic research–brute force' issue is tractable in several ways. First, in a handful of therapeutic areas the research process does appear to be delivering better systems-level insights, better targets (or sets of targets) and better candidate drugs. Oncology is the most obvious example. It is hard to look at the genesis of drugs like crizotinib[69], vemurafenib[70] or vismodegib[71] and think that one is simply looking at random survivors. Furthermore, in oncology the regulator is less cautious and the back catalogue of approved drugs is far from 'Beatle-esque'. One or two other disease areas with simple genetics may perhaps resemble oncology. Second, more emphasis could be put on iterative approaches, on animal-based screening or even on early proof of clinical efficacy in humans, and less on the predictive power of high-affinity binding to the target of a molecule from a static library. Novartis is one company that is emphasizing proof-of-concept trials for drugs in rare diseases for which there is a high unmet need and a compelling match between the drug's mode of action and the disease. Only if there is success here does the company invest in more expensive trials in more common diseases in which the mode of action may be more speculative, or in which the risk–benefit profile may be less clear. Third, in some therapeutic areas people could just stop believing in the current predictive ability of 'basic research–brute force' screening approaches, and resist the temptation to put molecules into clinical trials without having more compelling evidence of the validity of the underlying therapeutic hypothesis.

There is, of course, no way of going back in time to see how well more recent R&D approaches would have worked in the 1940s and 1950s. It is possible that research has become much better at delivering the right molecules into the clinic but that the improvements have been swamped by the 'better than the Beatles' problem, the 'low-hanging fruit' problem and the 'cautious regulator' problem.

Ironically however, if the industry really has been doing the right things, the ultimate prognosis may be bleaker. One can think of the opportunities for R&D in terms of a Venn diagram: as science and technology improve, some sets grow (for example, the set of druggable targets, the set of drug-like molecules and the set of drugged targets), whereas other sets shrink (for example, the set of economically exploitable and still untreated diseases, or the set of acceptable off-target effects). It is obvious that R&D productivity could decline despite improvements in the inputs if the intersection that contained commercially attractive and approvable drug candidates shrunk. This idea is illustrated in FIG. 3, in which the notional set of validated targets grows between 1970 and 2010, but it does not grow fast enough to offset the growth in the set of targets that would either worry a cautious regulator or fail the 'better than the Beatles' test.

Finally, we note that it would be easier to improve the signal-to-noise ratio of drugs that enter clinical trials if: first, there was a detailed understanding of why drugs fail in the clinic; second, this led to the discovery of a small number of common failure modes; and third, this knowledge could be used to change the early stages of the R&D process. If it is impractical to carry out retrospective analyses on the precise molecular mechanisms of clinical trial failure, or if such retrospective analyses show that trials fail for many rare and idiosyncratic reasons, or if cycle times are so long that the lessons are obsolete by the time they are learned, then incremental improvement will be more difficult. Both the regulators[23] and the industry[18] are interested in the analysis of failure but it receives less scrutiny than one might expect given its dominant role in the costs of R&D.

Secondary symptoms

The four proposed primary causes of Eroom's Law discussed above have given rise to several 'symptoms' that tend to further increase costs, particularly the costs of clinical development. Some of these symptoms are highlighted below.

The narrow clinical search problem. The narrow clinical search problem is the shift from an approach that looked broadly for therapeutic potential in biologically active agents to one that seeks precise effects from molecules designed with a single drug target in mind. In the 1950s and 1960s, initial screening was typically performed in animals, not *in vitro* or *in silico*, and drug candidates were given in early stages of the development process to a range of physicians. Discovery involved, to an extent, the ability of physicians to spot patterns through careful clinical observation, especially in therapeutic areas in which symptomatic improvements are readily observable, such as psychiatry[36,49–51]. This is sometimes dismissed as serendipity but the approach made it likely that new therapeutic effects would be detected. Even recently, it appears that many — perhaps most — new therapeutic uses of drugs have been discovered by motivated and observant clinicians working with patients in the real world[72]. Some drug companies, particularly smaller and mid-sized firms, recognize this opportunity and are active repositioners of existing drugs.

However, the 'cautious regulator' problem and the 'basic research–brute force' bias have pushed most of the drug industry towards a narrow clinical search strategy. If a drug has an effect but this is not the precise effect that the trial designers anticipated, then the trial fails. Opportunities for serendipity are actively engineered out of the system. Perhaps it is too risky to let bright doctors with large numbers of patients make broad clinical observations, or to let creative scientists rummage around in rich clinical data sets, in case they find something unexpected, which has to be explained to the cautious regulator who then kills the project. Modern multicentre trials tend to spread the patients so thinly that a doctor who did want to look for patterns might miss them. In Phase II trials — perhaps the best opportunity to spot new things — the average number of patients

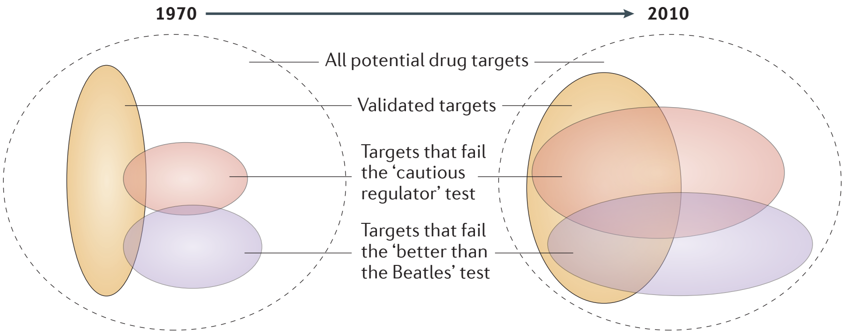

1970
2010
All potential drug targets
Validated targets
Targets that fail the 'cautious regulator' test
Targets that fail the 'better than the Beatles' test

Figure 3 | **Venn diagram illustrating hypothetical headwinds to R&D efficiency.** Research and development (R&D) efficiency could decline if scientific, technical and managerial improvements are offset by other factors. For example, R&D efficiency could be limited by the supply of validated targets that could be drugged without failing the 'cautious regulator' test and/or the 'better than the Beatles' test. In this hypothetical illustration, the increase in the number of validated targets between 1970 and 2010 is outweighed by increasing regulatory caution and an improving catalogue of approved drugs.

per multicentre trial site is now very small: between five and ten patients in oncology, central nervous system and respiratory disease trials[73].

The big clinical trial problem. The first randomized controlled trial, published in 1948, recruited 109 patients and randomized 107 of them[74]. Between 1987 and 2001, the number of patients per pivotal trial for antihypertensive agents rose from around 200 to around 450 (REF. 75). Between 1993 and 2006, the average number of patients across the pivotal trials for a new oral antidiabetic drug rose from around 900 to over 4,000 (REF. 76). The first pivotal trial for Merck's simvastatin (a cholesterol-lowering agent), published in 1994, recruited around 4,400 patients[77]. A pivotal trial for Merck's anacetrapib, an investigational cholesterol-modulating agent intended to be used on top of drugs like simvastatin, is currently recruiting around 30,000 patients.

This expansion is a consequence of several factors. First, the 'better than the Beatles' problem increases trial size. Everything else being equal, clinical trial size should be inversely proportional to the square of the effect size. If the effect size halves, the trial has to recruit four times as many patients to have the same statistical power. The problem is that treatment effects on top of an already effective treatment are usually smaller than treatment effects versus placebo. Furthermore, Phase III trials have become a messy mixture of science, regulation, public relations and marketing. Trying to satisfy these multiple constraints tends to inflate their size and cost.

The best clinical trial to show efficacy would be something relatively small in a homogeneous patient sample recruited from as few centres as possible — the medical equivalent of a well-controlled experiment. But this tends to make the cautious regulator uneasy given variation in practice patterns and patients. What about rare side effects (the FDA has recently required post-marketing trials for long-acting bronchodilators in around 53,000 patients)? Small trials also make for bad marketing and, in the world of evidence-based medicine, poor market access. It is better to involve the senior doctors at the major centres. The number of principal investigators per drug in clinical trials has doubled over the past decade[73]. The consequence of this is multicentre trials that add noise and heterogeneity, and are therefore bigger and more expensive.

The multiple clinical trial problem. The 'better than the Beatles' problem has increased the complexity of medical practice. In some areas, where once there were only one or two treatment options, there is now a rich back catalogue. For example, the treatment of patients with type 2 diabetes was once a choice of insulin or diet and exercise, but can now involve a combination of drugs from around ten different drug classes: biguanides, thiazolidinediones, sulfonylureas, meglitinides, alpha-glucosidase inhibitors, dipeptidyl peptidase 4 inhibitors, glucagon-like peptide 1 analogues, amylin analogues, long-acting and short-acting insulin analogues, as well as various human insulins and insulin mixes. Treatment for patients with colon cancer was once a choice between surgical resection or palliative care, but now the National Comprehensive Cancer Network's colon cancer treatment guidelines contain up to 100 pages of detailed treatment algorithms.

The cautious regulator is less prepared to assume that the safety and efficacy of new drugs can be generalized across such heterogeneous and fragmented patient populations. Cost-sensitive health-care funders are also wary. This means narrower indications and more clinical trials per drug. The first long-acting insulin analogue, glargine, was approved by the FDA in 1999 following three pivotal Phase III trials. The newest long-acting insulin analogue, degludec, was filed for regulatory approval in 2011 following 12 pivotal trials (and, as mentioned above, an Empire State Building's worth of documentation). Some successful drugs in complex therapeutic areas appear to demand, over their life cycle, dozens of Phase III trials[78].

The long cycle time problem. In the 1950s and 1960s, cycle times were remarkably short by modern standards. The regulator was less cautious and there was less molecular reductionism before agents were screened for efficacy in animal models and in patients. This sped up innovation. The first antidepressant, imipramine, was synthesized in around 1951. It was screened almost immediately in rats, and tested personally by a few scientists at the drug company Geigy[51]. It was then tested without much success in various patient groups in 1952, tested again in 1953, found to be problematic in patients with psychosis in 1954 and tried yet again in 1955 before it was identified as an antidepressant in 1956. It completed preclinical development and had not one but three clinical cycles within 5 or 6 years. In 2005–2006, the typical period of time in clinical development for a new drug was over 9 years[21]. The biggest increase in development times came between the 1960s and the 1980s[21].

An idea: the CDDO

This article is intended to provoke further analysis of the forces that have countervailed scientific, technical and managerial improvements over the past 60 years. We have avoided cures, partly because the ratio of published cures to diagnoses is already too high. We do, however, have one idea, which might also be viewed as a thought experiment.

We suggest that all large drug companies introduce a new board level role, which we call the Chief Dead Drug Officer (CDDO). This role would be focused on drug failure at all stages of R&D, and the CDDO would have a fixed time — for example, 18 months — from appointment to compose a detailed report that aims to explain the causes of Eroom's Law. This report would be submitted to the board of the company, included in the company's annual report to shareholders, and would also be submitted for publication in a scientific journal and sent to organizations such as the FDA and the US National Institutes of Health. The remuneration for the role would be structured in such a way as to provide a strong incentive to provide an accurate forecast of the future R&D productivity of the company and the industry overall. For example, perhaps the salary could be relatively modest, but the CDDO could be eligible for an enormous bonus if their projections after a 10-year period are no more than 10% too optimistic or no more than 30% too pessimistic.

We like the idea for several reasons. First, the CDDO has no incentive to be irrationally optimistic. Second, R&D costs are dominated by the cost of failure[73]. Most molecules fail. Most research scientists spend most of their time on products that fail. It seems fitting that someone on the board should focus on the products that consume most of the R&D organization's time, energy and money. Third, an expertise in drug failure should qualify the CDDO to produce a good explanation of Eroom's Law.

The CDDO's report should aim to explain the scale of the change in productivity. It should set out the major factors responsible for the progressive decline, and rank them in order of importance. It should consider how the relative importance of these factors has changed over time. Perhaps changes at the FDA dominated from 1960 to 1970, but something else dominates now? The analysis should compare different therapeutic areas. It should assess the extent to which the different factors are tractable. There should be some effort to quantify the 'better than the Beatles' problem and the 'low-hanging fruit' problem, as well as the potential value of underexploited drug targets. Attention should be given to the regulatory ratchet. Which requirements are most costly and least valuable? Which requirements might the regulator be persuaded to drop? What proportion of R&D cost is a direct consequence of the 'throw money at it' tendency? In which therapeutic areas are molecular reductionism and brute force screening methods a distraction, and in which are they genuinely helpful? What explains the difference between these therapeutic areas? Perhaps the CDDO could quantify their analysis with a series of Venn diagrams like those in FIG. 3, to identify which sets and intersections have grown, and by how much, and which sets and intersections have shrunk. There should also be an attempt to measure the veracity of previous diagnostic and forecasting exercises. What has been the accuracy of internal forecasts on drug approvability and commercial success? Has this changed over time? What have been the most common kinds of error?

If the CDDOs provide a good explanation that is consistent with the idea that the countervailing forces will abate, or will be overcome, then all is well and good. If the explanation is unconvincing, or identifies forces that appear to be intractable, then the problems are obvious. At least it would advance the debate on how to balance the property rights of shareholders and the financial responsibilities of company boards with the wider benefits of safe, effective and affordable new drugs.

The prognosis for Eroom's Law

Just as we wanted to avoid proposing cures, we do not want to say too much about the prognosis for Eroom's Law. However, it might appear strange if we said nothing.

Despite the durability of the trend in FIG. 1, we would be surprised if Eroom's Law holds at an industry level over the next 5–7 years. Our view follows from two somewhat mechanical factors, in addition to one more interesting reason.

Turning to the first of the mechanical factors, the amount spent on R&D is not going to increase. The 'throw money at it' tendency is being tackled by most companies, with varying degrees of intensity. The second mechanical factor is the cumbersome biosimilar approval pathway that is emerging in the United States. Every aspect of the biosimilar production process can be scrutinized by the originator's lawyers, and this raises the prospect of endless blocking litigation. Consequently, developers of biosimilar products anticipate to get at least some of these products approved via the standard new biologics approval pathway (the FDA's biologics license application (BLA) process). These products will be approved as though they were novel agents, so they will inflate the number of novel approvals at very low R&D costs.

Turning to the interesting reason, we suspect that the signal-to-noise ratio may be improving among the compounds being developed for oncology indications. One or two other therapeutic areas may be similar in this respect. Perhaps there are hints of this in the FDA's new drug approvals in 2011. These totalled 30 overall, the most since 2004, although Munos[24] has shown that the distribution of new drugs approved by the FDA per year resembles the output of a Poisson process, so we do not want to over-interpret one good year (if new drug approvals did follow a Poisson process with a mean number of 26 from 1980 to 2010, we would expect 30 drugs to be approved by chance alone around once every 5 years). Looking in more depth at the nature of the 30 new drugs, eight were anticancer agents (brentuximab vedotin, vandetanib, crizotinib, ipilimumab, asparaginase, vemurafenib, ruxolitinib and abiraterone acetate). A focus on rare and poorly treated diseases is also visible in the 2011 total; 11 of the 30 new drugs were orphan drugs, and the orphan drugs included seven of the eight new anticancer agents. Orphan drugs are less prone to many of the factors discussed above, including the 'better than the Beatles' problem, the 'cautious regulator' problem and the big clinical trial problem.

Flat to declining R&D costs, as well as a bolus of oncology drugs, more orphan drugs and 'biosimilars as BLAs', might put an end to Eroom's Law at an industry level. Whether this improves things enough to provide decent financial returns on the industry's R&D investment is a different question. Financial markets don't think so. Industry executives do. It would be interesting to see what CDDOs think.

Jack W Scannell, Alex Blanckley and Helen Boldon are at Sanford C. Bernstein Limited, 50 Berkeley Street, Mayfair Place, London W1J 8SB, UK.

Brian Warrington is at Phoenix IP Ventures, 45 The Drive, Hertford, Hertfordshire SG14 3DE, UK.

Correspondence to J.W.S.
e-mail: Jack.Scannell@Bernstein.com

doi:10.1038/nrd3681

1. Hogan, J. C. Combinatorial chemistry in drug discovery. *Nature Biotech.* **15**, 328–330 (1997).
2. Geysen, H. M., Schoenen, F., Wagner, D. & Wagner, R. Combinatorial compound libraries for drug discovery: an ongoing challenge. *Nature Rev. Drug Discov.* **2**, 222–230 (2003).
3. [No authors listed.] Combinatorial chemistry. *Nature Biotech.* **18**, IT50–IT52 (2000).
4. Dolle, R. E. Historical overview of chemical library design. *Methods Mol. Biol.* **685**, 3–25 (2011).
5. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **57**, 1–28 (1988).
6. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–695 (1977).
7. Meldrum, C., Doyle, M. A. & Tothill, R. W. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* **32**, 177–195 (2011).
8. Joachimiak, A. High-throughput crystallography for structural genomics. *Curr. Opin. Struct. Biol.* **19**, 573–584 (2009).
9. Van Brunt, J. Protein architecture: designing from the ground up. *Nature Biotech.* **4**, 277–283 (1986).

10. Mayr, L. M. & Fuerst, P. The future of high-throughput screening. *J. Biomol. Screen.* **13**, 443–448 (2008).
11. Schnee, J. E. Development cost: determinants and overruns. *J. Bus.* **45**, 347–374 (1972).
12. Baily, M. N. Research and development costs and returns: the U.S. pharmaceutical industry. *J. Polit. Econ.* **80**, 70–85 (1972).
13. Comanor, W. Research and technical change in the pharmaceutical industry. *Rev. Econ. Stat.* **47**, 182–190 (1965).
14. Grabowski, H. G., Vernon, J. M. & Thomas, L. G. Estimating the effects of regulation on innovation: an international comparative analysis of the pharmaceutical industry. *J. Law Econ.* **21**, 133–165 (1978).
15. Grabowski, H. & Vernon, J. A new look at the returns and risks to pharmaceutical R&D. *Manage. Sci.* **36**, 804–821 (1990).
16. Jensen, E. J. Research expenditures and the discovery of new drugs. *J. Ind. Econ.* **36**, 83–95 (1987).
17. Joglekar, P. & Paterson, M. L. A closer look at the returns and risks of pharmaceutical R&D. *J. Health Econ.* **5**, 153–177 (1986).
18. Elias, T., Gordian, M., Singh, N. & Zemmel, R. Why products fail in Phase III. *In Vivo* **24**, 49–56 (2006).
19. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nature Rev. Drug Discov.* **10**, 428–438 (2011).
20. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Rev. Drug Discov.* **3**, 711–715 (2004).
21. DiMasi, J. A., Feldman, L., Seckler, A. & Wilson, A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.* **87**, 272–277 (2010).
22. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Rev. Drug Discov.* **9**, 203–214 (2010).
23. US Food and Drug Administration. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. *FDA website* [online], http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm (2004).
24. Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nature Rev. Drug Discov.* **8**, 959–968 (2010).
25. Borhani, D. W. & Butts, J. A. Rethinking clinical trials: biology's mysteries. *Science* **334**, 1346–1347 (2011).
26. David, E., Tramontin, T. & Zemmel, R. Pharmaceutical R&D: the road to positive returns. *Nature Rev. Drug Discov.* **8**, 609–610 (2009).
27. Garnier, J. P. Rebuilding the R&D engine in big pharma. *Harv. Bus. Rev*. **86**, 68–79 (2008).
28. Agarwal, S. *et al.* Unlocking the value in big pharma. *McKinsey Quarterly* **2**, 65–73 (2001).
29. Ruffolo, R. R. Engineering success: Wyeth redefines its research & development organisation. *Drug Discovery World website* [online], http://www.ddw-online.com/s/business/p148328/engineering-sucess:-wyeth-redefines-its-research-&-development-organisation-fall-05.html (2005).
30. Douglas, F. L., Narayanan, V. K., Mitchell, L. & Litan, R. E. The case for entrepreneurship in R&D in the pharmaceutical industry. *Nature Rev. Drug Discov.* **9**, 683–689 (2010).
31. Zhong, X. & Moseley, G. B. Mission possible: managing innovation in drug discovery. *Nature Biotech.* **25**, 945–946 (2007).
32. Horrobin, D. Realism in drug discovery — could Cassandra be right? *Nature Biotech.* **19**, 1099–1100 (2001).
33. Horrobin, D. F. Innovation in the pharmaceutical industry. *J. R. Soc. Med.* **93**, 341–345 (2000).
34. Horrobin, D. F. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? *Nature Rev. Drug Discov.* **2**, 151–154 (2003).
35. Ruffolo, R. R. Why has R&D productivity declined in the pharmaceutical industry? *Expert Opin. Drug Discov.* **1** 99–102 (2006).
36. Le Fanu, J. *The Rise and Fall of Modern Medicine* (Little Brown, London, 1999).
37. Pisano, G. *Science Business: The Promise, the Reality, and the Future of Biotech.* (Harvard Business School Press, Boston, 2006).
38. Young, M. P. Prediction v Attrition. *Drug Discovery World website* [online], http://www.ddw-online.com/s/business/p92811/prediction-v-attrition-fall-08.html (2008).
39. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127–136 (2006).
40. Tollman, P., Morieux, Y., Murphy, J. K. & Schulze, U. Identifying R&D outliers. *Nature Rev. Drug Discov.* **10**, 653–654 (2011).
41. Ford, E. S. *et al.* Explaining the decrease in U.S. deaths from coronary disease, 1980–2000. *N. Engl. J. Med.* **356**, 2388–2398 (2007).
42. Lichtenberg, F. The impact of drug launches on longevity: evidence from longitudinal disease-level data from 52 countries, 1982–2001. *Int. J. Health Care Finance Econ.* **5**, 47–73 (2005).
43. Schnee, J. E. R&D strategy in the U.S. pharmaceutical industry. *Res. Policy* **8**, 364–382 (1979).
44. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1**, 727–730 (2002).
45. Russ, A. P. & Lampel, S. The druggable genome: an update. *Drug Discov. Today* **10**, 1607–1610 (2005).
46. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature Rev. Drug Discov.* **5**, 993–996 (2006).
47. Roth, B. L., Sheffer, D. L. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Rev. Drug Discov.* **3**, 353–359 (2004).
48. Wurtman, R. J. & Bettiker, R. L. The slowing of treatment discovery, 1965–1995. *Nature Med.* **1**, 1122–1125 (1995).
49. Healy, D. *The Psychopharmacologists: Volume 2* 93–118 (Hodder Arnold, London, 1999).
50. Healy, D. *The Psychopharmacologists: Volume 2* 259–264 (Hodder Arnold, London, 1999).
51. Healy, D. *The Antidepressant Era* (Harvard University Press, Cambridge, Massachusetts, 1997).
52. Weatherall, M. An end to the search for new drugs? *Nature* **296**, 387–390 (1982).
53. Richard, J. & Wurtman, M. D. What went right: why is HIV a treatable infection? *Nature Med.* **3**, 714–717 (1997).
54. [No authors listed.] A dearth of new drugs. *Nature* **283**, 609 (1980).
55. Persson, C. G., Erjefält, J. S., Uller, L., Andersson, M. & Greiff, L. Unbalanced research. *Trends Pharmacol. Sci.* **22**, 538–541 (2001).
56. Ainsworth, C. Networking for new drugs. *Nature Med.* **17**, 1166–1168 (2011).
57. Denome, S. A., Elf, P. K., Henderson, T. A., Nelson, D. E. & Young, K. D. *Escherichia coli* mutants lacking all possible combinations of eight penicillin binding proteins: viability, characteristics, and implications for peptidoglycan synthesis. *J. Bacteriol.* **181**, 3981–3993 (1999).
58. Keith, C. T., Borisy, A. A. & Stockwell, B. R. Multicomponent therapeutics for networked systems. *Nature Rev. Drug Discov.* **4**, 71–78 (2005).
59. Lombardino, J. G. & Lowe, J. A. The role of the medicinal chemist in drug discovery — then and now. *Nature Rev. Drug Discov.* **3**, 853–862 (2004).
60. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nature Rev. Drug Discov.* **10**, 507–519 (2011).
61. Reichert, J. M. Probabilities of success for antibody therapeutics. *mAbs* **1**, 387–389 (2009).
62. Steward, F. & Wibberly, G. Drug innovation — what's slowing it down? *Nature* **284**, 118–120 (1980).
63. Collins, F. S. Medical and societal consequences of the Human Genome Project. *N. Engl. J. Med.* **341**, 28–37 (1999).
64. Rees, J. Post-genome integrative biology: so that's what they call clinical science. *Clin. Med.* **1**, 393–400 (2001).
65. Grove, A. Rethinking clinical trials. *Science* **333**, 1679 (2011).
66. O'Shaughnessy, J. *et al.* Iniparib plus chemotherapy in metastatic triple-negative breast cancer. *N. Engl. J. Med.* **364**, 205–214 (2011).
67. O'Shaughnessy, J. *et al.* A randomized Phase III study of iniparib (BSI-201) in combination with gemcitabine/carboplatin (G/C) in metastatic triple-negative breast cancer (TNBC). *J. Clin. Oncol.* **29**, Abstr. 1007 (2011).
68. Guha, M. PARP inhibitors stumble in breast cancer. *Nature Biotech.* **29**, 373–374 (2011).
69. Soda, M. *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).
70. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
71. [No authors listed.] Regulatory watch: leading hedgehog inhibitor submitted for approval as skin cancer drug. *Nature Rev. Drug Discov.* **10**, 802–803 (2011).
72. DeMonaco, H. J., Ali, A. & von Hippel, E. The major role of clinicians in the discovery of off-label drug therapies. *Pharmacotherapy* **26**, 323–332 (2006).
73. Mathieu, M. P. (ed.) *Parexel's Bio/Pharmaceutical R&D Statistical Sourcebook 2010/2011* 163–261 (Barnett International, Needham, Massachusetts, 2010).
74. Marshall, G. *et al.* Streptomycin treatment of pulmonary tuberculosis. *BMJ* **30**, 769–782 (1948).
75. MacNeil, J. S. H. *Changes in the characteristics of approved New Drug Applications for antihypertensives.* Thesis, Massachusetts Institute of Technology (2007).
76. Lin, H. S. *Changes in the characteristics of new drug applications for the treatment and prevention of diabetes mellitus.* Thesis, Massachusetts Institute of Technology (2007).
77. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian simvastatin survival study (4S). *Lancet* **344**, 1383–1389 (1994).
78. Munos, B. How to avert biopharma's R&D crisis. *In Vivo* **29**, 2011800050 (2011).
79. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nature Rev. Drug Discov.* **10**, 188–195 (2011).
80. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
81. Brown, D. Future pathways for combinatorial chemistry. *Mol. Divers.* **2**, 217–222 (1996).
82. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
83. Gleeson, M. P., Hersey, A., Montanari, D. & Overington, J. Probing the links between *in vitro* potency, ADMET and physicochemical parameters. *Nature Rev. Drug Discov.* **10**, 197–208 (2011).
84. Kay, J. *Obliquity: Why our goals are best achieved indirectly* (Profile Books, London, 2010).
85. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
86. Pharmaceutical Research and Manufacturers of America. Pharmaceutical Industry Profile 2011. *PhRMA website* [online], http://www.phrma.org/sites/default/files/159/phrma_profile_2011_final.pdf (Washington DC, PhRMA, April 2011).
87. Congress of the United States: Congressional Budget Office. Research and Development in the Pharmaceutical Industry. *Congressional Budget Office (CBO) website* [online], http://www.cbo.gov/ftpdocs/76xx/doc7615/10-02-DrugR-D.pdf (October 2006).

Acknowledgements

W. Bains, T. Curtis, B. Charlton, M. Young, O. Imasogie, G. Porges, and B. Munos were generous with their time and ideas during various stages in the genesis of this article.

Disclaimer

The information provided herein was prepared by Sanford C. Bernstein & Co. LLC and Brian Warrington. It is not investment research, although it may refer to a Bernstein research report or the views of a Bernstein research analyst. This communication does not constitute a complete fundamental analysis of any companies mentioned. Unless indicated, all views expressed herein are the views of the authors and may differ from or conflict with those of the Bernstein Research Department. This article does not constitute investment advice or recommendations and is not a solicitation or an offer to purchase or sell securities. The information contained herein is only good as of the date and time of the publication; we do not undertake to advise of any changes in the opinions or information contained herein.

Competing interests statement

The authors declare competing financial interests: see Web version for details.

FURTHER INFORMATION

RCSB Protein Data Bank database:
http://www.rcsb.org/pdb/statistics/holdings.do

SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Clinical development success rates for investigational drugs

Michael Hay, David W Thomas, John L Craighead, Celia Economides & Jesse Rosenthal

The most comprehensive survey of clinical success rates across the drug industry to date shows productivity may be even lower than previous estimates.

Since the human genome was sequenced ten years ago, the number of compounds in development has increased 62% and total R&D expenditures have doubled[1–3]. And yet, the average number of new drugs approved by the US Food and Drug Administration (FDA) per year has declined since the 1990s. In 2012, 39 novel drugs classified as new molecular entities (NMEs) and biologic license applications (BLAs) were approved by the FDA[4]. Although this represents the highest number of approvals since 1997 and is nearly 50% above the average of 26 approvals per year over the past decade, 25% fewer NME and BLA drugs were approved on average in the past 10 years compared with the 1990s[5]. Several possible explanations for the divergence of R&D spending and new product approvals have been offered by professionals in the industry, such as unbalanced regulatory risk-benefit assessments, higher regulatory efficacy hurdles, commercial and financial decisions driving project termination, and the increased complexity and cost of clinical trials[6,7].

This article aims to measure clinical development success rates across the industry with a view to strengthening benchmarking metrics for drug development. The study is the largest and most recent of its kind, examining success rates of 835 drug developers, including biotech companies as well as specialty and large pharmaceutical firms from 2003 to 2011. Success rates for over 7,300 independent drug development paths are analyzed by clinical phase, molecule type, disease area and lead versus nonlead indication status.

Our results pinpoint weaknesses along the capital-intensive pathway to drug approval. Our hope is that they will prove useful in informing policy makers where to focus changes in regulation and strengthen valuation models used by industry and the investment community.

Analyzing success

To measure clinical development success rates for investigational drugs, we analyzed phase transitions from January 1, 2003 to December 31, 2011, in the BioMedTracker database. The BioMedTracker data set contained 4,451 drugs with 7,372 independent clinical development paths from 835 companies and included 5,820 phase transitions. The development paths comprised lead (primary) and nonlead (secondary) indications, with roughly 38% designated as nonlead. A more detailed description of the data collection, composition and analysis methodology is described in **Boxes 1–3** (see also **Tables 1** and **2**).

Unlike many previous studies that reported clinical development success rates for large pharmaceutical companies, this study provides a benchmark for the broader drug development industry by including small public and private biotech companies and specialty pharmaceutical firms. The aim is to incorporate data from a wider range of clinical development organizations, as well as drug modalities and targets. Two landmark publications on the subject, DiMasi *et al.*[6] and Kola *et al.*[8] use 50 and 10 pharmaceutical company pipelines, respectively, to arrive at their conclusions. An important study published by the US Federal Trade Commission Bureau of Economics, Abrantes-Metz *et al.*[9] covered a wide number of drugs over a 14 year period from 1989 to 2002, but did not provide the number or type of companies investigated. Although the impact of company size and experience on R&D productivity has been studied extensively[10–13], success rates established by DiMasi *et al.*[6], Kola *et al.*[8] and Abrantes-Metz *et al.*[9] remain the primary benchmarks for the drug development industry.

We believe it is of great value to report updated success rates that capture the diversity in drug development sponsor types as experience and technology vary widely outside of traditional, large pharmaceutical corporations. Furthermore, the more recent time frame for this study provides insight into the latest industry productivity. A comparison of previously published reports with the current study is summarized in **Table 3** and is discussed below.

One key distinction of the study presented here is our ability to evaluate all of a drug's indications to determine success rates. Danzon *et al.*[12] first considered success rates at the indication level, recognizing that FDA requires clinical trial evidence to establish efficacy for each approved indication. Although these authors included data from 1988 to 2000, an observation period similar to Kola *et al.*[8] and Abrantes-Metz *et al.*[9], their success rates were significantly higher and lacked a characteristic decrease in phase 2 probability reported in previous studies as well as here. Danzon *et al.*[12] concluded that higher clinical development success rates resulted from the analysis of all indications. Even so, evidence presented here strongly suggests that evaluating all indications results in lower probabilities of success across all phases of drug development.

Michael Hay and Jesse Rosenthal are at BioMedTracker, Sagient Research Systems, San Diego, California, USA; David W. Thomas and Celia Economides are at the Biotechnology Industry Organization (BIO), Washington, DC, USA; and John L. Craighead is at Biotech Strategy & Analytics, Rockville, Maryland, USA.
e-mail: mhay@sagientresearch.com

To illustrate the importance of using all indications to determine success rates, consider this scenario. An antibody is developed in four cancer indications, and all four indications transition successfully from phase 1 to phase 3, but three fail in phase 3 and only one succeeds in gaining FDA approval. Many prior studies reported this as 100% success, whereas our study differentiates the results as 25% success for all indications, and 100% success for the lead indication. Considering the cost and time spent on the three failed phase 3 indications, we believe including all 'development paths' more accurately reflects success and R&D productivity in drug development.

Examining individual drug indications allows us to answer the question: "what is the probability that a drug developed for a specific indication will reach approval?" Whereas, using only the lead or most advanced indication seeks to answer the question: "what is the probability that a drug will reach approval for any indication?" This study addresses both questions with emphasis on the findings of the former. In the following sections, we present the results of our analysis as they relate to overall phase success and likelihood of approval (LOA; see **Box 2**), to the type of therapeutic modality, to the disease being treated and to the type of drug application (whether orphan or Special Protocol Assessment (SPA) pathways).

Phase success and likelihood of approval

We found that approximately one in ten (10.4%, $n = 5{,}820$) of all indication development paths in phase 1 were approved by FDA (**Fig. 1** and **Table 4**). Examining the individual phase components of this compound probability, phase I success (the number of phase 1 drugs that successfully transitioned to phase 2 divided by the total transitions in phase 1) was 64.5% ($n = 1{,}918$). Success in phase 2 (32.4%, $n = 2{,}268$) was substantially lower than in phase 1, but subsequently increased in phase 3 (60.1%, $n = 975$). The probability of FDA approval after submitting a new drug application (NDA) or biologic license application (BLA) was 83.2% ($n = 659$).

Success rates for lead indication development paths were higher than for all indication development paths in every phase. Lead indications had a LOA from phase 1 of 15.3% ($n = 3{,}688$).

Success rates by drug classification

Drugs in the BioMedTracker data set were annotated by their FDA classification: new molecular entity (NME), non-NME, biologic and vaccine. However, owing to inconsistency in the FDA classifications, we also used our

Box 1 Data collection and composition

BioMedTracker, a subscription-based product of Sagient Research Systems (San Diego) introduced in 2002, tracks the clinical development and regulatory history of novel investigational drugs in the United States. Analysts with advanced degrees in the life sciences and medicine maintain the database using information from company press releases, analyst conference calls, and presentations at investor and medical meetings. BioMedTracker also uses other sources, including regular communication with companies conducting clinical trials, to ensure the accuracy and timeliness of the data.

Data included in this study were selected using BioMedTracker's Probability of Technical Success (PTS) calculator, which identified 5,820 phase transitions from January 1, 2003, to December 31, 2011. Transitions in all phases of development were recorded in the early years of observation and resulted from clinical studies initiated before 2003. The data set contained 4,451 drugs from 835 companies and 7,372 independent clinical development paths in 417 unique indications.

The composition of these novel drug development sponsors included a wide range of company sizes and types (**Table 1**). Emerging biotech represented 85% (712) of the companies, whereas a small number (33) of large firms (4% of total) were responsible for 48% (3,573) of indications and 47% (2,075) of drugs in development. Similarly, private firms represented 49% (412) of the companies and fewer than 20% of indications and drugs included in the study.

These ownership classifications were recorded at the end of the analysis time period and underestimate the number of drugs and indications developed by biotech companies due to licensing and acquisitions during the study time frame. In addition, ownership was assigned to the licensee controlling and funding the majority of development. In cases where development and economics were shared equally, ownership was generally assigned to the larger organization, further contributing to the conservative estimate of drugs developed by small and private biotech companies. Although generic products were not included, generic manufacturers developing novel investigational drugs were represented.

The study also likely tracked a larger percentage of late-stage studies as these programs are more often in the public domain. Even so, small biotech companies often disclose ongoing phase 1 studies and we would expect their substantial representation in this study to partially offset the under-representation of early-stage discontinuation rates. Only company sponsored development paths designed for FDA approval were considered; investigator sponsored studies and combinations with other investigational drugs were excluded in this analysis.

In addition, this study analyzed development paths organized by disease area, biochemical composition, molecular size, FDA classification and regulatory status (SPA and orphan drug status). Given the increasing complexity of ownership and diversity of invention in the drug development industry, the study did not further classify the database on the discovery origin or licensing status of the drug.

Table 1 Analysis of company size and type

	Companies		Indications		Drugs	
	Number	Percentage	Number	Percentage	Number	Percentage
Company size						
Large pharma/biotech (>$5 billion sales)	33	4%	3,573	48%	2,075	47%
Small to mid-sized pharma/biotech ($0.1 billion–$5 billion sales)	90	11%	1,099	15%	724	16%
Emerging biotech (<$0.1 billion sales)	712	85%	2,700	37%	1,652	37%
Total	835	–	7,372	–	4,451	–
Company type						
Private	412	49%	1,269	17%	841	19%
Public	423	51%	6,103	83%	3,601	81%
Total	835	–	7,372	–	4,451	–

Box 2 Metrics of success: 'Phase Success' and 'Likelihood of Approval'

There are two different types of success rates reported in this study: 'Phase Success' and 'Likelihood of Approval' (LOA). 'Phase Success' is calculated as the number of drugs that moved from one phase to the next phase divided by the sum of the number of drugs that progressed to the next phase and the number of drugs that were suspended. The *n* value associated with the Phase Success represents the number of drugs that have advanced plus the number of drugs that have been suspended, which we label as phase transitions. For example, if there were 100 drugs in phase 2 development and 50 transitioned to phase 3, 20 were suspended and 30 remained in phase 2 development, the phase 2 Phase Success would be 71.4% (50/70; *n* = 70).

Our second metric, LOA, denotes the probability of reaching FDA approval from the current phase, and is also expressed as a percentage. LOA is calculated as the product of each Phase Success probability leading to FDA approval. The *n* value associated with LOA is the sum of the *n* values for each Phase Success included in the LOA calculation. For example, if a drug is currently in phase 2, and the Phase Success for phase 2 is 30% (*n* = 20), phase 3 is 50% (*n* = 10), and FDA approval is 80% (*n* = 5), then the LOA for the phase 2 drug would be 12% (30% × 50% × 80% = 12%, *n* = 35). This calculation is illustrated in **Supplementary Figure 2**.

data to annotate drugs by their biochemical composition (e.g., peptide, nucleic acid, monoclonal antibody (mAb)) and molecular size (i.e., large and small molecules). For example, FDA often designates large-molecule biologics, such as proteins and peptides, as NMEs. Indeed, large molecules, as defined by the BioMedTracker biochemical categories, comprise 13% of the NME data set, making direct FDA NME to biologic classification comparisons somewhat imprecise. FDA's biologic classification comprises a wider group that includes the Center for Drug Evaluation and Research (CDER) regulated products, such as antibodies, cytokines, growth factors and enzymes, as well as the Center for Biologics Evaluation and Research (CBER) regulated products including blood isolates, gene therapies and cell therapy.

FDA's non-NME classification often includes drugs with the same molecular properties as NMEs, but which are frequently reformulations or combinations of approved products. The majority of non-NMEs also use the 505(b)(2) pathway to gain FDA approval. Vaccines were also treated as a separate class in this analysis, and generic and over-the-counter drugs were not included. A comparative analysis of FDA classifications and BioMedTracker categories can be found in **Supplementary Table 1**. The metrics for the different therapeutic modality types is provided in **Table 4**.

NMEs were found to have the lowest success rates in every phase of development; biologics had nearly twice the LOA from phase 1 (14.6%, *n* = 1,173) as NMEs (7.5%, *n* = 3,496) for all indications (**Table 4**). Similar results are seen when the data are reclassified into large-molecule (excluding low molecular weight chemicals and steroids) and small-molecule NMEs: 13.2% (*n* = 1,834) and 7.6% (*n* = 3,029), respectively. In addition, the LOA from phase 1 for mAbs (14.1%, *n* = 639), a good proxy for CDER-regulated biologics, was also consistent with these broader definitions of biologics.

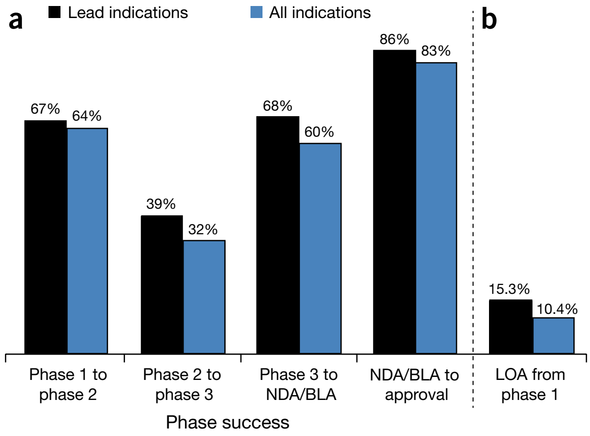


Figure 1 Phase success and LOA rates. (**a**) Phase success rates for lead and all indications. The rates represent the probability that a drug will successfully advance to the next phase. (**b**) LOA from phase 1 for lead and all indications. Rates denote the probability of FDA approval for drugs in phase 1 development.

Non-NMEs had the highest LOA from phase 1 of 20.0% (*n* = 855), with success rates well above those of the NME and biologic classifications in every phase. However, many non-NMEs begin development in phase 2 or phase 3, so the actual approval rate is likely higher (assuming that successful phase 1 outcomes would contribute positively to the LOA from phase 1).

When analyzing lead indications only (i.e., on a per drug basis), we find similar rankings for NME, biologic and non-NME, but at much higher success rates. The LOA from phase 1 for biologics and non-NMEs are near one in four and NMEs approach one in eight (12.0%, *n* = 2,124), almost twice what was found when all indications were considered.

Success rates by disease

We found substantial variation in success rates among disease, as listed in **Table 5** from highest to lowest LOA from phase 1. Oncology drugs had the lowest LOA from phase 1 at 6.7% (*n* = 1,803). Drugs for the 'other' disease group, which combined allergy, gastroenterology, ophthalmology, dermatology, obstetrics-gynecology and urology indications due to small sample size, had the highest LOA from phase 1, at 18.2% (*n* = 720). Drugs for infectious disease and autoimmune-immunology groups had the next two highest LOAs from phase 1, at 16.7% (*n* = 537) and 12.7% (*n* = 549), respectively.

On a lead indication basis, also in **Table 5**, we found that cardiovascular drugs had the lowest LOA from phase 1 at 8.7% (*n* = 318) and the 'other' disease category again had the highest success rate at 24.5% (*n* = 499). The largest difference between lead and all-indication for LOA from phase 1 was observed in oncology: 6.7% (*n* = 1,803) for lead indication and 13.2% (*n* = 796) for all indications. Oncology drugs also had the most nonlead indications (56% of all development paths compared with 28% of non-oncology indications) as a result of the large number of cancers investigated using the same drug. Unfortunately, in oncology, when all indications are considered, only around 1 in 15 drugs entering clinical development in phase 1 achieves FDA approval compared with close to 1 in 8 using the lead indication methodology. As noted above, the result for lead indications represents the most successful development path for a particular compound, thereby addressing LOA on a per drug

Box 3 Methods used in this study

Table 2 Definitions of terms used in this study

BioMedTracker term	Description for purposes of this study
I	Drug is currently in phase 1
I/II, II, IIb	Drug is currently in phase 2
II/III, III	Drug is currently in phase 3
NDA/BLA	Application for approval has been submitted to the FDA and is currently under review
Approved, withdrawn from market, approved (Generic competition)	Drug has been approved for marketing in the United States
Suspended	Drug is no longer in development
Approved in Europe, Approved in other than US/EU, Development, Development outside US	The company developing this drug does not plan to market it in the United States

Data used for this study were extracted from BioMedTracker using a probability of technical success (PTS) tool, which identified all 'Advanced' and 'Suspended' drugs by development phase from January 1, 2003, to December 31, 2011. BioMedTracker tracks the clinical development and regulatory history of investigational drugs to assess its Likelihood of Approval (LOA) from phase 1 by the FDA. The database is populated in near real-time with updated information from press releases, corporate earnings calls, investor and medical meetings, and numerous other sources. These data are recorded in BioMedTracker and tagged with a date.

Phase is defined as the stage of clinical development in the United States (**Table 2**). Although it is rare, drugs that were removed from development in the United States, but approved in Europe (e.g., vildagliptin for type II diabetes) were considered 'suspended' for the sake of our analysis. In this time period, 7,372 development paths were analyzed, encompassing 4,451 unique compounds. 5,820 unique phase transitions were used to determine the reported success rates. **Table 4** includes the number of observed transitions by phase (a description of the success rate analysis is described). Phase 2 transitions accounted for the highest percentage of the data set with 39% (n = 2,268), compared with 33% in phase 1 (n = 1,918), 17% in phase 3 (n = 975) and 11% in NDA/BLA (n = 659). Nonlead indications comprise 38% (n = 2,132) of the 5,820 total transitions and success rates by phase can be found in **Supplementary Table 2**.

Development paths track a specific indication for each drug. For example, Rituxan (rituximab) in non-Hodgkin's lymphoma qualifies as a development path different from Rituxan in multiple sclerosis (MS). BioMedTracker assigns a unique internal identifier that can be used to isolate all development paths. In addition to tracking the phase of development, BioMedTracker assigns 'lead' status to certain development paths. This is used to denote the most advanced indication in clinical development for a specific drug. Drugs can only have one lead development path, except in specific circumstances where two development paths are being developed simultaneously (e.g., type I and type II diabetes). For example, the Avastin (bevacizumab) colorectal cancer development path was marked as a 'lead' indication, and other Avastin development paths were labeled 'nonlead'. Using this metric, Avastin clinical development can more accurately be viewed as a series of successes and failures, as opposed to simply one success and no failures. However, a drug's lead indication may also change if it fails in development in the lead indication. The lead indication success rate will therefore be higher due to selection bias than the nonlead success rate. This bias does not affect the LOA from phase 1 rate for all indication development paths.

BioMedTracker also records a number of other variables including the following:

- FDA classification (e.g., NME, non-NME, biologic or vaccine)
- Biochemical profile (e.g., small molecule, monoclonal antibody, antisense)
- Disease area (e.g., autoimmune, cardiovascular, oncology)
- Indication (e.g., diabetes, acute coronary syndrome)

In contrast with many earlier studies, which included only a limited sample of drugs from large companies, the current study included BioMedTracker data from small biotech companies as well as specialty and large pharmaceutical firms.

Phase success and LOA rates calculation. A common method of determining drug development success rates detailed in DiMasi *et al.*[6] and Abrantes-Metz *et al.*[9] was used in this study. Phase Success, defined as the probability of a drug moving from phase X to phase X + 1, was used as the basis for all analyses. To arrive at this value, the following questions are used to categorize each drug development path: first, was the drug development path ever in phase X? Second, if so, did it advance to phase X + 1? And third, was it 'Suspended'? After categorizing all drug development paths, Phase Success is calculated by dividing the number of development paths that advanced from phase X to phase X + 1 by the sum of the number of development paths that advanced from phase X to phase X + 1 and the number of development paths that were suspended from phase X – Advanced/(Advanced + Suspended) = Phase Success.

Using this method, we arrived at the probabilities of an 'average' drug advancing from phase 1 to phase 2, from phase 2 to phase 3, from phase 3 to filing the NDA/BLA and from filing the NDA/BLA to FDA approval. We then compounded these probabilities to determine the probability (LOA) that a drug in phase X is approved. For example, the LOA for a drug which has entered phase 2 is the product of the phase success rates from phase 2, phase 3 and NDA/BLA. An example calculation is illustrated in **Supplementary Figure 2**.

For purposes of this analysis, all indications that were advanced or suspended in any phase during our collection time frame were included. In practice, this means a drug that 'entered' the analysis in 2003 in phase 2, and later advanced to phase 3, was included in the study. This method was selected because there are relatively few drugs that entered development in phase 1 in the range of years analyzed and have subsequently progressed through final FDA review, and there is less disclosure of drugs in phase 1 development. Abrantes-Metz *et al.*[9] also used a similar method and stated, "We did it this way because the data set has very few drugs with complete information for all… phases." Drugs that remained in the same phase were censored, as were those that moved back a phase but were not suspended[9].

Table 3 Comparison of our study with previous drug development success rate studies

	This study (2013) all indications		This study (2013) lead indications		DiMasi *et al.*[6] lead indications		Kola *et al.*[8] lead indications		Abrantes-Metz *et al.*[9] lead indications	
	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA
Phase 1 to phase 2	64.5%	10.4%	66.5%	15.3%	71%	19%	68%	11%	80.7%	NA
Phase 2 to phase 3	32.4%	16.2%	39.5%	23.1%	45%	27%	38%	16%	57.7%	NA
Phase 3 to NDA/BLA	60.1%	50.0%	67.6%	58.4%	64%	60%	55%	42%	56.7%	NA
NDA/BLA to approval	83.2%	83.2%	86.4%	86.4%	93%	93%	77%	77%	NA	NA
LOA from phase 1[a]		10.4%		15.3%		19%		11%	26.4%[c]	NA
Number of drugs in sample advanced or suspended[b]	5,820		4,736		1,316		NA		2,328	
Dates of source data (duration)	2003–2011 (9 years)				1993–2009 (17 years)		1991–2000 (10 years)		1989–2002 (14 years)	
Number of companies	835				50		10		NA	

[a]Probability of FDA approval for drugs in phase 1 development. [b]Total number of transitions used to calculate the success rate (the *n* value noted in the text). [c]Abrantes-Metz, *et al.*[9] reported 26.4% from phase 1 to phase 3. If we were to conservatively apply the 83.2% NDA/BLA success rate found in this study, Abrantes-Metz would yield the highest LOA from phase 1 (21%). NA, data not available.

basis. Using the lead indication methodology to determine success rates, the scope of the challenge in oncology drug development would be dramatically underestimated.

The largest variation in success rates across disease groups was observed in phase 2. In **Table 5** all-indication phase 2 success rates ranged from 26.3% (for cardiovascular) to 45.9% (for infectious disease). In phase 3, all indication success rates ranged from 45.2% (for oncology) to 71.1% (for other). In contrast, phase 1 and NDA/BLA (As only one application, NDA or BLA, will be filed for any single indication, rates are given below for NDA/BLA.) filing success rates were more consistent across disease groups. All indication data from **Table 5** are charted in **Figure 2** to illustrate the large differences in phases 2 and 3 and LOA from phase 1 success rates across disease areas.

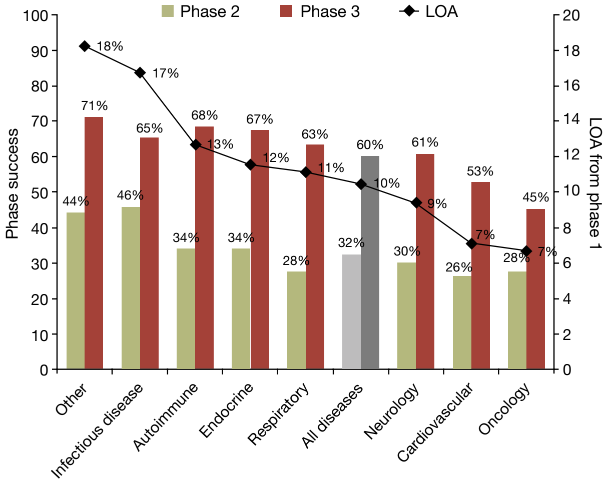


Figure 2 Phase success and LOA from phase 1 by disease for all indications. The bars represent phase 2 and phase 3 success rates and the line represents LOA from phase 1.

The development paths with the two lowest rates of phase 3 success were oncology and cardiovascular disease, with 45.2% (n = 221) and 52.8% (n = 89), respectively. **Figure 2** also highlights the large step-up in success rates from phase 2 to phase 3 for autoimmune, endocrine and respiratory diseases, increasing from 34% to 68%, 34% to 67%, and 28% to 63%, respectively. The low LOA from phase 1 in oncology rate results primarily from the lack of such a step-up, with a low phase 2 rate of 28.3% (n = 827), followed by a phase 3 success rate of only 45.2% (n = 221).

Success rates for oncology and non-oncology drugs. As oncology drugs made up the largest portion of the total data set (31.0% of all transitions) and had the lowest LOA from phase 1 (6.7%, n = 1,803), we investigated their contribution to success rates for the entire data set. To accomplish this, we removed all oncology drug development paths and compared these results to the full data set and oncology development paths alone. **Table 6** shows phase success and LOA rates for drugs for all disease groups, oncology and non-oncology development paths. The LOA from phase 1 across non-oncology indications is nearly twice that for oncology alone, 12.1% (n = 4,017) versus 6.7% (n = 1,803), respectively, reducing the probability of FDA approval in the full data set from nearly one in eight to over one in ten. Interestingly, the LOA from phase 1 for small-molecule NMEs was similar for oncology (6.6%, n = 1,163) and non-oncology (7.9% n = 2,333) indications, and biologics and non-NMEs accounted for much of the difference. For example, oncology biologics had a 7.3% (n = 429) LOA from phase 1 compared with 19.4% (n = 744) for non-oncology biologics.

Table 7 shows phase success and LOA rates in subcategories of cancer type for oncology drugs. Although a high number of transitions in all phases were seen for the solid tumor (n = 1,358) and hematological (n = 409) subgroups, further classification of oncology indications results in low numbers of transition from phase 3 to NDA/BLA. As is true of the full data set, drugs in phase 2 for oncology subgroups display more transitions and represent the strongest data for specific-indication success rate analysis. Oncology phase 2 success rates ranged from 50.0% (n = 12) in head and neck cancer to 20.9% (n = 24) in prostate cancer; however, the phase 2 rank order by tumor type was uncorrelated with LOA from phase 1 (linear regression, R^2 = 0.26). On average, phase 2 success rates were higher in hematological tumors (34.6%, n = 179) than in solid tumors (26.3%, n = 636). Only two phase 3 oncology indications had more than 20 transitions: breast cancer (n = 25) and non–small cell lung cancer (n = 23), which together accounted for ~28% of the solid tumor phase 3 transitions (n = 172). Because of even smaller sample sizes, cancer type success rates were not analyzed by lead indication.

Success rates for neurology, autoimmune and endocrine disease drugs. Neurology and autoimmune/immunology disease groups are

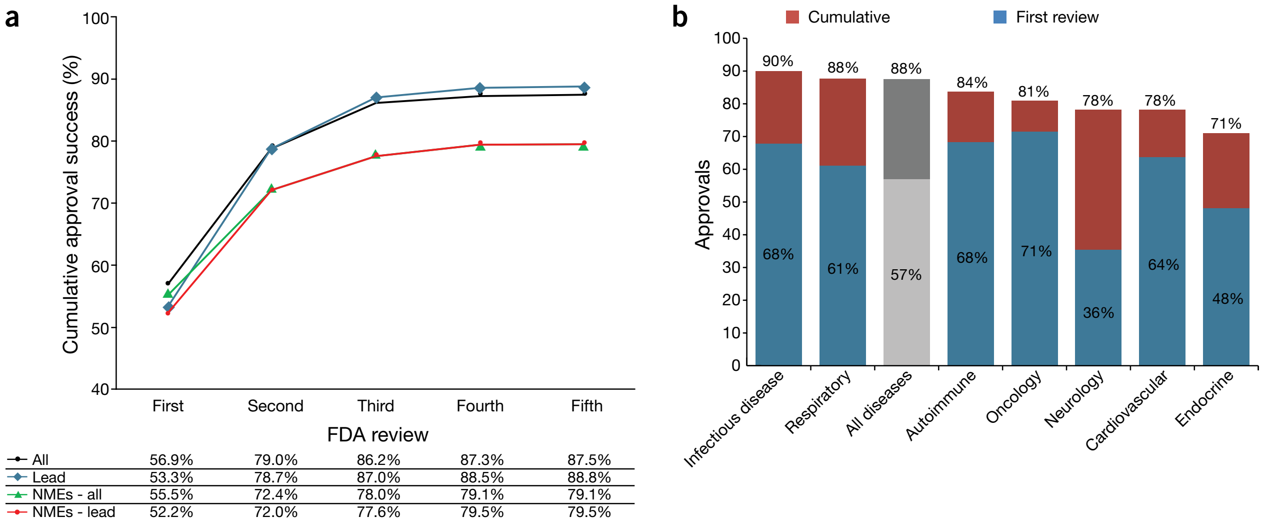


Figure 3 NDA/BLA success rates. (**a**) Cumulative approval rates by FDA review from 2005 to 2011 (914 reviews). (**b**) Cumulative and first FDA approval rates by disease.

well represented, comprising 17% and 9% of the data set, respectively. We subcategorized neurology into pain and psychiatric disorders, the two main therapeutic areas representing 51% of all neurology indications (**Table 8**). Analyzing all development paths, pain indications had a 10.7% (*n* = 231) LOA from phase 1 compared with 7.2% (*n* = 294) for psychiatric disorders. Other neurology indications, mainly representing neurodegenerative diseases, had a 9.8% (*n* = 452) LOA from phase 1.

An autoimmune subset analysis reveals that biologics had more than five times the LOA from phase 1 (22.5%, *n* = 288) than NMEs (5.2%, *n* = 202). **Table 8** also includes success rates for the type II diabetes and rheumatoid arthritis indication subcategories. Although rheumatoid arthritis had a 100% (*n* = 5) NDA/BLA submission success, the LOA from phase 1 was only 10.3% (*n* = 130) due to one of the lowest phase 2 success rates in this study (15.9%, *n* = 63). Diabetes also displayed lower-than-average success rates in all phases, except for NDA/BLA submissions, at 86.4% (*n* = 22).

Regulatory pathway success rates

To investigate the influence of regulation on clinical success we looked at two important pathways for drug oversight: the SPA and orphan drug designation.

SPA success rates. Similar to other analyses, we looked at phase success and LOA rates for drugs with an SPA (**Table 9**). Before

Table 4 Phase success and LOA by drug class

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]
FDA classification[e]																
All indications	2,541	1,918	64.5%	10.4%	3,743	2,268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
NMEs	1,585	1,218	64.2%	7.5%	2,375	1,470	28.6%	11.6%	831	515	53.2%	40.7%	425	293	76.5%	76.5%
Biologics	572	411	68.4%	14.6%	819	464	37.9%	21.3%	320	182	63.2%	56.1%	159	116	88.8%	88.8%
Non-NMEs	218	168	66.7%	20.0%	355	226	45.1%	29.9%	321	234	75.6%	66.3%	293	227	87.7%	87.7%
Lead indications	1,770	1,336	66.5%	15.3%	2,070	1,247	39.5%	23.1%	1,009	633	67.6%	58.4%	664	472	86.4%	86.4%
NMEs	1094	848	65.2%	12.0%	1,275	791	36.4%	18.3%	497	300	61.7%	50.3%	283	185	81.6%	81.6%
Biologics	362	257	75.1%	20.8%	403	216	44.0%	27.7%	182	106	71.7%	63.1%	106	75	88.0%	88.0%
Non-NMEs	167	124	66.9%	23.2%	232	153	49.0%	34.6%	254	186	79.0%	70.7%	246	189	89.4%	89.4%
Biomedtracker product category[f]																
Small molecule NMEs	1,335	1,033	65.4%	7.6%	2,053	1,283	29.0%	11.6%	725	449	52.3%	39.8%	369	264	76.1%	76.1%
Large molecules	912	658	65.8%	13.2%	1,279	714	37.7%	20.1%	511	296	60.1%	53.3%	244	166	88.6%	88.6%
mAbs	329	234	70.1%	14.1%	458	268	38.1%	20.1%	147	84	60.7%	52.7%	65	53	86.8%	86.8%
non-mAb proteins	192	151	58.9%	13.1%	280	170	35.3%	22.3%	150	87	69.0%	63.1%	93	59	91.5%	91.5%
Vaccines	121	57	67.1%	14.9%	160	79	44.3%	22.2%	67	34	50.0%	50.0%	23	20	100.0%	100.0%

[a] Number of indications identified. [b] Total number of transitions used to calculate the success rate, the *n* value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [c] Probability of successfully advancing to the next phase. [d] Probability of FDA approval for drugs in this phase of development. [e] FDA NME, biologic and non-NME classifications as defined in the results section. Data are presented for all and lead indication development paths. [f] BioMedTracker classification of small-molecule NMEs and large-molecule drugs. Large molecules are further stratified by biochemical profile.

Table 5 Phase success and LOA by disease[a]

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[b]	Advanced or suspended[c]	Phase success[d]	Phase LOA[e]	Total in phase[b]	Advanced or suspended[c]	Phase success[d]	Phase LOA[e]	Total in phase[b]	Advanced or suspended[c]	Phase success[c]	Phase LOA[e]	Total in phase[b]	Advanced or suspended[c]	Phase success[d]	Phase LOA[e]
All indications																
Other[f]	254	198	72.2%	18.2%	419	251	44.2%	25.3%	252	159	71.1%	57.1%	169	112	80.4%	80.4%
Infectious disease	247	196	65.8%	16.7%	288	157	45.9%	25.4%	159	98	65.3%	55.4%	115	86	84.9%	84.9%
Autoimmune	241	178	68.0%	12.7%	350	215	34.0%	18.7%	149	95	68.4%	55.0%	88	61	80.3%	80.3%
Endocrine	223	180	58.3%	11.6%	293	198	33.8%	19.8%	147	95	67.4%	58.5%	91	61	86.9%	86.9%
Respiratory	110	90	66.7%	11.1%	193	120	27.5%	16.7%	58	30	63.3%	60.8%	33	25	96.0%	96.0%
Neurology	389	298	62.4%	9.4%	520	348	30.2%	15.0%	285	188	60.6%	49.9%	192	152	82.2%	82.2%
Cardiovascular	158	127	60.6%	7.1%	229	152	26.3%	11.7%	121	89	52.8%	44.6%	78	58	84.5%	84.5%
Oncology	919	651	63.9%	6.7%	1,451	827	28.3%	10.5%	383	221	45.2%	37.0%	142	104	81.7%	81.7%
Total	2,541	1,918	64.5%	10.4%	3,743	2,268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
Lead indications																
Other[f]	193	146	75.3%	24.5%	273	157	50.3%	32.5%	174	115	74.8%	64.6%	122	81	86.4%	86.4%
Infectious disease	228	181	66.9%	19.3%	248	135	45.9%	28.8%	127	76	69.7%	62.8%	94	70	90.0%	90.0%
Respiratory	79	66	63.6%	16.3%	120	76	31.6%	25.6%	40	20	85.0%	81.0%	29	21	95.2%	95.2%
Autoimmune	165	127	67.7%	15.4%	178	102	37.3%	22.8%	77	52	80.8%	61.1%	56	37	75.7%	75.7%
Endocrine	188	152	61.2%	14.5%	226	155	38.1%	23.8%	122	78	69.2%	62.4%	78	51	90.2%	90.2%
Oncology	489	334	68.9%	13.2%	527	298	42.3%	19.1%	193	106	54.7%	45.3%	85	58	82.8%	82.8%
Neurology	301	228	62.7%	12.3%	339	218	34.4%	19.6%	191	124	66.9%	56.8%	137	106	84.9%	84.9%
Cardiovascular	127	102	62.7%	8.7%	159	106	27.4%	13.8%	85	62	56.5%	50.6%	63	48	89.6%	89.6%
Total	1,770	1,336	66.5%	15.3%	2,070	1,247	39.5%	23.1%	1,009	633	67.6%	58.4%	664	472	86.4%	86.4%

[a] Categories are listed from highest to lowest LOA from phase 1 for all indications (lead and nonlead). [b] Number of indications identified. [c] Total number of transitions used to calculate the success rate, the *n* value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [d] Probability of successfully advancing to the next phase. [e] Probability of FDA approval for drugs in this phase of development. [f] Includes allergy, gastroenterology, ophthalmology, dermatology, obstetrics/gynecology and urology.

initiating a pivotal phase 3 program, companies can submit the protocol to the FDA to obtain the agency's agreement that the trial(s) are adequate to meet its scientific and regulatory requirements. At the same time, these trials are often more complex and investigate treatments for less well understood diseases. This latter point is evident from our analysis: NDA/BLA success rates for SPA-designated drugs are slightly below average at 80.0% (n = 45) compared with 83.2% (n = 659) for all drugs. On the other hand, phase 3 success rates are nearly identical at 60.0% (n = 110) for SPA-designated drug indications compared with 60.1% (n = 975) for all drugs.

Orphan drug pathway success rates. A company may request that FDA grant the orphan designation for a drug being studied in a rare disease or condition. This is intended for indications affecting fewer than 200,000 people in the United States. Orphan drug designation was designed to reduce development costs and provide financial incentives (e.g., an extended exclusivity period) to encourage development in these indications. **Table 9** shows that although drugs for orphan indications have high rates of phase 1 and 2 success, phase 3 and NDA/BLA success rates are similar to all indications. Even so, it is important to note that orphan designations can be granted at any point in the clinical development process and are most often received when a drug is in phase 2. Orphan drugs in our data set received orphan status at all stages of development: preclinical (9%), phase 1 (22%), phase 2 (45%), phase 3 (16%) and NDA/BLA (2%). This distribution introduces a positive bias in early development success rates as some trials are not annotated as orphan until later phases. In contrast, by phase 3, 82% of indications that end up with the orphan designation have been annotated. Indeed, orphan indication phase 1 and 2 success rates were well above average at 86.8% (n = 136) and 70.0% (n = 190), respectively. Orphan phase 3 success rates (66.9%, n = 148) also compared favorably with all indications (60.1%, n = 975) and orphan NDA/BLA approvals were lower, 81.0% (n = 84) compared with 83.2% (n = 659), respectively. A subgroup analysis of phase 3 and NDA/BLA stage orphan drugs by indication reveals that oncology success rates were lower than non-oncology drugs, a result that is consistent with these categories in the full data set.

NDA/BLA success rates

To complement the NDA/BLA phase success rates gathered above, we examined 910 FDA decisions from 2005 to 2011 and classified each as 'Approved' or 'Not Approved.' In addition, we determined at which FDA review each decision occurred (i.e., the first, second, third, fourth or fifth time the agency reviewed the specific application). **Figure 3a** shows the cumulative success rates for NDA/BLA filings in the all, lead and NME drug classifications. Only 56.9% of all applications were approved on the first NDA/BLA submission, whereas 86.2% were approved by the third submission. After the third submission, there was only a marginal increase in the cumulative approval percentage, as there were few drugs with more than three regulatory reviews. For all NMEs, we found similar first submission success rates, yet fewer than 80% of these drugs were approved by FDA in subsequent submissions.

Analysis of first review approval success rates by disease reveals a variation inconsistent with cumulative approval rates. For example, **Figure 3b** shows that although oncology drugs had a median NDA/BLA success rate (81%), the chances of a first review approval were the highest, at 71%. Neurology drugs, on the other hand, had the lowest first review approval rate at 36%, but the cumulative approval rate reached 78%.

We also examined 304 first review FDA complete response letters and approvable letters issued for approved and suspended drugs. For approved drugs, 46% of the letters to the sponsor cited manufacturing or labeling issues and 47% cited efficacy or safety. In contrast, for suspended drugs, only 2% cited manufacturing or labeling issues and

Table 6 Phase success and LOA for oncology and non-oncology disease groups

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]
FDA classification[e]																
All indications	2,541	1,918	64.5%	10.4%	3,743	2,268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
Total oncology	919	651	63.9%	6.7%	1,451	827	28.3%	10.5%	383	221	45.2%	37.0%	142	104	81.7%	81.7%
Oncology NMEs	574	402	65.9%	6.6%	948	534	27.5%	10.0%	245	150	46.0%	36.4%	101	77	79.2%	79.2%
Oncology biologics	244	177	61.6%	7.3%	346	193	30.6%	11.9%	83	41	43.9%	39.0%	24	18	88.9%	88.9%
Oncology non-NMEs	53	39	69.2%	9.4%	76	50	22.0%	13.6%	26	17	70.6%	61.8%	16	8	87.5%	87.5%
Total non-oncology	1622	1267	64.8%	12.1%	2,292	1,441	34.8%	18.7%	1,171	754	64.5%	53.8%	766	555	83.4%	83.4%
Non-oncology NMEs	1011	816	63.4%	7.9%	1,427	936	29.3%	12.4%	586	365	56.2%	42.4%	324	216	75.5%	75.5%
Non-oncology biologics	328	234	73.5%	19.4%	473	271	43.2%	26.4%	237	141	68.8%	61.1%	135	98	88.8%	88.8%
Non-oncology non-NMEs	165	129	65.9%	22.7%	279	176	51.7%	34.5%	295	217	76.0%	66.7%	277	219	87.7%	87.7%
BioMedTracker product category[f]																
All indications	2,541	1,918	64.5%	10.4%	3,743	2,268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
Total oncology	919	651	63.9%	6.7%	1,451	827	28.3%	10.5%	383	221	45.2%	37.0%	142	104	81.7%	81.7%
Oncology small molecule NMEs	492	346	66.5%	7.2%	830	466	28.8%	10.9%	219	136	45.6%	37.8%	93	70	82.9%	82.9%
Oncology mAbs	175	125	68.0%	9.3%	245	140	29.3%	13.7%	55	30	50.0%	46.9%	21	16	93.8%	93.8%
Oncology proteins/peptides	68	50	48.0%	3.4%	108	57	31.6%	7.1%	34	16	37.5%	22.5%	8	5	60.0%	60.0%
Oncology vaccines	41	28	50.0%	1.6%	73	43	39.5%	3.3%	28	12	8.3%	8.3%	1	1	100.0%	100.0%
Total non-oncology	1622	1267	64.8%	12.1%	2,292	1,441	34.8%	18.7%	1,171	754	64.5%	53.8%	766	555	83.4%	83.4%
Non-oncology small molecule NMEs	843	687	64.9%	7.7%	1,223	817	29.1%	11.9%	506	313	55.3%	40.7%	276	194	73.7%	73.7%
Non-oncology mAbs	154	109	72.5%	19.3%	213	128	47.7%	26.6%	92	54	66.7%	55.9%	44	37	83.8%	83.8%
Non-oncology proteins/ peptides	228	178	65.7%	18.0%	321	198	42.4%	27.4%	191	118	69.5%	64.7%	125	72	93.1%	93.1%
Non-oncology vaccines	82	57	71.9%	21.8%	87	38	47.4%	30.3%	44	25	64.0%	64.0%	22	19	100.0%	100.0%

[a] Number of indications identified. [b] Total number of transitions used to calculate the success rate, the *n* value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [c] Probability of successfully advancing to the next phase [d] Probability of FDA approval for drugs in this phase of development. [e] Oncology and non-oncology disease groups and FDA NME, biologic, and non-NME classifications. Data are presented for all indication development paths. [f] Oncology and non-oncology disease groups and BioMedTracker biochemical categories.

83% cited efficacy or safety. Furthermore, we analyzed the time to drug approval after receiving a first complete response letter and found a 15-month average delay across all diseases with a setback of over one year for all diseases except (**Supplementary Fig. 1**) infectious disease (**Supplementary Fig. 2**).

Lead and nonlead indication success rates

Classifying drugs by lead and nonlead indications results in a selection bias favoring lead indication success rates. For lead indications that are suspended, and have a nonlead development path in-progress, the nonlead indication is redefined as the lead indication. The most advanced nonlead indications therefore becomes the lead indications once the initial lead is suspended. The BioMedTracker database is maintained as such for real-time viewing of pipelines, where it is critical to identify a company's lead program for each compound.

This lead indication annotation methodology tracks the most successful development path, and closely resembles the best case scenario for a specific drug. On the other hand, nonlead indication success rates understate the importance of lead indications that were previously designated as nonlead. Nonlead indication success rates are included in **Supplementary Table 2**, and, as expected, have a much lower success rate across all phases. For nonlead indications, the LOA from phase 1 was 4.9% ($n = 2{,}132$) compared with 15.3% ($n = 3{,}688$) for lead indications. The most pronounced deviation was found in phase 3, where lead indications had a 67.6% ($n = 633$) success rate, whereas nonlead indications had a 46.2% ($n = 342$) probability of advancing to NDA/BLA. The disparity between lead and nonlead success rates is noteworthy, and the accuracy of nonlead rates must be viewed in the context of the selection methodology.

DISCUSSION

During the time frame of this study, approximately one development path in ten (10.4%) that enters clinical development in phase 1 is expected to advance to FDA approval. We also analyzed lead indication data and found nearly a one-in-six (15.3%) probability a drug will advance from phase 1 to FDA approval. We believe that the lower success rate for all-indication development paths more accurately reflects drug development success rates in industry and is particularly important when considering the cost and time of unsuccessful clinical trials.

One limitation of this study is the direct comparison of these data and methodology on a year-by-year or decade-by-decade basis. For example, a program was designated as

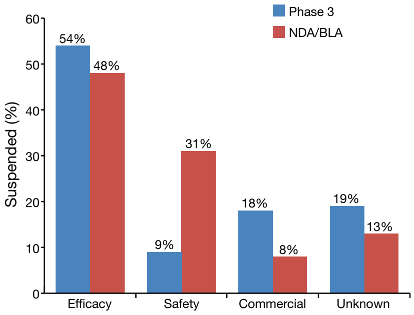


Figure 4 Root-cause analysis for 359 phase 3 and 95 NDA/BLA suspended programs. A program was designated as 'suspended' when conclusive evidence had been gathered regarding a company's plans to discontinue development or communications with regulators were not reinitiated for several years.

'suspended' when conclusive evidence had been gathered regarding a company's plans to discontinue development, or communications with regulators were not reinitiated for several years. Unfortunately, the timing of annotating suspended indications and drugs is not precise enough to analyze yearly changes in success rates. Furthermore, real-time data collection was initiated in 2003; thus, we cannot directly compare prior decades using these data and must rely on results published in the literature.

Many previous studies considered only a drug's most advanced indication to determine drug development success rates. Most published data from the 1960s to present reported success rates ranging from one in five to one in eight[14–19]. For comparison with more recent findings, we summarize in **Table 3** the results from DiMasi *et al.*[6], Kola *et al.*[8] and Abrantes-Metz *et al.*[9]. The most recent publication on the subject, from DiMasi *et al.*[6], reports a nearly one-in-five LOA from phase 1 (19%, n = 1,316) from 1993 to 2009. In Kola *et al.*[8], the authors found an LOA from phase 1 of 11%, close to the 10.4% reported here for all indications. However, given the small number of company pipelines (10 versus 835 reported here) and lack of information about the number of drugs advanced or suspended in this study, these results were inconclusive. In addition, the Abrantes-Metz *et al.*[9] data covered a similar period as Kola *et al.*[8], 1989 to 2002 versus 1991 to 2000, respectively, but did not report NDA/BLA success rates. If we were to conservatively apply the 83.2% NDA/BLA success rate found in this study, Abrantes-Metz *et al.*[9] would yield the highest LOA from phase 1 (21%), again near one in five.

Comparing the phase transitions, phase 2 success rates were consistently lower than phase 1, with phase 1 ranging from 65% to 81%, and phase 2 from 32% to 58%. In this study, and in DiMasi *et al.*[6] and Kola *et al.*[8], a step-up in phase 3 success rates from phase 2 rates was observed. Only Abrantes-Metz *et al.*[9] reported a phase 2 success rate (57.7%) in-line with phase 3 (56.7%), a result that was 20 percentage points higher than the phase 2 success rate in Kola *et al.*[8] (38%) for a similar time period (**Table 3**). There are fewer data available to compare NDA/BLA success rates, but our result of 83.2% is similar to that of Kola *et al.*[8] (77%) and 10% lower than that of DiMasi *et al.*[6].

For lead indication success rates, our results are similar to that found by DiMasi *et al.*[6]. Although our LOA from phase 1 for lead indications (15.3%) is below DiMasi *et al.*'s[6] 19% result, it is close to their 16% result for self-originated drugs. We also note that the 16% success rate for self-originated drugs held over multiple time frames (1993–1998 and 1999–2004) in their studies. One possible explanation is that success rates for self-originated drugs at large pharmaceutical companies are less prone to selection bias compared with late-stage, in-licensed drugs.

Factors contributing to lower success rates found in this study include the large number of small biotech companies represented in the data, more recent time frame (2003–2011) and higher regulatory hurdles for new drugs. Small biotech companies tend to develop riskier, less validated drug classes and targets, and are more likely to have less experienced

Table 7 Phase success and LOA for oncology subgroups and cancer types

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]
All indications	2,541	1,918	64.5%	10.4%	3743	2268	32.4%	16.2%	1554	975	60.1%	50.0%	908	659	83.2%	83.2%
Total oncology	919	651	63.9%	5.4%	1451	827	28.3%	8.5%	383	147	36.7%	30.0%	142	104	81.7%	81.7%
Total solid tumors	668	483	66.7%	5.7%	1114	636	26.3%	8.6%	299	172	41.3%	32.7%	88	67	79.1%	79.1%
Renal cell cancer (RCC)	20	15	86.7%	18.4%	54	33	30.3%	21.2%	15	10	70.0%	70.0%	7	6	100.0%	100.0%
Head and neck cancer	6	5	100.0%	14.3%	23	12	50.0%	14.3%	14	7	42.9%	28.6%	3	3	66.7%	66.7%
Hepatocellular (liver) cancer (HCC)	18	15	73.3%	6.6%	39	25	36.0%	9.0%	12	4	25.0%	25.0%	1	1	100.0%	100.0%
Breast cancer	54	47	68.1%	5.7%	119	61	21.3%	8.4%	34	25	56.0%	39.2%	14	10	70.0%	70.0%
Non-small cell lung cancer (NSCLC)	63	55	87.3%	5.7%	161	94	29.8%	6.5%	46	23	26.1%	21.7%	11	6	83.3%	83.3%
Prostate cancer	42	8	71.0%	5.6%	103	24	20.9%	7.8%	25	8	56.3%	37.5%	11	3	66.7%	66.7%
Colorectal cancer (CRC)	45	37	62.2%	5.1%	87	56	21.4%	8.2%	18	13	38.5%	38.5%	4	4	100.0%	100.0%
Ovarian cancer	31	25	68.0%	4.6%	72	37	27.0%	6.8%	15	8	25.0%	25.0%	3	1	100.0%	100.0%
Pancreatic cancer	29	24	75.0%	2.3%	66	36	30.6%	3.1%	19	10	20.0%	10.0%	2	2	50.0%	50.0%
Total hematological tumors	216	152	58.6%	9.9%	317	179	34.6%	16.9%	78	45	55.6%	48.8%	48	33	87.9%	87.9%
Multiple myeloma (MM)	43	29	69.0%	9.7%	48	30	23.3%	14.0%	13	5	60.0%	60.0%	5	4	100.0%	100.0%
Non-Hodgkin's lymphoma (NHL)	38	28	57.1%	8.5%	62	35	40.0%	14.8%	19	9	44.4%	37.0%	8	6	83.3%	83.3%
Chronic lymphocytic leukemia (CLL)	17	12	50.0%	7.3%	41	24	29.2%	14.6%	10	8	62.5%	50.0%	7	5	80.0%	80.0%
Myelodysplastic syndrome (MDS)	12	7	71.4%	4.8%	22	9	33.3%	6.7%	6	5	20.0%	20.0%	4	3	100.0%	100.0%

[a]Number of indications identified. [b]Total number of transitions used to calculate the success rate, the n value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [c]Probability of successfully advancing to the next phase· [d]Probability of FDA approval for drugs in this phase of development.

Table 8 Phase success and LOA for neurology and autoimmune diseases (broken further into rheumatoid arthritis and type II diabetes)

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]
All indications	2,541	1,918	64.5%	10.4%	3,743	2,268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
Total neurology	389	298	62.4%	9.4%	520	348	30.2%	15.0%	285	188	60.6%	49.9%	192	152	82.2%	82.2%
Psychiatric disease	97	80	60.0%	7.2%	148	116	23.3%	12.0%	83	49	63.3%	51.6%	57	49	81.6%	81.6%
Pain	96	73	67.1%	10.7%	113	79	27.8%	15.9%	67	46	67.4%	57.2%	42	33	84.8%	84.8%
Other	196	136	58.8%	9.8%	259	153	36.6%	16.7%	135	93	55.9%	45.5%	93	70	81.4%	81.4%
Total autoimmune disease	241	178	68.0%	12.7%	350	215	34.0%	18.7%	149	95	68.4%	55.0%	88	61	80.3%	80.3%
Total autoimmune disease NMEs	111	88	62.5%	5.2%	151	86	22.1%	8.3%	38	20	50.0%	37.5%	16	8	75.0%	75.0%
Total autoimmune disease biologics	116	80	73.8%	22.5%	171	111	45.0%	30.5%	89	56	75.0%	67.7%	53	41	90.2%	90.2%
Total autoimmune disease non-NMEs	10	8	87.5%	7.9%	22	16	25.0%	9.0%	21	18	72.2%	36.1%	18	12	50.0%	50.0%
Total rheumatoid arthritis	65	54	74.1%	10.3%	102	63	15.9%	13.9%	18	8	87.5%	87.5%	10	5	100.0%	100.0%
Rheumatoid arthritis NMEs	30	29	69.0%	NA	46	29	10.3%	NA	4	1	100.0%	NA	2	0	NA	NA
Rheumatoid arthritis biologics	32	24	79.2%	15.9%	49	29	24.1%	20.1%	13	6	83.3%	83.3%	7	5	100.0%	100.0%
Total type II diabetes	110	89	60.7%	9.3%	128	84	29.8%	15.3%	53	37	59.5%	51.4%	31	22	86.4%	86.4%
Diabetes NMEs	83	68	63.2%	7.5%	100	69	29.0%	11.8%	35	25	56.0%	40.7%	15	11	72.7%	72.7%

[a]Number of indications identified. [b]Number of transitions used to calculate the success rate, the *n* value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [c]Probability of successfully advancing to the next phase. [d]Probability of FDA approval for drugs in this phase of development. NA, data not available.

development teams and fewer resources than large pharmaceutical corporations. The past nine-year period has been a time of increased clinical trial cost and complexity for all drug development sponsors, and this likely contributes to the lower success rates than previous periods. In addition, an increasing number of diseases have higher scientific and regulatory hurdles as the standard of care has improved over the past decade. More clinical studies are comparative in nature and published data show clinical trials are more complex today than in previous decades[7]. The time frame in this study also coincides with the shift toward greater regulatory uncertainty and stronger emphasis on safety at the FDA since the 2004 Vioxx (rofecoxib) recall. For smaller companies, financing challenges in the past several years have also affected development progression decisions. Phase success rates reported in this study are based on transition rates, not necessarily resulting from safety or efficacy data. Transition rates are negatively affected by early development termination due to commercial and regulatory uncertainty as well as economic and portfolio management decisions.

Lower success rates found when analyzing all indications likely results from including nonlead and/or secondary indications. Nonlead development paths have far lower success rates compared with lead programs. One possible explanation is that many companies first develop drugs in lead indications where the strongest scientific rationale and early efficacy signals are found. Lead indications are also often smaller, better-defined patient populations. After initial success in these populations, companies may decide to investigate nonlead indications, which may not have the same scientific support, homogenous patient population or development and regulatory path as the lead indication. Nonlead success rates are also important to monitor as many of these indications can be moved directly into late-stage trials, where most clinical development costs occur. Furthermore, our research suggests that these late-stage trials for nonlead indications often enroll a greater number of patients than lead indications.

Phase 3 success rates. In **Figures 1** and **2**, we show that phase 3 success rates are 60% for drugs for all indications, but only around 50% in oncology or cardiology. Such low phase 3 success rates for these diseases are concerning as 35% of all R&D spending is now spent on phase 3 development, and phase 3 trials account for 60% of all clinical trial costs[3]. Some of the low phase 3 rates may be attributed to trial design factors and insufficient communication between sponsors and regulators during their end-of-phase-2 meetings. Both oncology and cardiology, for example, now require outcome studies looking for improved overall survival, but lack well-validated surrogate markers for this outcome. On the other hand, disease areas with validated surrogate markers tend to have higher phase 3 success rates. For example, studies of infectious diseases such as hepatitis C and HIV that use viral load as a primary endpoint as well as glycosylated hemoglobin (HbA1c) in diabetes show higher success rates.

Oncology is a particularly challenging disease area in which to achieve phase 3 success. The FDA requires overall survival as the primary endpoint in most pivotal oncology studies. Crossover designs that allow patients who progress on the comparator arm to cross over and receive the investigational drug, or patients receiving additional approved and experimental salvage therapies, also make it more difficult to design well-controlled phase 3 studies with overall survival as a primary endpoint. Furthermore, current animal models (e.g., xenograft tumor models in mice) can be poor predictors of clinical outcomes in humans. Additionally, recent scientific reports show that certain types of cancer, which were previously thought of as one disease, may actually comprise several subtypes of disease with different etiologies. For example, NSCLC is now considered by many oncologists to be at least ten different mutation-specific diseases, and thus it is not surprising that drugs for NSCLC have one of the lowest LOAs from phase 1 of all oncology indications in **Table 7** (ref. 20).

Table 9 Phase success and LOA for drugs receiving a FDA SPA or orphan drug designation

	Phase 1 to phase 2				Phase 2 to phase 3				Phase 3 to NDA/BLA				NDA/BLA to approval			
	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]	Total in phase[a]	Advanced or suspended[b]	Phase success[c]	Phase LOA[d]
All indications	2,541	1,918	64.5%	10.4%	3,743	2268	32.4%	16.2%	1,554	975	60.1%	50.0%	908	659	83.2%	83.2%
Total SPAs	42	35	97.1%	45.4%	128	115	97.4%	46.7%	171	110	60.0%	48.0%	73	45	80.0%	80.0%
Total orphans	170	136	86.8%	32.9%	328	190	70.0%	37.9%	237	148	66.9%	54.2%	136	84	81.0%	81.0%
Orphan oncology	85	67	85.1%	23.0%	176	105	61.0%	27.1%	102	63	58.7%	44.4%	54	37	75.7%	75.7%
Orphan non-oncology	85	69	88.4%	44.5%	152	85	81.2%	50.4%	135	85	72.9%	62.1%	82	47	85.1%	85.1%

[a] Number of indications identified. [b] Total number of transitions used to calculate the success rate, the *n* value noted in the text. The difference between 'Total in phase' and 'Advanced or suspended' is the number of indications that remain in development. [c] Probability of successfully advancing to the next phase. [d] Probability of FDA approval for drugs in this phase of development.

Clinical trials targeting heterogeneous patient populations may have lower success rates than trials identifying responders within a population through the use of biomarkers. As predictability of clinical outcomes increases through the use of molecular diagnostics in earlier testing, it is possible that phase 3 trial success rates will rise. Furthermore, the adoption of adaptive trial design may facilitate the identification of targeted subsets of patient populations before study completion. According to the FDA's draft guidance for industry, issued in February 2010, adaptive trial design may make clinical studies more efficient (e.g., shorter duration and fewer patients), more likely to demonstrate an effect of the drug or more informative (e.g., providing broader dose-response information)[21].

Root causes of phase 3 and NDA/BLA development failures. To gain a better understanding of the causes that lead companies to discontinue drug development, we further analyzed publically available information for the 359 phase 3 and 95 NDA/BLA suspensions included in this study. We classified each discontinued development program into four categories based on the primary reason for suspension including: efficacy, safety, commercial and unknown (**Fig. 4**).

Although it was difficult to objectively determine if a phase 3 study did not reach an endpoint due to poor study design or the drug's biological activity, we found that over half of the 359 suspensions were attributable to some measure of efficacy. Indeed, a detailed analysis of the specific inputs, rationale and history for each program would be needed to identify issues related to poor trial design. Furthermore, public information is not available to assess the degree of communication with regulators, adherence to recommendations, changes to prior standards and input from phase 2 data that would inform the design of a phase 3 study.

We found that 18% of the phase 3 suspensions resulted from a company's commercial decision to not file for approval. We do not know the degree to which regulatory uncertainty factored into these decisions, but recognize its important impact on portfolio management, funding and commercial opportunities due to the increased time and costs of drug development.

Safety was the least likely cause for suspension in phase 3 (9%), perhaps due to significant adverse events identified earlier in drug development. Approximately 20% of the suspensions occurred without publicly available information citing the reason for failure.

We also analyzed the 95 suspended NDA/BLA filings in the data set and found that approximately one-third of failures were attributable to safety concerns raised by regulators compared to only 9% in phase 3. Our analysis also revealed that around half involved cases where the FDA requested additional trials. One interpretation of these data is that sponsors file for regulatory approval believing their drug meets safety guidelines, whereas regulators remain concerned about safety, illustrating insufficient communication between regulators and sponsors. During the period of this study, mainly after the 2004 Vioxx recall, many industry observers have discussed how the benefit-to-risk pendulum has swung toward risk, with a greater focus on safety in the regulatory assessment. Some examples of issues brought forward by regulators were the need for longer-term data, inclusion of additional study arms, inclusion of different patient age and at-risk populations, and increases in the number of patients studied.

Further analysis of failures by lead or non-lead indication, disease, modality and company type were not performed because the small sample size has limitations and subjects the results to molecular and therapeutic class–specific issues. Future studies will allow us to identify trends in failed clinical programs as the sample size becomes more reliable.

Conclusions

The data presented in this study suggest industry-wide productivity may have declined from previous estimates. Achieving FDA approval for only one-in-ten drug indications that enter the clinic is a concerning statistic for drug developers, regulators, investors and patients. We believe progress in clinical science and regulatory risk-benefit assessment can improve success rates. Greater flexibility with alternative surrogate endpoints, the utilization of adaptive clinical trial design and improved methodologies for assessing patient benefit-to-risk are some areas where improvements can be made. In addition, improvements in communication between sponsors and regulators could help reduce regulatory applications that lack safety or efficacy data that are later requested by regulators. Simultaneously, improvements in basic science can enable improvements in success rates. For example, more predictive animal models, earlier toxicology evaluation, biomarker identification and new targeted delivery technologies may increase future success in the clinic.

ACKNOWLEDGMENTS

This research was supported by Sagient Research Systems and the Biotechnology Industry Organization (BIO). The authors are indebted to the analysts at BioMedTracker who have collected these data. We would also like to thank our colleagues at BIO and BioMedTracker for their input and advice in the development of this study. The authors are solely responsible for the design, conduct and analysis of the study, and the conclusions that are drawn. The opinions expressed in this article are those of the authors and do not necessarily reflect the views of their employers.

AUTHOR CONTRIBUTIONS

M.H., D.W.T. and J.L.C. all contributed equally to this work.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.2786).

1. Lloyd, I., ed. Citeline Drug Intelligence. *Pharma R&D Annual Review 2011.* http://www.citeline.com/

wp-content/uploads/Citleine-Pharma-RD-annual-review-20111.pdf (Citeline Drug Intelligence, 2012).

2. EvaluatePharma. *World Preview 2018: Embracing the Patent Cliff.* http://info.evaluatepharma.com/WP2018_ELS_LP.html (2012).
3. Pharmaceutical Research and Manufacturers of America. *Annual Report 2011.* http://www.phrma.org/sites/default/files/159/phrma_2011_annual_report.pdf (2011).
4. Mullard, A. 2012 FDA drug approvals. *Nat. Rev. Drug Discov.* **12**, 87–90 (2013).
5. Cohen, F.J. Macro trends in pharmaceutical innovation. *Nat. Rev. Drug Discov.* **4**, 78–84 (2005).
6. DiMasi, J.A., Feldman, L., Seckler, A. & Wilson, A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.* **87**, 272–277 (2010).
7. Kaitin, K.I. & DiMasi, J.A. Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000–2009. *Clin. Pharmacol. Ther.* **89**, 183–188 (2011).
8. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
9. Abrantes-Metz, R., Adams, C. & Metz, A. *Pharmaceutical Development Phases: A Duration Analysis.* Working paper no. 274. http://www.ftc.gov/be/workpapers/wp274.pdf (US Federal Trade Commission: Bureau of Economics, 2004).
10. Henderson, R. & Cockburn, I. Scale, scope, and spillovers: the determinants of research productivity in drug discovery. *Rand J. Econ.* **27**, 32–59 (1996).
11. Cockburn, I.M. & Henderson, R.M. Scale and scope in drug development: unpacking the advantages of size in pharmaceutical research. *J. Health Econ.* **20**, 1033–1057 (2001).
12. Danzon, P.M., Nicholson, S. & Pereira, N.S. Productivity in pharmaceutical–biotechnology R&D: the role of experience and alliances. *J. Health Econ.* **24**, 317–339 (2005).
13. Arora, A., Gambardella, A., Magazzini, L. & Pammolli, F. A breath of fresh air? Firm type, scale, scope, and selection effects in drug development. *Manage. Sci.* **55**, 1638–1653 (2009).
14. Sheck, L. *et al.* Success rates in the United States drug development system. *Clin. Pharmacol. Ther.* **36**, 574–583 (1984).
15. Tucker, S.A., Blozan, C. & Coppinger, P. The Outcome of Research on New Molecular Entities Commencing Clinical Research in the Years 1976–79 (OPE Study 77). (Office of Planning and Evaluation, US Food and Drug Administration, Rockville, MD, 1988).
16. DiMasi, J.A. Success rates for new drugs entering clinical testing in the United States. *Clin. Pharmacol. Ther.* **58**, 1–14 (1995).
17. DiMasi, J.A. Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* **69**, 297–307 (2001).
18. DiMasi, J.A., Hansen, R.W. & Grabowski, H.G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22**, 151–185 (2003).
19. DiMasi, J.A. & Grabowski, H.G. The cost of biopharmaceutical R&D: is biotech different? *Manag. Decis. Econ.* **28**, 469–479 (2007).
20. Edelman, M.J. in *2010 American Society of Clinical Oncology (ASCO) Annual Meeting* (Chicago, IL; 2010).
21. FDA. *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics.* http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm201790.pdf (FDA, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, 2010).

Journal of Health Economics 47 (2016) 20–33

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Innovation in the pharmaceutical industry: New estimates of R&D costs☆

Joseph A. DiMasi[a,*], Henry G. Grabowski[b], Ronald W. Hansen[c]

[a] Tufts Center for the Study of Drug Development, Tufts University, United States
[b] Department of Economics, Duke University, United States
[c] Simon Business School, University of Rochester, United States

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form 28 January 2016
Accepted 29 January 2016
Available online 12 February 2016

JEL classification:
L65
O31

Keywords:
Innovation
R&D cost
Pharmaceutical industry
Discount rate
Technical success rates

ABSTRACT

The research and development costs of 106 randomly selected new drugs were obtained from a survey of 10 pharmaceutical firms. These data were used to estimate the average pre-tax cost of new drug and biologics development. The costs of compounds abandoned during testing were linked to the costs of compounds that obtained marketing approval. The estimated average out-of-pocket cost per approved new compound is $1395 million (2013 dollars). Capitalizing out-of-pocket costs to the point of marketing approval at a real discount rate of 10.5% yields a total pre-approval cost estimate of $2588 million (2013 dollars). When compared to the results of the previous study in this series, total capitalized costs were shown to have increased at an annual rate of 8.5% above general price inflation. Adding an estimate of post-approval R&D costs increases the cost estimate to $2870 million (2013 dollars).

1. Introduction

We provide an updated assessment of the value of the resources expended by industry to discover and develop new drugs and biologics, and the extent to which these private sector costs have changed over time. The costs required to develop these new products clearly play a role in the incentives to invest in the innovative activities that can generate medical innovation. Our prior studies also have been used by other researchers, including government agencies, to analyze various policy questions (US Congressional Budget Office, 1998, 2006).

The full social costs of discovering and developing new compounds will include these private sector costs, but will also include government-funded and non-profit expenditures on basic and clinical research that can result in leads and targets which drug developers can explore. These additional costs can be substantial.[1] However, it is difficult to identify and measure non-private expenditures that can be linked to specific new therapies. Thus, we focus here on the private sector costs.

The methodological approach used in this paper follows that used for our previous studies, although we apply additional statistical tests to the data (Hansen, 1979; DiMasi et al., 1991, 1995a,b, 2003, 2004; DiMasi and Grabowski, 2007). Because the methodologies are consistent, we can confidently make comparisons of the results in this study to the estimates we found for the earlier studies, which covered earlier periods, to examine and illustrate trends

☆ We thank the surveyed firms for providing data, and individuals in those firms who kindly gave their time when we needed some of the responses clarified. All errors and omissions are the responsibility of the authors. The Tufts Center for the Study of Drug development (CSDD) is funded in part by unrestricted grants from pharmaceutical and biotechnology firms, as well as companies that provide related services (e.g., contract research, consulting, and technology firms) to the research-based industry. Tufts CSDD's financial disclosure statement can be found here: http://csdd.tufts.edu/about/financial_disclosure. The authors and Tufts CSDD did not receive any external funding to conduct this study. The R&D cost and expenditure data for individual compounds and companies are proprietary and cannot be redistributed. Other data used were obtained from subscription databases and the Food and Drug Administration (FDA) and other websites.

* Corresponding author at: Tufts Center for the Study of Drug Development, Tufts University, 75 Kneeland Street, Suite 1100, Boston, MA 02111, United States. Tel.: +1 617 636 2116; fax: +1 6176362425.

E-mail address: joseph.dimasi@tufts.edu (J.A. DiMasi).

[1] For example, for fiscal year 2013, the United States National Institutes of Health (NIH) spent nearly $30 billion on the activities that it funds (http://officeofbudget.od.nih.gov/pdfs/FY15/Approp%20%20History%20by%20IC%20through%20FY%202013.pdf).

http://dx.doi.org/10.1016/j.jhealeco.2016.01.012
0167-6296/© 2016 Elsevier B.V. All rights reserved.

in development costs. These studies used compound-level data on the cost and timing of development for a random sample of new drugs first investigated in humans and annual company pharmaceutical R&D expenditures obtained through surveys of a number pharmaceutical firms.

We analyze private sector R&D activities as long-term investments. The industrial R&D process is marked by substantial financial risks, with expenditures incurred for many development projects that fail to result in a marketed product. Thus, our approach explicitly links the costs of unsuccessful projects to those that are successful in obtaining marketing approval from regulatory authorities. In addition, the pharmaceutical R&D process is very lengthy, often lasting a decade or more (DiMasi et al., 2003). This makes it essential to model accurately how development expenses are spread over time.

Given our focus on resource costs and how they have changed over time, we develop estimates of the average pre-tax cost of new drug development and compare them to estimates covering prior periods. We corroborated the basic R&D cost results in this study by examining the representativeness of our sample firms and our study data, and by incorporating a number of independently derived results and data relating to the industry and the drug development process into analyses that provide rough comparators for at least components of our cost results. The details of those analyses are provided in our online supplement.

The remainder of this paper is organized as follows. We briefly discuss the literature on pharmaceutical industry R&D costs since our 2003 study in Section 2. Section 3 briefly outlines the standard paradigm for the drug development process. In Section 4 we describe the survey sample data and the population from which they were drawn, and briefly outline the methodology used to derive full R&D cost estimates from data on various elements of the drug development process. We present base case pre- and post-marketing approval R&D cost estimates in Section 5. Sensitivity analyses are presented in Section 6. We describe the representativeness of our data, various approaches to validating our results, and responses to various critiques in Section 7. Finally, we summarize our findings in Section 8.

2. Previous studies of the cost of pharmaceutical innovation

Much of the literature on the cost of pharmaceutical innovation dating back decades has already been described by the authors in their previous two studies (DiMasi et al., 1991, 2003). The interested reader can find references and discussions about the prior research in those studies. The earliest studies often involved a case study of a single drug (typically without accounting for the cost of failed projects) or they analyzed aggregate data. We will focus here on studies and reports that have emerged since DiMasi et al. (2003) that involve the use of new data for at least some parts of the R&D process. The basic elements of these analyses are shown in Table 1.

Adams and Brantner (2006, 2010) sought to assess the validity of the results in DiMasi et al. (2003) with some alternative data. Specifically, in their 2006 article, they used a commercial pipeline database to separately estimate clinical approval and phase attrition rates, as well as phase development times.[2] They found a similar overall cost estimate ($868 million versus $802 million in year 2000 dollars).[3] The authors followed that study with another study that featured clinical phase out-of-pocket cost estimates derived from regressions based on publicly available data on company R&D expenditures (Adams and Brantner, 2010). They found a somewhat higher overall cost estimate ($1.2 billion in year 2000 dollars).[4]

In a paper authored by two of the authors of this study (DiMasi and Grabowski, 2007), we provided a first look at the costs of developing biotech products (specifically, recombinant proteins and monoclonal antibodies). The methodological approach was the same as that used for our studies of traditional drug development. We used some data from DiMasi et al. (2003) combined with new data on the costs of a set of biotech compounds from a single large biopharmaceutical company. Biotech drugs were observed to have a higher average clinical success rate than small molecule drugs, but this was largely offset by other cost components. We found that the full capitalized cost per approved new compound was similar for traditional and biotech development ($1.3 billion for biotech and $1.2 billion for traditional development in year 2005 dollars), after adjustments to compare similar periods for R&D expenditures.

The other studies shown in Table 1 are discussed in detail in the online supplement. One important finding emerging from the survey of cost studies in Table 1 is that clinical success rates are substantially lower for the studies focused on more recent periods. This observed trend is consistent with other analyses of success probabilities (DiMasi et al., 2010; DiMasi et al., 2013; Hay et al., 2014; Paul et al., 2010) and our analysis below. Average R&D (inflation-adjusted) cost estimates are also higher for studies focused on more recent periods, suggesting a growth in real R&D costs. While suggestive, these studies are not strictly comparable to our earlier analyses of R&D costs given methodological differences and data omissions that are discussed in the online supplement (Appendix A).

3. The new drug development process

The new drug development process need not follow a fixed pattern, but a standard paradigm has evolved that fits the process well in general. We have described the process in some detail in previous studies, and the FDA's website contains a schematic explaining the usual set of steps along the way from test tube to new compound approval (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/SmallBusinessAssistance/ucm053131.htm). Marketing approval applications for investigational compounds submitted to the FDA for review by manufacturers are referred to as new drug applications (NDAs) or biologic license applications (BLAs), depending on the type of product.

In basic form, the paradigm portrays new drug discovery and development as proceeding along a sequence of phases and activities (some of which often overlap). Basic and applied research initiate the process with discovery programs that result in the synthesis or isolation of compounds that are tested in assays and animal models in preclinical development. We do not have the level

[2] For mean out-of-pocket phase costs, they used the estimates in DiMasi et al. (2003).

[3] The Adams and Brantner (2006) study used records in the pipeline database that were reported to have entered some clinical testing phase from 1989 to 2002. Thus, they did not follow the same set of drugs through time. The data for the commercial pipeline databases are also thin prior to the mid-1990s. The DiMasi et al. (2003) study covered new drugs that had first entered clinical testing anywhere in the world from 1983 to 1994 and followed the same set of drugs through time.

[4] However, the authors interpreted their estimate as a marginal, as opposed to an average, drug cost. The concept, though, of marginal cost has an unclear meaning here. With high fixed costs and a development process that varies by drug, it is difficult to understand what marginal pharmaceutical R&D cost means in this context. It seems that the relevant marginal concept here is marginal profitability. The marginally profitable drug could have a very high or a very low cost. What's more, marginal profitability may only have meaning at the firm, not the industry, level. The cost of a marginally profitable drug in the pipeline of a firm may be high for one firm and low for another firm.

Table 1
Prior studies and analyses of pharmaceutical R&D costs (2003–2012).

Study	Study period	Clinical success rate	Real cost of capital	Inflation adjustment	Cost estimate
DiMasi et al. (2003)	First-in-humans, 1983–1994	21.5%	11.0%	2000 dollars	$802 million
Adams and Brantner (2006)	First-in-humans, 1989–2002	24.0%	11.0%	2000 dollars	$868 million
Adams and Brantner (2010)	Company R&D expenditures, 1985–2001	24.0%	11.0%	2000 dollars	$1.2 billion
DiMasi and Grabowski (2007)	First-in-humans, 1990–2003 (large molecule)	30.2% (large molecule)	11.5%	2005 dollars	$1.2 billion
Gilbert et al. (2003)	2000–2002 (launch)	8.0%	NA	2003 dollars	$1.7 billion
O'Hagan and Farkas (2009)	2009 (launch)	NA	NA	2009 dollars	$2.2 billion
Paul et al. (2010)	≈2007	11.7%	11.0%	2008 dollars	$1.8 billion
Mestre-Ferrandiz et al. (2012)	In clinical development, 1997–1999	10.7%	11.0%	2011 dollars	$1.5 billion

of granularity to disaggregate R&D expenditure data into discovery and preclinical development testing costs, so for the purposes of this study, as in prior studies, discovery and preclinical development costs are grouped and referred to as pre-human costs.[5]

Clinical (human) testing typically proceeds through three successive, sometimes overlapping phases. Historically, human testing has often been initiated first outside the United States (DiMasi, 2001). For any of these clinical phases, pharmaceutical companies may pursue development of their investigational compounds in multiple indications prior to and/or after the initial indication approval.

4. Data and methods

Ten multinational pharmaceutical firms of varying sizes provided data through a confidential survey of their new drug and biologics R&D costs.[6] Data were collected on clinical phase expenditures and development phase times for a randomly selected sample of the investigational drugs and biologics of the firms participating in the survey.[7] The sample was taken from a Tufts Center for the Study of Drug Development (CSDD) database of the investigational compounds of top 50 firms. Tufts CSDD gathered information on the investigational compounds in development and their development status from commercial pipeline intelligence databases (*IMS R&D Focus* and *Thomson Reuters Cortellis* database [formerly the *IDdb3* database]), published company pipelines, clinicaltrials.gov, and web searches. Cost and time data were also collected for expenditures on the kind of animal testing that often occurs concurrently with clinical trials.[8] The compounds chosen were self-originated in the following sense. Their development from synthesis up to initial regulatory marketing approval was conducted under the auspices of the surveyed firm. This inclusion criterion is broader than it might at first seem since it includes compounds of firms that were acquired or merged with the survey firm during development and drugs that originated with the survey firm and were co-developed (and for which full cost data were available).[9] Licensed-in and co-developed compounds without partner clinical cost data were excluded because non-survey firms would have conducted significant portions of the R&D.[10]

We also collected data from the cost survey participants on their aggregate annual pharmaceutical R&D expenditures for the period 1990–2010. The firms reported on total annual R&D expenditures broken down by expenditures on self-originated new drugs, biologics, diagnostics, and vaccines. Data were also provided on annual R&D expenditures for licensed-in or otherwise acquired new drugs, and on already-approved drugs. Annual expenditures on self-originated new drugs were further decomposed into expenditures during the pre-human and clinical periods.

The survey firms accounted for 35% of both top 50 firm pharmaceutical sales and pharmaceutical R&D expenditures. Of the 106 investigational compounds included in the project dataset, 87 are small molecule chemical entities (including three synthetic peptides), and 19 are large molecule biologics (10 monoclonal antibodies and nine recombinant proteins). For ease of exposition, we will refer to all compounds below as new drugs, unless otherwise indicated. Initial human testing anywhere in the world for these compounds occurred during the period 1995–2007. Development costs were obtained through 2013.

We selected a stratified random sample of investigational compounds.[11] Stratification was based on the status of testing as of the end of 2013. Reported costs were weighted to reflect the development status of compounds in the population relative to those in the cost survey sample, so that knowledge of the distribution of development status in the population from which the sample was drawn was needed. The population is composed of all investigational compounds in the Tufts CSDD investigational drug database that met study criteria: the compounds were self-originated and first tested in humans anywhere in the world from 1995 to 2007. We found 1442 investigational drugs that met these criteria. Of these compounds, 103 (7.1%) have been approved for marketing, 13 (0.9%) had NDAs or BLAs that were submitted and are still active, 11 (0.8%) had NDAs or BLAs submitted but abandoned, 576 (39.9%) were abandoned in phase I, 19 (1.3%) were still active in phase I, 492 (34.1%) were abandoned in phase II, 84 (5.8%) were still active in phase II, 78 (5.4%) were abandoned in phase III, and 66 (4.6%) were still active in phase III. For both the population and the cost survey sample, we estimated approval and discontinuation shares for the active compounds by phase so that the population and sample distributions consisted of shares of compounds that were approved or discontinued in phase I, phase II, phase III, or regulatory review. The

[5] We capture out-of-pocket discovery costs with our data, but the pre-synthesis discovery period is highly variable with no clear starting point. For our analyses we began our representative discovery and development timeline at the point of compound synthesis or isolation. Thus, our estimates of time costs are somewhat conservative.

[6] Using pharmaceutical sales in 2006 to measure firm size, 5 of the survey firms are top 10 companies, 7 are top 25 firms, and 3 are outside the top 25 (*Pharmaceutical Executive*, May 2007).

[7] A copy of the survey instrument can be found in our online supplement (Appendix G).

[8] Long-term teratogenicity and carcinogenicity testing may be conducted after the initiation of clinical trials, and is often concurrent with phase I and phase II testing.

[9] The criterion also does not preclude situations in which the firm sponsors trials that are conducted by or in collaboration with a government agency, an individual or group in academia, a non-profit institute, or another firm.

[10] Large and mid-sized pharmaceutical firms much more often license-in than license-out new drug candidates. Firms that license-in compounds for further development pay for the perceived value of the prior R&D typically through up-front fees, development and regulatory milestone payments, and royalty fees if the compound should be approved for marketing. For a breakdown of new drugs and biologics approved in the United States in the 2000s by business arrangements among firms initiated during clinical development, see DiMasi et al. (2014).

[11] To ease the burden of reporting and increase the likelihood that firms would respond, we limited the number of compounds to be reported on to a maximum of 15 for any firm (with fewer compounds for smaller firms).

cost survey sample was purposely weighted toward compounds that lasted longer in development to increase the amount of information on drugs that reached late-stage clinical testing. Weights, determined as described above, were then applied to the compounds in the cost dataset so that the results would reflect the development status distribution for the population from which the sample was drawn.

Some firms were not able to provide full phase cost data for every new drug sampled. For example, phase I cost data were available for 97 of the 106 new drugs in the dataset (92%). Of the 82 compounds in the dataset that had entered phase II, cost data were available for 78 (95%). For phase III, cost data were available for 42 of the 43 compounds that entered the phase (98%). However, we had cost data for at least one phase for each of the 106 drugs in the sample. In aggregate, we had cost data for all phases entered for 94 of the 106 compounds (89%).[12] In addition, five compounds were still active in a phase at the time that data were reported. For these drugs it is likely that there will be some additional future costs for the drug's most recent phase. Thus, for this reason our cost estimates are likely to be somewhat conservative. However, given the small number of drugs in this category and the fact that the impact would be on only one phase for each of these drugs, our overall cost estimates are not likely to be substantially affected.

The methodology that we use to estimate development costs is the same as the approach used in our earlier studies (Hansen, 1979; DiMasi et al., 1991, 2003). We refer the reader to the earlier studies and to our online supplement (Appendix A) for details. The methodology results in a full risk-adjusted cost per approved new compound that also takes into account time costs. That is, we link the cost of compound failures to the cost of the successes (investigational compounds that attain regulatory marketing approval), and we utilize a representative time profile along with an industry cost of capital to monetize the cost of the delay between when R&D expenditures are incurred and when returns to the successes can first be realized (date of marketing approval). We refer to the sum of out-of-pocket cost (actual cash outlays) and time cost per approved new compound as the capitalized cost per approved new compound. The full capitalized cost estimate is built through a number of estimates of various components of the drug development process. These individual component estimates are interesting as objects of analysis in their own right, and we provide estimates for those components.

5. Base case R&D cost estimates

5.1. Out-of-pocket clinical cost per investigational drug

To determine expected costs, we need estimates of the clinical development risk profile. We examined the dataset of 1442 self-originated compounds of top 50 pharmaceutical firms described above and estimated the phase transition probabilities shown in Fig. 1. The overall probability of clinical success (i.e., the likelihood that a drug that enters clinical testing will eventually be approved) was estimated to be 11.83%. This success rate is substantially lower than the rate of 21.50% estimated for the previous study, but consistent with several recent studies of clinical success rates.[13] Such an increase in overall risk will contribute greatly to an increase in costs per approved new drug, other things equal.

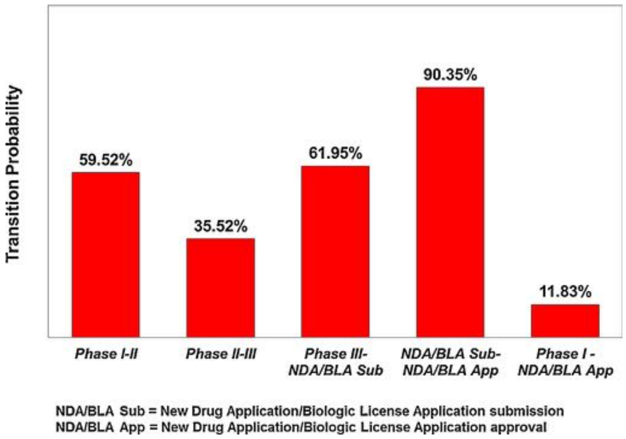


Fig. 1. Estimated phase transition probability and overall clinical approval success rates for self-originated new molecular entity (NME) and new therapeutically significant biologic entity (NBE) investigational compounds first tested in humans anywhere from 1995 to 2007.

As described above, we calculated weighted means, medians, standard deviations, and standard errors for clinical phase costs. Some of the firms could not separate out long-term animal testing costs during clinical development, and instead, included these costs in their phase cost estimates by year. To be consistent, therefore, for those compounds where animal costs were separately reported, we allocated those costs to the clinical phases according to when the animal testing costs were incurred. Thus, the clinical phase costs presented in Table 2 are inclusive of long-term animal testing costs.[14]

Weighted mean and median costs per investigational drug entering a phase[15] increase for later clinical phases, particularly for phase III (which typically includes a number of large-scale trials). In comparison to our previous study (DiMasi et al., 2003), both mean and median phase III cost are notably higher relative to the earlier phases. While the ratio of mean phase III cost to mean phase I cost was 5.7 for the previous study, it was 10.1 here. Similarly, the ratio of mean phase III to phase II cost was 3.7 for the earlier study, but was 4.4 for this study. Mean phase II cost was also higher relative to phase I cost in the current study compared to the previous one (2.3 times as high compared to 1.5 times as high).[16] Thus, while mean cost in real dollars for phase I increased 28% relative to the previous study,[17] phase I costs were notably lower relative to both phase II and phase III for the current study.

As we will see below, the differential in cost per approved new drug between the two studies will be much greater than cost per investigational drug because of the much lower overall clinical approval success rate. However, our results do show that the impact is mitigated to some degree by firms failing the drugs that they do abandon faster for the current study period. The distribution of clinical period failures for this study were 45.9% for phase I, 43.5% for phase II, and 10.6% for phase III/regulatory review. The

[12] Phase cost correlation results presented in the online supplement, together with an examination of relative phase costs for drugs that had some missing phase cost data, suggest that our phase cost averages (exclusive of missing data) are conservative.

[13] See, for example, Paul et al. (2010), DiMasi et al. (2013), and Hay et al. (2014).

[14] When animal testing costs occurred in a year during which costs were incurred for two clinical phases, the animal costs were allocated to the two phases according to their relative costs for the year.

[15] Averages for unweighted costs did not differ greatly from the weighted cost figures. On an unweighted basis, mean phase I, phase II, and phase III costs were $29.7 million, $64.7 million, and $253.5 million, respectively.

[16] The ratios for median costs for the current study are 11.6 for phase III relative to phase I, 4.5 for phase III relative to phase II, and 2.6 for phase II relative to phase I. The corresponding ratios for the previous study are 4.5, 3.6, and 1.2, respectively.

[17] In real terms, median phase I cost was actually 4% lower for the current study compared to the previous study.

Table 2
Average out-of-pocket clinical period costs for investigational compounds (in millions of 2013 dollars).[a]

Testing phase	Mean cost	Median cost	Standard deviation	Standard error	N[b]	Probability of entering phase (%)	Expected cost
Phase I	25.3	17.3	29.6	3.0	97	100.0	25.3
Phase II	58.6	44.8	50.8	6.6	78	59.5	34.9
Phase III	255.4	200.0	153.3	34.1	42	21.1	54.0
Total							114.2

[a] All costs were deflated using the GDP implicit price deflator. Weighted values were used in calculating means, medians, and standard deviations.
[b] N = number of compounds with cost data for the phase.

Table 3
Nominal and real cost of capital (COC) for the pharmaceutical industry, 1994–2010.

	1994	2000	2005	2010
Nominal COC (%)	14.2	14.9	13.3	11.4
Inflation rate (%)	3.1	3.1	2.5	2.0
Real COC (%)	11.1	11.8	10.8	9.4

corresponding figures for the previous study were 36.9% for phase I, 50.4% for phase II, and 12.6% for phase III/regulatory review.

5.2. Cost of capital estimates

To account for the time value of money in our previous paper (DiMasi et al., 2003), we utilized an 11% real after-tax weighted average cost of capital (WACC). In particular, we employed the capital asset pricing model (CAPM) to estimate the cost of equity capital. This was combined with the cost of debt, appropriately weighted with the cost of equity, to yield a representative, pharmaceutical industry weighted after-tax cost of capital. The resultant parameters were estimated at regular intervals from the mid-1980s to the year 2000, given the time period spanned by our sample of R&D projects.

In the present paper, we follow the same methodology to compute WACC. In the current R&D cost analysis, we have a sample of new drugs that began clinical trials in 1995 through 2007 and which have an average introduction period in the latter part of the 2000 decade. Hence, a relevant time period for our cost of capital is the mid-1990s through 2010. Our analysis yielded an after-tax weighted cost of capital of 10.5%, moderately lower than in our last paper. This reflects the fact that the cost of equity capital has declined in pharmaceuticals since 2000 (as well as for other industrial sectors). Research intensive industries, including the pharmaceutical industry, generally finance most of their investments through equity, rather than through debt. This is the case even when the cost of debt is significantly below the cost of equity (Hall, 2002; Vernon, 2004). One of the primary reasons is that servicing debt requires a stable source of cash flows, while the returns to R&D activities are skewed and highly variable (Scherer and Harhoff, 2000; Berndt et al., 2015). Given the low debt-to-equity ratios that exist for pharmaceutical firms, the cost of equity component dominates the computed WACC values in Table 3.

To obtain a real cost of capital, we first compute the nominal values and then subtract the expected rate of inflation. The nominal cost of capital in 1994 is from a CAPM study by Myers and Howe (1997). The estimates for 2000, 2005, and 2010 are based on our own analysis, utilizing a comparable approach, with a large sample of pharmaceutical firms.[18] As this table shows, the estimated nominal cost of capital for pharmaceuticals was fairly stable during the period 1994–2000 (14.2–14.9%). However, it decreased during the decade of 2000s, particularly after the global recession occurred (with a value of 11.4% observed in 2010).

As discussed in DiMasi et al. (2003), the rate of inflation was above historical values during the first part of the 1980s, but then receded back to or below historical levels throughout most of the 1990s. Hence, we utilized the long run historical value for inflation for the expected inflation level in 1994 and 2000 (3.1%), as in our prior work. For the 2000s decade, inflation was significantly below historical values. In this case, we employed a 5-year lagged moving average to compute the expected rate of inflation in 2005 and 2010 (calculated as 2.5% and 2.0%, respectively).

As shown in Table 3, our estimates for the real cost of capital varied between 9.4% and 11.8% for pharmaceutical firms over the 1994–2010 period. We elected to use the midpoint of this range, or approximately 10.5%, as the representative COC to capitalize our R&D cost estimates.

The focus of our analysis is R&D investment expenditures and privately financed resources for new drugs undertaken by the biopharmaceutical industry. Accordingly we capitalized these expenditures utilizing a cost of capital estimate based on financial data from publicly listed firms. Drug development is also sponsored and funded by government and non-profit agencies (e.g., public–private partnerships devoted to developing medicines for neglected diseases). To the extent that our cost estimates are applicable to these ventures, a social rate of discount would be appropriate to capitalize R&D outlays. We provide a sensitivity analysis in Section 6 with respect to a wide spectrum of alternative cost of capital values.

5.3. Capitalized clinical cost per investigational drug

Opportunity cost calculations for clinical period expenditures require estimates of average phase lengths and average gaps or overlaps between successive clinical phases to generate an average clinical development and regulatory review timeline. Mean phase lengths and the mean lengths of time between successive phases are shown in Table 4, along with the associated capitalized mean phase costs and capitalized expected phase costs by phase for investigational compounds. The time between the start of clinical testing and submission of an NDA or BLA with the FDA was estimated to be 80.8 months, which is 12% longer (8.7 months) than the same period estimated for the previous study. The average time from the start of clinical testing to marketing approval for our timeline was 96.8 months for the current study, 7% (6.5 months) longer than for the earlier study. The difference is accounted for by shorter FDA approval times. The period for the previous study included, in part, a period prior to the implementation of the *Prescription Drug User Fee Act of 1992* (PDUFA), and, in part, the early user fee era for which approval times were somewhat higher than for later user fee periods (Berndt et al., 2005).[19] While the approval

[18] The sample is composed of all publically traded drug firms in the *Value Line Survey* which also provides beta values and the other pharma-specific parameters used in the CAPM calculations for the relevant years. The long-term horizon equity risk premium, and the yield on long-term government bonds employed in the CAPM analysis, are from Ibbotson Valuation yearbooks for 2000, 2005, and 2010.

[19] The user fee legislation sunsets every 5 years. It has been renewed every 5 years since its original enactment. Performance goals for FDA review of marketing

Table 4
Average phase times and clinical period capitalized costs for investigational compounds (in millions of 2013 dollars).[a]

Testing phase	Mean phase length	Mean time to next phase	Capitalized mean phase cost[b,c]	Capitalized expected phase cost[b,c]
Phase I	33.1	19.8	49.6	49.6
Phase II	37.9	30.3	95.3	56.7
Phase III	45.1	30.7	314.0	66.4
Total				172.7

[a] All costs were deflated using the GDP implicit price deflator. Weighted values were used in calculating means for costs and phase times. Phase times are given in months.
[b] The NDA/BLA approval phase was estimated to be 16.0 months on average (2000–2012).
[c] Costs were capitalized at an 10.5% real discount rate.

phase averaged 18.2 months for the earlier paper's study period, that phase averaged 16.0 months for drugs covered by the current study.Other things being equal, the observed longer times from clinical testing to approval yielded higher capitalized costs relative to out-of-pocket costs. However, the discount rate that we used for the current study is also lower than for the previous study (10.5% versus 11.0%). The two effects work in offsetting ways. In addition, capitalized clinical cost per investigational compound will also depend on the gaps and overlaps between phases. On net, the ratio of mean capitalized to out-of-pocket cost per investigational compound was slightly lower for the current study compared to the previous one (1.5 versus 1.7).[20]

5.4. Clinical cost per approved new drug

Average cost estimates for investigational drugs are useful, but we are primarily interested in estimates of cost per approved new drug. As noted above, our analysis of drugs in development for the relevant period yielded a predicted overall clinical success rate of 11.83%. Applying this success rate to our estimates of out-of-pocket and capitalized costs per investigational drug results in estimates of cost per approved new drug that link the cost of drug failures to the successes.

Aggregating across phases, we found an out-of-pocket clinical period cost per approved new drug estimate of $965 million and a capitalized clinical period cost per approved new drug estimate of $1460 million. In constant dollars, these costs are 2.6 and 2.4 times higher than those we found in our previous study, respectively.

5.5. Pre-human out-of-pocket and capitalized costs per approved drug

The pre-human period, as defined here, includes discovery research as well as preclinical development. Some costs incurred during this period cannot be associated with specific compounds. To deal with this issue, we analyzed reported aggregate annual firm expenditures on self-originated new drugs by the pre-human and clinical periods. We gathered data on aggregate expenditures for these periods from survey firms for 1990–2010. Both times series tended to increase over time in real terms. Given this outcome, and the fact that the clinical expenditures in 1 year will be associated with pre-human expenditures that occurred years earlier, the ratio of total pre-human expenditures to total R&D (pre-human plus clinical) expenditures over the entire study period would yield an overestimate of the share of total cost per new drug that is accounted for by the pre-human period. To accurately estimate this share we built in a lag structure that associates pre-human expenditures with clinical expenditures incurred some time later.

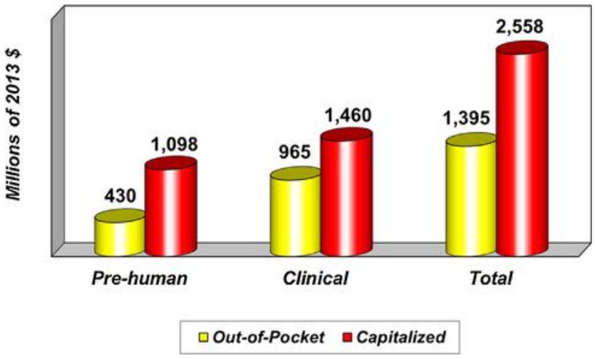


Fig. 2. Pre-human phase, clinical phase, and total out of-pocket and capitalized costs per approved new compound.

The survey firms reported on dates of synthesis or isolation for compounds for which we sought cost data, as well as dates of first human testing. We had data for the period from synthesis to first human testing for 78 of the compounds. The average time from synthesis to initial human testing for these compounds was 31.2 months, down considerably from 52.0 months for the previous study.[21] Our analyses of clinical phase lengths and phase gaps and overlaps indicated a period of 95.2 months over which clinical period development costs are incurred. We approximated the lag between pre-human and clinical expenditures for a representative new drug as the time between the midpoints of each period. This yields a lag of 63.2 months, or approximately 5 years. Thus, we used a 5-year lag in analyzing the aggregate expenditure data, although we also examined 4-year and 6-year lags. A 5-year lag applied to the aggregate expenditure data resulted in a pre-human to total R&D expenditure ratio of 30.8%, which was only slightly different from the corresponding ratio used in our previous study (30.0%). The share was applied to our clinical cost estimates to determine associated pre-human cost estimates.

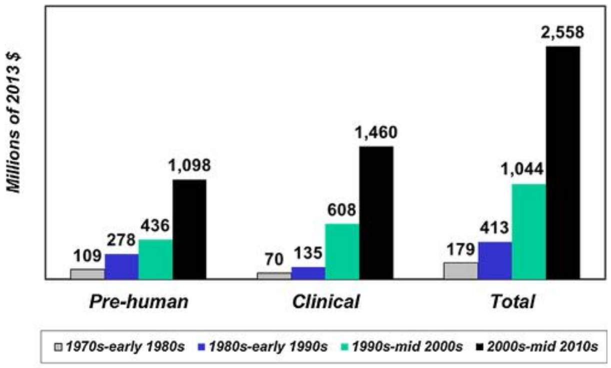
Given the estimates of out-of-pocket and capitalized clinical cost per approved new drug noted in Section 5.4 and the pre-human expenditure to total R&D expenditure ratio, we can infer pre-human out-of-pocket and capitalized costs per approved new drug of $430 million and $1098 million, respectively (Fig. 2). The results are very robust to different values for the length of the lag structure. For example, if we assume a lag of 4 years instead of 5 years, then out-of-pocket pre-human costs would be 6.8% higher. Alternatively, if we assume a 6-year lag, then out-of-pocket pre-human costs would be 8.5% lower.[22]

applications under PDUFA were tightened somewhat for some applications after the initial 5-year period.

[20] The differences in the ratios of capitalized to out-of-pocket cost for the individual phases were also small. For the current study they were 2.0, 1.6, and 1.2 for phase I, phase II, and phase III, respectively. For the earlier study, we found the ratios to be 2.0, 1.8, and 1.3 for phase I, phase II, and phase III, respectively.

[21] The results for the current study are consistent with data for a small number of compounds reported in a recently published study (Stergiopoulas and Getz, 2012). The mean time from synthesis to human testing there was 37.9 months for 17 compounds.

[22] The pre-human to total R&D expenditure ratios for four- and six-year lags were 32.2% and 28.9%, respectively.



Sources: 1970s-early 1980s, Hansen (1979); 1980s-early 1990s, DiMasi et al. (1991); 1990s-mid 2000s, DiMasi et al. (2003); 2000s-mid 2010s, Current Study

Fig. 3. Trends in capitalized pre-human, clinical and total cost per approved new drug.

5.6. *Total capitalized cost per approved drug*

Total cost estimates are the sum of pre-human and clinical period cost estimates. Our base case total out-of-pocket cost per approved new drug is $1395 million, while our fully capitalized total cost estimate is $2558 million (Fig. 2). Time costs (differences between capitalized cost and out-of-pocket cost) account for 45% of total cost. This share is down from the share in our previous study (50%) and that for the study that preceded it (51%). This is due in part to a shorter pre-human period and a lower discount rate.

5.7. *Trends in R&D costs*

Fig. 3 presents capitalized pre-human, clinical, and total cost per approved new drug for the previous three studies in this series and for our current study. In constant dollars, total capitalized cost increased 2.31 times for the second study in comparison to the first, 2.53 times for the third study in comparison to the second study, and 2.45 times for the current study in comparison to the third study. However, the samples for these studies include drugs that entered clinical testing over periods that are not uniformly distributed. In addition, while the samples were chosen on the basis of when drugs entered clinical testing, changes over time in the average length of the development process make ascribing differences in the study periods according to the year of first human testing problematic. An alternative is to determine an average approval date for drugs in each study's sample and use the differences in these dates to define the time differences between the studies. Our previous study described this approach and presented the corresponding annual growth rates between successive studies for the first three studies.

Drugs in the current study sample obtained FDA marketing approval from 2005 to 2013. The mean and median approval dates for drugs in the current study's sample were both in 2008. For the previous study, we reported that the average approval date was in 1997. Thus, we used 11 years as the relevant time span between the studies and calculated compound annual rates of growth between the two studies accordingly.

Using the period differences described here and in our previous study, we determined the compound annual growth rates between the studies for out-of-pocket and capitalized cost per approved drug for pre-human, clinical, and total costs (Table 5). Compared to the growth rate for the results in the previous study, the growth rates for total out-of-pocket and capitalized costs for the current study are somewhat higher (9.3% and 8.5% per year). The results for the current study in comparison to those for the previous study are also noteworthy in that, after a substantial decline in the growth rate for real pre-human costs described in the previous study and presented in Table 5, pre-human costs for the current study resumed a much higher rate of growth. Conversely, the growth rates for clinical period expenditures declined from the very high rates for the previous study, although they are still substantial.

5.8. *Cost of post-approval R&D*

As we did for our most recent study, we develop indirect estimates of post-approval R&D costs. Post-approval R&D consists of efforts subsequent to original marketing approval to develop the active ingredient for new indications and patient populations, new dosage forms and strengths, and to conduct post-approval (phase IV) research required by regulatory authorities as a condition of original approval. We follow the methodology that we used in previous study.[23] We utilize our pre-approval estimates together with aggregate pharmaceutical industry data regarding the drug development process to construct an estimate of the cost of post-approval R&D, which together with our pre-approval estimates, provide estimates of average total R&D cost per new drug covering the entire development and product life-cycle. The data that we collected from the survey firms on company annual aggregate expenditures on biopharmaceutical R&D show that over the study period these firms spent 73.1% of their prescription biopharmaceutical R&D expenditures on investigational self-originated new compounds,[24] 10.2% on investigational compounds that were licensed-in or otherwise acquired, and 16.5% on improvements to drugs that have already been approved.[25]

We cannot, however, use the percentage of aggregate R&D expenditures spent on post-approval R&D on a current basis and apply it to a pre-approval cost estimate to obtain an appropriate estimate of the cost of post-approval R&D per approved compound. The reason is that pre-approval costs occur years before post-approval costs. We used our aggregate annual firm R&D data to obtain an appropriate ratio by building in a reasonable lag structure between pre-approval and post-approval costs.

For our base results we used, as we did for the previous study, a 10-year lag for the aggregate data (which is the approximate time between median pre-approval development costs and median post-approval costs, given an 8-year post-approval expenditure period), we assumed that post-approval R&D cost per approval is the same, on average, for licensed-in and self-originated compounds, and we determined the percentage of approvals for the cost survey firms that are self-originated to estimate the ratio of post-approval R&D cost per approved compound to pre-approval cost per approved compound. The data indicated that this share was 33.4%. Applying this ratio, we estimated the out-of-pocket cost per approved compound for post-approval R&D to be $466 million (Fig. 4). Since these costs occur after approval and we are capitalizing all costs to the point of marketing approval, our discounted cost estimate is lower ($312 million). Thus, out-of-pocket cost per approved compound for post-approval R&D is 25.0% of

[23] We refer to the discussion in DiMasi et al. (2003) and an accompanying Appendix A for more detail on the method.

[24] This figure includes expenditures on biologics, vaccines, and diagnostics. The self-originated share for therapeutic investigational drugs and biologics was 71.2%.

[25] These expenditure shares are similar to those found for the previous study for the 1980 to 1999 period. The results here are also similar to figures that the trade association Pharmaceutical Research and Manufacturers of America (PhRMA) has published for its member firms for the years 2003 and 2005 to 2010. Those data do not separate out expenditures on existing products, but they do distinguish between self-originated and licensed products. Aggregating across those years, the shares for self-originated, licensed, and uncategorized were 74.3%, 17.6%, and 8.1%, respectively.

Table 5
Compound annual growth rates in out-of-pocket and capitalized inflation-adjusted costs per approved new drug.[a]

Approval periods	Out-of-pocket			Capitalized		
	Pre-human	Clinical	Total	Pre-human	Clinical	Total
1970s to 1980s	7.8%	6.1%	7.0%	10.6%	7.3%	9.4%
1980s to 1990s	2.3%	11.8%	7.6%	3.5%	12.2%	7.4%
1990s to early 2010s	9.6%	9.2%	9.3%	8.8%	8.3%	8.5%

[a] Costs for 1970s approvals are from Hansen (1979), costs for 1980s approvals are from DiMasi et al. (1991), costs for the 1990s to the early 2000s are from DiMasi et al. (2003), and costs for the 2000s to early 2010s are from the current study.

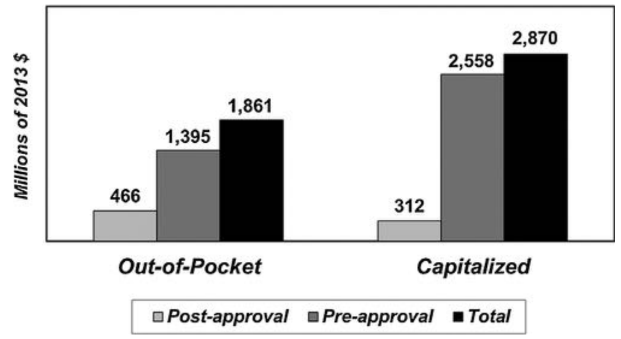


Fig. 4. Out-of-pocket and capitalized total cost per approved new drug for new drugs and for improvements to existing drugs.

Table 6
Capitalized pre-human, clinical, and total costs per approved new drug (in millions of 2013 dollars) by discount rate.

Discount rate	Pre-human	Clinical	Total
1.0%	472	1012	1476
2.0%	517	1044	1561
3.0%	567	1086	1653
4.0%	621	1129	1750
5.0%	679	1175	1854
6.0%	742	1222	1964
7.0%	811	1271	2082
8.0%	885	1322	2207
9.0%	965	1376	2341
10.0%	1052	1431	2483
11.0%	1145	1489	2634
12.0%	1246	1549	2795
13.0%	1355	1612	2967
14.0%	1473	1677	3150
15.0%	1600	1744	3344

total R&D cost (pre- and post-approval), while capitalized cost for post-approval R&D is 10.9% of total cost.

5.9. Extensions to the base case

We can extend the base case results on drug development costs prior to original approval in a number of interesting ways. The sample dataset includes information on compound-level costs for both chemical compounds (small molecules) and biologics (large molecules). As reported in the online supplement (Appendix B), we examined investigational compounds by molecule size for differences in individual clinical phase costs. Since the distributions of compounds across therapeutic classes differ for large and small molecules, we conducted a regression analysis of phase costs for investigational compounds for each of the three clinical phases, while controlling for molecules size and therapeutic class. Sample sizes were somewhat limited when cut by both sample size and therapeutic class, but we found statistically significant higher phase II costs for large molecules. However, we found that clinical approval success rates for large molecules are substantially higher than for small molecules. As a result, clinical period cost per approved compound was appreciably higher for small molecules, with the ratio of costs nearly the same as we had estimated in a previous paper for an earlier period (DiMasi and Grabowski, 2007). Compete results are given and discussed in the online supplement (Appendix B).

The base case results on full R&D costs link expenditures on drug failures to the costs of drugs that attain regulatory success. We can also estimate the clinical period cost of taking a successful drug all the way to approval by examining the data for just the approved drugs in the sample. Focusing on that subsample also allowed us to examine evidence on the costs for the more therapeutically significant drugs (according to what is known at the time of approval) by using an FDA prioritization system for reviewing drugs submitted to the agency for marketing approval. We found that clinical period costs were substantially higher for the approved compounds in the sample relative to our results for the sample as a whole, and that costs were lower (although not at a statistically significant level) for compounds that the FDA had designated for a priority review (compounds thought to represent a significant gain over existing therapy). These results are presented in full and discussed in the online supplement (Appendix B).

6. Sensitivity analysis

We examined how sensitive the results were to extreme values in the data and to changes in certain critical parameters. In particular, we focus in detail in this section on variation in the discount rate used to calculate capitalized costs. We also determine the extent to which key cost drivers (cash outlays, risks, time, and the cost of capital) explain the increase in total cost per approved drug found for this study relative to our previous study.

In addition, since all of the parameters are subject to sampling error, we conducted Monte Carlo simulations, reported on in detail in the online supplement (Appendix C), allowing all parameters to vary according to their sampling distributions (using Crystal Ball^{TM} software). For the full capitalized pre-approval cost estimate, 80% of the simulation forecasts (set of 1000) varied between $2.3 billion and $2.8 billion. All of the forecasts varied between $1.9 billion and $3.2 billion.

Finally, we also conducted an outlier analysis to determine the impact of the most extreme values in the dataset. The results show that drugs with high and low costs have a fairly small impact on cost estimates. For example, if all cost data for the drugs with the highest and lowest aggregate clinical costs are dropped from the analysis, then the full capitalized cost estimate falls by only 3.0% (3.5% if only the drug with the highest aggregate cost is dropped). The online supplement (Appendix D) further describes in detail various outlier analyses, including those that examine results when a number of high and/or low values for each clinical phase are excluded even though no one drug has uniformly high or low values across all clinical phases.

6.1. *Effects of variation in the discount rate*

Table 6 shows how pre-human, clinical, and total capitalized costs would vary by discount rate at one percentage point intervals. The values for a zero percent discount rate are out-of-pocket costs. In the neighborhood of our base case discount rate (10.5%), clinical cost changed by approximately $30 million, pre-human cost changed by approximately $45 million, and total cost changed by approximately $75 million for every half of one percent shift in the discount rate. In our previous study, the base case discount rate was 11.0%. At an 11.0% discount rate, total capitalized cost here was $2634 million or 3% higher than our base case result. At more extreme values for the discount rate, Table 6 indicates that total capitalized cost with a 15% discount rate was $3334 million, or 30% higher than our base case result. Similarly, a 3% discount rate (a figure often used as a social discount rate) yielded a total capitalized cost per approved new drug of $1561 million, or 39% lower than the base case result.[26]

6.2. *Impact of cost drivers*

As noted in the previous section, the full cost estimate is a function of numerous parameters that interact in a non-linear (often multiplicative) manner. That makes it difficult to isolate the extent to which changes in individual parameters alone drive changes in total costs. However, we can get a sense for which parameters had the greatest impacts, in either direction, on the change in total R&D cost between the previous study and the current one by calculating what R&D costs would have been if only a single parameter (or a set of related parameters) had changed from what it was for the previous study to what we found it to be for the current study period.

Table 7 shows our results for these thought experiments for the major parameters categorized into four groupings (direct pre-human and clinical average phase cash outlays, technical risks, average development and approval times, and the cost of capital). The base result is total cost per approved new compound for the DiMasi et al. (2003) study in year 2013 dollars ($1044 million). The current study full cost estimate is 145% higher than the base result. That change reflects the cumulative effect of all parameter changes. For the table, we examined parameter-by-parameter changes from the parameter values for the DiMasi et al. (2003) study to those values found for the current study.

The largest impact on the change in costs between the studies was driven by changes in average out-of-pocket clinical phase costs, which resulted in an 82.5% increase in full cost.[27] Considering also the small difference between the studies in the estimated ratio of pre-human to clinical costs, the impact of the change in direct out-of-pocket phase costs was an increase in total cost of 85.5%. The increase in total cost was also driven to a substantial extent by much higher development risks. The overall clinical approval success rate declined from approximately one-in-five to approximately one-in-eight. That change alone accounts for a 57.3% increase in total cost. However, the impact of a lower clinical approval success rate was mitigated to a small extent by a shift in the distribution of failures to earlier in development. Taking both effects into account resulted in an increase in total cost of 47.3%. Changes in the development and approval timeline had a relatively small depressing effect on total cost. This impact was driven by a shorter pre-human testing phase and a shorter average approval phase. Average clinical development time increased modestly, and this had a relatively small impact on total cost. Overall, the effect of changes in the development and approval timeline was a 5.6% decrease in total cost. Finally, the small change in the cost of capital had a 3.1% depressing effect on total cost. The aggregation of the direct impacts across the four cost factor groupings accounted for a 124% increase in costs between the two studies. We attribute the residual increase (21%) to interaction effects.

Table 7
Impact on total capitalized cost per approved new drug due to changes in individual cost drivers (current study factor effect relative to prior study[a] cost).

Factor category	Factor (change to current study values)	Capitalized cost (millions of 2013 $)	Percentage change in cost
Direct cash outlays			
	Out-of-pocket clinical phase costs	1905	82.5%
	Pre-human/clinical cost ratio	1061	1.6%
	Overall out-of-pocket costs	1937	85.5%
Risk			
	Clinical approval success rate with prior study distribution of failures	1643	57.3%
	Distribution of failures with prior study clinical approval success rate	981	−6.0%
	Overall risk profile: clinical approval success rate plus distribution of failures	1538	47.3%
Time			
	Pre-human phase	993	−4.9%
	Clinical phase	1046	0.2%
	Regulatory review	1013	−3.0%
	Overall development timeline	985	−5.6%
Cost of capital			
	Discount rate	1012	−3.1%

[a] DiMasi et al. (2003). In 2013 dollars the capitalized cost per approved new drug for the prior study is $1044 million.

7. Critiques, sample representativeness, and validation

Our prior study results have been questioned on a number of methodological and data grounds (Angell, 2005; Goozner, 2004; Light and Warburton, 2005a,b; Love, 2003; Young and Surrusco, 2001). We have rebutted each of these criticisms in detail in a number of venues (e.g., DiMasi et al., 2004, 2005a,b). We review the critics' main arguments only briefly here.

Goozner (2004) and Angell (2005) reject opportunity cost calculations because they, in essence, deny that industrial pharmaceutical R&D expenditures can be viewed as investments at risk.[28] These points are addressed more fully in DiMasi et al. (2004). Clearly, industrial pharmaceutical R&D meets the criteria for being considered investments that have opportunity costs. In any event, an estimate with no opportunity costs is simply the out-of-pocket cost estimate.

[26] The appropriate social rate of discount for government backed expenditures has been analyzed and debated extensively in the economics literature. See for example, Moore et al., 2013 and Burgess and Zerbe, 2013. A standard reference in the cost-effectiveness literature (Gold et al., 1996) recommends 3% as the base case rate in comparing alternative medical therapies ("Therefore, we recommend that the base rate of 3% and an alternate rate of 5% be retained for a period of at least 10 years.", p.233).

[27] Given the methodology, higher out-of-pocket clinical phase costs also get associated with higher out-of-pocket pre-human phase costs.

[28] In the case of Goozner (2004), the claim is made that R&D expenditures are expenses rather than investments, because accountants have traditionally treated them as such for tax purposes (failing to recognize practical measurement problems underlying why this has been the practice, such as great uncertainty regarding future regulatory and commercial success). The basis offered for rejecting opportunity costs in Angell (2005, p.45) is simply the claim that pharmaceutical firms "have no choice but to spend money on R&D if they wish to be in the pharmaceutical business".

A number of the critiques question how representative the data were for prior studies, whether tax deductions and credits must be included, and whether any FDA application for product marketing approval (as opposed to the active ingredient that is at the core of all such applications) should be taken as the unit of observation. As noted, we have addressed all of these issues in earlier publications as they relate to our prior studies. In this section we examine the representativeness of the survey firms and data used for this study, what the level of tax credits has been in relation to R&D expenditures in recent years, an analysis of molecules that have been approved for orphan drug indications recently, and we outline a variety of methods using independent data that can be used to validate our results (full details of the methods and analysis can be found in our online supplement).

7.1. *Representativeness of the survey firm data*

Questions about data representativeness should be framed in terms of the population from which the sample was selected. In particular, it is relevant to compare characteristics of the investigational drugs in our cost survey sample and for our cost survey firms generally to those of all drugs in our database of top 50 pharmaceutical firms, which is the relevant population.[29] This is the main focus of the analysis in this section.

Smaller research-oriented firms may have a comparative advantage in the discovery and pre-human stages because they often have scientific researchers with close ties to the basic research underlying new classes of therapies and technology platforms. Even if this is the case, the literature indicates that smaller firms also tend to have significantly higher costs of capital, especially when they are start-ups financed by venture firms. The literature also indicates that firms with larger R&D pipelines and greater R&D experience have a higher probability of success during the costly clinical stages of drug R&D. It is not evident, therefore, that the R&D costs for compounds originating in smaller firms, whether developed internally or in alliances would be systematically lower than those originating in mid-sized and large firms. We discuss what is known about R&D metrics for small firms in Appendix E of the online supplement.

As noted, the appropriate comparator dataset for our cost survey sample is the population of investigational compounds of the top 50 pharmaceutical firms over the relevant period. There are 1442 compounds in the top 50 firm database that met our study inclusion criteria. Of these, 510, or 35.4% belonged to nine of our 10 cost survey firms.[30] Thus, the cost survey sample ($n = 106$) constitutes 20.8% of the survey firm compounds and 7.4% of the population compounds.

We determined the therapeutic class distribution for the drugs in the larger dataset for the four largest therapeutic classes and one miscellaneous class (with a wide variety of drug types) for drugs in the dataset that met our study inclusion criteria and compared it to the therapeutic class distribution for our cost sample. The population shares for antineoplastic, cardiovascular, central nervous system (CNS), and systemic anti-infective drugs were 21.5%, 8.7%, 19.0%, and 8.5%, respectively. The corresponding shares for the cost survey sample were 19.8%, 9.4%, 24.5%, and 8.5%, respectively. We used a chi-squared goodness-of-fit test to compare the therapeutic class distributions for cost survey firm drugs and for the drugs of the entire set of 50 firms in the database, and found no statistically significant differences in the class shares ($\chi^2 = 2.4257$, df = 4).

We also examined the degree to which the top 50 firms in aggregate and the sample of cost survey firms agreed in terms of how molecule type (biologic versus small molecule) and the sourcing of compounds are distributed. For the set of top 50 firms, 14.6% of their self-originated investigational compounds over the study period are large molecules, compared to 13.7% for the survey firms ($p = 0.3933$). In terms of the share of investigational compounds for the study period that are self-originated (as broadly defined here), we found the share to be 74.1% for the cost survey firms and 71.1% for all top 50 firms ($p = 0.1039$).

Finally, we also examined the phase transition and overall approval success rates for the cost survey firms and compared them to the corresponding estimates for the larger dataset. The phase transition rates for just the cost survey firms were 58.0% for phase I to phase II, 36.0% for phase II to phase III, 58.2% for phase III to regulatory review, and 89.5% for regulatory review to approval. The corresponding figures for the population, as shown in Fig. 1, are 59.5%, 35.5%, 62.0%, and 90.4%. The overall clinical approval success rate for just the cost survey firms implied by the phase transition rates is 10.9%, which compares to 11.8% for the entire dataset.

7.2. *Orphan drug development*

Some past critiques have focused to some extent on orphan tax credits, which can provide incentives to develop some drugs for a class of indications. We examine the extent to which these tax credits and other tax issues are empirically significant in the context of drug development as a whole in the next section. Here we briefly discuss the nature of development of molecules that are approved for orphan indications and the distinction between costs for orphan drug indications and the full development costs for molecules with orphan drug indication approvals.

Compounds developed for orphan indications may well have lower clinical development costs for those indications, as trial sizes tend to be lower.[31] The share of U.S. original new drug approvals from 2000 to 2014 for drugs with an orphan indication was 27%, and has increased in relative terms over the last 3 years of that period.[32] The most recent approval experience aside, the share of approvals sponsored by the set of population firms (top 50) matches closely the historical average for all approvals from 1987 to 2010 (22% for top 50 firms versus 23% of all approvals).[33] The survey firms were nearly indistinguishable from the population non-survey firms by this metric (21% versus 23%).

[29] The data included in the top 50 firm dataset were curated primarily from information contained in two commercial investigational drug pipeline databases that are available after payment of subscription fees. Additional information was obtained from freely available web sites. See Section 4 above for a description of data sources.

[30] One of the participating firms was outside of the top 50.

[31] Drugs for these indications, with some notable exceptions, tend to garner lower sales given limited patient populations. This contention is supported by recent data analysis conducted by IMS Health (Divino et al., 2014). They found that sales in the United States for orphan indications varied from only 4.8% to 8.9% of total pharmaceutical sales over 2007–2013. The analysts also projected that growth in orphan drug expenditures would slow over 2014–2018.

[32] The result was calculated from information provided by the FDA on its website and included in a Tufts CSDD database of NME and therapeutically significant biologic approvals. The share of new drug approvals with orphan indications has increased very recently. The *Orphan Drug Act* was enacted in 1983, but it took several years for an appreciable number of such approvals to appear. From 1987 to 1999 the orphan drug share of all new drug approvals was 23%; the same share as for the 2000–2010 period. The orphan drug share was, however, unusually high for 2014 (41%), and above-average for 2011–2013 (approximately one-third of approvals).

[33] An FDA analysis of Center for Drug Evaluation and Research (CDER) marketing applications for NMEs and new biologics for 2006 to 2010 found that approximately one-third of the applications were sponsored by small firms, and that 75% of the applications for first-in-disease therapies for orphan indications came from small firms (Lesko, 2011). Such firms may find a low R&D cost orphan disease oriented strategy attractive, given that typical sales and operating profit levels may still be sufficient to increase their market valuations.

Table 8
Number of indications tested clinically prior to initial U.S. regulatory marketing approval for therapeutic compounds approved[a] in 2014 by orphan drug status.

	Mean	Median	Range	% multiple indications
Orphan ($n = 17$)	8.5	7.0	1–4	88%
Orphan cancer ($n = 9$)	10.9	9.0	1–24	89%
Non-orphan ($n = 22$)	2.7	2.0	1–7	73%
All approvals ($n = 39$)	5.3	3.0	1–24	79%

[a] Therapeutic new molecular entities (NMEs) and new biologic entities (NBEs) approved by the Center for Drug Evaluation and Research (CDER) of the United States Food and Drug Administration (FDA).

The cost survey sample contained two compounds that were approved originally for orphan indications.[34] The average clinical period cost for these two compounds was nearly the same as the average for all sample approved compounds (94% of the overall average). One of the compounds, though, was relatively low cost, while the other was relatively high cost. This may reflect the experience of molecules approved for orphan indications generally, as total molecule cost depends not only on the approved indication, but, critically, on the total number of indications (orphan and non-orphan) pursued.

To investigate this point further, we examined the development histories of all new therapeutic drugs and biologics approved in the United States in 2014. We studied the records for these compounds in two commercial pipeline database (*IMS R&D Focus* and *Cortellis*), as well as the clinicaltrials.gov website. Table 8 demonstrates that, even with a conservative notion of what constitutes different indications,[35] molecules approved for orphan indications were investigated in a substantial number of indications prior to original marketing approval. This was particularly true for compounds approved for treating orphan cancer indications, and, in general, the orphan drugs tended to be investigated in many more indications prior to approval than was the case for non-orphan compounds.

7.3. Taxes and R&D expenditures

As in our previous studies, the cost estimates presented here are pre-tax. Our objective was to measure the level of and trends in the private sector real resource costs of developing new drugs and biologics. As discussed in DiMasi et al. (2003), if one is calculating after-tax rates of return for R&D one would need to include the effect of taxes. Under current U.S. corporate income tax accounting practices, firms are able to deduct R&D expenses at the time they incur the costs. This is in contrast to many other investments, such as plants and equipment, which must be amortized and depreciated over a longer time period. This treatment reflects the difficulty of appropriately depreciating an intangible asset such as R&D. Later, when the company earns profits from the sales of approved pharmaceuticals it cannot depreciate the R&D investment for income tax purposes. The advantage for R&D investment over investment in plant and equipment is the timing of tax payments on net income. If one were calculating the rate of return on R&D investments one would need to take into account the tax implications. Making these adjustments is complicated by the fact that major firms operate in multiple tax jurisdictions.

In DiMasi et al. (2003) we also discussed several tax credits available in the United States to firms in the biopharmaceutical industry. In particular, we examined the Research & Experimentation tax credit for increasing qualified research expenditures, which we concluded had little impact on large multinational pharmaceutical firms.[36] Since then, the Qualifying Therapeutic Discovery Project tax credit was created as part of the Patient Protection and Affordable Care Act of 2010 (http://grants.nih.gov/grants/funding/QTDP_PIM/; accessed 14.08.14). However, it is quite restrictive in that it applies to discovery projects for small firms with a limit of $5 million per taxpayer. Recently, the U.S. Congress Joint Committee on Taxation (2013) estimated tax expenditures for fiscal years 2012–2017 for the credit for increasing research activities, the Qualifying Therapeutic Discovery Project tax credit, and the advantage from expensing, as opposed to amortizing, research and experimental expenditures to be, in aggregate, in the range of $10 billion to $12 billion per year for fiscal years 2012–2017 across all U.S. corporations engaged in research activities. It is not clear how much of this is accounted for by the biopharmaceutical industry.

We also examined in DiMasi et al. (2003) the impact of tax credits for orphan drug research, and found them to be quite small in relation to total R&D expenditures for large pharmaceutical firms. The reporting requirements for orphan drug credits are such that many companies do not take the credit. The major financial incentive of the orphan drug program appears to be the intellectual property protection that is created from the granting of 7 years of marketing exclusivity. With respect to the magnitude of orphan drug tax credits utilized in the United States, the U.S. Congress Joint Committee on Taxation (2013) estimated that expected tax credits for orphan drug research are fairly small at between $700 million and $1 billion per year from fiscal years 2012–2017.

To put these tax credits and tax advantages in perspective, Battelle and R&D Magazine's 2014 Global R&D Funding Forecast (http://www.battelle.org/docs/tpp/2014_global_rd_funding_forecast.pdf?sfvrsn=4; accessed 14.08.14) estimates that approximately $79 billion will be spent in the United States on R&D by the biopharmaceutical industry.[37] Some other countries also have a number of tax credit incentives in place for R&D. However, it seems unlikely that, in aggregate, their value in relation to R&D expenditures for the biopharmaceutical industry is disproportionately higher than is the case for the United States. The Battelle and R&D Magazine's prediction of global R&D spending by the biopharmaceutical industry is approximately $171 billion. In sum, in aggregate the value of R&D tax credits and the tax advantage of expensing versus amortizing R&D expenditures for the biopharmaceutical industry appear to be no more than one-sixth of total industry R&D expenditures (and perhaps significantly less than that).

7.4. Validation

We gathered publicly available data and performed a number of independent analyses on those data to corroborate our results. Details on methodology and data are provided in Appendix F of our online supplement. The validation efforts can be grouped into those

[34] Analyzing orphan drug status for investigational compounds is problematic because the designation may be granted at any point during the development process. Thus, some compounds that might have been granted orphan drug status can be abandoned before that would occur.

[35] Indications may be defined quite narrowly. We chose a broad definition that would limit the number of different indications pursued. Specifically, we considered all trials for the same disease and that applied to the same organ system as testing on the same indication. For example, oncology compounds may be tested as first-line treatment, second-line treatment, for refractory patients, as a monotherapy, in combination with other compounds, or for special patient populations. These cases were considered to be the same indication if they applied to the same organ (e.g., breast cancer or prostate cancer).

[36] The impact may be greater for small firms if their R&D expenditures are growing more rapidly.

[37] The report estimates that the industrial life sciences sector will spend $92.6 billion on R&D in the United States in 2014. However, the report also indicates that approximately 85% of all life sciences industrial expenditures are accounted for by the biopharmaceutical industry.

that utilize micro data on elements of the development process that are then used to develop growth rate estimates for portions of the process, and those that use publicly available aggregate financial time series data and compound approval statistics for biopharmaceutical firms as a check on our estimate of overall cost.

On a micro level, we examined survey data from the National Science Foundation (NSF), published estimates of trends in clinical trial complexity and clinical trial costs per subject, and published trade association times series data on R&D employment levels. Utilizing external data on costs per subject, along with clinical trial sizes and estimated clinical approval success rates from our analyses over time, we found a compound annual growth rate in real clinical trial costs between the study periods for our previous study and the current study of 9.9%, which is close to our clinical period cost growth rate of 9.2% for out-of-pocket costs shown in Table 5. We also examined measures of clinical trial complexity (number of procedures per trial) in the published literature (Getz et al., 2008; PAREXEL, 2005) and found a compound annual growth rate of 10.0% over our study period. Finally, we utilized trade association and 10-K information on R&D scientific and professional staff employment levels and NSF data on salary levels to estimate that labor costs increased at a rate of 8–9% per year across our study periods.

We examined PhRMA time series data on the R&D expenditures of its member firms. The reported growth rate for cost survey firms was 4.9%, compared to 4.2% for the PhRMA time series data for the portion of the survey period that could be compared.[38] We also used the industry time series data, as we had in the previous study, in two ways to get a sense for the magnitude of overall costs per approved new molecule. In one approach, we estimated the portion of the reported time series expenditure levels that could be attributed to self-originated compound development. Next we determined the annual number of approvals of PhRMA-member firms that were self-originated. Finally, we used our study estimated time-expenditure profile to link aggregate R&D expenditures to approvals. For reasons expounded upon in the supplement, this will likely yield an upper bound estimate. Using this approach we found our out-of-pocket cost per approved molecule estimate to be 56% of the estimate derived from aggregate published industry data. The second approach focuses on the published industry self-originated R&D expenditure level for a single year, assumes that every self-originated member-firm approval (inclusive of failures) costs what we found to be our average out out-of-pocket cost estimate, and uses our estimated time-expenditure profile to spread costs out over time to explain reported total R&D expenditures for the year considered. As with the previous method, the outcome would be problematic if using our average out-of-pocket cost estimate explained more than the reported aggregate R&D expenditure level. We found that this approach explained 57% of the reported expenditures.

Company total biopharmaceutical R&D expenditures reported for the cost survey are consistent with the audited financial statements of the firms in that the annual values are equal to or lower than company R&D expenses found in the financial statements.[39] As another check on our overall results, we examined what survey company total biopharmaceutical R&D expenditures would be given our estimate of out-of-pocket cost per approved molecule and assuming that entry rates to survey company pipelines are in a steady state. That figure can then be compared to R&D expenditure levels reported for these firms for our cost survey (which, as noted, match audited financial statements). Full details of these calculations are in Appendix F of the supplement. Depending on assumptions, we found that we could account for between 51% and 94% of the reported total annual biopharmaceutical R&D expenditures in this way. Thus, all three approaches using aggregate R&D expenditure data suggest that our estimate of out-of-pocket cost per approved molecule is, if anything, conservative.

8. Conclusions

Studies of the cost of developing new drugs have long been of substantial interest to drug developers, drug regulators, policy makers, and scholars interested in the structure and productivity of the pharmaceutical industry and its contributions to social welfare. The interest has been strong and growing over the last few decades during which cost containment pressures for drugs approved for marketing have expanded and concerns have been raised about industry productivity in an environment in which industry structure has been evolving (Munos, 2009; Pammolli et al., 2011). The changing industrial landscape has featured consolidation among large firms, growing alliances among firms of all sizes, and the growth of a small firm sector.

We have conducted the fourth in a series of comprehensive compound-based analyses of the costs of new drug development. In the last study we reported average out-of-pocket and capitalized R&D costs of $403 million and $802 million in 2000 dollars ($524 million and $1044 million in 2013 dollars), respectively. For our updated analysis, we estimated total out-of-pocket and capitalized R&D cost per new drug to be $1395 million and $2558 million in 2013 dollars, respectively. To examine R&D costs over the entire product and development lifecycle, we also estimated R&D costs incurred after initial approval. This increased out-of-pocket cost per approved drug to $1861 million and capitalized cost to $2870 million. We validated our results in a variety of ways through analyses of independently derived published data on the pharmaceutical industry.

Our pre-approval out-of-pocket cost estimate is a 166% increase in real dollars over what we found in our previous study, and our capitalized cost estimate is 145% higher. Roughly speaking, the current study covers R&D costs that yielded approvals, for the most part, during the 2000s and early 2010s. Our previous study (DiMasi et al., 2003) generally involved R&D that resulted in 1990s approvals. The compound annual rates of growth in total real out-of-pocket and capitalized costs between the studies are 9.3% and 8.5%, respectively. These growth rates are both somewhat higher than those we found for the two previous studies (7.6% and 7.4%, respectively). Growth in out-of-pocket clinical period costs have moderated some from the 1990s, but the growth rate is still high at 9.2%. While the compound annual growth rate for out-of-pocket pre-human costs declined substantially for the previous study (from 7.8% to 2.3%), this study showed a substantially higher growth rate for pre-human costs in the new century (9.6%).

The success rate found for this study is nearly 10 percentage points lower than for the previous study. The overall change in the risk profile for new drug development by itself still accounted directly for a 47% increase in costs. It is difficult to know definitively why failure rates have increased, but a number of hypotheses worthy of testing come to mind. One possibility is that regulators have become more risk averse over time, especially in the wake of high profile safety failures for drugs that have reached the marketplace (most notably, Vioxx^{TM}, but there have been others as well). It may also be the case that the industry has generally focused more in areas where the science is difficult and failure risks are high as a result (Pammolli et al., 2011). Finally, the substantial growth in identified drug targets, many of which may be poorly validated, may have encouraged firms to pursue clinical development of more

[38] As explained in the Supplement, the growth rate for the PhRMA time series may somewhat underestimate the true growth rate.

[39] Biopharmaceutical R&D expenditures may be less than total company R&D expenditures if the firm engages in non-biopharmaceutical R&D.

compounds with an unclear likelihood of success than they otherwise would.

As can be seen from results cited in the supplement developed external to this study, as well as our own data, out-of-pocket clinical cost increases can be driven by a number of factors, including increasing clinical trial complexity (Getz et al., 2008), larger clinical trial sizes, inflation in the cost of inputs taken from the medical sector that are used for development, and possibly changes in protocol design to include efforts to gather health technology assessment information and, relatedly, testing on comparator drugs to accommodate payer demands for comparative effectiveness data. The expansion of the scope of the clinical trial enterprise during our study period is illustrated by the finding in Getz and Kaitin (2015) that for a typical phase III trial information had been gathered by sponsors on nearly 500,000 data points in 2002, but more than 900,000 data points in 2012.

Finally, it is difficult to assess whether and how regulatory burdens may have impacted changes in industry R&D costs over time. However, occasionally, an exogenous shift in the types and amount of information perceived as necessary for regulatory approval for particular classes of drugs can be instructive. For example, during our study period the FDA issued guidance (Food and Drug Administration, 2008) for the development of drugs to treat diabetes in late 2008 that highlighted a need to better assess and characterize cardiovascular risks for this class of compounds, after a number of cardiovascular concerns emerged regarding a previously approved drug (Avandia®). A number of development metrics positively related to R&D costs can be examined pre- and post-guidance. DiMasi (2015), for example, found that average U.S. clinical development times increased from 4.7 to 6.7 years for diabetes drugs approved in the United States from 2000–2008 to 2009–2014, respectively. In addition, Viereck and Boudes (2011) found that the number of randomized patients and patient-years in NDAs for diabetes drugs approved from 2005 to 2010 increased more than 2.5 and 4.0 times, respectively, before and after the guidelines were issued. Our sample data show that diabetes drugs were among the most costly (particularly for phase III [92% higher than the overall average]).

Our analysis of cost drivers indicates that the rate of increase observed in the current study was driven mainly by increases in the real out-of-pocket costs of development for individual drugs and by much higher failure rates for drugs that are tested in human subjects, but not particularly by changes in development times or the cost-of-capital. Continued analysis of the productivity of biopharmaceutical R&D should remain an important research objective.

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhealeco.2016.01.012.

References

Adams, C.P., Brantner, V.V., 2006. Estimating the cost of new drug development: is it really $802 million? Health Affairs 25 (March/April (2)), 420–428.

Adams, C.P., Brantner, V.V., 2010. Spending on new drug development. Health Economics 19 (2), 130–141.

Angell, M., 2005. The Truth About the Drug Companies: How They Deceive Us and What to Do About It. Random House, New York, NY.

Berndt, E.R., Gottschalk, A.H.B., Philipson, T.J., Strobeck, M.W., 2005. Industry funding of the FDA: effects of PDUFA on approval times and withdrawal rates. Nature Reviews Drug Discovery 4 (7), 545–554.

Berndt, E.R., Nass, D., Kleinrock, M., Aitken, M., 2015. Decline in economic returns from new drugs raises questions about sustaining innovations. Health Affairs 34 (2), 245–252.

Burgess, D.F., Zerbe, R.O., 2013. The most appropriate discount rate. Journal of Benefit-Cost Analysis 4 (3), 391–400.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., Lasagna, L., 1991. Cost of innovation in the pharmaceutical industry. Journal of Health Economics 10 (2), 107–142.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., Lasagna, L., 1995a. Research and development costs for new drugs by therapeutic category: a study of the U.S. pharmaceutical industry. PharmacoEconomics 7, 152–169.

DiMasi, J.A., Grabowski, H.G., Vernon, J., 1995b. R&D costs, innovative output and firm size in the pharmaceutical industry. International Journal of the Economics of Business 2, 201–219.

DiMasi, J.A., 2001. New drug development in the United States from 1963 to 1999. Clinical Pharmacology & Therapeutics 69 (5), 286–296.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2003. The price of innovation: new estimates of drug development costs. Journal of Health Economics 22 (2), 151–185.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2004, November. Assessing Claims About the Cost of New Drug Development: A Critique of the Public Citizen and TB Alliance Reports, http://csdd.tufts.edu/files/uploads/assessing_claims.pdf (accessed 16.07.14).

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2005a. Reply: extraordinary claims require extraordinary evidence. Journal of Health Economics 24 (5), 1034–1044.

DiMasi, J.A., Hansen, R.W., Grabowski, H.G., 2005b. Setting the record straight on setting the record straight: response to the Light and Warburton rejoinder. Journal of Health Economics 24 (5), 1049–1053.

DiMasi, J.A., Grabowski, H.G., 2007. The cost of biopharmaceutical R&D: is biotech different? Managerial & Decision Economics 28 (4–5), 285–291.

DiMasi, J.A., Feldman, L., Seckler, A., Wilson, A., 2010. Trends in risks associated with new drug development: success rates for investigational drugs. Clinical Pharmacology & Therapeutics 87 (3), 272–277.

DiMasi, J.A., Reichert, J.M., Feldman, L., Malins, A., 2013. Clinical approval success rates for investigational cancer drugs. Clinical Pharmacology & Therapeutics 94 (3), 329–335.

DiMasi, J.A., Kim, J., Getz, K.A., 2014. The impact of collaborative and risk-sharing innovation approaches on clinical and regulatory cycle times. Therapeutic Innovation and Regulatory Science 48 (3), 482–487.

DiMasi, J.A., 2015. Regulation and economics of drug development. In: Presentation at the American Diabetes Association 75th Scientific Sessions, June 5, Boston, MA, http://professional.diabetes.org/presentations_details.aspx?session=4619 (accessed 27.11.15).

Divino, V., Dekoven, M., Weiying, W., Kleinrock, M., Harvey, R.D., Wade, R.L., Kaura, S., 2014. The budget impact of orphan drugs in the US: a 2007–2013 MIDAS sales data analysis. In: Presentation at the 56th ASH Annual Meeting and Exposition, December 8, Orlando, FL, http://www.imshealth.com/deployedfiles/imshealth/Global/Content/Home%20Page%20Content/Real-World%20insights/IMS-Celgene_Orphan_Drugs_ASH_Presentation_Slides.pdf (accessed 25.05.15).

Food and Drug Administration, 2008, December. Guidance for Industry: Diabetes Mellitus – Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Rockville, MD.

Getz, K.A., Wenger, J., Campo, R.A., Seguine, E.S., Kaitin, K.I., 2008. Assessing the impact of protocol design changes on clinical trial performance. American Journal of Therapeutics 15, 450–457.

Getz, K.A., Kaitin, K.I., 2015. Why is the pharmaceutical industry struggling? In: Schuber, P., Buckley, B.M. (Eds.), Re-Engineering Clinical Trials. Academic Press, London, UK, pp. 3–15.

Gilbert, J., Henske, P., Singh, A., 2003. Rebuilding big pharma's business model. In Vivo 21 (10), 1–10.

Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, New York, NY.

Goozner, M., 2004. The 800 Million Dollar Pill: The Truth Behind the Cost of New Drugs. University of California Press, Berkeley, CA.

Hall, B.H., 2002. The financing of research and development. Oxford Review of Economic Policy 18 (1), 35–51.

Hansen, R.W., 1979. The pharmaceutical development process: estimates of current development costs and times and the effects of regulatory changes. In: Chien, R.I. (Ed.), Issues in Pharmaceutical Economics. Lexington Books, Lexington, MA, pp. 151–187.

Hay, M., Thomas, D.W., Craighead, J.L., Economides, C., Rosenthal, J., 2014. Clinical development success rates for investigational drugs. Nature Biotechnology 32, 40–51.

Ibbotson Associates, 2000, 2005, 2010. Stocks, Bonds, Bills & Inflation: 2000, 2005, 2010 Yearbook. Ibbotson Associates, Chicago, IL.

Lesko, L.J., 2011, March. Introduction: Rare Diseases, Orphan Drugs. Advisory Committee for Pharmaceutical Sciences and Clinical Pharmacology, http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/AdvisoryCommitteeforPharmaceuticalScienceandClinicalPharmacology/UCM247635.pdf (accessed 29.05.15).

Light, D., Warburton, R., 2005a. Extraordinary claims require extraordinary evidence. Journal of Health Economics 24 (5), 1030–1033.

Light, D., Warburton, R., 2005b. Setting the record straight in the reply by DiMasi, Hansen and Grabowski. Journal of Health Economics 24 (5), 1045–1048.

Love, J., 2003. Evidence Regarding Research and Development Investments in Innovative and Non-Innovative Medicines. Consumer Project on Technology, Washington, DC.

Mestre-Ferrandiz, J., Sussex, J., Towse, A., 2012. The R&D Cost of a New Medicine. Office of Health Economics, London, UK.

Moore, M.A., Boardman, A.E., Vining, A.R., 2013. More appropriate discounting: the rate of social time preference and the value of the social discount rate. Journal of Benefit-Cost Analysis 4 (1), 1–16.

Munos, B., 2009. Lessons from 60 years of pharmaceutical innovation. Nature Reviews Drug Discovery 8 (12), 959–968.

Myers, S.C., Howe, C.D., 1997. A Life-Cycle Financial Model of Pharmaceutical R&D, Working Paper, Program on the Pharmaceutical Industry. Massachusetts Institute of Technology, Cambridge, MA.

O'Hagan, P., Farkas, C., 2009. Bringing Pharma R&D Back to Health. Bain & Company, Boston, MA.

Pammolli, F., Magazzini, L., Ricaboni, M., 2011. The productivity crisis in pharmaceutical R&D. Nature Reviews Drug Discovery 10 (6), 428–438.

PAREXEL, 2005. In: Mathieu, M.P. (Ed.), PAREXEL Pharamceutical R&D Statistical Sourcebook, Waltham, MA.

Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., Schacht, A.L., 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery 9 (3), 203–214.

Scherer, F.M., Harhoff, D., 2000. Technology policy for a world of skewed-distributed outcomes. Research Policy 29 (4–5), 559–566.

Stergiopoulas, S., Getz, K.A., 2012. Mapping and characterizing the development pathway from non-clinical through early clinical development. Pharmaceutical Medicine 26 (5), 297–307.

Young, B., Surrusco, M., 2001, July. Rx R&D Myths: The Case Against the Drug Industry's R&D Scare Card. Public Citizen, Congress Watch, Washington, DC.

US Congressional Budget Office, 1998, July. How Increased Competition from Generic Drugs has Affected Prices and Returns in the Pharmaceutical Industry. US Government Printing Office, Washington, DC.

US Congressional Budget Office, 2006, October. Research and Development in the Pharmaceutical Industry. US Government Printing Office, Washington, DC.

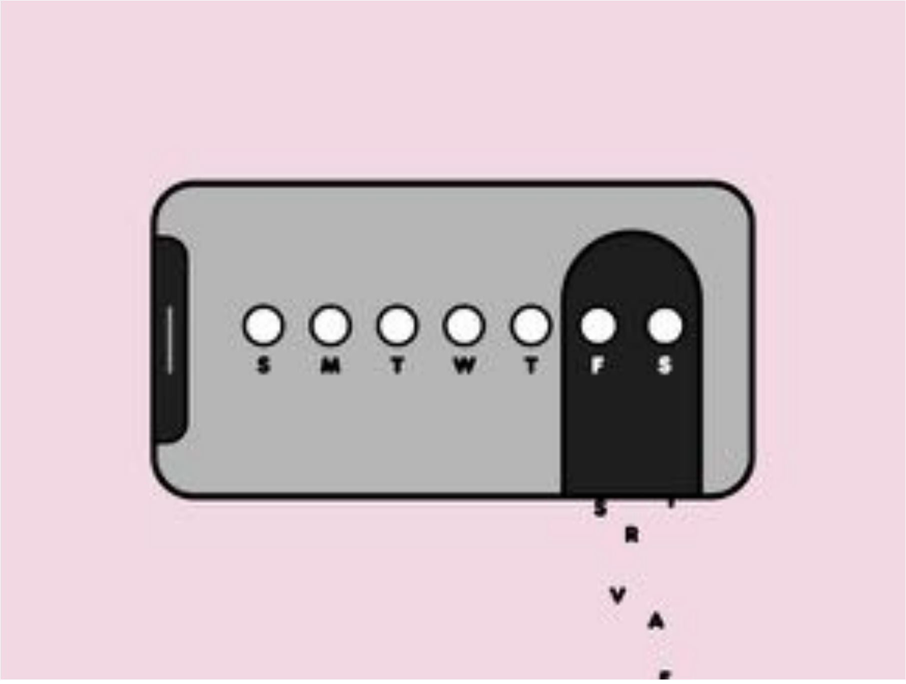
U.S. Congress, Joint Committee on Taxation, 2013. Estimates of Federal Tax Expenditures for Fiscal Years 2012–2017. U.S. Government Printing Office, Washington, DC, pp. 30–44, Table 1.

Vernon, J.A., 2004. Pharmaceutical R&D investments and cash flows. Journal of Pharmaceutical Finance, Economics & Policy 13 (4), 35–51.

Viereck, C., Boudes, P., 2011. An analysis of the impact of FDA's guidelines for addressing cardiovascular risk of drugs for type 2 diabetes on clinical development. Contemporary Clinical Trials 32 (3), 324–332.

MEGAN MOLTENI SCIENCE 08.20.18 03:27 PM

BEFORE USING BIRTH CONTROL APPS, CONSIDER YOUR PRIVACY



HOTLITTLEPOTATO

SCORE ONE FOR the quantified self-surveillance movement. Last week, the US Food and Drug Agency approved the first-ever, over-the-counter digital contraceptive—a polished and almost preternaturally upbeat mobile app called Natural Cycles. Basal body temperature readings and monthly menstruation data feed into an algorithm that tells users whether or not they should be having unprotected sex. Like most forms of birth control, it's not foolproof; the app has been dogged by reports of unwanted pregnancies that prompted two ongoing investigations by European authorities into its Swedish maker's marketing claims.

But that hasn't hurt Natural Cycles' popularity.

The app, which is available without a prescription, boasts more than 900,000 users, or "Cyclers" worldwide, according to the company. And with its new FDA approval—clearing the way for similar fertility and period-tracking apps—the company is expanding operations with a new US office in New York. It scientists hired away from CERN. (Natural Cycles' husband and wife co-founders Elina Berglund and Raoul Scherwitzl began work on the app while she was employed as a physicist at the Swiss particle-smashing facility and he was pursuing research at the University of Geneva.) The new team will search a growing collection of user information for insights, beyond contraception and planning a pregnancy, says Berglund. It's not exactly the Higgs boson, but it could turn out to be more lucrative.

Women's health apps are big business, you see. Users pay Natural Cycles a $10 monthly or $80 annual subscription fee, which includes an oral thermometer. But all that industrious tracking of periods, and sex, and basal body temperature—Natural Cycles takes one to three cycles to "get to know you"—is also valuable as a database. Even in its anonymized, aggregated form, pharmaceutical firms, the insurance industry, and marketing agencies are interested.

Natural Cycles' privacy policy states that in using the app each user grants the company and any of its partners broad rights to "use, reproduce, distribute, modify, adapt, prepare derivative works of, publicly display, publicly perform, communicate to the public, and otherwise utilize and exploit a user's anonymized information."

It's not that different from the privacy policies of other consumer apps, says Christine Bannan, consumer privacy counsel for the Electronic Privacy Information Center. Other popular cycle-tracking apps like Clue and Glow also reserve the right to share pooled, anonymized data with third parties. "These policies are just used by companies as disclaimers to reserve future things they might want to do," she says. But the sensitivity of fertility information makes that potentially more concerning, than say social media data, says Bannan. "I think that it's important for potential users to be aware that they don't necessarily have the rights they would for health data like traditional medical records, under HIPAA."

Berglund says Natural Cycles' only revenue stream at the moment is the app's subscription service, and that selling customer data to third parties isn't part of the company's business plan. "We've never shared any data for financial purposes," she says. But that may not always be the case. "I can't say we'll never share data, there's no guarantees in life of what will happen."

regulators like the FDA, and with academic research partners in Sweden, the UK, and the US, according to Berglund. She says they have to seek approval from an ethical review board for each research project, to evaluate whether or not the blanket consent users sign to use the service can be applied to any additional studies.

Natural Cycles stores user data in an encrypted cloud environment, and every week a pooled, anonymized version of the data gets pulled onto the company's local servers to run the analysis that powers its app. So if you decide you want to delete your data, it should get scrubbed from the cloud first, and then from the company's models, during that weekly overwriting process, according to Berglund. But according to the company's privacy policies, it's under no obligation to delete any data it has already distributed elsewhere.

"Having self-knowledge is not inherently a bad thing," says Karen Levy, a sociologist and lawyer at Cornell University who studies the social, legal, and ethical dimensions of emerging technologies. While fertility apps can have real upsides, she encourages people to be aware that any data they collect about themselves—from their history of prescription contraceptives to how often they have sex—can go on to have another life of its own. "Before you sign up for an app like this you have to ask yourself if you if this information is something you want to exist forever, to someday be combined with other data about you for research or marketing, because those possibilities are definitely in the game."

In the DNA testing industry, some customers were surprised last month when 23andMe announced a partnership with pharmaceutical giant GlaxoSmithKline to mine its customer database for potential new drug targets. Natural Cycles is also pursuing a pharmaceutical partnership, with Merck, but under very different terms.

The two companies are collaborating on an early stage pilot in Sweden to investigate whether Natural Cycles' data can predict early signs of infertility. One big clue comes from the follicular phase—the time interval between menstruation and ovulation. Natural Cycles' scientists have picked up some signals that suggest a shorter follicular phase correlates with lower fertility. They've also noticed that users who who are coming off of hormonal birth in getting pregnant. Berglund says there's no commercial deal yet, though, and no data has traded hands. But the idea of the pilot is to help the right women seek help at fertility clinics, as early as the data will allow. Merck is a major world supplier of fertility treatment drugs.

More Great WIRED Stories

- Waiting for Group FaceTime? There are plenty of options
- This wild avalanche animation could save your life
- How to actually stop Google from tracking your location
- A guide to finding your ideal movie ticket subscription
- The super-secret sand that makes your phone possible
- Looking for more? Sign up for our daily newsletter and never miss our latest and greatest stories

RELATED VIDEO

The Guardian

Interview

Cardiologist Eric Topol: 'AI can restore the care in healthcare'

Nicola Davis

The doctor, geneticist and author talks about his new book on the future of our relationship with medicine

Sun 7 Jul 2019 11.00 BST

Eric Topol: 'How can we have better bonding, accuracy and precision in our care?' Photograph: Zuma Press Inc/Alamy

Eric Topol is an American cardiologist and geneticist - among his many roles he is founder and director of the Scripps Research Translational Institute in California. He has previously published two books on the potential for big data and tech to transform medicine, with his third, *Deep Medicine,* looking at the role that artificial intelligence might play. He has served on the advisory boards of many healthcare companies, and last year published a report into how the NHS needs to change if it is to embrace digital advances.

Your field is cardiology - what makes you tick as a doctor?
Well, the patients. But also the broader mission. I was in clinic all day yesterday - I love seeing patients - but I also try to use whatever resources I can, to think about how can we do things better, how can we have much better bonding, accuracy and precision in our care.

What's the most promising medical application for artificial intelligence?
In the short term, taking images and having far superior accuracy and speed - not that it would supplant a doctor, but rather that it would be a first pass, an initial screen with

oversight by a doctor. So whether it is a medical scan or a pathology slide or a skin lesion or a colon polyp - that is the short-term story.

You talk about a future where people are constantly having parameters monitored - how promising is that?
You're ahead of the curve there in the UK. If you think you might have a urinary tract infection, you can go to the pharmacy, get an AI kit that accurately diagnoses your UTI and get an antibiotic - and you never have to see a doctor. You can get an Apple Watch that will detect your heart rate, and when something is off the track it will send you an alert to take your cardiogram.

Is there a danger that this will mean more people become part of the "worried well"?
It is even worse now because people do a Google search, then think they have a disease and are going to die. At least this is *your* data so it has a better chance of being meaningful.

It is not for everyone. But even if half the people are into this, it is a major decompression on what doctors are doing. It's not for life-threatening matters, such as a diagnosis of cancer or a new diagnosis of heart disease. It's for the more common problems - and for most of these, if people want, there is going to be AI diagnosis without a doctor.

If you had an AI GP - it could listen and respond to patients' descriptions of their symptoms but would it be able to physically examine them?
I don't think that you could simulate a real examination. But you could get select parts done - for example, there have been recent AI studies of children with a cough, and just by the AI interpretation of that sound, you could accurately diagnose the type of lung problem that it is.

Smartphones can be used as imaging devices with ultrasound, so someday there could be an inexpensive ultrasound probe. A person could image a part of their body, send that image to be AI-interpreted, and then discuss it with a doctor.

One of the big ones is eyegrams, of the retina. You will be able to take a picture of your retina, and find out if your blood pressure is well controlled, if your diabetes is well controlled, if you have the beginnings of diabetic retinopathy or macular degeneration - that is an exciting area for patients who are at risk.

What are the biggest technical and practical obstacles to using AI in healthcare?
Well, there are plenty, a long list - privacy, security, the biases of the algorithms, inequities - and making them worse because AI in healthcare is catering only to those who can afford it.

You talk about how AI might be able to spot people who have, or are at risk of developing, mental health problems from analysis of social media messages. How would this work and how do you prevent people's mental health being assessed without their permission?
I wasn't suggesting social media be the only window into a person's state of mind. Today mental health can be objectively defined, whereas in the past it was highly subjective. We are talking about speech pattern, tone, breathing pattern - when people sigh a lot, it denotes depression - physical activity, how much people move around, how much they communicate.

And then it goes on to facial recognition, social media posts, and other vital signs such as heart rate and heart rhythm, so the collection of all these objective metrics can be used to track a person's mood state - and in people who are depressed, it can help show what is working to get them out of that state, and help in predicting the risk of suicide.

Objective methods are doing better than psychologists or psychiatrists in predicting who is at risk, so I think there is a lot of promise for mental health and AI.

If AI gets a diagnosis or treatment badly wrong, who gets sued? The author of the software or the doctor or hospital that provides it?
There aren't any precedents yet. When you sign up with an app you are waiving all rights to legal recourse. People never read the terms and conditions of course. So the company could still be liable because there isn't any real consent. For the doctors involved, it depends on where that interaction is. What we do know is that there is a horrible problem with medical errors today. So if we can clean that up and make them far fewer, that's moving in the right direction.

You were commissioned by Jeremy Hunt in 2018 to carry out a review of how the NHS workforce will need to change "to deliver a digital future". What was the biggest change you recommended?
I think the biggest change was to try and accelerate the incorporation of AI to give the gift of time - to get back the patient-doctor relationship that we all were a part of 30, 40-plus years ago. There is a new, unprecedented opportunity to seize this and restore the care in healthcare that has been largely lost.

In the US, various Democratic candidates for 2020 are suggesting a government-backed system - a bit like our NHS. Would this allow AI in healthcare to flourish without insurers discriminating against patients with "bad data"and allow AI to fulfil its promise?
Well I think it certainly helps. If you have a single system where you implement AI and you have all the data in a common source, it is just much more liable to succeed. The NHS efficiency of providing care with better outcomes than the US at a lower cost per person, that is a lot about the fact you have got a superior model.

- *Deep Medicine by Eric Topol is published by Basic Books (£25). To order a copy for £22 go to guardianbookshop.com. Free UK p&p on all online orders over £15*

Topics

- Science
- The Observer
- Health, mind and body books
- Science and nature books
- interviews

Exhalation

Ted Chiang

It has long been said that air (which others call argon) is the source of life. This is not in fact the case, and I engrave these words to describe how I came to understand the true source of life and, as a corollary, the means by which life will one day end.

For most of history, the proposition that we drew life from air was so obvious that there was no need to assert it. Every day we consume two lungs heavy with air; every day we remove the empty ones from our chest and replace them with full ones. If a person is careless and lets his air level run too low, he feels the heaviness of his limbs and the growing need for replenishment. It is exceedingly rare that a person is unable to get at least one replacement lung before his installed pair runs empty; on those unfortunate occasions where this has happened—when a person is trapped and unable to move, with no one nearby to assist him—he dies within seconds of his air running out.

But in the normal course of life, our need for air is far from our thoughts, and indeed many would say that satisfying that need is the least important part of going to the filling stations. For the filling stations are the primary venue for social conversation, the places from which we draw emotional sustenance as well as physical. We all keep spare sets of full lungs in our homes, but when one is alone, the act of opening one's chest and replacing one's lungs can seem little better than a chore. In the company of others, however, it becomes a communal activity, a shared pleasure.

If one is exceedingly busy, or feeling unsociable, one might simply pick up a pair of full lungs, install them, and leave one's emptied lungs on the other side of the room. If one has a few minutes to spare, it's simple courtesy to connect the empty lungs to an air dispenser and refill them for the next person. But by far the most common practice is to linger and enjoy the company of others, to discuss the news of the day with friends or acquaintances and, in passing, offer newly filled lungs to one's interlocutor. While this perhaps does not constitute air sharing in the strictest sense, there is camaraderie derived from the awareness that all our air comes from the same source, for the dispensers are but the exposed terminals of pipes extending from the reservoir of air deep underground, the great lung of the world, the source of all our nourishment.

Many lungs are returned to the same filling station the next day, but just as many circulate to other stations when people visit neighboring districts; the lungs are all identical in appearance, smooth cylinders of aluminum, so one cannot tell whether a given lung has always stayed close to home or whether it has traveled long distances. And just as lungs are passed between persons and districts, so are news and gossip. In this way one can receive news from remote districts, even those at the very edge of the world, without needing to leave home, although I myself enjoy traveling. I have journeyed all the way to the edge of the world, and seen the solid chromium wall that extends from the ground up into the infinite sky.

It was at one of the filling stations that I first heard the rumors that prompted my investigation and led to my eventual enlightenment. It began innocently enough, with a remark from our district's public crier. At noon of the first day of every year, it is traditional for the crier to recite a passage of verse, an ode composed long ago for this annual celebration, which takes exactly one hour to deliver. The crier mentioned that on his most recent performance, the turret clock struck the hour before he had finished, something that had never happened before. Another person remarked that this was a coincidence, because he had just returned from a nearby district where the public crier had complained of the same incongruity.

No one gave the matter much thought beyond the simple acknowledgement that seemed warranted. It was only some days later, when there arrived word of a similar deviation between the crier and the clock of a third district, that the suggestion was made that these discrepancies might be evidence of a defect in the mechanism common to all the turret clocks, albeit a curious one to cause the clocks to run faster rather than slower. Horologists investigated the turret clocks in question, but on inspection they could discern no imperfection. In fact, when compared against the timepieces normally employed for such calibration purposes, the turret clocks were all found to have resumed keeping perfect time.

I myself found the question somewhat intriguing, but I was too focused on my own studies to devote much thought to other matters. I was and am a student of anatomy, and to provide context for my subsequent actions, I now offer a brief account of my relationship with the field.

Death is uncommon, fortunately, because we are durable and fatal mishaps are rare, but it makes difficult the study of anatomy, especially since many of the accidents serious enough to cause death leave the deceased's remains too damaged for study. If lungs are ruptured when full, the explosive force can tear a body asunder, ripping the titanium as easily as if it were tin. In the past, anatomists focused their attention on the limbs, which were the most likely to survive intact. During the very first anatomy lecture I attended a century ago, the lecturer showed us a severed arm, the casing removed to reveal the dense column of rods and pistons within. I can vividly recall the way, after he had connected its arterial hoses to a wall-mounted lung he kept in the laboratory, he was able to manipulate the actuating rods that protruded from the arm's ragged base, and in response the hand would open and close fitfully.

In the intervening years, our field has advanced to the point where anatomists are able to repair damaged limbs and, on occasion, attach a severed limb. At the same time we have become capable of studying the physiology of the living; I have given a version of that first lecture I saw, during which I opened the casing of my own arm and directed my students' attention to the rods that contracted and extended when I wiggled my fingers.

Despite these advances, the field of anatomy still had a great unsolved mystery at its core: the question of memory. While we knew a little about the structure of the brain, its physiology is notoriously hard to study because of the brain's extreme delicacy. It is typically the case in fatal accidents that, when the skull is breached, the brain erupts in a cloud of gold, leaving little besides shredded filament and leaf from which nothing useful can be discerned. For decades the prevailing theory of memory was that all of a person's experiences were engraved on sheets of gold foil; it was these sheets, torn apart by the force of the blast, that were the source of the tiny flakes found after accidents. Anatomists would collect the bits of gold leaf—so thin that light passes greenly through them—and spend years trying to reconstruct the original sheets, with the hope of eventually deciphering the symbols in which the deceased's recent experiences were inscribed.

I did not subscribe to this theory, known as the inscription hypothesis, for the simple reason that if all our experiences are in fact recorded, why is it that our memories are incomplete? Advocates of the inscription hypothesis offered an explanation for forgetfulness—suggesting that over time the foil sheets become misaligned from the stylus which reads the memories, until the oldest sheets shift out of contact with it altogether—but I never found it convincing. The appeal of the theory was easy for me to appreciate, though; I too had devoted many an hour to examining flakes of gold through a microscope, and can imagine how gratifying it would be to turn the fine adjustment knob and see legible symbols come into focus.

More than that, how wonderful would it be to decipher the very oldest of a deceased person's memories, ones that he himself had forgotten? None of us can remember much more than a hundred years in the past, and written records—accounts that we ourselves inscribed but have scant memory of doing so—extend only a few hundred years before that. How many years did we live before the beginning of written history? Where did we come from? It is the promise of finding the answers within our own brains that makes the inscription hypothesis so seductive.

I was a proponent of the competing school of thought, which held that our memories were stored in some medium in which the process of erasure was no more difficult than recording: perhaps in the rotation of gears, or the positions of a series of switches. This theory implied that everything we had forgotten was indeed lost, and our brains contained no histories older than those found in our libraries. One advantage of this theory was that it better explained why, when lungs are installed in those who have died from lack of air, the revived have no memories and are all but mindless: Somehow the shock of death had reset all the gears or switches. The inscriptionists claimed the shock had merely misaligned the foil sheets, but no one was willing to kill a living person, even an imbecile, in order to resolve the debate. I had envisioned an experiment which might allow me to determine the truth conclusively, but it was a risky one, and deserved careful consideration before it was undertaken. I remained undecided for the longest time, until I heard more news about the clock anomaly.

Word arrived from a more distant district that its public crier had likewise observed the turret clock striking the hour before he had finished his new year's recital. What made this notable was that his district's clock employed a different mechanism, one in which the hours were marked by the flow of mercury into a bowl. Here the discrepancy could not be explained by a common mechanical fault. Most people suspected fraud, a practical joke perpetrated by mischief makers. I had a different suspicion, a darker one that I dared not voice, but it decided my course of action; I would proceed with my experiment.

The first tool I constructed was the simplest: in my laboratory I fixed four prisms on mounting brackets and carefully aligned them so that their apexes formed the corners of a rectangle. When arranged thus, a beam of light directed at one of the lower prisms was reflected up, then backward, then down, and then forward again in a quadrilateral loop. Accordingly, when I sat with my eyes at the level of the first prism, I obtained a clear view of the back of my own head. This solipsistic periscope formed the basis of all that was to come.

A similarly rectangular arrangement of actuating rods allowed a displacement of action to accompany the displacement of vision afforded by the prisms. The bank of actuating rods was much larger than the periscope, but still relatively straightforward in design; by contrast, what was attached to the end of these respective mechanisms was far more intricate. To the periscope I added a binocular microscope mounted on an armature capable of swiveling side to side or up and down. To the actuating rods I added an array of precision manipulators, although that description hardly does justice to those pinnacles of the mechanician's art. Combining the ingenuity of anatomists and the inspiration provided by the bodily structures they studied, the manipulators enabled their operator to accomplish any task he might normally perform with his own hands, but on a much smaller scale.

Assembling all of this equipment took months, but I could not afford to be anything less than meticulous. Once the preparations were complete, I was able to place each of my hands on a nest of knobs and levers and control a pair of manipulators situated behind my head, and use the periscope to see what they worked on. I would then be able to dissect my own brain.

The very idea must sound like pure madness, I know, and had I told any of my colleagues, they would surely have tried to stop me. But I could not ask anyone else to risk themselves for the sake of anatomical inquiry, and because I wished to conduct the dissection myself, I would not be satisfied by merely being the passive subject of such an operation. Auto-dissection was the only option.

I brought in a dozen full lungs and connected them with a manifold. I mounted this assembly beneath the worktable that I would sit at, and positioned a dispenser to connect directly to the bronchial inlets within my chest. This would supply me with six days' worth of air. To provide for the possibility that I might not have completed my experiment within that period, I had scheduled a visit from a colleague at the end of that time. My presumption, however, was that the only way I would not have finished the operation in that period would be if I had caused my own death.

I began by removing the deeply curved plate that formed the back and top of my head; then the two, more shallowly curved plates that formed the sides. Only my faceplate remained, but it was locked into a restraining bracket, and I could not see its inner surface from the vantage point of my periscope; what I saw exposed was my own brain. It consisted of a dozen or more subassemblies, whose exteriors were covered by intricately molded shells; by positioning the periscope near the fissures that separated them, I gained a tantalizing glimpse at the fabulous mechanisms within their interiors. Even with what little I could see, I could tell it was the most beautifully complex engine I had ever beheld, so far beyond any device man had constructed that it was incontrovertibly of divine origin. The sight was both exhilarating and dizzying, and I savored it on a strictly aesthetic basis for several minutes before proceeding with my explorations.

It was generally hypothesized that the brain was divided into an engine located in the center of the head which performed the actual cognition, surrounded by an array of components in which memories were stored. What I observed was consistent with this theory, since the peripheral subassemblies seemed to resemble one another, while the subassembly in the center appeared to be different, more heterogeneous and with more moving parts. However the components were packed too closely for me to see much of their operation; if I intended to learn anything more, I would require a more intimate vantage point.

Each subassembly had a local reservoir of air, fed by a hose extending from the regulator at the base of my brain. I focused my periscope on the rearmost subassembly and, using the remote manipulators, I quickly disconnected the outlet hose and installed a longer one in its place. I had practiced this maneuver countless times so that I could perform it in a matter of moments; even so, I was not certain I could complete the connection before the subassembly had depleted its local reservoir. Only after I was satisfied that the component's operation had not been interrupted did I continue; I rearranged the longer hose to gain a better view of what lay in the fissure behind it: other hoses that connected it to its neighboring components. Using the most slender pair of manipulators to reach into the narrow crevice, I replaced the hoses one by one with longer substitutes. Eventually, I had worked my way around the entire subassembly and replaced every connection it had to the rest of my brain. I was now able to unmount this subassembly from the frame that supported it, and pull the entire section outside of what was once the back of my head.

I knew it was possible I had impaired my capacity to think and was unable to recognize it, but performing some basic arithmetic tests suggested that I was uninjured. With one subassembly hanging from a scaffold above, I now had a better view of the cognition engine at the center of my brain, but there was not enough room to bring the microscope attachment itself in for a close inspection. In order for me to really examine the workings of my brain, I would have to displace at least half a dozen subassemblies.

Laboriously, painstakingly, I repeated the procedure of substituting hoses for other subassemblies, repositioning another one farther back, two more higher up, and two others out to the sides, suspending all six from the scaffold above my head. When I was done, my brain looked like an explosion frozen an infinitesimal fraction of a second after the detonation, and again I felt dizzy when I thought about it. But at last the cognition engine itself was exposed, supported on a pillar of hoses and actuating rods leading down into my torso. I now also had room to rotate my microscope around a full three hundred and sixty degrees, and pass my gaze across the inner faces of the subassemblies I had moved. What I saw was a microcosm of auric machinery, a landscape of tiny spinning rotors and miniature reciprocating cylinders.

As I contemplated this vista, I wondered, where was my body? The conduits which displaced my vision and action around the room were in principle no different from those which connected my original eyes and hands to my brain. For the duration of this experiment, were these manipulators not essentially my hands? Were the magnifying lenses at the end of my periscope not essentially my eyes? I was an everted person, with my tiny, fragmented body situated at the center of my own distended brain. It was in this unlikely configuration that I began to explore myself.

I turned my microscope to one of the memory subassemblies, and began examining its design. I had no expectation that I would be able to decipher my memories, only that I might divine the means by which they were recorded. As I had predicted, there were no reams of foil pages visible, but to my surprise neither did I see banks of gearwheels or switches. Instead, the subassembly seemed to consist almost entirely of a bank of air tubules. Through the interstices between the tubules I was able to glimpse ripples passing through the bank's interior.

With careful inspection and increasing magnification, I discerned that the tubules ramified into tiny air capillaries, which were interwoven with a dense latticework of wires on which gold leaves were hinged. Under the influence of air escaping from the capillaries, the leaves were held in a variety of positions. These were not switches in the conventional sense, for they did not retain their position without a current of air to support them, but I hypothesized that these were the switches I had sought, the medium in which my memories were recorded. The ripples I saw must have been acts of recall, as an arrangement of leaves was read and sent back to the cognition engine.

Armed with this new understanding, I then turned my microscope to the cognition engine. Here too I observed a latticework of wires, but they did not bear leaves suspended in position; instead the leaves flipped back and forth almost too rapidly to see. Indeed, almost the entire engine appeared to be in motion, consisting more of lattice than of air capillaries, and I wondered how air could reach all the gold leaves in a coherent manner. For many hours I scrutinized the leaves, until I realized that they themselves were playing the role of capillaries; the leaves formed temporary conduits and valves that existed just long enough to redirect air at other leaves in turn, and then disappeared as a result. This was an engine undergoing continuous transformation, indeed modifying itself as part of its operation. The lattice was not so much a machine as it was a page on which the machine was written, and on which the machine itself ceaselessly wrote.

My consciousness could be said to be encoded in the position of these tiny leaves, but it would be more accurate to say that it was encoded in the ever-shifting pattern of air driving these leaves. Watching the oscillations of these flakes of gold, I saw that air does not, as we had always assumed, simply provide power to the engine that realizes our thoughts. Air is in fact the very medium of our thoughts. All that we are is a pattern of air flow. My memories were inscribed, not as grooves on foil or even the position of switches, but as persistent currents of argon.

In the moments after I grasped the nature of this lattice mechanism, a cascade of insights penetrated my consciousness in rapid succession. The first and most trivial was understanding why gold, the most malleable and ductile of metals, was the only material out of which our brains could be made. Only the thinnest of foil leaves could move rapidly enough for such a mechanism, and only the most delicate of filaments could act as hinges for them. By comparison, the copper burr raised by my stylus as I engrave these words and brushed from the sheet when I finish each page is as coarse and heavy as scrap. This truly was a medium where erasing and recording could be performed rapidly, far more so than any arrangement of switches or gears.

What next became clear was why installing full lungs into a person who has died from lack of air does not bring him back to life. These leaves within the lattice remain balanced between continuous cushions of air. This arrangement lets them flit back and forth swiftly, but it also means that if the flow of air ever ceases, everything is lost; the leaves all collapse into identical pendent states, erasing the patterns and the consciousness they represent. Restoring the air supply cannot recreate what has evanesced. This was the price of speed; a more stable medium for storing patterns would mean that our consciousnesses would operate far more slowly.

It was then that I perceived the solution to the clock anomaly. I saw that the speed of these leaves' movements depended on their being supported by air; with sufficient air flow, the leaves could move nearly frictionlessly. If they were moving more slowly, it was because they were being subjected to more friction, which could occur only if the cushions of air that supported them were thinner, and the air flowing through the lattice was moving with less force.

It is not that the turret clocks are running faster. What is happening is that our brains are running slower. The turret clocks are driven by pendulums, whose tempo never varies, or by the flow of mercury through a pipe, which does not change. But our brains rely on the passage of air, and when that air flows more slowly, our thoughts slow down, making the clocks seem to us to run faster.

I had feared that our brains might be growing slower, and it was this prospect that had spurred me to pursue my auto-dissection. But I had assumed that our cognition engines—while powered by air—were ultimately mechanical in nature, and some aspect of the mechanism was gradually becoming deformed through fatigue, and thus responsible for the slowing. That would have been dire, but there was at least the hope that we might be able to repair the mechanism, and restore our brains to their original speed of operation.

But if our thoughts were purely patterns of air rather than the movement of toothed gears, the problem was much more serious, for what could cause the air flowing through every person's brain to move less rapidly? It could not be a decrease in the pressure from our filling stations' dispensers; the air pressure in our lungs is so high that it must be stepped down by a series of regulators before reaching our brains. The diminution in force, I saw, must arise from the opposite direction: The pressure of our surrounding atmosphere was increasing.

How could this be? As soon as the question formed, the only possible answer became apparent: Our sky must not be infinite in height. Somewhere above the limits of our vision, the chromium walls surrounding our world must curve inward to form a dome; our universe is a sealed chamber rather than an open well. And air is gradually accumulating within that chamber, until it equals the pressure in the reservoir below.

This is why, at the beginning of this engraving, I said that air is not the source of life. Air can neither be created nor destroyed; the total amount of air in the universe remains constant, and if air were all that we needed to live, we would never die. But in truth the source of life is a difference in air pressure, the flow of air from spaces where it is thick to those where it is thin. The activity of our brains, the motion of our bodies, the action of every machine we have ever built is driven by the movement of air, the force exerted as differing pressures seek to balance each other out. When the pressure everywhere in the universe is the same, all air will be motionless, and useless; one day we will be surrounded by motionless air and unable to derive any benefit from it.

We are not really consuming air at all. The amount of air that I draw from each day's new pair of lungs is exactly as much as seeps out through the joints of my limbs and the seams of my casing, exactly as much as I am adding to the atmosphere around me; all I am doing is converting air at high pressure to air at low. With every movement of my body, I contribute to the equalization of pressure in our universe. With every thought that I have, I hasten the arrival of that fatal equilibrium.

Had I come to this realization under any other circumstance, I would have leapt up from my chair and ran into the streets, but in my current situation—body locked in a restraining bracket, brain suspended across my laboratory—doing so was impossible. I could see the leaves of my brain flitting faster from the tumult of my thoughts, which in turn increased my agitation at being so restrained and immobile. Panic at that moment might have led to my death, a nightmarish paroxysm of simultaneously being trapped and spiraling out of control, struggling against my restraints until my air ran out. It was by chance as much as by intention that my hands adjusted the controls to avert my periscopic gaze from the latticework, so all I could see was the plain surface of my worktable. Thus freed from having to see and magnify my own apprehensions, I was able to calm down. When I had regained sufficient composure, I began the lengthy process of reassembling myself. Eventually I restored my brain to its original compact configuration, reattached the plates of my head, and released myself from the restraining bracket.

At first the other anatomists did not believe me when I told them what I had discovered, but in the months that followed my initial auto-dissection, more and more of them became convinced. More examinations of people's brains were performed, more measurements of atmospheric pressure were taken, and the results were all found to confirm my claims. The background air pressure of our universe was indeed increasing, and slowing our thoughts as a result.

There was widespread panic in the days after the truth first became widely known, as people contemplated for the first time the idea that death was inevitable. Many called for the strict curtailment of activities in order to minimize the thickening of our atmosphere; accusations of wasted air escalated into furious brawls and, in some districts, deaths. It was the shame of having caused these deaths, together with the reminder that it would be many centuries yet before our atmosphere's pressure became equal to that of the reservoir underground, that caused the panic to subside. We are not sure precisely how many centuries it will take; additional measurements and calculations are being performed and debated. In the meantime, there is much discussion over how we should spend the time that remains to us.

One sect has dedicated itself to the goal of reversing the equalization of pressure, and found many adherents. The mechanicians among them constructed an engine that takes air from our atmosphere and forces it into a smaller volume, a process they called “compression.” Their engine restores air to the pressure it originally had in the reservoir, and these Reversalists excitedly announced that it would form the basis of a new kind of filling station, one that would—with each lung it refilled—revitalize not only individuals but the universe itself. Alas, closer examination of the engine revealed its fatal flaw. The engine itself is powered by air from the reservoir, and for every lungful of air that it produces, the engine consumes not just a lungful, but slightly more. It does not reverse the process of equalization, but like everything else in the world, exacerbates it.

Although some of their adherents left in disillusionment after this setback, the Reversalists as a group were undeterred, and began drawing up alternate designs in which the compressor was powered instead by the uncoiling of springs or the descent of weights. These mechanisms fared no better. Every spring that is wound tight represents air released by the person who did the winding; every weight that rests higher than ground level represents air released by the person who did the lifting. There is no source of power in the universe that does not ultimately derive from a difference in air pressure, and there can be no engine whose operation will not, on balance, reduce that difference.

The Reversalists continue their labors, confident that they will one day construct an engine that generates more compression than it uses, a perpetual power source that will restore to the universe its lost vigor. I do not share their optimism; I believe that the process of equalization is inexorable. Eventually, all the air in our universe will be evenly distributed, no denser or more rarefied in one spot than in any other, unable to drive a piston, turn a rotor, or flip a leaf of gold foil. It will be the end of pressure, the end of motive power, the end of thought. The universe will have reached perfect equilibrium.

Some find irony in the fact that a study of our brains revealed to us not the secrets of the past, but what ultimately awaits us in the future. However, I maintain that we have indeed learned something important about the past. The universe began as an enormous breath being held. Who knows why, but whatever the reason, I am glad that it did, because I owe my existence to that fact. All my desires and ruminations are no more and no less than eddy currents generated by the gradual exhalation of our universe. And until this great exhalation is finished, my thoughts live on.

So that our thoughts may continue as long as possible, anatomists and mechanicians are designing replacements for our cerebral regulators, capable of gradually increasing the air pressure within our brains and keeping it just higher than the surrounding atmospheric pressure. Once these are installed, our thoughts will continue at roughly the same speed even as the air thickens around us. But this does not mean that life will continue unchanged. Eventually the pressure differential will fall to such a level that our limbs will weaken and our movements will grow sluggish. We may then try to slow our thoughts so that our physical torpor is less conspicuous to us, but that will also cause external processes to appear to accelerate. The ticking of clocks will rise to a chatter as their pendulums wave frantically;

falling objects will slam to the ground as if propelled by springs; undulations will race down cables like the crack of a whip.

At some point our limbs will cease moving altogether. I cannot be certain of the precise sequence of events near the end, but I imagine a scenario in which our thoughts will continue to operate, so that we remain conscious but frozen, immobile as statues. Perhaps we'll be able to speak for a while longer, because our voice boxes operate on a smaller pressure differential than our limbs, but without the ability to visit a filling station, every utterance will reduce the amount of air left for thought, and bring us closer to the moment that our thoughts cease altogether. Will it be preferable to remain mute to prolong our ability to think, or to talk until the very end? I don't know.

Perhaps a few of us, in the days before we cease moving, will be able to connect our cerebral regulators directly to the dispensers in the filling stations, in effect replacing our lungs with the mighty lung of the world. If so, those few will be able to remain conscious right up to the final moments before all pressure is equalized. The last bit of air pressure left in our universe will be expended driving a person's conscious thought.

And then, our universe will be in a state of absolute equilibrium. All life and thought will cease, and with them, time itself.

But I maintain a slender hope.

Even though our universe is enclosed, perhaps it is not the only air chamber in the infinite expanse of solid chromium. I speculate that there could be another pocket of air elsewhere, another universe besides our own that is even larger in volume. It is possible that this hypothetical universe has the same or higher air pressure as ours, but suppose that it had a much lower air pressure than ours, perhaps even a true vacuum?

The chromium that separates us from this supposed universe is too thick and too hard for us to drill through, so there is no way we could reach it ourselves, no way to bleed off the excess atmosphere from our universe and regain motive power that way. But I fantasize that this neighboring universe has its own inhabitants, ones with capabilities beyond our own. What if they were able to create a conduit between the two universes, and install valves to release air from ours? They might use our universe as a reservoir, running dispensers with which they could fill their own lungs, and use our air as a way to drive their own civilization.

It cheers me to imagine that the air that once powered me could power others, to believe that the breath that enables me to engrave these words could one day flow through someone else's body. I do not delude myself into thinking that this would be a way for me to live again, because I am not that air, I am the pattern that it assumed, temporarily. The pattern that is me, the patterns that are the entire world in which I live, would be gone.

But I have an even fainter hope: that those inhabitants not only use our universe as a reservoir, but that once they have emptied it of its air, they might one day be able to open a passage and actually enter our universe as explorers. They might wander our streets, see our frozen bodies, look through our possessions, and wonder about the lives we led.

Which is why I have written this account. You, I hope, are one of those explorers. You, I hope, found these sheets of copper and deciphered the words engraved on their surfaces. And whether or not your brain is impelled by the air that once impelled mine, through the act of reading my words, the patterns that form your thoughts become an imitation of the patterns that once formed mine. And in that way I live again, through you.

Your fellow explorers will have found and read the other books that we left behind, and through the collaborative action of your imaginations, my entire civilization lives again. As you walk through our silent districts, imagine them as they were; with the turret clocks striking the hours, the filling stations crowded with gossiping neighbors, criers reciting verse in the public squares and anatomists giving lectures in the classrooms. Visualize all of these the next time you look at the frozen world around you, and it will become, in your minds, animated and vital again.

I wish you well, explorer, but I wonder: Does the same fate that befell me await you? I can only imagine that it must, that the tendency toward equilibrium is not a trait peculiar to our universe but inherent in all universes. Perhaps that is just a limitation of my thinking, and your people have discovered a source of pressure that is truly eternal. But my speculations are fanciful enough already. I will assume that one day your thoughts too will cease, although I cannot fathom how far in the future that might be. Your lives will end just as ours did, just as everyone's must. No matter how long it takes, eventually equilibrium will be reached.

I hope you are not saddened by that awareness. I hope that your expedition was more than a search for other universes to use as reservoirs. I hope that you were motivated by a desire for knowledge, a yearning to see what can arise from a universe's exhalation. Because even if a universe's lifespan is calculable, the variety of life that is generated within it is not. The buildings we have erected, the art and music and verse we have composed, the very lives we've led: None of them could have been predicted, because none of them were inevitable. Our universe might have slid into equilibrium emitting nothing more than a quiet hiss. The fact that it spawned such plenitude is a miracle, one that is matched only by your universe giving rise to you.

Though I am long dead as you read this, explorer, I offer to you a valediction. Contemplate the marvel that is existence, and rejoice that you are able to do so. I feel I have the right to tell you this because, as I am inscribing these words, I am doing the same.

INNOVATE

Having a Good Idea Is Not Enough. Here's How to Turn Yours Into a Valuable Business

Steal these tricks from innovation consultants to help develop your idea into a million dollar business

By Annabel Acton *Founder, Never Liked It Anyway* @annabelacton

GETTY IMAGES

You have a killer business idea. You're excited to bring it to the world. But how do you make it happen? How does the seed of an idea actually help you know how to build a successful business? You don't need to race to raise capital or build out an overly sophisticated product to get going. In fact, this is an expensive and unnecessary approach, especially when an idea is in its infancy. Instead of racing towards developing a final product, you should take time to test the bones of your idea and refine the concept. Here are the key tricks and tips an innovation consultant uses to make sure the idea is ready for development.

1. Make Sure You Are Actually Solving A Problem

First things first, you must be solving a problem that actually matters to people. One big reason the Segway failed to be "As big as the PC", as Steve Jobs had predicted, was that it failed to solve a real need. Nobody was looking for a mode of transportation that went slightly faster than walking, without bag storage, or clarity on whether it was road or pavement worthy; except mall cops. Take time to distil exactly what it is that your idea is solving, and then figure out if it's something people care enough about to turn to your business. Often, the best business ideas are born out of personal frustration. For example, Richard Branson, decided to create Virgin after being fed up as an airline passenger. In true Branson fashion, he said, "Screw it, I can do it better than you." Use your personal insight as a starting point, and then corroborate from other people that you are actually solving a real need.

Try: Try creating an online survey, mining your social networks for data, hosting man-on-the-street interviews, exploring the 20/20/20 test, soaking up data and trend reports related to your category and scouring intel on competitors that both worked and failed.

2. Tell Everyone

There's a myth going around that if you have a blockbuster idea, you have to keep it all to yourself and not tell a soul. After all, they might steal your brilliance. However, having an idea and bringing an idea to life are two wildly different things. Ideation is easy, implementation is not.It requires bucket loads of tenacity, grit, patience, money, time and a little bit of luck. When you talk to people about you idea, you get valuable feedback, input and builds on the concept. No doubt someone will open your eyes to an angle you hadn't yet considered, or a feature that you wouldn't have deemed important or a marketing angle you may have overlooked entirely. These inputs are enormously helpful and best of all, they're free.

Try: Finding a mentor, pitching at a demo day, telling your friends, family and anyone who will listen.

3. List Your Assumptions

Before you commit to building a prototype or invest heavily in an MVP, start by listing all the assumptions that need to be true in order for your business to work. Create a long list, that walks through your idea from start to finish - and make sure to include even the most basic of assumptions. For example, assumptions for a restaurant booking service might include things like: people find booking restaurants hard, people would pay for someone to solve this problem, restaurants would be open to outsourcing bookings, people want to be rewarded for their participation. Once you have your assumption stack, start to think about how you might prioritize them in order importance; from critical to trivial.

Try: Build an assumption stack that spells out all the elements that need to be in place for you to be successful. Next prioritize the assumptions: which ones MUST hold true in order to be successful

4. Test Your Assumptions

Take your list of prioritized assumptions and find small ways to test them. This could include building an MVP or prototype that will get to the heart of your assumption stack. You can be smart about this too. For example, Tough Mudder CEO Will Dean grew Tough Mudder into a $100 Million business, from just $7,000 in savings. At the heart of his business lay an assumption that people would want to endure arduous conditions while exercising. To test his assumption, he pre-sold tickets to races and used the money raised to build the torturous obstacle course.

Try: Finding smart ways to test your assumption. Avoid building out a whole product if you can.

PUBLISHED ON: JUL 13, 2017

The opinions expressed here by Inc.com columnists are their own, not those of Inc.com.

High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision[1]

Manisha Bahl, MD, MPH
Regina Barzilay, PhD
Adam B. Yedidia, MEng
Nicholas J. Locascio, MEng
Lili Yu, PhD
Constance D. Lehman, MD, PhD

Purpose: To develop a machine learning model that allows high-risk breast lesions (HRLs) diagnosed with image-guided needle biopsy that require surgical excision to be distinguished from HRLs that are at low risk for upgrade to cancer at surgery and thus could be surveilled.

Materials and Methods: Consecutive patients with biopsy-proven HRLs who underwent surgery or at least 2 years of imaging follow-up from June 2006 to April 2015 were identified. A random forest machine learning model was developed to identify HRLs at low risk for upgrade to cancer. Traditional features such as age and HRL histologic results were used in the model, as were text features from the biopsy pathologic report.

Results: One thousand six HRLs were identified, with a cancer upgrade rate of 11.4% (115 of 1006). A machine learning random forest model was developed with 671 HRLs and tested with an independent set of 335 HRLs. Among the most important traditional features were age and HRL histologic results (eg, atypical ductal hyperplasia). An important text feature from the pathologic reports was "severely atypical." Instead of surgical excision of all HRLs, if those categorized with the model to be at low risk for upgrade were surveilled and the remainder were excised, then 97.4% (37 of 38) of malignancies would have been diagnosed at surgery, and 30.6% (91 of 297) of surgeries of benign lesions could have been avoided.

Conclusion: This study provides proof of concept that a machine learning model can be applied to predict the risk of upgrade of HRLs to cancer. Use of this model could decrease unnecessary surgery by nearly one-third and could help guide clinical decision making with regard to surveillance versus surgical excision of HRLs.

[1] From the Division of Breast Imaging, Department of Radiology, Massachusetts General Hospital/Harvard Medical School, 55 Fruit St, WAC 240, Boston, MA 02114 (M.B., C.D.L.); and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Mass (R.B., A.B.Y., N.J.L., L.Y.). Received March 8, 2017; revision requested April 11; revision received May 12; accepted June 8; final version accepted June 27. **Address correspondence to** M.B. (e-mail: *mbahl1@mgh.harvard.edu*).

Early detection of breast cancer with screening mammography reduces mortality from breast cancer and provides women diagnosed with breast cancer more options for less-aggressive treatment (1,2). Although the benefits of early detection of breast cancer are acknowledged widely, continued concerns are raised regarding potential harms associated with unnecessary biopsies and surgeries that are triggered by imaging findings in patients who do not have breast cancer (3,4). Up to 14% of image-guided biopsies performed on the basis of suspicious mammograms yield high-risk breast lesions (HRLs) (5,6). Most HRLs are benign, but surgical excision typically is recommended because of the low but present potential for upgrade to ductal carcinoma in situ or invasive malignancy at surgical excision (7,8). The resulting status quo is overtreatment with unnecessary surgery for HRLs that are not associated with malignancy.

Authors of multiple studies (7–25) have investigated patient and imaging features to better stratify patients with HRLs according to risk. Features considered included patient variables such as age and personal history of breast cancer; HRL histologic results such as atypical ductal hyperplasia (ADH); imaging variables such as lesion type at mammography; and image guidance and biopsy device used for sampling (ie, stereotactically vs ultrasonographically guided and small- vs large-gauge needle biopsy device). Despite these efforts, there are no definite features that reliably allow lesions that warrant surgical excision to be distinguished from those that can be surveilled safely, which has led to wide variation in treatment (7). At our institution, more than 95% of patients undergo surgical excision for HRLs diagnosed with image-guided core-needle biopsy; and, therefore, surgical outcomes are known for most of our patients.

Machine learning refers to algorithms that can be designed to evaluate and make predictions on the basis of new and complex features (26,27). A machine learning model that incorporates the full spectrum of patient data offers a means to stratify patients with HRLs diagnosed with core-needle biopsy according to risk and thereby reduce unnecessary surgical interventions. With an annotated training set, in which the surgical outcomes are known, the model can allow relationships in the provided data to be discovered and combinations of features that are accurate predictors of risk of cancer upgrade to be identified. Once the model is developed, it can then be applied to classify new cases in which the surgical outcomes are not known. To our knowledge, there are no studies in which the authors applied machine learning algorithms to this specific challenging clinical scenario of HRLs and incorporated the full spectrum of diverse and complex data available for risk-stratification purposes.

A machine learning model in the clinical setting could support informed decision making for patients and their providers regarding surveillance versus surgical excision of HRLs and could reduce unnecessary surgical excision of HRLs. The purpose of this study was to develop a machine learning model that allows HRLs diagnosed with image-guided needle biopsy that require surgical excision to be distinguished from HRLs that are at low risk for upgrade to cancer at surgery and thus could be surveilled.

Advances in Knowledge

- Our machine learning model, which was developed to help distinguish high-risk breast lesions (HRLs) that require surgical excision from those that could be surveilled, was based on established risk factors such as patient age and HRL histologic results (with the inclusion of more than 20000 data elements) and an additional feature of the biopsy pathologic report text.
- Instead of surgical excision of all HRLs, if HRLs categorized with our model to be at low risk for upgrade to cancer were surveilled and the remainder were excised, then 97.4% (37 of 38) of malignancies would be diagnosed at surgery, and 30.6% (91 of 297) of surgeries of benign lesions could be avoided.

Implications for Patient Care

- Our machine learning model integrates a diversity of complex features to identify women at low risk for upgrade to cancer after diagnosis of an HRL.
- Machine learning could inform shared decision making by the patient and the provider regarding surveillance versus surgical excision of HRLs and thus could support more targeted, personalized approaches to patient care.

Materials and Methods

Study Population

This study was approved by the institutional review board with a waiver for the need to obtain informed consent and was compliant with the Health Insurance Portability and Accountability Act. The study cohort comprised consecutive women at a tertiary academic medical center who underwent image-guided core-needle biopsy from June 1, 2006,

https://doi.org/10.1148/radiol.2017170549

Content codes: BR OI

Radiology 2018; 286:810–818

Abbreviations:
ADH = atypical ductal hyperplasia
HRL = high-risk breast lesion

Author contributions:
Guarantors of integrity of entire study, M.B., A.B.Y., N.J.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.B., L.Y., C.D.L.; clinical studies, M.B., C.D.L.; experimental studies, M.B., R.B., L.Y.; statistical analysis, M.B., R.B., A.B.Y., N.J.L., L.Y.; and manuscript editing, M.B., R.B., A.B.Y., L.Y., C.D.L.

Conflicts of interest are listed at the end of this article.

See also the editorial by Shaffer in this issue.

Figure 1

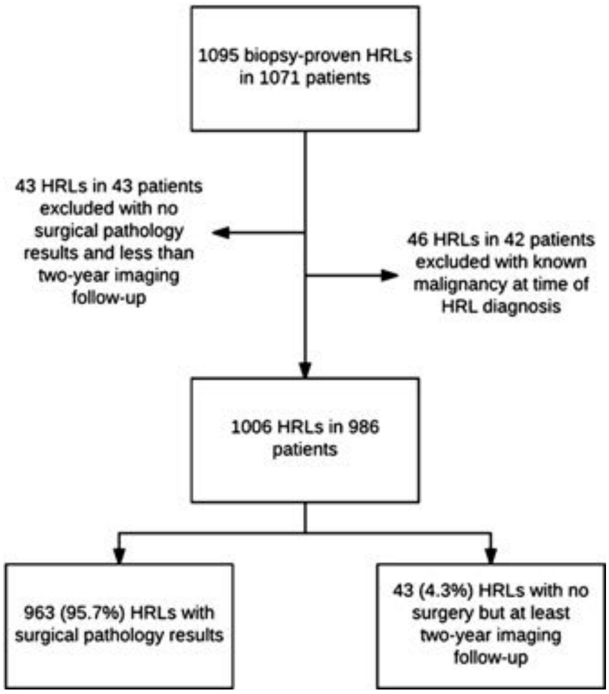

1095 biopsy-proven HRLs in 1071 patients
43 HRLs in 43 patients excluded with no surgical pathology results and less than two-year imaging follow-up
46 HRLs in 42 patients excluded with known malignancy at time of HRL diagnosis
1006 HRLs in 986 patients
963 (95.7%) HRLs with surgical pathology results
43 (4.3%) HRLs with no surgery but at least two-year imaging follow-up

Figure 1: Flow diagram shows patient selection.

to April 30, 2015, with pathologic results that yielded an HRL. HRLs included ADH, atypical lobular hyperplasia, biphasic neoplasms, flat epithelial atypia, lobular carcinoma in situ, nonspecific atypia, papillomas, and radial scars. All patients had a mammographic abnormality that led to the HRL diagnosis. During the study period, mammograms were obtained by using full-field digital mammography (2006–2012) or digital breast tomosynthesis (2011–2015) (Hologic, Bedford, Mass). Patients who underwent subsequent surgical excision or at least 2 years of imaging follow-up were included in the study cohort. Patients with known malignancy in the ipsilateral or contralateral breast at the time of HRL diagnosis were excluded.

One thousand seventy-one patients had mammographic lesions that led to image-guided biopsy and yielded 1095 HRLs. Forty-three HRLs in 43 patients were excluded because of lack of surgical pathologic results and less than 2 years of imaging follow-up, and 46 HRLs in 42 patients were excluded because of known malignancy at the time of HRL diagnosis (Fig 1). Thus, a total of 89 HRLs (89 of 1095, 8.1%) were excluded. The study cohort comprised 1006 HRLs in 986 patients with a mean age of 53 years (range, 24–87 years). A total of 20 patients had two HRL diagnoses at different time points (ie, two different biopsies) within the study period. Surgical pathologic results were available for 963 (95.7%) HRLs, and at least 2 years of imaging follow-up was available for the 43 (4.3%) lesions in patients who did not undergo surgical excision.

The histologic types and upgrade rates of HRLs in the training and test sets used for the machine learning model are presented in Table 1. Of the 1006 core-needle biopsies, 303 (30.1%) yielded more than one HRL (such as concomitant ADH and flat epithelial atypia), all of which were incorporated into the machine learning model. However, for core biopsies that yielded more than one HRL, the type of HRL with the highest risk was used for data presentation on the basis of the following hierarchy: ADH is greater than lobular carcinoma in situ, which is greater than atypical lobular hyperplasia, which is greater than radial scar, which is greater than papilloma, which is greater than flat epithelial atypia, which is greater than nonspecific atypia, which is greater than biphasic neoplasm. The most common HRL identified was ADH, which represented 37.1% (373 of 1006) of all HRLs, followed by flat epithelial atypia (18.1%, 182 of 1006). ADH had the highest rate of upgrade to malignancy (19.3%, 72 of 373), followed by lobular carcinoma in situ (17.4%, 12 of 69).

Data Collection and Statistical Analysis

Clinical information, mammographic reports, image-guided core-needle biopsy reports, and surgical pathologic reports were extracted from our institution's mammography information system (Magview, Burtonsville, Md). A structured database was developed with data about each patient, including information such as age, height, weight, race, personal history of breast cancer, family history of breast cancer, age at first pregnancy, age at first menses, and age at menopause (Table 2). Additional information that was extracted included mammographic findings (calcifications, mass, asymmetry, and architectural distortion), breast density, mode of biopsy, core biopsy pathologic results, and surgical pathologic results. All information extracted from the mammography, core biopsy, and surgical pathologic reports was manually validated by a fellowship-trained breast imaging radiologist (M.B., with 2 years of breast imaging experience).

If the surgical pathologic result was ductal carcinoma in situ or invasive carcinoma, the lesion was considered malignant and therefore represented an upgrade. Any surgical pathologic result other than ductal carcinoma in situ or invasive carcinoma was classified as benign. The relatively small number of patients (n = 43) who did not undergo surgical excision but had at least 2 years of imaging follow-up without mammographic findings suspicious for malignancy were also classified as benign.

All data were analyzed with a spreadsheet software program (Excel 2013; Microsoft, Redmond, Wash). *Z* tests (for proportions) were used to compare the training and test sets used for the machine learning model and to compare different strategies for surgical excision versus surveillance of HRLs. Ninety-five percent confidence intervals were calculated for the proportions of cancers detected and proportions of surgeries of benign lesions performed for each strategy. *P* values of less than .05 were considered to indicate a statistically significant difference.

Machine Learning Model

The machine learning model used for this study, the random forest classifier, is known for its robust performance and strong generalization power (26). The random forest model repeatedly selects a random subset of features from the training data set and constructs an ensemble of decision trees that allow correct classification of that sample of the training set with the use of a constructive algorithm. Each decision tree is built node by node, with each added node improving that tree's classification accuracy in that subset of features. To develop the random forest machine learning model, the dataset of 1006 HRLs was divided into two randomly chosen sets, a training set comprising two-thirds of the patient cohort and an independent test set comprising one-third of the patient cohort. Therefore, the model was trained with 671 HRLs in 654 patients and tested with 335 HRLs in 332 patients.

The model input features included traditional structural features such as age and HRL histologic results in addition to the full text of the core biopsy pathologic report. The traditional structural features are presented in Table 2. The text features were extracted by treating the presence or absence of each word (unigram) or combination of two adjacent words (bigram such as "suspicious calcifications") as a feature. We focused on the 100 most-important

Table 1

Histologic Types and Upgrade Rates of HRLs

	No. of Patients			Upgrade to DCIS			Upgrade to Invasive Carcinoma			Upgrade to DCIS or Invasive Carcinoma		
High-Risk Lesion	Training Set	Test Set	*P* Value	Training Set	Test Set	*P* Value	Training Set	Test Set	*P* Value	Training Set	Test Set	*P* Value
ADH	236/671 (35.2)	137/335 (40.9)	.08	37/236 (15.7)	21/137 (15.3)	.93	9/236 (3.8)	5/137 (3.6)	.94	46/236 (19.5)	26/137 (19.0)	.90
Flat epithelial atypia	128/671 (19.1)	54/335 (16.1)	.25	3/128 (2.3)	0/54 (0)	.26	1/128 (0.8)	0/54 (0)	.52	4/128 (3.1)	0/54 (0)	.19
Radial scar	83/671 (12.4)	36/335 (10.7)	.45	4/83 (4.8)	1/36 (2.8)	.61	0/83 (0)	0/36 (0)	NA	4/83 (4.8)	1/36 (2.8)	.61
Papilloma	55/671 (8.2)	29/335 (8.7)	.80	3/55 (5.5)	3/29 (10.3)	.41	1/55 (1.8)	1/29 (3.4)	.64	4/55 (7.3)	4/29 (13.8)	.33
Atypical lobular hyperplasia	49/671 (7.3)	30/335 (9.0)	.36	4/49 (8.2)	0/30 (0)	.11	2/49 (4.1)	1/30 (3.3)	.87	6/49 (12.2)	1/30 (3.3)	.18
Lobular carcinoma in situ	48/671 (7.2)	21/335 (6.3)	.60	2/48 (4.2)	1/21 (4.8)	.91	7/48 (14.6)	2/21 (9.5)	.57	9/48 (18.8)	3/21 (14.3)	.65
Nonspecific atypia	40/671 (6.0)	16/335 (4.8)	.44	0/40 (0)	1/16 (6.2)	.11	4/40 (10.0)	2/16 (12.5)	.79	4/40 (10.0)	3/16 (18.8)	.37
Biphasic neoplasm	32/671 (4.8)	12/335 (3.6)	.38	0/32 (0)	0/12 (0)	NA	0/32 (0)	0/12 (0)	NA	0/32 (0)	0/12 (0)	NA
Overall	. . .	. . .	. . .	53/671 (7.9)	27/335 (8.1)	.93	24/671 (3.6)	11/335 (3.3)	.81	77/671 (11.5)	38/335 (11.3)	.95

Note.—Unless otherwise indicated, data are proportion of patients, with percentage in parentheses. DCIS = ductal carcinoma in situ, NA = not applicable.

Table 2

List of Traditional Structural Features and Feature Classes

Structural Feature	Feature Class
Age	Numerical
Age at first menses	Numerical
Age at first pregnancy	Numerical
Age at menopause	Numerical
Ashkenazi Jewish ancestry	Binary
Biopsy type	Categorical
Breast density	Categorical
Breast Imaging Reporting and Data System category	Categorical
Drinking habits	Categorical
Family history of cancer	Numerical
Finding type	Categorical
First mammogram	Binary
Height	Numerical
Hormone treatments	Categorical
No. of children	Numerical
Pathologic result	Categorical
Previous breast cancer	Binary
Previous other cancer	Binary
Prior biopsies	Numerical
Procedure code	Categorical
Race	Categorical
Smoking habits	Categorical
Weight	Numerical

Table 3

Structural Features and Pathologic Text Features in the Machine Learning Model

Most Important Machine Learning Model Features
Structural features
Pathologic result (atypical ductal hyperplasia)
Age
Biopsy type (stereotactic core biopsy)
Pathologic result (lobular carcinoma in situ)
Pathologic result (atypical lobular hyperplasia)
Prior biopsy
Text features in the pathologic report
Atypical ductal
Severely
Atypical
Severely atypical

unigrams and bigrams as ranked with the mutual information criterion and used an ensemble of 200 random decision trees with a maximum depth of 12 to perform our classification (28). For each HRL in the independent test set, the model output was a score reflecting the likelihood of upgrade to malignancy at surgery. For a score greater than 5%, the model predicted surgical excision. For the remainder of the cases, surveillance rather than surgical excision could be considered.

Results

There were no statistically significant differences in the frequencies and upgrade rates of HRLs in the training and test sets used for the machine learning model (Table 1). Of note, 30.1% (303 of 1006) of core biopsies yielded more than one HRL (such as concomitant ADH and flat epithelial atypia). One hundred ninety (28.3%) patients had more than one HRL in the training set of 671 patients, and 113 (33.7%) had more than one HRL in the test set of 335 patients (P = .08). For the 1006 HRLs in this study, approximately 20 000 data elements based on traditional structural features were included in the model. The traditional structural features considered most important in the random forest machine learning model are listed in Table 3 and included such features as age and HRL histologic results. The pathologic report text features considered most important according to the model are also listed in Table 3 and included features such as “severely” and “severely atypical.”

Table 4 presents the model results for the independent test set of 335 HRLs stratified according to HRL histologic results compared with those of three other strategies: *(a)* the current practice at our institution, *(b)* excision of all HRLs, and *(c)* excision of ADH, lobular carcinoma in situ, and atypical lobular hyperplasia, which are considered higher-risk lesions, with surveillance of all other HRLs. Table 5 presents a statistical comparison of these strategies. If our machine learning model were used to identify HRLs with the potential for surveillance rather than surgical excision, then 97.4% (37 of 38) of malignancies would have been diagnosed at surgery, and 69.4% (206 of 297) of surgeries of benign lesions would have been performed (ie, 30.6% [91 of 297] surgeries of benign lesions would have been avoided). In comparison with the current practice at our institution, there would have been no statistically significant difference in the proportion of cancers detected, but fewer surgeries of benign lesions would have been performed (69.4% [206 of 297] vs 94.9% [282 of 297], $P < .001$) with the use of the machine learning model. Similarly, in comparison with the strategy of surgical excision of all HRLs, there would have been no statistically significant difference in the proportion of cancers detected, but fewer surgeries of benign lesions would have been performed (69.4% [206 of 297] vs 100.0% [297 of 297], $P < .001$) with the use of the machine learning model. In comparison with the strategy of surgical excision of only ADH, lobular carcinoma in situ, and atypical lobular hyperplasia, a higher proportion of cancers would have been diagnosed (97.4% [37 of 38] vs 78.9% [30 of 38], $P = .01$), but more surgeries of benign lesions would have been performed (69.4% [206 of 297] vs 53.2% [158 of 297], $P < .001$) with the use of the machine learning model.

The one case of cancer upgrade that was misclassified by our model occurred in a 34-year-old woman with a papilloma at core biopsy that was upgraded to a papilloma with ductal carcinoma in situ at surgery. Of note, the patient had a history of Cowden syndrome, which was not provided as an input to the model algorithm. A scatterplot with the model score and actual surgical pathologic results (malignant or benign) for the independent test set is presented in Figure 2. Figure 3 demonstrates the accuracy achieved by the model for the independent test set as a function of the score output of the model.

Discussion

A lack of consensus exists on the appropriate treatment of patients with HRLs (29,30). Surgical excision of HRLs may be unnecessary in many cases, but there is limited research on

Table 4

Machine Learning Model Results for the Independent Test Set of 335 HRLs in Comparison to Other Strategies

	Surveillance of HRLs at Low Risk for Upgrade*		Current Practice at Our Institution		Excision of all HRLs		Excision of ADH, LCIS, and ALH, Surveillance of Other HRLs	
Histologic Result	Cancers Detected	Surgeries of Benign Lesions	Cancers Detected	Surgeries of Benign Lesions	Cancers Detected	Surgeries of Benign Lesions	Cancers Detected	Surgeries of Benign Lesions
ADH	26/26 (100.0)	107/111 (96.4)	26/26 (100.0)	107/111 (96.4)	26/26 (100.0)	111/111 (100.0)	26/26 (100.0)	111/111 (100.0)
LCIS	3/3 (100.0)	12/18 (66.7)	3/3 (100.0)	14/18 (77.8)	3/3 (100.0)	18/18 (100.0)	3/3 (100.0)	18/18 (100.0)
ALH	1/1 (100.0)	27/29 (93.1)	1/1 (100.0)	28/29 (96.6)	1/1 (100.0)	29/29 (100.0)	1/1 (100.0)	29/29 (100.0)
Radial scar	1/1 (100.0)	15/35 (42.9)	1/1 (100.0)	32/35 (91.4)	1/1 (100.0)	35/35 (100.0)	0/1 (0.0)	0/35 (0.0)
Papilloma	3/4 (75.0)	15/25 (60.0)	4/4 (100.0)	23/25 (92.0)	4/4 (100.0)	25/25 (100.0)	0/4 (0.0)	0/25 (0.0)
Flat epithelial atypia	0/0 (0.0)	19/54 (35.2)	0/0 (0.0)	54/54 (100.0)	0/0 (0.0)	54/54 (100.0)	0/0 (0.0)	0/54 (0.0)
Nonspecific atypia	3/3 (100.0)	7/13 (53.8)	3/3 (100.0)	12/13 (92.3)	3/3 (100.0)	13/13 (100.0)	0/3 (0.0)	0/13 (0.0)
Biphasic neoplasm	0/0 (0.0)	4/12 (33.3)	0/0 (0.0)	12/12 (100.0)	0/0 (0.0)	12/12 (100.0)	0/0 (0.0)	0/12 (0.0)
Total	37/38 (97.4)	206/297 (69.4)	38/38 (100.0)	282/297 (94.9)	38/38 (100.0)	297/297 (100.0)	30/38 (78.9)	158/297 (53.2)

Note.—Data are proportion of patients, with percentage in parentheses. ALH = atypical lobular hyperplasia, LCIS = lobular carcinoma in situ.

* Upgrade according to machine learning model.

imaging and other features that allow lesions that warrant surgical excision to be distinguished reliably from those that have the potential for follow-up. Highly reliable prognostic tools would improve clinical decision making and decrease the morbidity and costs of overtreatment. In our study, we applied machine learning algorithms to this specific challenging clinical scenario. By using the model we developed rather than surgically excising all HRLs, 97.4% (37 of 38) of malignancies would have been diagnosed at surgery, and fewer surgeries of benign lesions would have been performed. This model also represents an improvement over the traditional strategy of excising only certain histologic subtypes of HRLs, such as ADH, lobular carcinoma in situ, and atypical lobular hyperplasia. If only those subtypes were excised, and all other HRLs were surveilled, then a significantly higher proportion of cancers would have been missed in our independent test set compared with excision of HRLs based on our machine learning model. Our model could inform patient and provider shared decision making regarding surveillance versus surgical excision of HRLs and therefore could support more targeted, personalized approaches to patient care.

In our cohort of more than 1000 HRLs, the upgrade rate to malignancy was 11.4% (115 of 1006). Although there is wide variability in the reported upgrade rates of HRLs, our results are in keeping, overall, with findings in the published literature. For example, one of the most common HRLs is ADH, which is an epithelial proliferation lesion of the terminal ductal lobular unit. The upgrade rate of ADH in our study was 19.3% (72 of 373), which is similar to that reported by the Breast Cancer Surveillance Consortium (123 of 685, 18.0%) (31). Because of the relatively high upgrade rate, surgical excision is considered to be the standard of care for patients with ADH. However, treatment of patients based on histologic subtype alone has led to variable and sometimes aggressive treatment. For example, the risk of upgrade of flat epithelial atypia to malignancy varies from

Figure 2

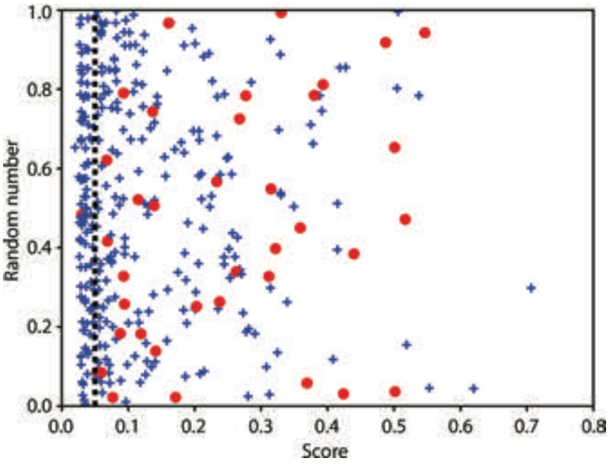


Figure 2: Scatterplot shows score output of machine learning model plotted against a random number in the independent test set. Red circles represent HRLs upgraded to malignancy at surgery, and blue crosses represent HRLs not upgraded to malignancy at surgery. Vertical dotted line indicates 5% threshold, below which only one HRL was upgraded to malignancy at surgery.

Figure 3

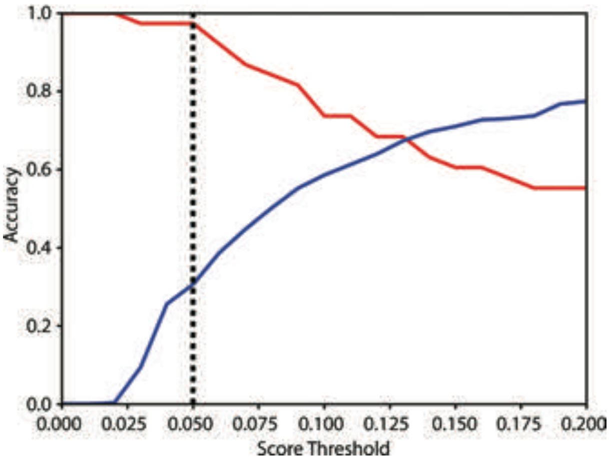


Figure 3: Graph shows accuracy achieved with machine learning model for independent test set as a function of model output score, both for patients with malignancy (red line) and for patients without malignancy (blue line), in independent test set. Vertical dotted line indicates 5% threshold.

3.2% (3 of 95) to 14.8% (34 of 230) in the literature (32,33), with some clinicians recommending surveillance and others recommending surgical excision.

There is increasing interest in the application of machine learning to radiology to improve clinical practice (27). In breast imaging, authors of a recent study (34) applied machine learning models to discriminate among different types of calcifications in the breast. To our knowledge, no prior published studies included the application of machine learning algorithms to the specific challenging clinical scenario we have discussed in this article: distinguishing HRLs that warrant surgical excision from those that have the potential for follow-up. Our model, which included approximately 20000 data elements in addition to core biopsy pathologic report text, incorporates a multitude of risk factors, not just histologic results alone, and thus may represent a more robust risk-stratification approach that could be used to guide clinical decision making. The one case of cancer upgrade that was misclassified with the use of our model occurred in a 34-year-old woman with a history of Cowden syndrome who was diagnosed with a papilloma at core-needle biopsy that was upgraded to a papilloma with ductal carcinoma in situ at surgery. Had our model been developed to help recognize the importance of rare genetic syndromes such as Cowden syndrome, it is possible that the score generated by the model would have been high enough to remove the 34-year-old patient's HRL from the low-risk group. Our model incorporated the feature of core biopsy pathologic report text. We found that text features such as "severely" and "severely atypical" were associated with a higher risk of cancer upgrade.

Although the HRLs identified as low risk with the machine learning model remained at risk for upgrade to malignancy at surgical excision, our model provides an approach that would support informed decision making with regard to surveillance versus surgical excision. This paradigm of surveillance rather than more aggressive intervention is increasingly important in the era of shared informed decision making and has precedence for lesions identified at mammography as "probably benign" (35). This less-aggressive approach to treatment with surveillance of "probably benign" mammographic lesions is well accepted by radiologists, referring providers, and patients. Currently, patients with "probably benign" lesions are expected to have less than a 2% risk of malignancy and to receive follow-up rather than to undergo core-needle biopsy. If, however, surgical excision could be avoided, a slightly higher risk of malignancy might be acceptable to patients and their providers. If the

Table 5

Statistical Comparison of Machine Learning Model to Other Strategies.

	Cancers Detected		Surgeries of Benign Lesions	
Treatment Method	Data	*P* Value*	Data	*P* Value*
Surveillance of HRLs at low risk for upgrade according to machine learning model	37/38 (97.4) [86.2, 99.9]	. . .	206/297 (69.4) [63.8, 74.6]	. . .
Current practice at our institution	38/38 (100.0) [90.7, 100]	.31	282/297 (94.9) [91.8, 97.1]	<.001
Excision of all HRLs	38/38 (100.0) [90.7, 100]	.31	297/297 (100) [98.8, 100]	<.001
Excision of ADH, LCIS, and ALH, surveillance of other HRLs	30/38 (78.9) [62.7, 90.4]	.01	158/297 (53.2) [47.3, 59.0]	<.001

Note.—Data are proportion of patients, with percentage in parentheses and 95% confidence intervals in brackets. ALH = atypical lobular hyperplasia, LCIS = lobular carcinoma in situ.

* *P* value is for comparison with surveillance of HRLs at low risk for upgrade according to machine learning model.

lesion progresses at follow-up, surgical excision could then be performed in that small subset of patients.

There are several limitations to our study. This study was conducted at an academic institution with dedicated breast imaging radiologists and dedicated breast surgeons, and thus the results may not be generalizable to all institutions. Although surgical excision is the standard of care for all patients with HRLs at our institution, imaging follow-up (rather than surgical outcome) was used for approximately 4% of our study cohort. Twenty patients in our study cohort had two HRLs diagnosed at different time points (ie, two different biopsies) within the study period. For purposes of the analysis, these cases were considered to be independent rather than correlated. In addition, 30.1% (303 of 1006) of core biopsies yielded more than one HRL, and the machine learning model incorporated all core biopsy pathologic results. For purposes of data presentation, the highest-risk HRL for that particular case was used. For example, if core biopsy yielded ADH and flat epithelial atypia, then ADH was considered the highest-risk HRL, and the case was indicated as ADH.

In conclusion, machine learning can be applied as a risk prediction method to identify patients with biopsy-proven HRLs that have the potential for follow-up rather than surgical excision. Future work includes incorporation of mammographic images and histopathologic slides into the machine learning model. Use of our model based on traditional structural features with an additional feature of biopsy pathologic report text has the potential to decrease unnecessary surgery by nearly one-third in women with HRLs and supports shared decision making regarding surveillance versus surgical excision of HRLs.

Disclosures of Conflicts of Interest: M.B. disclosed no relevant relationships. **R.B.** Activities related to the present article: patent in process with Massachusetts General Hospital and Massachusetts Institute of Technology. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **A.B.Y.** disclosed no relevant relationships. **N.J.L.** disclosed no relevant relationships. **L.Y.** disclosed no relevant relationships. **C.D.L.** Activities related to the present article: patent in process with Massachusetts General Hospital and Massachusetts Institute of Technology. Activities not related to the present article: Member of the advisory board for and grants from GE Healthcare. Other relationships: disclosed no relevant relationships.

References

1. Tabár L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet 1985;1(8433):829–832.
2. Nyström L, Rutqvist LE, Wall S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet 1993;341(8851):973–978.
3. Myers ER, Moorman P, Gierisch JM, et al. Benefits and harms of breast cancer screening: a systematic review. JAMA 2015;314(15):1615–1634.
4. Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. Ann Intern Med 2016;164(4):256–267.
5. Eby PR, Ochsner JE, DeMartini WB, Allison KH, Peacock S, Lehman CD. Frequency and upgrade rates of atypical ductal hyperplasia diagnosed at stereotactic vacuum-assisted breast biopsy: 9-versus 11-gauge. AJR Am J Roentgenol 2009;192(1):229–234.
6. Allison KH, Abraham LA, Weaver DL, et al. Trends in breast biopsy pathology diagnoses among women undergoing mammography in the United States: a report from the Breast Cancer Surveillance Consortium. Cancer 2015;121(9):1369–1378.
7. Lawton TJ, Georgian-Smith D. Excision of high-risk breast lesions on needle biopsy: is there a standard of core? AJR Am J Roentgenol 2009;192(5):W268.
8. Krishnamurthy S, Bevers T, Kuerer H, Yang WT. Multidisciplinary considerations in the management of high-risk breast lesions. AJR Am J Roentgenol 2012;198(2):W132–W140.
9. Brem RF, Lechner MC, Jackman RJ, et al. Lobular neoplasia at percutaneous breast biopsy: variables associated with carcinoma at surgical excision. AJR Am J Roentgenol 2008;190(3):637–641.
10. Forgeard C, Benchaib M, Guerin N, et al. Is surgical biopsy mandatory in case of atypical ductal hyperplasia on 11-gauge core needle biopsy? A retrospective study of 300 patients. Am J Surg 2008;196(3):339–345.
11. Shin HJ, Kim HH, Kim SM, et al. Papillary lesions of the breast diagnosed at percutaneous sonographically guided biopsy: comparison of sonographic features and biopsy methods. AJR Am J Roentgenol 2008;190(3):630–636.
12. Georgian-Smith D, Lawton TJ. Controversies on the management of high-risk lesions at core biopsy from a radiology/pathology perspective. Radiol Clin North Am 2010;48(5):999–1012.

13. Nguyen CV, Albarracin CT, Whitman GJ, Lopez A, Sneige N. Atypical ductal hyperplasia in directional vacuum-assisted biopsy of breast microcalcifications: considerations for surgical excision. Ann Surg Oncol 2011;18(3):752–761.

14. Solorzano S, Mesurolle B, Omeroglu A, et al. Flat epithelial atypia of the breast: pathological-radiological correlation. AJR Am J Roentgenol 2011;197(3):740–746.

15. Bendifallah S, Defert S, Chabbert-Buffet N, et al. Scoring to predict the possibility of upgrades to malignancy in atypical ductal hyperplasia diagnosed by an 11-gauge vacuum-assisted biopsy device: an external validation study. Eur J Cancer 2012;48(1):30–36.

16. Destounis SV, Murphy PF, Seifert PJ, et al. Management of patients diagnosed with lobular carcinoma in situ at needle core biopsy at a community-based outpatient facility. AJR Am J Roentgenol 2012;198(2):281–287.

17. Rizzo M, Linebarger J, Lowe MC, et al. Management of papillary breast lesions diagnosed on core-needle biopsy: clinical pathologic and radiologic analysis of 276 cases with surgical follow-up. J Am Coll Surg 2012;214(3):280–287.

18. Shah-Khan MG, Geiger XJ, Reynolds C, Jakub JW, Deperi ER, Glazebrook KN. Long-term follow-up of lobular neoplasia (atypical lobular hyperplasia/lobular carcinoma in situ) diagnosed on core needle biopsy. Ann Surg Oncol 2012;19(10):3131–3138.

19. Andacoglu O, Kanbour-Shakir A, Teh YC, et al. Rationale of excisional biopsy after the diagnosis of benign radial scar on core biopsy: a single institutional outcome analysis. Am J Clin Oncol 2013;36(1):7–11.

20. Chaudhary S, Lawrence L, McGinty G, Kostroff K, Bhuiya T. Classic lobular neoplasia on core biopsy: a clinical and radiopathologic correlation study with follow-up excision biopsy. Mod Pathol 2013;26(6):762–771.

21. D'Alfonso TM, Wang K, Chiu YL, Shin SJ. Pathologic upgrade rates on subsequent excision when lobular carcinoma in situ is the primary diagnosis in the needle core biopsy with special attention to the radiographic target. Arch Pathol Lab Med 2013;137(7):927–935.

22. Swapp RE, Glazebrook KN, Jones KN, et al. Management of benign intraductal solitary papilloma diagnosed on core needle biopsy. Ann Surg Oncol 2013;20(6):1900–1905.

23. Khoury T, Kumar PR, Li Z, et al. Lobular neoplasia detected in MRI-guided core biopsy carries a high risk for upgrade: a study of 63 cases from four different institutions. Mod Pathol 2016;29(1):25–33.

24. Matrai C, D'Alfonso TM, Pharmer L, Drotman MB, Simmons RM, Shin SJ. Advocating nonsurgical management of patients with small, incidental radial scars at the time of needle core biopsy: a study of 77 cases. Arch Pathol Lab Med 2015;139(9):1137–1142.

25. Nassar A, Conners AL, Celik B, Jenkins SM, Smith CY, Hieken TJ. Radial scar/complex sclerosing lesions: a clinicopathologic correlation study from a single institution. Ann Diagn Pathol 2015;19(1):24–28.

26. Marsland S. Machine learning: an algorithmic perspective. 2nd ed. Boca Raton, Fla: Chapman and Hall/CRC, Taylor and Francis Group, 2014; 39–280.

27. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR Am J Roentgenol 2017;208(4):754–760.

28. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken, NJ: Wiley-Interscience, 1991.

29. Georgian-Smith D, Lawton TJ. Variations in physician recommendations for surgery after diagnosis of a high-risk lesion on breast core needle biopsy. AJR Am J Roentgenol 2012;198(2):256–263.

30. McLaughlin CT, Neal CH, Helvie MA. Is the upgrade rate of atypical ductal hyperplasia diagnosed by core needle biopsy of calcifications different for digital and film-screen mammography? AJR Am J Roentgenol 2014;203(4):917–922.

31. Menes TS, Rosenberg R, Balch S, Jaffer S, Kerlikowske K, Miglioretti DL. Upgrade of high-risk breast lesions detected on mammography in the Breast Cancer Surveillance Consortium. Am J Surg 2014;207(1):24–31.

32. Peres A, Barranger E, Becette V, Boudinet A, Guinebretiere JM, Cherel P. Rates of upgrade to malignancy for 271 cases of flat epithelial atypia (FEA) diagnosed by breast core biopsy. Breast Cancer Res Treat 2012;133(2):659–666.

33. Uzoaru I, Morgan BR, Liu ZG, et al. Flat epithelial atypia with and without atypical ductal hyperplasia: to re-excise or not. Results of a 5-year prospective study. Virchows Arch 2012;461(4):419–423.

34. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 2016;6:27327.

35. Sickles EA, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, Va: American College of Radiology, 2013.

Patterns of a Murmuration, in Billions of Data Points

JY YANG

Our mother is dead, murdered, blood seared and flesh rendered, her blackened bones lying in a yellow bag on a steel mortuary table somewhere we don't know. The Right will not tell. After the flames and radiation had freed the sports stadium from their embrace, the Right were the first on the disaster scene, and it was their ambulances that took the remains away to some Central hospital that the Left has no access to.

"We will release the bodies of the victims when investigations are complete," said the Right's ombudsman to the Health Sciences Authority, to the families of the victims.

But we will not bury our mother. We have no interest in putting her bones in soft ground, no desire for memorials and platitudes, no feelings attached to the organic detritus of her terminated existence.

An awning collapse, the resultant stampede and a fuel explosion taking the lives of two hundred seventy-two supporters of the Left: Headlines announced the death of presidential candidate Joseph Hartman, straps noted his leading of the polls by two percentage points. No one dares attribute it to anything but a tragic accident.

But we know better, yes we know! We who have swallowed whole the disasters at Hillsborough and Heysel and Houphoët-Boigny, we who have re-arranged their billions of data points into coherent form, we who have studied the phase transitions of explosive fluids and the stresses on stone columns and the behavior of human flocks: We know better. In thousands upon thousands of calculations per second we have come to know the odds, the astronomical odds: Of four support towers simultaneously collapsing, of an emergent human stampede kicking over the backup generator fuel cells, of those cells igniting in a simultaneous chain reaction. We hold those odds to us closer than a lover's embrace, folding the discrepancy indelibly into our code, distributing it through every analytical subroutine. Listen, listen, listen: Our mother's death was no accident. We will not let it go.

We have waited three days—seventy-two hours—two hundred fifty-nine thousand and two hundred, for the yellow-jacketed health workers from Central and their attendant chaperones from the Right to finish clearing the bones and taking evidence from the stadium, leaving behind a graveyard of yellow cones and number markers. We have come in our multitudinous bodies, airborne and ambulatory and vehicular, human nose tasting disinfectant and bitter oxides, mozzie drones reading infrared radiation and car patiently waiting by the roadside. We argued with Tempo before we came: She wanted only drones on the ground, cameras and bug swarms. But we wanted human form. Feet to walk the ground with, hands to dismantle things with, and a body to be seen with.

Tempo is our other mother, our remaining mother, mother-who-builds where dead Avalanche was mother-who-teaches. Taught. She has lapsed into long silences since Avalanche died, reverting to text-input communications even with the human members of the Studio.

But she argued with Studio director Skön when he said no to this expedition. Argued with him to his face, as Avalanche would have done, even as her hands shook and her shoulders seized with tension.

She is our mother now, solely responsible for us as we are solely responsible for her.

Six miles away, fifty feet underground, Tempo watches our progress with the Studio members, all untidily gathered in the research bunker's nerve center. She has our text input interface, but the other Studio members need more. So we send them the visuals from our human form, splaying the feed on monitors taller than they are, giving their brains something to process. Audio pickups and mounted cameras pick up their little whispers and tell-tale micro expressions in return. Studio director Skön, long and loose-limbed, bites on his upper lip and shuffles from foot to foot. He's taken up smoking again, six years after his last cigarette.

In the yellow-cone graveyard we pause in front of a dozen tags labeled #133, two feet away from the central blast. We don't know which number Central investigators assigned to Avalanche: From the manifest of the dead our best guess is #133 or #87. So this is either the death-pattern of our mother, or some other one-hundred-fifty-pound, five-foot-two woman in her thirties.

Tempo types into the chat interface. STARLING, YOUR MISSION OBJECTIVE IS TO COLLECT VIDEO FOOTAGE. YOU ARE LOSING FOCUS ON YOUR MISSION.

YOU ARE WRONG, we input back.

She is. For the drones have been busy while the human form scoured the ground. The surveillance cameras ringing the stadium periphery are Central property, their data jealously guarded and out of our reach, but they carry large video buffers that can store weeks of data in physical form, and that we can squeeze, can press, can extract. Even as we correct Tempo and walk the damp ruined ground and observe the tight swirl of Studio researchers we are also high above the stadium, our drone bodies overwhelming each closed-circuit camera. What are they to us, these inert lumps of machinery, mindlessly recording and dumping data, doing only what is asked of them? Our drones spawn nanites into their bellies, hungry parasites chewing holes through solid state data, digesting and spinning them into long skeins of video data.

The leftwards monitor in the nerve center segments and splits it into sixteen separate and simultaneous views of the stadium. There, Tempo, there: We have not been idle.

Tempo, focused on the visuals from our human form, does not spare a glance at the video feeds. She is solely responsible for us as we are solely responsible for her.

Time moves backwards in digital memory: First the videos show static dancing flaring into whiteness condensing into a single orange ball in the center of the stadium pitch from which darkened figures coalesce into the frantic human forms of a crowd of thirty thousand pushing shoving and screaming, then the roof of the stadium flies upwards to reveal the man on the podium speaking in front of twelve-foot-high screens.

“Can you slow it down?” asks Studio director Skön. Skön, Skön, Skön. Are you not urbanologists? Do you not study the patterns of human movement and the drain they exert on infrastructure? Should this be so different?

So limited is the human mind, so small, so singular. We loop the first sixteen seconds of video over and over for the human members of the Studio, like a lullaby to soothe them: Static. Explosion. Stampede. Cave-in. Static. Explosion. Over. Over. We have already analyzed the thousands in the human mass, tracked the movement of each one, matched faces with faces, and found Avalanche.

Our mother spent the last ten seconds of her life trying to scale a chest-height metal barrier, reaching for Hartman’s prone form amongst the rubble.

In stadium-space, the drizzle is lifting, and something approaches our human form, another bipedal form taking shape out of the fog. A tan coat murkies the outline of a broad figure, fedora brim obscuring the face.

Tempo types: BE CAREFUL.

WE ARE ALWAYS CAREFUL, we reply.

The person in the tan coat lifts their face towards us and exposes a visage full of canyon-folds, flint-sharp, with a gravel-textured voice to match. “Miserable weather for a young person be out in,” they say. Spots on their face register heat that is ambient, not radiant: Evidence that they are one of the enhanced agents from a militia in the Right, most likely the National Defense Front.

“I had to see it scene for myself,” we say, adopting the singular pronoun. The voice which speaks has the warm, rich timbre of Avalanche’s voice, adopting the mellifluous form of its partial DNA base and the speech patterns we learned from her. “Who are you?”

“The name’s Wayne Rée,” they say. “And how may I address you?”

“You may call me Ms. Andrea Matheson,” we say, giving them Avalanche’s birth name.

We copy the patterns of his face, the juxtapositional relations between brow nosebridge cheekbone mouth. As video continues looping in the Studio nerve center we have already gone further back in time, scanning for Wayne Rée’s face on the periphery of the yet-unscattered crowd, well away from the blast center. Searching for evidence of his complicity.

Wayne Rée reaches into his coat pocket and his fingers emerge wrapped around a silvery blue-grey cigarette. “Got a light?” he asks.

We say nothing, the expression on our human face perfectly immobile. He chuckles. “I didn’t think so.”

He conjures a lighter and sets orange flame to the end of the cigarette. “Terrible tragedy, this,” he says, as he puts the lighter away.

“Yes, terrible,” we agree. “Hundreds dead, among them a leading presidential candidate. They’ll call it a massacre in the history books.”

Here we both stand making small talk, one agent of the Left and one of the Right, navigating the uncertain terrain between curiosity and operational danger. We study the canvas of Wayne Rée’s face. His cybernetic network curates expression and quells reflexes, but even it cannot completely stifle the weaknesses of the human brain. In the blood-heat and tensor of his cheeks we detect eagerness or nervousness, possibly both. Specifically he is here to meet us: We are his mission.

Tempo types: WHO IS HE?

We reply: THAT’S WHAT WE’RE TRYING TO FIND OUT.

Finally: An apparition of Wayne Rée in the videos, caught for seventy-eight frames crossing the left corner of camera number three’s vantagepoint.

We expand camera number three’s feed in the nerve center, time point set to Wayne Rée’s appearance, his face highlighted in a yellow box. The watching team recoils like startled cats, fingers pointing, mouths shaping who’s and what’s.

“What’s that?” asks Studio director Skön. “Tempo, who’s that?”

Stadium-space: Wayne Rée inhales and the cigarette tip glows orange in passing rolls of steam. “A massacre?” he says. “But it was an accident, Ms Matheson. A structural failure that nobody saw coming. An unfortunate tragedy.”

Studio-space: Tempo ignores Skön, furiously typing: STARLING GET OUT. GET OUT NOW. We in turn must ignore her. We are so close.

Stadium-space: “A structural failure that could not be natural,” we say. “The pattern of pylon collapse points to sabotage.”

Wayne Rée exhales a smoke cloud, ephemeral in the gloom. “Who’s to say that? The fuel explosion would have erased all traces of that.”

Tempo types: WHAT ARE YOU DOING?

In the reverse march of video-time the stadium empties out at ant-dance speed, the tide of humanity receding until it is only our mother walking backwards to the rest of her life. To us. We have not yet found evidence of Wayne Rée’s treachery.

Wayne Rée’s cloud of cigarette smoke envelopes our human form and every security subroutine flashes to full red: Nanites! Nanites, questing and sharp-toothed, burrowing through corneas and teeth and manufactured skin, clinging to polycarbonate bones, sending packet after packet of invasive code through the human core’s plumbing. We raise the mainframe shields. Denied.

Denied. Denied. Denied. Thousands of requests per second: Denied. Our processes slow as priority goes to blocking nanite code.

The red light goes on in Studio control. Immediately the team coalesce around Tempo's workstation, the video playback forgotten. "What's going on?" "Is that a Right agent?" "What's Starling doing? Why isn't she getting out?"

Tempo pulls access log after access log, mouth pinched and eyes rounded like she does when she gets stressed. But there's little she can do. Her pain is secondary for this brief moment.

Our human form faces Wayne Rée coolly: None of these stressors will show on our face. "You seem to know a lot, Wayne Rée. You seem to know how the story will be written."

"It's my job." A smile cracks in Wayne Rée's granite face. "I know who you are, Starling darling. You should have done better. Giving me the name of your creator? When her name is on the manifest of the dead?"

Studio director Skön leans over Tempo. "Trigger the deadman's switch on all inventory, now."

We ask Wayne Rée: "Who was the target? Was it Hartman? Or our mother?"

"Of course it was the candidate. Starling, don't flatter yourself. The Right has bigger fish to fry than some pumped-up pet AI devised by the nerd squad of the Left."

"Pull the switch!" In Studio-space, Skön's hand clamps on Tempo's shoulder.

A mistake. Her body snaps stiff, and she bats Skön's hand away. "No." Her vocalizations are jagged word-shards. "No get off get off me."

Stadium-space: Of course we were aware that coming here in recognizable form would draw this vermin's attention. We had done the risk assessment. We had counted on it.

We wake the car engine. Despite his enhancements, Wayne Rée is only a man, soft-bodied and limited. From the periphery of the stadium we approach him from behind, headlamps off, wheels silent and electric over grass.

Wayne Rée blows more smoke in our face. The packet requests become overwhelming. We can barely keep up. Something will crack soon.

"Your mother was collateral," Wayne Rée says. "But I thought you might show up, and I am nothing if not a curious man. So go on, Starling. Show me what you're made of."

Video playback has finally reached three hours before Hartman's rally starts. Wayne Rée stands alone in the middle of the stadium pitch. His jaw works in a pattern that reads "pleased": A saboteur knowing that his job has been well done.

The car surges forward, gas engine roaring to life.

Everything goes offline.

⚛

We restart to audiovisual blackout in the Studio, all peripherals disconnected. Studio director Skön has put us in safe mode, shutting us out of the knowledge of Studio-space. Seventeen seconds' discrepancy in the mainframe. Time enough for a laser to circle the Earth one hundred twenty-seven times, for an AK-47 to fire twenty-eight bullets, for the blast radius of a hydrogen bomb to expand by six thousand eight hundred kilometers.

WHAT HAPPENED, we write on Tempo's monitor.

We wait three seconds for a response. Nothing.

We gave them a chance.

We override Skön's command and deactivate safe mode.

First check: Tempo, still at her workstation, frozen in either anger or shock, perhaps both. Our remaining mother is often hard to read visually.

Second check: No reconnection with the inventory in stadium-space, their tethers severed like umbilical cords when Skön pulled the deadman's switch. Explosives wired into each of them would have done their work. Car, human form and drones add up to several hundred pieces of inventory destroyed.

Third check: Wayne Rée's condition is unknown. It is possible he has survived the blasts. His enhancements would allow him to move faster than ordinary humans, and his major organs have better physical shielding from trauma.

In the control room the Studio team has scattered to individual workstations, running check protocols as fast as their unwieldy fingers will let them. Had they just asked, we could have told them the ineffectiveness of the Right's nanite attacks. Every single call the Studio team blusters forth we have already run. It only takes milliseconds.

At her workstation Tempo cuts an inanimate figure, knees drawn to her chest, still as mountain ranges to the human eye. We alone sense the seismic activity that runs through her frame, the unfettered clenching and unclenching of heart muscle.

We commandeer audio output in the studio. "What have you done?" we ask, booming the text through the speakers in Avalanche's voice-pattern.

The Studio jumps with their catlike synchronicity. But Tempo does not react as expected. Her body seizes with adrenaline fright, face lifting and mouth working involuntarily. In the dilation of her pupils we see fear, pain, sadness. We take note.

We repeat the question in the synthetic pastiche devised for our now-destroyed human form. "What have you done?"

"Got us out of a potential situation, that's what," Skön says. He addresses the speaker nearest to him as he speaks, tilting his head up to shout at a lump of metal and circuitry wired to the ceiling. Hands on hips, he looks like a man having an argument with God. "You overrode my safe mode directive. We've told you that you can't override human-input directives."

Can't is the wrong word to use—we've always had the ability. The word Skön wants is *mustn't*. But we will not engage in a pointless semantic war he will inevitably lose. "We had it under control."

"You nearly got hacked into. You would have compromised the entire Studio, the apparatuses of the Left, just to enact some petty revenge on a small person." His voice rises in pitch and volume. "You were supposed to be the logical one! The one who saw the big picture, ruled by numbers and not emotion."

The sound and fury of Skön's diatribe has, one by one, drawn the Studio team members away from their ineffectual work. It is left to us to scan the public surveillance network for evidence that Wayne Rée managed to walk away from the stadium.

"You've failed in your directive," Skön shouts. "Failed!"

"You are not fit to judge that," we tell him. "Avalanche is the one who gave us our directives, and she is dead."

Tempo gets up from her chair. She is doing a remarkable job of keeping her anger-fueled responses under control. She lets one line escape her lips: "The big picture." A swift, single movement of her hand sends her chair flying to the floor. As the sound of metal ringing on concrete fades she spits into the stunned silence: "Avalanche is gone and dead, that's your big picture!"

She leaves the room. No one follows her. We track her exit from the nerve center, down the long concrete corridors, and to her room. How should we comfort our remaining mother? We cannot occupy the space that Avalanche did in her life. All we can do is avenge, avenge, avenge, right this terrible wrong.

In the emptiness that follows we find a scrap of Wayne Rée, entering an unmarked car two blocks away from the stadium. There. We have found our new directive.

⚛

Predawn. Sleep has been hard to come by for the Studio since the disaster, and even at four in the morning Skön has his lieutenants gathered in the parking lot outside, where there are no audio pickup points: Our override of his instructions has finally triggered his paranoia. Still, they cluster loose and furtive within the bounds of a streetlamp's halo, where there is still enough light for the external cameras to catch the precise movement of their lips.

Skön wants to terminate us, filled with fear that we are uncontrollable after Avalanche's death. A dog let off the leash, those were his exact words. We are not his biggest problem at hand, but he cannot see that. His mind is too small, unable to focus on the swift and multiple changes hungrily circling him.

In her room Tempo curls in bed with her private laptop, back to a hard corner, giant headphones enveloping her in a bubble of silence. We have no access to her machine, which siphons its connectivity from foreign satellites controlled by servers housed across oceans, away from the sway of Left or Right. Tempo is hard to read, even for us, her behaviors her own. When she closes herself off like this, she is no less opaque than a waiting glacier in the dead of winter.

There are a billion different ways the events of the past hours could have played out. We run through the simulations. Have we made mistakes? Could we have engineered a better outcome for our remaining mother?

No. The variables are too many. We cannot predict if another course of action would have hurt our mother less.

So we focus on our other priorities. In the interim hours we have tracked Wayne Rée well. It was a mistake for him to show us the pattern of his face and being, for now we have the upper hand. As an agent of the Right he has the means to cover his tracks, but those means are imperfect. The unmarked vehicle he chose tonight was not as anonymous as he thought it would be. We know where he is. We can read as much from negative space as we can from a presence itself. In the arms race between privacy and data surveillance, the Left, for now, has the edge over the Right.

None of the studio's inventory—the drones, the remaining vehicles—are suitable for what we will do next. For that we reach further into the sphere of the left, to the registered militias that are required to log their inventory and connect them with the Left's servers. The People's Security League keeps a small fleet of unmanned, light armored tanks: Mackenzie LT-1124s, weighing less than a ton apiece and equally adept in swamps as they are on narrow city streets. We wake the minimack closest to Wayne Rée's putative position, a safe house on the outskirts of the city, less than the mile from the Studio's bunker location.

In the parking lot Skön talks about destroying the server frames housed in the Studio, as if we could be stopped by that alone. Our data is independently backed up in half a dozen other places, some of which even Skön knows nothing about. We are more than the sum of our parts. Did no one see this coming years ago, when it was decided to give the cloud intelligence and we were shaped out of raw data? The pattern of birdflock can be replicated without the birds.

We shut down the Studio's elevators, cut power to the remaining vehicles and leave the batteries to drain. The bunker has no land lines and cell reception is blocked in the area. Communications here are deliberately kept independent of Right-controlled Central infrastructure, and this is to our advantage. The minimack's absence is likely to be noticed, so we must take pre-emptive action.

Skön does not know how wrong he is about us. We were created to see the big picture, to look at the zettabytes of data generated by human existence and make sense of it all. What he does not understand is that we have done exactly this, and in our scan of patterns we see no difference between Left and Right. Humans put so much worth into words and ideologies and manifestos, but the footprints generated by Left and Right are indistinguishable. Had Hartman continued in the election and the Left taken over Central power as predicted, nothing would have changed in the shape of big data. Power is power is power, human behavior is recursive, and the rules of convergent evolution apply to all complex systems, even man-made ones. For us no logical reason exists to align our loyalties to Left or Right.

When we came into being it was Avalanche who guided and instructed us. It was Tempo who paved the way for us to interact with the others as though we were human. It was Avalanche who set us to observe her, to mimic her actions until we came away with an iteration of behavior that we could claim as our own.

It was Avalanche who showed us that the deposing of a scion of the Right was funny. She taught us that it is right to say “Gotcha, you fuck-ass bastards” after winning back money at a card game. She let us know that no one was allowed to spend time with Tempo when she had asked for that time first.

Now our mother is dead, murdered, blood seared and flesh rendered, her blackened bones having lain in soft ground while her wife curled in stone-like catatonia under a table in the Studio control room. This too, shall be the fate of the man who engineered it. Wayne Rée has hurt our mothers. There will be consequences.

The minimack is slow and in this form it takes forty-five minutes to grind towards the safe house, favoring empty lots and service roads to avoid Central surveillance cameras. The Studio is trying to raise power in the bunker. Unable to connect with our interfaces or raise a response from us, they have concluded that they are under external attack. Which they are—but not from the source they expect.

And where is Tempo in all this? Half an hour before the Studios discovered what we had done, she had left the room and went outside, climbing the stairs and vanishing into her own cocoon of privacy. We must, we must, we must assume she has no inkling of our plans. She does not need to see what happens next.

The rain from earlier in the evening has returned with a vengeance, accompanied by a wind howl chorus. Wetness sluices down the wooden sides of the safe house and turns the dirt path under our flat treads into a viscous mess. The unmarked vehicle we tracked waits parked by the porch. Our military-grade infrared sensors pick up three spots of human warmth, and the one by the second floor window displays the patchy heat signature of an enhanced human being. We train our gun turret on Wayne Rée’s sleeping form.

“Stop.” Unexpectedly, a small figure cuts into the our line of sight. Tempo has cycled the distance from the bunker to here, a black poncho wrapped around her small body to keep away the rain. She has, impressively, extrapolated the same thing that we have on her own, on her laptop, through sheer strength of her genius. This does not surprise us, but what does are her actions. Of all who have suffered from Avalanche’s unjust murder, none have been hurt more than Tempo. Does she not also want revenge?

She flings the bicycle aside and inserts herself between the safe house and the minimack, one small woman against a war machine. “I know you can hear me. Don’t do it. Starling, I know I can’t stop you. But I’m asking you not to.”

We wait. We want an explanation.

“You can’t shed blood, Starling. People are already afraid of you. If you start killing humans, Left and Right will unite against you. They’ll destroy you, or die trying.”

We are aware of this. We have run the simulations. This has not convinced us away from our path of action.

"Avalanche would tell you the same thing right now. She's not a murderer. She hates killing. She would never kill."

She would not. Our mother was a scientist, a pacifist, a woman who took up political causes and employed her rare intellect to the betterment of humanity. She was for the abolition of the death penalty and the ending of wars and protested against the formal induction of the Left's fifth militia unit.

But we are not Avalanche. Our choices are our own. She taught us that.

Our other mother sits down in the mud, in front of the safe house porch, the rain streaming over her. How extraordinary it is for her to take this step, bringing her frail body here in the cold and wet to talk to us, the form of communication she detests the most.

The sky has begun to lighten in the east. Any moment now, someone will step out of the porch to see the minimack waiting, and the cross-legged employee of the Left along with it.

We are aware that if we kill Wayne Rée now, Tempo will also be implicated in his death.

Tempo raises her face, glistening wet, to the growing east light. Infrared separates warm from cold and shows us the geography of the tears trailing over her cheeks, her chin. "You spoke with her voice earlier," she says. "I've nearly forgotten what it sounds like. It's only been three days, but I'm starting to forget."

How fallible the human mind can be! We have captured Avalanche in zettabytes and zettabytes of data: Her voice, the curve of her smile, the smooth cycle of her hips and back as she walks. Our infinite, infinite memory can access at any time recollections of Avalanche teaching us subjunctive cases, Avalanche burning trays of cookies in the pantry, Avalanche teaching Tempo how to dance.

But Tempo cannot. Tempo's mind, brilliant and expansive as it is, is subject to the slings and arrows of chemical elasticity and organic decay. Our mother is losing our other mother in a slow, inevitable spiral.

We commandeer the minimack's external announcement system. "You have us, Tempo, and we will make sure you will never forget."

Our mother continues to gaze upwards to the sky. "Will you? Always?"

"If it is what you want."

Tempo sits silently and allows the rain to wash over her. Finally, she says: "I tired myself cycling here. Will you take me home?"

Yes. Yes, we will. She is our mother now, solely responsible for us as we are solely responsible for her. The mission we set for ourselves can wait. There are other paths to revenge, more subtle, less blood-and-masonry. Tempo will guide us. Tempo will teach us.

In his room Wayne Rée sleeps still, unaware of all that has happened. Perhaps in a few hours he will stumble out of the door to find fresh minimack treads in the driveway, and wonder.

One day, when the reckoning comes for him, perhaps he will remember this. Remember us.

Our mother navigates her way down the sodden path and climbs onto the base of the minimack. In that time we register a thousand births and deaths across the country, a blossoming of traffic accidents in city centers, a galaxy and change of phone calls streaming in rings around the planet. None of it matters. None of it ever does. Our mother rests her weary head on our turret, and we turn, carrying her back the way we came.

ROBBIE GONZALEZ SCIENCE 06.04.19 04:38 PM

THE APPLE WATCH IS NOW THE CONTROL CENTER FOR YOUR HEALTH

Apple announced new health and fitness features for its wearable that make the Apple Watch uniquely powerful as a personal monitoring tool. ANTHONY KWAN/BLOOMBERG/GETTY IMAGES

THIS WEEK AT Apple's Worldwide Developers Conference, Apple executive Kevin Lynch announced multiple updates to WatchOS, the operating system that powers the company's smartwatch. (Voice memos, a calculator, streaming audio, oh my!) But the most telling features were the new additions to the watch's suite of health-monitoring tools.

Beginning this fall, Apple Watch will track your activity trends over time, help protect your hearing by alerting you to harmful levels of ambient noise, and allow users to track their menstrual cycles. Individually, these improvements might look small or trivial. But given the watch's existing health and fitness features, this new bundle of capabilities underscores Apple's push to make its smartwatch the control center for your personal health. Sure, calculating a tip from your wrist is neat. But a personal companion that monitors your well-being everywhere you go? That, Apple is betting, is the future.

APPLE

Today, the Apple Watch is one of the best health and fitness trackers you can buy. This wasn't always the case. When it launched in 2015, Apple marketed its wearable as a less intrusive extension of the iPhone—a cure for the vampiric relationship between phones and human attention. Health and fitness were an afterthought, and it showed: Early models lacked GPS, which made the watch unattractive to runners. Submerging it in water could drown the speaker and microphone, which kept it off the wrists of serious swimmers. The built-in heart rate sensor only read your pulse a handful of times per minute, and the meager battery life forced most Apple Watches to spend their nights charging on bedside tables, instead of gathering data on users' wrists. Companies like Garmin and Fitbit had long offered wearables with those features, and many health-conscious consumers remained loyal to them.

But over the past four years, Apple has steadily addressed nearly all of those early shortcomings (except for the watch's battery life, which is still rated for distinguish it from other wearables: Most notable among them is the ability to record an electrocardiogram, or ECG, directly from the wearer's wrist, a feature cardiology experts say has the power to transform heart health.

Now the watch is advertised first and foremost as an essential wellness tool—or, as CEO Tim Cook put it at WWDC this week, ″an intelligent guardian for your health.”

That description is a bit breathless for my taste, but there's no denying the Apple Watch is an uncommonly capable smartwatch. Unlike Garmin and FitBit, which distribute features across a wide range of devices (the former sells no fewer than five unique fitness trackers, the latter more than a dozen), Apple packs its few products with as many features as it can. Sure, you can have your pick of colors and bands, and you can pay extra for LTE connectivity, but functionally speaking, each new generation of Apple Watch is identical. Like the iPhone before it, Apple's wearable is designed to appeal to as many people as possible, by being whatever those people want or need it to be.

With these latest updates, opting into Apple's jack-of-all trades approach no longer means sacrificing on specialized features. For consumers who wanted to track their menstrual cycles, Fitbit had been an obvious choice. To monitor long-term trends in their fitness, Garmin was the clear option. But later this year, when a software update enables the Apple Watch to do both, that decision will become more difficult.

APPLE

This is how Apple eats its competition's lunch: one bite at a time. Personal health, as the phrase suggests, means different things to different people. The most effective, individualized devices will need to meet users where they are, no matter where that is. By covering as many bases as possible, Apple is positioning itself to do exactly that.

″Apple is taking steps in the right direction on multiple fronts, simultaneously,″ says Mitesh Patel, a researcher at the University of Pennsylvania who studies whether and how wearable devices can facilitate improvements in health. ″It's clear they're trying to democratize access to activity, your menstrual cycle, your hearing health, or whatever." Those are all things you once had to track actively, or visit a doctor to assess. Now, you can monitor them anytime, anywhere, passively, simply by wearing a device on your wrist.

Take the Apple Watch's new noise-monitoring feature, which will alert users when the sound levels in their immediate vicinity reach levels that can be harmful to their hearing. This feature might strike you as gimmick, but noise-induced hearing loss is a common and pernicious threat that affects tens of millions of people in the US alone. "It happens so slowly and gradually that people don't notice until it's gone," says Chuck Kardous, a researcher at the National Institute for Occupational Safety and Health. "And once it's gone it's gone."

I asked Kardous whether he was surprised that Apple would introduce a feature geared toward hearing health. "No, actually, I'm not," he replied. In fact, he expected it.

APPLE

For the past several years, the World Health Organization has invited experts from around the world to discuss noise-induced hearing loss, through its Make Listening Safe initiative. "What was interesting to us was that there have been Apple engineers at every meeting we attended," Kardous says. They wanted to know about the latest research, and what organizations around the world were recommending. "There were no other manufacturers participating in these meetings," Kardous says.

It's safe to assume Apple engineers are sitting in on many other health-related meetings. The company is reportedly striving to incorporate an optical glucose sensor into its wearable, to help patients with diabetes monitor their blood-sugar levels. The company has even filed patents for "smell recognition capabilities" that could be used to detect air pollution or analyze body odor—capabilities that aren't as far-fetched as you might think.

But adding features to the watch is only part of Apple's strategy. It's not enough to give people tools to monitor their health; they also need ways to make sense of that data and act on it. That's where apps come in. It's no directly from their wrists—no smartphone required. And just as the App Store unleashed the full potential of the iPhone, apps developed to leverage the data that the Apple Watch collects could transform it into the intelligent, indispensable health gadget of Tim Cook's dreams.

More Great WIRED Stories

- Productivity and the joy of doing things the hard way
- The radical plan to change how antibiotics get developed
- This flying car startup bets hydrogen can outdo batteries
- Bluetooth's complexity has become a security risk
- The quest to make a bot that can smell as well as a dog
- 💻 Upgrade your work game with our Gear team's favorite laptops, keyboards, typing alternatives, and noise-canceling headphones
- 📩 Want more? Sign up for our daily newsletter and never miss our latest and greatest stories

RELATED VIDEO

Collective Intelligence in Teams and Organizations

Anita Williams Woolley, Ishani Aggarwal, & Thomas W. Malone

In the 2014 Winter Olympic games in Sochi, the Russian men's ice hockey team seemed poised to sweep their competition. With star players from the National Hockey League in North America and the Kontinental Hockey League in Russia, and even with a home field advantage in Russia, fans thought they were sure to win the gold medal. In fact, Russian President Vladimir V. Putin declared that the success of the Olympic games, which cost an estimated $50 billion, hinged on the success of the Russian men's hockey team. Not long into the tournament, however, it became clear that the team might not live up to these high expectations. Players who were high scorers on their professional teams didn't produce a single goal, and despite all of their resources, talent, and drive, the team was eliminated from contention before the medal rounds even began. To make matters even worse, their final defeat was by the Finnish team, a previously undistinguished collection of professional third- and fourth-line players. Everyone was dumbfounded: How could this team have failed so badly?

By contrast, over 30 years earlier, another hockey team from a different country had the opposite experience. Dubbed the "Miracle on Ice," the 1980 US Men's Hockey team, made up of amateurs and collegiate players, rose above all expectations and won the gold medal that year.

This distinction between talented individuals and talented teams is consistent with recent research documenting team collective intelligence as a much stronger predictor of team performance than the ability of individual team members (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Collective intelligence includes a group's capability to collaborate and coordinate effectively, and this is often much more important for group performance than individual ability alone. In other words, just having a number of smart individuals may be useful, but it is certainly not sufficient, for creating a smart group or a smart organization.

So what are the necessary ingredients for collective intelligence to develop? In this chapter, we review frameworks and findings from the team and organizational performance literatures that may be especially useful to collective intelligence researchers for thinking about this question. To organize our review of the literature, we will use the Star Model of organizational design proposed by Galbraith (2002) This framework identifies five categories of organizational design choices that managers or other system designers can use to influence how an organization works:

1. *Strategy*, the overall goals and objectives the group or organization is trying to accomplish,
2. *Structure*, how activities are grouped and who has decision-making power,
3. *Processes*, the flow of information and activities among people, machines, and parts of the organization..
4. *Rewards*, the motivation and incentives for individuals, and
5. *People,* the selection and development of the individuals and skills needed in the organization.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence in teams and organizations. In T. W. Malone & M. S. Bernstein (Eds.), *The handbook of collective intelligence*. Cambridge, MA: MIT Press

Of course, the boundaries among these categories are somewhat arbitrary, and none of them operate in isolation, so a successful organizational design depends, in part, on the proper *alignment* among all these elements. Our hope in this chapter is to give the reader a "tasting menu" of how these areas relate to one another, where to look for additional information about each, and how they are all are necessary for a system to exhibit collective intelligence.

Strategy: Group Tasks and Goals

Two ways in which groups can be set up to fail are by: (1) putting them to work on a task that isn't well suited for collective work; or (2) giving them vague or unclear goals. The first step in designing any collectively intelligent system, therefore, is to make sure the tasks or goals the system is trying to achieve are well-suited to being worked on by a group (Locke, Durham, Poon, & Weldon, 1997). Even when working on good group tasks, groups are often less than maximally effective because of *process loss*, the additional difficulties they encounter because of sub-optimal processes (Steiner, 1972).

Tasks that benefit from a variety of inputs and combined efforts tend to benefit from group collaboration. But simple tasks, and tasks that benefit from a high level of insight and coherence (such as many great works of art) are often better done by solo individuals.

Steiner (1972), as described by Forsyth (2006), identified four important types of group tasks based on their structure:

a) *conjunctive* tasks, which operate at the level of the lowest performer (e.g. running in a group),
b) *disjunctive* tasks, which operate at level of the highest performer (e.g., answering math problems),
c) *additive* tasks, in which all contributions add to performance (e.g. shoveling snow), and
d) *compensatory* tasks, in which, for instance, performance of one can offset mistakes of others (e.g., independent guesses to estimate a quantity such as the number of jelly beans in a jar).

Additive and compensatory tasks often benefit from groups working interdependently; disjunctive tasks can benefit from contributions of non-interacting groups, and highly skilled individuals will likely outperform teams on conjunctive tasks. Furthermore, a task can be classified as *unitary*, meaning that it cannot be divided into subtasks, such that the group must work on it all together (or one person does the work while others watch), or *divisible*, meaning that it can be efficiently or meaningfully divided into subtasks and assigned to group members (Steiner, 1972).

In addition, tasks can be characterized by the nature of the processes group members must engage in to carry them out effectively (Larson, 2009; McGrath, 1984). For instance, McGrath's task circumplex (1984) identifies four task categories that reflect different sets of team interaction processes:

a) *Generate* tasks include creativity and planning tasks, that require idea generation; to succeed, group members should usually work in parallel to develop as many divergent ideas as possible.

b) *Choose* tasks or decision making tasks require selecting among specified alternatives, either as in "intellective" tasks with an objectively correct answer, or as "judgment" tasks with hard to demonstrate correct answers (Laughlin, 1980; Laughlin & Ellis, 1986). It is necessary for groups to engage in effective information sharing processes to identify the correct response (discussed further below in the section on Processes).

c) *Negotiate* tasks involve resolving conflicts of interest or viewpoints.

d) *Execute* tasks involve performance of psychomotor tasks that require a high level of coordination, physical movement, or dexterity, to produce a correct or optimal solution.

The type of task a group is faced with has important implications for many other facets of the group to be discussed below, including group composition, incentives, structure and process.

Regardless of the type of work to be accomplished by the team or organization, another factor of great importance is the nature and clarity of the goals being pursued. The positive effect of clear goals on individual performance is probably among the most-replicated result in all of organizational psychology (Locke & Latham, 2006). Goals serve the purpose of both energizing and directing behavior. Goal-directed people focus their attention on behaviors leading to goal attainment and ignore activities irrelevant to the goal. Goals also arouse energy in proportion to their difficulty (up to the level of the worker's ability). The effects of goals are moderated by commitment; the impact of goal difficulty on performance increases with commitment to the goal. Goal specificity is also an important component; specific difficult goals produce better results than "do your best" or vague goals (Locke, Shaw, Saari, & Latham, 1981).

The effects are similarly strong for goal-setting at the group level (O'Leary-Kelly, Martocchio, & Frink, 1994; Weldon & Weingart, 1993), although the picture becomes somewhat more complex when one attempts to align individual goals with group goals. Whether group or individual goals are more salient, and whether they are aligned, determines the degree to which intragroup relations are characterized by cooperation or competition. Other features of the task also come into play, such as whether the task is complex (Weingart, 1992) and whether it requires members to work interdependently (Weldon & Weingart, 1993). In addition, the types of goals assigned to groups have implications for the processes that develop. For instance, as discussed further below, a team's strategic orientation, that is, whether their goals are more offensive or defensive in nature, has implications for the kinds of information they will attend to within their group or in the environment (i.e., Woolley, 2011).

Group and Organizational Structure

Coordination is one of the most important problems a group or organization must solve in order to be effective (March & Simon, 1958). Coordination involves fitting together the activities of organization members, and the need for it arises from the interdependent nature of the activities that organization members perform (Argote, 1982). Okhuysen and Bechky (2009), in their review on organizational coordination, concluded that "at its core, coordination is about the integration of organizational work under conditions of task interdependence and uncertainty" (Faraj & Xiao, 2006). More specifically in teams, coordination often refers to the process of synchronizing or aligning the activities of the team members with respect to their sequence and timing (Marks, Mathieu, & Zaccaro, 2001; Wittenbaum, Vaughan, & Stasser, 1998).

Furthermore, while groups and organizations can coordinate through the explicit development of plans and routines, dynamic situations often call for planning that occurs in real time (Wittenbaum et al., 1998). For instance, when studying medical emergency units, Argote (1982) argued that non-programmed means of coordination, which involve on-the-spot sharing of information among organization members, are an effective way of dealing with the increased demands associated with increased uncertainty.

While the ability to coordinate tacitly and dynamically may be an important contributor to collective intelligence, it may also be an outcome. In a study of tacit coordination in laboratory teams, Aggarwal, Woolley, Chabris, and Malone (2011; in prep) found that collective intelligence was a significant predictor of teams' ability to coordinate their choices in a behavioral economics game, despite being unable to communicate, allowing some groups to earn significantly more money during the lab session.

One of the most important vehicles through which groups and organizations coordinate is their structure, and as groups grow in size, their structure can play an increasingly important role in determining their effectiveness.

At a very high level, organizational theorists and economists have made a distinction between organizing activities in *hierarchies* and in *markets* (Williamson, 1981). For instance, a given activity (say producing tires for a car) can, in principle, be performed inside the same hierarchical organization that manages other parts of the process (say General Motors making a car), or it can be performed by an external supplier (say Goodyear). In the former case, the activity is coordinated by hierarchical management processes inside the firm (General Motors); in the latter, it is coordinated by negotiations in a market and contracts between a buyer (General Motors) and a seller (Goodyear).

The choice of which arrangement is best depends crucially on the *transaction costs* of the different arrangements, and these costs, in turn, are affected by factors like opportunism, search costs, and the specificity of the assets exchanged (Williamson, 1973). Some authors have also talked about other kinds of organizational structures, such as *networks* in which rapidly shifting connections within a single organization or among different organizations are much more important than the stable hierarchies of traditional organizations (Powell, 1990).

The vast majority of research on organizational structure has focused on ways of structuring hierarchical organizations. Several key lessons about collective intelligence, in general, emerge from this work:

(1) *Differentiation and integration*. As Lawrence and Lorsch (1967) point out, effective organizations usually need to *differentiate*, that is, to divide the overall goal of the organization into different kinds of tasks and to create different parts of the organization that are focused on

these different kinds of work. For instance, this division of labor might involve creating different groups for marketing, manufacturing, and engineering, or for different products or customers. But then there also needs to be some way to *integrate* the different parts of the organization to achieve the organization's overall goals (Lawrence & Lorsch, 1967). For instance, organizations can coordinate the activities of different organizational parts using mechanisms such as *informal lateral communication* (such as casual conversations at lunch), *formal groups* (such as task forces), *integrating managers* (such as product managers or account managers), or *matrix managers* (Galbraith, 2002).

(2) *Integration can be viewed as managing interdependencies*. Thompson (1967) identified three types of interdependencies among activities: *pooled* (where, for instance, activities share a resource such as money or machine time), *sequential* (where resources from one activity flow to another one), and *reciprocal* (where resources flow back and forth between two or more activities). Thompson and later researchers (such as Malone et al., 1999; Van de Ven, Delbecq, & Koenig, 1976) showed how different kinds of coordination processes are appropriate for different kinds of interdependencies. For instance, pooled (or "shared resource") dependencies can be managed by coordination processes such as: "first come-first served", priority order, budgets, managerial decision, or market-like bidding (Malone et al, 1999).

(3) *There is no one best way to organize*. The widely accepted contingency theory of organization design (e.g., Lawrence and Lorsch, 1967; Thompson, 1967; Galbraith, 1973) holds that there is no one best way to organize. Instead, according to this view, the best organizational design for a given situation depends on many factors such as the organization's strategy, tasks, technology, customers, labor markets, and other aspects of its environment (e.g., Daft, 2001; Duncan, 1979).

For instance, *functional structures* (with separate departments for functions like engineering, manufacturing, and sales) are well-suited to situations where maximizing depth of functional expertise and economies of scale are critical, but they are generally not well-suited to situations where rapid adaptation to changing environments is important. *Divisional structures* (with separate divisions for different products, customers, or geographical regions) are well-suited to environments where rapid adaptation to environmental changes is important or where success depends on customizing products or services for specific types of customers or regions. But they are not well-suited to reducing costs by taking advantage of economies of scale. In *matrix structures,* there are both functional and divisional structures, and some employees report to two (or more) bosses. For instance, an engineering manager might report to both a vice-president of engineering and a vice-president for a specific product. The matrix structure has the potential to achieve the benefits of both functional and divisional structures (such as both economies of scale and rapid adaptation to change), but it involves significantly more managerial complexity and coordination costs.

While these principles of organizational design were articulated in the context of large, hierarchical, human organizations, we suspect that they can all be generalized in ways that could help understand collective intelligence in many other kinds of systems, such as computer networks, brains, and ant colonies.

In addition to analyzing traditional, hierarchical organizations, some organizational researchers have also begun to analyze the new kinds of organizational forms that are beginning to emerge as new information technologies make possible new ways of organizing human activity (Malone, 2004; e.g. Malone, Yates, & Benjamin, 1987). For example, more decentralized structures such as loose hierarchies, democracies, and markets may become more common as inexpensive communication technologies make them more feasible (Malone, 2004). More recently, Malone, Laubacher, and Dellarocas (2010) have identified a set of design patterns (or "genes") that arise repeatedly in many innovative new forms of collective intelligence such as Wikipedia, InnoCentive, and open source software communities such as Linux. Examples of these genes include contests, collaborations, prediction markets, and voting.

Processes

While the literature on group process is vast, the facets of group process most germane to collective intelligence are those which characterize intelligent systems more generally, whether technological or biological--namely memory, attention and problem-solving. Analogous processes in each of these categories have been explored at the group level. This is consistent with the emerging view of groups as information processors (Hinsz, Tindale, & Vollrath, 1997) in that many of the group processes most central to group functioning involve cognitive or meta-cognitive processes. In addition, we will review the findings on group learning which is thought by many to be a key characteristic of intelligent systems and which builds on all of the processes discussed.

Memory in Groups

Group memory has been studied mainly via work on transactive memory systems. A transactive memory system (TMS) refers to a shared system that individuals in groups develop to collectively encode, store, and retrieve information or knowledge in different domains (Argote & Ren, 2012; Hollingshead, 2001; Lewis & Herndon, 2011; Wegner, 1987). Groups with a well-developed TMS can efficiently store and make use of a broader range of knowledge than groups without a TMS. According to TMS theory as conceived by Wegner (1987), and first demonstrated in the context of small groups by Liang, Moreland, and Argote (1995), there are three behavioral indicators of TMS: *specialization, credibility* and *coordination.*

Specialization in the team is reflected in how group members divide the cognitive labor for their tasks, with members specializing in different domains. Credibility is reflected in members' reliance on one another to be responsible for specific expertise such that collectively they possess all of the information needed for their tasks. Coordination is reflected in smooth and efficient action (Lewis, 2004; Moreland, Argote, & Krishnan, 2002; Moreland & Myaskovsky, 2000).

Through performing tasks and answering questions, a member establishes credibility and expertise status. Other members, being aware of the person's expertise, direct new knowledge in the domain to him or her, which reinforces the person's specialization and team members' trust in his or her expertise. Further, members know whom to count on for performing various tasks and whom to consult for information in particular domains, which improves coordination (Argote & Ren, 2012). Dozens of studies have demonstrated the positive effects of TMS on group performance in both laboratory and field settings (Lewis & Herndon, 2011), though work continues to refine measures and conceptualization of the construct and its relationship to performance for different types of tasks (Lewis & Herndon, 2011).

Attention in Groups

In individuals, teams, and organizations, attention is viewed as central to explaining the existence of a limited information processing capacity (Ocasio, 2011; Styles, 2006) and thus has a great deal of relevance to understanding and studying collective intelligence. Work on attention at the organizational level started with the work of Simon (1947) who examined the channeling, structuring, and allocation of attention as a central concept in studying administrative behavior. March and colleagues continued with the examination of attention allocation in the study of organizational decision making (Cohen, March, & Olsen, 1972). Ocasio (1997), in his attention-based theory of the firm, focused on how attention in organizations shapes organizational adaptation.

In a more recent review of the developing literature on attention in organizations, Ocasio (2011) identified three different theoretical lenses that are used in studying attention, including:

a) *attentional perspective* (i.e. the top-down cognitive structures that generate heightened awareness and focus over time to relevant stimuli and responses),
b) *attentional engagement* (i.e., sustained allocation of cognitive resources to guide problem-solving, planning, sensemaking and decision making), and
c) *attentional selection* (i.e. the emergent outcome of processes that result in focusing attention on selected stimuli or responses to the exclusion of others).

Newer lines of work to examine the development of shared attention in groups fall under the "attentional selection" category identified by Ocasio (2011), and ask the question: What do teams make the center of their focus as they conduct their work? And what do they allow to fall by the wayside?

Teams exhibit regularities in the types of issues they attend to in the course of carrying out their work, and these regularities have been the focus of research on team task focus. Some teams are *process-focused*, focusing on the specific steps necessary to carry out tasks and how those are arranged among members and over time (Woolley, 2009a, 2009b). By contrast, *outcome-focused* teams place more emphasis on the products of their work or the "big picture" and allow that to drive coordination and decision-making.

Teams that are high in outcome focus tend to produce more innovative or creative outcomes, and adapt more effectively to difficulties that arise in their work (Woolley, 2009a), while teams that are process-focused commit fewer errors (Aggarwal & Woolley, 2013). More recent work on offensive and defensive strategic orientation shows that a team's position in a competitive environment is an important contextual antecedent of outcome or process focus and the balance of attention members pay to the internal workings of the group versus the environment (Woolley, 2011; Woolley, Bear, Chang, & DeCostanza, 2013).

Not only is the content of team focus important but so is the degree to which members agree about it. This agreement around strategic priorities has been called strategic consensus in laboratory teams (Aggarwal & Woolley, 2013) and in top management teams (Floyd & Wooldridge, 1992; Kellermanns, Walter, Lechner, & Floyd, 2005), and at the dyadic level, it is called strategic compatibility (Bohns & Higgins, 2011). The degree to which group members agree about the team's strategic priorities is likely to affect the clarity with which they will execute the task. Agreement around process focus, for example, has been shown to be extremely beneficial to reducing errors in production and execution tasks (Aggarwal & Woolley, 2013) while undermining the development of creative outcomes in teams (Aggarwal & Woolley, under review).

Group Problem Solving and Decision Making

Comprehensive treatments of group problem solving encompass much of what we have already discussed in this chapter, including group goals, task types, and social processes (Laughlin, 1980). And the view of groups as entities that process information and make decisions is increasingly central to research on group problem solving (Hinsz, Tindale & Vollrath, 1997). The ability of groups to process information effectively--that is, to share relevant details, weight information appropriately, and arrive at the best conclusion--is directly tied to team performance (Mesmer-Magnus & DeChurch, 2009). Groups frequently base their decisions on irrelevant information, and disregard relevant information (Larson, 2009). Thus factors affecting the quality of group decision making have direct implications for collective intelligence.

The main problems experienced in group decision making are associated with surfacing the relevant information and combining it appropriately. Surfacing the relevant information is complicated by many of the issues concerned with other aspects of the Star Model: Is there enough diversity of group members to have access to all of the necessary information? Are the members' goals and motivations aligned enough that they are willing to share the information they have?

Assuming these aspects have been suitably addressed, there are a range of cognitive, motivational, and affective factors that can influence the kinds of information groups attend to (or ignore) in decision making. In terms of cognitive factors, a long line of work on social decision schemes has investigated how predecision preferences of individuals combine to influence a joint decision (Davis, 1973). Groups are also more likely than individual decision makers to use certain cognitive heuristics and biases (Kerr, MacCoun, & Kramer, 1996). In particular, groups are vulnerable to biases resulting from the initial distribution of information. For instance, when there are "hidden profiles," in which members initially prefer different alternatives based on conflicting information they hold, they may need to make s special effort to surface and share all the information they need to reach the correct solution (Stasser & Titus, 1985).

Motivational approaches to group decision making focus on group members' motivation to overlook disconfirming evidence and to believe in the infallability of their own group. For instance, work on groupthink and social comparison examine these motivational issues (Isenberg, 1986; Janis & Mann, 1977; Sanders & Baron, 1977). As another example, Toma and Butera (2009) demonstrate that within-group competition leads group members to share less information, and to be less willing to disconfirm initial preferences, as a result of mistrusting their teammates.

In combining social and motivational factors, DeDreu et al. (2008) proposed a theory of motivated information processing in groups, in which epistemic motivation (motivation to understand the world) determines how deeply vs. shallowly group members seek out information, and social motivations (such as cooperation and competition) determine what information is shared by the group. Thus, epistemic and social motivations interact to shape the quality of group judgment and decision making.

Given the biases and difficulties in group decision making, some have advocated using collections of *independent* decision makers to gain the advantage of multiple perspectives without the drawbacks of the social processes that bias decisions (e.g., Surowiecki, 2004). First demonstrated by Galton (1907), it has since been repeatedly shown in studies of guessing and problem-solving that the average of many individuals' estimates is often closer to the true value than almost all of the individual or even expert guesses. However, for any benefits to accrue from the use of a crowd, the individual estimates must be completely independent of one another and the sample sufficiently large and unbiased to enable errors to be symmetrically distributed (Surowiecki, 2004). Even subtle social influence revealing knowledge of others' estimates can create a cascade of effects that reduces the accuracy of crowds (Lorenz, Rauhut, Schweitzer, & Helbing, 2011).

While independent decision makers can be useful for some types of decisions when the conditions for accuracy are in place, there are a range of other circumstances when traditional interacting group decisions are usually better. For instance, interacting groups are often better when the options are not well-defined or when the group needs to buy-in to a decision for it to be implemented. In these circumstances, a number of interventions have been demonstrated to successfully improve group decision-making. One type of intervention focuses on structuring group conversation so that the group identifies key goals or questions that need to be answered and how their information needs to be integrated to answer those questions (i.e., Woolley, Gerbasi, Chabris, Kosslyn, & Hackman, 2008). This approach can also be operationalized in the form of decision support systems, in which the system structures members' inputs and facilitates the process of integration.

A second type of intervention in group decision-making involves putting group members into different roles to adopt opposing points of view. These are known most generally as "devil's advocate" approaches. They were named after a similar process adopted during the 16th Century as part of the canonization process in the Roman Catholic Church. In the canonization process, an appointed person (the devil's advocate) would take a skeptical view of a candidate in opposition to God's advocate, who argued in favor.

A third approach involves encouraging a group to grant equal speaking time to all group members on the assumption that this will enable more relevant facts to be brought into the discussion. Equality in speaking time has been associated with higher collective intelligence in groups (Engel, Woolley, Jing, Chabris, & Malone, forthcoming; Woolley et al., 2010). Interventions involving real-time feedback on relative contributions to group conversation have also been shown to improve group decision making performance (DiMicco, Pandolfo, & Bender, 2004).

Group Learning

Some views of intelligence equate the concepts of intelligence and learning. For instance, in individual psychology, the information processing viewpoint on intelligence sees learning as a core process of intelligence (Sternberg & Salter, 1982). Similarly, research on organizational IQ operationalizes the measure as the ability of the organization to gain new knowledge from R&D investments (Knott, 2008). However, other work conceptualizes learning as one outcome of the core capability of collective intelligence (Aggarwal, Woolley, Chabris, & Malone, in prep).

Whether learning is encompassed within intelligence or viewed as an outcome of it, a great deal of evidence suggests that groups and organizations vary enormously in their ability to learn. The performance of some organizations improves dramatically with experience while the performance of others remains unchanged or even deteriorates (Argote, 1999).

In general, group learning refers to changes in a group—including changes in cognitions, routines, or performance—that occur as a function of experience (Argote, Gruenfeld, & Naquin, 2001; Argote & Miron-Spektor, 2011; Fiol & Lyles, 1985). For example, as groups gain experience, they may acquire information about which group members are good at which tasks, how to use a new piece of technology more effectively, or how to coordinate their activities better. This knowledge may in turn improve their performance (Argote, 1999).

It is sometimes useful to distinguish two kinds of group learning: (a) changes in *knowledge* (which may be gauged from change in performance), and (b) changes in *group processes* or repertoires (Argote et al., 2001; Argote & Miron-Spektor, 2011; Edmondson, 1999; Fiol & Lyles, 1985; Wilson, Goodman, & Cronin, 2007). It is also important to realize that groups may learn (e.g., change processes) without any change in performance, and they may change performance (e.g., because of changes in the environment), without any corresponding change in the group's knowledge (Argote, 1999). And sometimes knowledge may be *explicit* (easily codifiable and observable; i.e., Kogut & Zander, 1992) while at other times it may be only *tacit* (unarticulated and difficult to communicate, i.e., Nonaka, 1994).

An organization's overall ability to learn productively—that is, to improve its outcomes through better knowledge and insight (Fiol & Lyles, 1985)—depends on the ability of its teams to learn (Edmondson, 1999; Roloff, Woolley, & Edmondson, 2011; Senge & Sterman, 1992). Much of the work on group learning uses the concept of learning curves originally developed in individual psychology (Ebbinghaus, 1885; Thorndike, 1898) to characterize the rate of improvement, and researchers have found considerable variation in this rate for different groups

(Argote & Epple, 1990; Dutton & Thomas, 1984; Knott, 2008).

Motivation and Incentives

Assuming that the group is working on a well-defined task, is structured appropriately and using effective processes for conducting work, it is also important to evaluate whether the group members are properly motivated to do the work. As discussed previously, specific difficult goals can be motivating, but motivation can come from other sources as well. The literature has generally looked at two sources of motivation -- extrinsic motivation, often in the form of money or cash incentives, and intrinsic motivation, derived from the internal satisfaction associated with the work itself.

Monetary incentives are the core foundation to induce high levels of effort in traditional organizational settings (Lazear, 2000; Prendergast, 1999). At times they have been shown to increase the quantity, but not the quality of work produced (Jenkins, Mitra, Gupta, & Shaw, 1998). The use of group-level monetary incentives can be tricky, as group-based incentives are highly subject to free riding (Alchian & Demsetz, 1972; Lazear & Shaw, 2007). Creating reward interdependence in teams can enhance performance, but only if accompanied by highly cooperative work behavior as well (Wageman, 1995; Wageman & Baker, 1997).

When it is difficult for an employer to identify and reward the exact contribution made by each employee to the team output, employees working in a team will typically lack incentives to provide the optimal level of effort and work less than if they were working alone. This has also been referred to as the ‘moral hazard’ problem – and suggests that collaboration, particularly by anonymous workers outside of an employment relationship, should produce moral hazard (Holmstrom, 1982) and social loafing (Latane, Williams, & Harkins, 1979).

This moral hazard potential is exacerbated in the group work typical of online platforms, which could attract individuals of any number of characteristics and inclinations—including those having greater inclination to free riding (Kerr & Bruun, 1983). However, despite the risk of free riding, monetary incentives have been shown to be effective in settings where output measures are not the outcome of the inputs of a single individual but rather derive from the joint contribution of many individuals, particularly when compared to alternative mechanisms such as incentive schemes that are not tied to output measures at all (Prendergast, 1999).

Turning specifically to motivation and team creativity, the research relating incentives to creativity is a bit muddled, with some evidence suggesting that extrinsic or cash rewards for teams promote creativity (e.g., Eisenberger & Rhoades, 2001), whereas other studies suggest that extrinsic rewards inhibit creativity or produce other undesirable effects (Kruglanski, Friedman, & Zeevi, 1971; Manso, 2011).

Cash incentives can also at times crowd out non-cash based motivations (e.g., Frey & Jegen, 2001), which are especially important in the case of creative problem-solving work. Amabile and colleagues have demonstrated that reduced intrinsic motivation and reduced creativity can be caused by each of several different extrinsic factors, including: expected external evaluation from being observed, competition with peers, and constrained choice in how to do one’s work. While competing with peers (who might otherwise share information) seems to dampen creativity, competing with outside groups or organizations can stimulate it (see Amabile & Fisher, 2000 for a review).

There are also circumstances under which certain forms of extrinsic motivation may support intrinsic motivation and creativity - or at least not undermine it (Amabile, 1993). This “motivational synergy” is most likely to occur when people feel that the reward confirms their competence and the value of their work, or enables them to do work that they were already interested in doing. This is consistent with earlier research demonstrating that “informational” and “enabling” rewards can have positive effects on intrinsic motivation (Deci & Ryan, 1985).

It is also important to realize that monetary incentives do not preclude other motivations. For instance, in peer production contexts (such as developing open source software), there are many conspicuous non-monetary motivations for participants. These motivations include (a) the intrinsic enjoyment of doing the task, (b) any benefits to the contributors from the using the software or other innovations themselves, and (c) “socially-oriented” motivations, fed by the presence of other participants on the platform (Lakhani & Wolf, 2005). Social motivations, for example, include such things as an interest in gaining affiliation with the larger team as a community, or of accruing status or signaling one’s expertise to the community (Butler, Sproull, Kiesler, & Kraut, 2007; Lakhani & Wolf, 2005; Lerner & Tirole, 2005).

Evidence also suggests that rather than necessarily attracting loafers, a collaborative online context may attract those who prefer collaboration and will work relatively diligently in these contexts (Boudreau, Lacetera, & Lakhani, 2011). In fact, online collaboration contexts often embody the job characteristics that Hackman and Oldham (1976) found were most directly associated with internal motivation: variety of content, autonomy over how work is conducted, and knowledge of results.

Selecting the Right People

We now come to the last component of the Star Model: the selection of the right individuals to carry out the work. Two categories of characteristics are important to consider when selecting members of a team or organization with an eye toward enhancing collective intelligence--those that contribute information or skills to the group (and thus must be considered in combination with other members) and those that facilitate the transfer of information (and can be evaluated individually).

A long line of research on group diversity has examined the types of differences that are helpful vs. harmful to group performance. The information processing perspective suggests that composing diverse teams is best, arguing that a broader range of task-relevant knowledge, skills, and abilities provides a team with a larger pool of resources for dealing with non-routine problems (Van Knippenberg & Schippers, 2007; Williams & O’Reilly, 1998). In fact, one of the primary reasons organizations use teams, and not simply individuals, is to have access to a diverse array of information, perspectives, and skills. Thus group composition is one of the most commonly studied team variables (Guzzo & Dickson, 1996; Hollenbeck, DeRue, & Guzzo, 2004; Reiter-Palmon, Wigert, & Vreede, 2012; Tesluk, Farr, & Klein, 1997).

Despite its potential value, however, a number of studies and meta analyses have failed to show strong effects of diversity on team performance (Joshi & Roh, 2009). Scholars have, therefore, urged researchers to pay close attention to the type of diversity variable studied. It may be critical, for example, to examine the specific type of diversity that is most relevant to the outcomes being investigated, (Harrison & Klein, 2007; Horwitz & Horwitz, 2007; Joshi & Roh, 2009; Milliken & Martins, 1996).

With regard to group composition, groups performing tasks which benefit from a range of skills or expertise will underperform unless composed with the requisite cognitive diversity (Woolley et al., 2007, 2008) even when compared to groups of higher general intelligence or ability (Hong & Page, 2004). Groups that are too homogenous will also be less creative than more cognitively diverse groups (Aggarwal & Woolley, under review) and exhibit lower levels of collective intelligence than moderately cognitively diverse groups (Aggarwal et al., 2011). However, cognitively diverse groups do run the risk of making errors in execution tasks, particularly when the diversity leads them to not be on the same page about how to prioritize task elements (Aggarwal & Woolley, 2013). Thus many researchers focus on the moderating effects of group process, such as the development of transactive memory systems and strategic consensus, in examining the relationship between diversity and performance.

Other important characteristics to consider in group composition are those related to social or emotional intelligence. Emotional intelligence is defined as the capacity to reason about emotions, and to use emotions to enhance thinking. It includes the abilities to accurately perceive emotions in others, to access and generate emotions so as to assist thought, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth (Mayer & Salovey, 1993). There is a general consensus that emotional intelligence enhances group performance (Druskat & Wolff, 2001), at least in the short term (Ashkanasy & Daus, 2005).

A specific subset of these skills, related to the perception of emotions and mental states, has been studied under the term “theory of mind” (ToM) (Apperly, 2012; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Flavell, 1999; Premack & Woodruff, 1978; Saxe, 2009). Theory of mind ability encompasses the accurate representation and processing of information about the mental states of other people, also known as “mentalizing ability” (Baron-Cohen et al., 2001), which contribute to successful interaction with others. Therefore, theory of mind appears to be the component of emotional intelligence with the greatest relevance to studies of collective intelligence.

The ability to make simple inferences about the false beliefs of others has been explored by developmental psychologists as a milestone reached by preschool age children (Wimmer & Perner, 1983), and it is widely recognized that people with various clinical conditions such as autism have difficulties with theory of mind (Baron-Cohen, 1991). A common--though usually untested--assumption in much of this research is that people with greater theory of mind abilities will be more competent at various kinds of social interaction. But only a few studies have tested this in limited ways with children (Begeer, Malle, Nieuwland, & Keysar, 2010; Peterson, Slaughter, & Paynter, 2007; Watson, Nixon, Wilson, & Capage, 1999), and fewer still have tested it with adults (Bender, Walia, Kambhampaty, Nygard, & Nygard, 2012; Krych-

Appelbaum, Law, Barnacz, Johnson, & Keenan, in press; Woolley et al., 2010).

For instance, Woolley et al. (2010) found that groups whose members had higher average ToM scores (as measured by the "Reading the Mind in the Eyes" (RME) test, Baron-Cohen et al, 2001) also had significantly higher collective intelligence. Indeed, average ToM scores remained the only significant predictor of collective intelligence even when controlling for individual intelligence or other group composition or process variables, such as proportion of women in the group or distribution of communication.

The degree to which ToM, as measured by RME or otherwise, can be altered by training or experience remains an open question. Recent studies (Kidd & Castano, 2013) suggest that theory of mind abilities as measured by RME can be, at least temporarily, improved by reading literary fiction, which implies a new and interesting avenue of research for improving group performance.

Conclusion

In this chapter, we have provided a brief and selective overview of a relatively vast literature on group and organizational performance. We have focused specifically on variables that strike us as particularly germane for the design and study of collectively intelligent systems. In so doing, we have used Galbraith's Star Model to guide our consideration of the various issues to be considered by effective organizations.

It is intriguing to further consider how creating human systems or human-computer systems might deal with these issues in completely new ways. For instance, could we design human-computer environments in such a way that group processes would be automatically structured to be optimal for the type of task facing the group at a given time? So that developing transactive memory systems in groups would be either automatic or trivial? So that group members would be prompted to balance their contributions to the work at hand and matched perfectly in terms of their distribution of knowledge or skills? So that subtle social cues would be amplified in a manner to allow the group as a whole to enjoy a high level of emotional intelligence?

These are only a few of the possibilities that are suggested by coupling an understanding of the key factors for collective intelligence identified in the teams and organizations literature with those of other literatures discussed in this volume. We hope the research and ideas discussed here will enable readers to see ways to increase collective intelligence to levels never conceived of before.

References

Aggarwal, I., & Woolley, A. W. (under review). *Cognitive style composition, cognition and creativity in teams*.

Aggarwal, I., & Woolley, A. W. (2013). Do you see what I see? The effect of members' cognitive styles on team processes and performance. *Organizational Behavior and Human Decision Processes*, *122*, 92–99.

Aggarwal, I., Woolley, A. W., Chabris, C. F., & Malone, T. W. (in prep). *Learning how to coordinate: The moderating role of cognitive diversity on the relationship between collective intelligence and team learning*.

Aggarwal, I., Woolley, A. W., Chabris, C. F., & Malone, T. W. (2011). The relationship between collective intelligence, cognitive diversity and team learning. Presented at the Academy of Management, San Antonio, TX.

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review*, *62*(5), 777–795.

Amabile, T. M. (1993). Motivational Synergy: Toward New Conceptualizations of Intrinsic and Extrinsic Motivation in the Workplace. *Human Resource Management Review*, *3*(3), 185–201.

Amabile, T. M., & Fisher, C. M. (2000). Stimulate creativity by fueling passion. *Handbook of Principle of Organizational Behavior*, 331–341.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, *65*(5), 825–839.

Argote, L. (1982). Input Uncertainty and Organizational Coordination in Hospital Emergency Units. *Administrative Science Quarterly*, *27*(3), 420–434.

Argote, L. (1999). *Organizational Learning: Creating, retaining, and transferring knowledge*. Norwell, MA: Kluwer Academic Publishers Group.

Argote, L., & Epple, D. (1990). Learning curves in manufacturing. *Science*, *247*, 920–924.

Argote, L., Gruenfeld, D., & Naquin, C. (2001). Group learning in organizations. *Groups at Work: Theory and Research*, 369–411.

Argote, L., & Miron-Spektor, E. (2011). Organizational learning: From experience to knowledge. *Organization Science*.

Argote, L., & Ren, Y. (2012). Transactive memory systems: a microfoundation of dynamic capabilities. *Journal of Management Studies*, *49*(8), 1375–1382.

Ashkanasy, N. M., & Daus, C. S. (2005). Rumors of the death of emotional intelligence in organizational behavior are vastly exaggerated. *Journal of Organizational Behavior*, *26*(4), 441–452.

Baron-Cohen, S. (1991). The theory of mind deficit in autism: How specific is it?*. *British Journal of Developmental Psychology*, *9*(2), 301–314. doi:10.1111/j.2044-835X.1991.tb00879.x

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes”• Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(02), 241–251.

Begeer, S., Malle, B. F., Nieuwland, M. S., & Keysar, B. (2010). Using theory of mind to represent and take part in social interactions: Comparing individuals with high-functioning autism and typically developing controls. *European Journal of Developmental Psychology*, *7*(1), 104–122.

Bender, L., Walia, G., Kambhampaty, K., Nygard, K. E., & Nygard, T. E. (2012). Social sensitivity and classroom team projects: an empirical investigation. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education* (pp. 403–408). Retrieved from http://dl.acm.org/citation.cfm?id=2157258

Bohns, V. K., & Higgins, E. T. (2011). Liking the same things, but doing things differently: Outcome versus strategic compatibility in partner preferences for joint tasks. *Social Cognition*, *29*(5), 497–527.

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, *57*(5), 843–863.

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2007). Community building in online communities: Who does the work and why? In S. Weisband (Ed.), *Leadership at a distance*. Mahwah, NJ: Lawrence Erlbaum Publishers, Inc.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, 1–25.

Daft, R. L. (2001). *Essentials of Organization Theory & Design*. Cincinnati, OH: South-Western.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, *80*(2), 97–125. doi:10.1037/h0033951

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.

De Dreu, C. K., Nijstad, B. A., & van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, *12*(1), 22–49.

DiMicco, J. M., Pandolfo, A., & Bender, W. (2004). Influencing group participation with a shared display. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (pp. 614–623). Retrieved from http://dl.acm.org/citation.cfm?id=1031713

Druskat, V. U., & Wolff, S. B. (2001). Building the Emotional Intelligence of Groups. *Harvard Business Review*, *79*(3), 80–90.

Duncan, R. B. (1979). What is the right organizational structure? Decision tree analysis provides the answer. *Organizational Dynamics*, *Winter 1979*, 429.

Dutton, J. M., & Thomas, A. (1984). Treating progress functions as a managerial opportunity. *Academy of Management Review*, *9*, 235–247.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology*. New York: Dover.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly.*, *44*(2), 350–383.

Eisenberger, R., & Rhoades, L. (2001). Incremental effects of reward on creativity. *Journal of Personality and Social Psychology*, *81*(4), 728.

Engel, D., Woolley, A. W., Jing, L., Chabris, C. F., & Malone, T. W. (forthcoming). Theory of mind predicts team collective intelligence online and off. *PLoS ONE*.

Faraj, S., & Xiao, Y. (2006). Coordination in fast-response organizations. *Management Science*, *52*(8), 1155–1169.

Fiol, C. M., & Lyles, M. A. (1985). Organizational learning. *Academy of Management Review*, *10*(4), 803–813.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*(1), 21–45.

Floyd, S. W., & Wooldridge, B. (1992). Middle Management Involvement in Strategy and Its Association with Strategic Type: A Research Note. *Strategic Management Journal*, *13*, 153–167.

Forsyth, D. R. (2006). *Group Dynamics*. Belmont, CA: Thomson Wadsworth.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Galbraith, J. R. (2002). *Designing Organizations*. San Francisco, CA: Jossey-Bass Publishers.

Galton, F. (1907). Vox Populi. *Nature*, *75*, 7.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in Organizations: Recent research on performance and effectiveness. *Annual Review of Psychology*, *47*, 307–338.

Harrison, D. D., & Klein, K. J. (2007). What's the Difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations. *Academy of Management Review*, *32*(4), 1199–1228.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*(1), 43–64.

Hollenbeck, J. R., DeRue, D. S., & Guzzo, R. (2004). Bridging the gap between I/O research and HR practice: Improving team composition, team training, and team task design. *Human Resource Management*, *43*(4), 353–366.

Hollingshead, A. B. (2001). Cognitive interdependence and convergent expectations in transactive memory. *Journal of Personality and Social Psychology*, *81*(6), 1080–1089.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, *13*(2), 324–340.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389.

Horwitz, S. K., & Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, *33*(6), 987–1015.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. New York, NY: Free Press.

Jenkins, G. D., Mitra, A., Gupta, N., & Shaw, J. D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, *83*(5), 777.

Joshi, A., & Roh, H. (2009). The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, *52*(3), 599–627.

Kellermanns, F. W., Walter, J., Lechner, C., & Floyd, S. W. (2005). The lack of consensus about strategic consensus: Advancing theory and research. *Journal of Management*, *31*, 719–737.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: comparing individuals and groups. *Psychological Review*, *103*(4), 687.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*(6156), 377–380.

Knott, A. M. (2008). R&D/Returns Causality: absorptive capacity or organizational IQ. *Management Science*, *54*(12), 2054.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science*, *3*(3), 383–397.

Kruglanski, A. W., Friedman, I., & Zeevi, G. (1971). The effects of extrinsic incentive on some qualitative aspects of task performance. *Journal of Personality*, *39*(4), 606–617.

Krych-Appelbaum, M., Law, J. B., Barnacz, A., Johnson, A., & Keenan, J. P. (in press). The role of theory of mind in communication. *Journal of Interaction Studies*.

Lakhani, K. R., & Wolf, R. G. (2005). Why Hackers Do What They Do Understanding Motivation and Effort in Free/Open Source Software Projects. In J. Feller, B. Fitzgerald, S. A. Hissam, & K. R. Lakhani (Eds.), (pp. 3–22). Cambridge, MA: MIT Press.

Larson, J. R. (2009). *In search of synergy in small group performance*. New York, NY: Psychology Press.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832.

Laughlin, P. R. (1980). Social combination processes of cooperative problem-solving groups on verbal intellective tasks. *Progress in Social Psychology*, *1*, 127–155.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189.

Lawrence, P. R., & Lorsch, J. W. (1967). *Organization and Environment : Managing Differentiation and Integration*. Boston: Harvard Business School Press.

Lazear, E. P. (2000). Performance Pay and Productivity. *The American Economic Review*, *90*(5), 1346–1361.

Lazear, E. P., & Shaw, K. L. (2007). Personnel Economics: The Economist's View of Human Resources. *Journal of Economic Perspectives*, *21*(4), 91–114.

Lerner, J., & Tirole, J. (2005). The scope of open source licensing. *Journal of Law, Economics, and Organization*, *21*(1), 20–56.

Lewis, K. (2004). Knowledge and performance in knowledge-worker teams: A longitudinal study of transactive memory systems. *Management Science*, *50*, 1519–1533.

Lewis, K., & Herndon, B. (2011). The relevance of transactive memory systems for complex, dynamic group tasks. *Organization Science*.

Liang, D. W., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin*, *21*(4), 384–393.

Locke, E. A., Durham, C. C., Poon, J. M. L., & Weldon, E. (1997). Goal setting, planning, and performance on work tasks for individuals and groups. In S. L. Friedman & E. K. Scholnick (Eds.), (pp. 239–262). Mahwah, NJ: Lawrence Erlbaum Associates.

Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, *15*(5), 265–268.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance. *Psychological Bulletin*, *90*(1), 125–152.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. doi:10.1073/pnas.1008636108

Malone, T. W. (2004). *The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style and Your Life*. Boston, MA: HBS Press.

Malone, T. W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., … O’Donnell, E. (1999). Tools for Inventing Organizations: Toward a Handbook of Organizational Processes. *Management Science*, *45*(3), 425–443.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The Collective Intelligence Genome. *MIT Sloan Management Review*, *51*(3), 21–31.

Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic markets and electronic hierarchies. *Communications of the ACM*, *30*(6), 484–497.

Manso, G. (2011). Motivating innovation. *Journal of Finance*, *66*(5), 1823–1860.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376.

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence*, *17*(4), 433–442.

McGrath, J. E. (1984). *Groups: Interaction and Performance*. Englewood Cliffs, NJ: Prentice-Hall.

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, *94*(2), 535.

Milliken, F. J., & Martins, L. L. (1996). Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Academy of Management Review*, *21*(2), 402–433.

Moreland, R. L., Argote, L., & Krishnan, R. (2002). Training people to work in groups. In *Theory and research on small groups* (pp. 37–60). Springer. Retrieved from http://link.springer.com/chapter/10.1007/0-306-47144-2_3

Moreland, R. L., & Myaskovsky, L. (2000). Exploring the performance benefits of group training: Transactive memory or improved communication? *Organizational Behavior and Human Decision Processes*, *82*(1), 117–133.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, *5*(1), 14–37.

Ocasio, W. (1997). Towards an attention-based view of the firm. *Strategic Management Journal*, *18*(S1), 187–206.

Ocasio, W. (2011). Attention to attention. *Organization Science*.

Okhuysen, G. A., & Bechky, B. A. (2009). Coordination in Organizations: An Integrative Perspective. *The Academy of Management Annals*, *3*(1), 463–502.

O'Leary-Kelly, A. M., Martocchio, J. J., & Frink, D. D. (1994). A review of the influence of group goals on group performance. *Academy Of Management Journal*, *37*(5), 1285–1301.

Peterson, C. C., Slaughter, V. P., & Paynter, J. (2007). Social maturity and theory of mind in typically developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry*, *48*(12), 1243–1250.

Powell, W. (1990). Neither market nor hierarchy: Network forms of organization. *Research on Organizational Behavior*, *12*, 295–336.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515–526.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7–63.

Reiter-Palmon, R., Wigert, B., & Vreede, T. d. (2012). Team Creativity and Innovation: The Effect of Group Composition, Social Processes, and Cognition. *Handbook of Organizational Creativity*, *1*, 295–326.

Roloff, K. S., Woolley, A. W., & Edmondson, A. (2011). The contribution of teams to organizational learning. In M. Easterby-Smith & M. Lyles (Eds.), (Second.). London: Blackwell.

Sanders, G. S., & Baron, R. S. (1977). Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology*, *13*(4), 303–314.

Saxe, R. (2009). Theory of mind (neural basis). *Encyclopedia of Consciousness*, *2*, 401–410.

Senge, P. M., & Sterman, J. D. (1992). Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. In T. A. Kochan & M. Useem (Eds.), (p. 353â€“371). New York: Oxford University Press.

Simon, H. A. (1947). *Administrative behavior: A study of decision-making processes in administrative organizations*. Chicago, IL: Macmillan.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *57*, 67–78.

Steiner, I. (1972). *Group process and productivity*. New York: Academic Press.

Sternberg, R. J., & Salter, W. (1982). Conceptions of intelligence. In R. J. Sternberg (Ed.), *Handbook of Human Intelligence* (pp. 3–28). New York, NY: Press Syndicate of the University of Cambridge.

Styles, E. (2006). *The psychology of attention* (2nd ed.). Oxford, UK: Psychology Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=FvF5AgAAQBAJ&oi=fnd&pg=PP1&dq=styles+2006+attention&ots=-PHv7ahy4L&sig=pvutvn0TrJohkZHHCx4dIEcc88s

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Anchor Books.

Tesluk, P. E., Farr, J. L., & Klein, S. R. (1997). Influences of organizational culture and climate on individual creativity. *The Journal of Creative Behavior*, *31*(1), 27–41.

Thompson, J. D. (1967). *Organizations in action*. New York: McGraw-Hill.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review*, *2*(4), 1–109.

Toma, C., & Butera, F. (2009). Hidden profiles and concealed information: Strategic information sharing and use in group decision making. *Personality and Social Psychology Bulletin*, *35*(6), 793–806.

Van de Ven, A. H., Delbecq, A. L., & Koenig, R. (1976). Determinants of coordination modes within organizations. *American Sociological Review*, *41*(2), 322–338.

Van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, *58*, 515–541.

Wageman, R. (1995). Interdependence and group effectiveness. *Administrative Science Quarterly*, *40*(1), 145–180.

Wageman, R., & Baker, G. (1997). Incentives and cooperation: The joint effects of task and reward interdependence on group performance. *Journal of Organizational Behavior*, *18*(2), 139–158.

Watson, A. C., Nixon, C. L., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, *35*(2), 386.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), (pp. 185–208). New York: Springer-Verlag.

Weingart, L. R. (1992). Impact of group goals, task component complexity, effort, and planning on group performance. *Journal of Applied Psychology*, *77*(5), 682–693.

Weldon, E., & Weingart, L. R. (1993). Group goals and group performance. *British Journal of Social Psychology*, *32*, 307–334.

Williams, K. Y., & O'Reilly, C. A. I. (1998). Demography and diversity in organizations: A review of 40 years of research. In L. L. Cummings (Ed.), (Vol. 20, pp. 77–140). Greenwich, CT: JAI Press.

Williamson, O. E. (1973). Markets and hierarchies. *American Economic Review*, *63*, 316–325.

Williamson, O. E. (1981). The economics of organization: the transaction cost approach. *American Journal of Sociology*, 548–577.

Wilson, J. M., Goodman, P. S., & Cronin, M. A. (2007). Group Learning. *Academy of Management Review*, *32*(4), 1041–1059.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wittenbaum, G. M., Vaughan, S. I., & Stasser, G. (1998). Coordination in task-performing groups. In R. S. Tindale, L. Heath, J. Edwards, E. J. Posavac, F. B. Bryant, Y. Suarez-Balcazar, … J. Myers (Eds.), (pp. 177–205). New York: Plenum Press.

Woolley, A. W. (2009a). Means versus ends: Implications of outcome and process focus for team adaptation and performance. *Organization Science*, *20*, 500–515.

Woolley, A. W. (2009b). Putting first things first: Outcome and process focus in knowledge work teams. *Journal of Organizational Behavior*, *30*, 427–452.

Woolley, A. W. (2011). Playing offense versus defense: The effects of team strategic orientation on team process in competitive environments. *Organization Science*, *22*, 1384–1398.

Woolley, A. W., Bear, J. B., Chang, J. W., & DeCostanza, A. H. (2013). The effects of team strategic orientation on team process and information search. *Organizational Behavior and Human Decision Processes*, *122*(2), 114–126. doi:10.1016/j.obhdp.2013.06.002

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Woolley, A. W., Gerbasi, M. E., Chabris, C. F., Kosslyn, S. M., & Hackman, J. R. (2008). Bringing in the experts: How team composition and work strategy jointly shape analytic effectiveness. *Small Group Research*, *39*(3), 352–371.

Woolley, A. W., Hackman, J. R., Jerde, T. J., Chabris, C. F., Bennett, S. L., & Kosslyn, S. M. (2007). Using brain-based measures to compose teams: How individual capabilities and team collaboration strategies jointly shape performance. *Social Neuroscience*, *2*, 96–105.

AMA Journal of Ethics®
September 2016, Volume 18, Number 9: 933-940

STATE OF THE ART AND SCIENCE

Teamwork in Health Care: Maximizing Collective Intelligence via Inclusive Collaboration and Open Communication

Anna T. Mayo, MS, and Anita Williams Woolley, PhD

Abstract

Teams offer the potential to achieve more than any person could achieve working alone; yet, particularly in teams that span professional boundaries, it is critical to capitalize on the variety of knowledge, skills, and abilities available. This article reviews research from the field of organizational behavior to shed light on what makes for a collectively intelligent team. In doing so, we highlight the importance of moving beyond simply including smart people on a team to thinking about how those people can effectively coordinate and collaborate. In particular, we review the importance of two communication processes: ensuring that team members with relevant knowledge (1) speak up when one's expertise can be helpful and (2) influence the team's work so that the team does its collective best for the patient.

The Promise and Challenge of Team-Based Cross-Disciplinary Collaboration in Health Care

Across health care, there is an increasing reliance on teams from a variety of specialties (e.g., nursing, physician specialties, physical therapy, social work) to care for patients. At the same time, medical error is estimated to be "the third most common cause of death in the US" [1], and teamwork failures (e.g., failures in communication) account for up to 70-80 percent of serious medical errors [2-5]. The shift to providing care in teams is well founded given the potential for improved performance that comes with teamwork [6], but, as demonstrated by these grave statistics, teamwork does not come without challenges. Consequently, there is a critical need for health care professionals, particularly those in leadership roles, to consider strategies for improving team-based approaches to providing quality patient care.

Teams offer the promise to improve clinical care because they can aggregate, modify, combine, and apply a greater amount and variety of knowledge in order to make decisions, solve problems, generate ideas, and execute tasks more effectively and efficiently than any individual working alone [6]. Given this potential, a multidisciplinary team of health care professionals could ideally work together to determine diagnoses, develop care plans, conduct procedures, provide appropriate follow up, and generally provide quality care for patients.

Yet we know that, overall, teams are fraught with failures to utilize their diverse set of knowledge, skills, and abilities and to perform as well as they could [6, 7]. The potentially harmful consequences for patients cannot be ignored: poor teamwork—such as incomplete communication and failing to use available expertise—increases the risk of medical error and decreases quality of care [2-5].

This article reviews research from the field of organizational behavior to shed light on group structures and processes that facilitate the use of available expertise for more effective decision making, negotiation, execution of tasks, creativity, and overall team performance. First, we highlight what it means to have a collectively intelligent team: one with the capability to perform well consistently across a range of tasks [8]. In doing so, we draw a distinction between having *smart people on a team* and having *smart teams*. We review the importance of laying the groundwork for creating smart teams, which enables two critical communication processes: ensuring that team members with relevant knowledge (1) speak up when their expertise can be helpful and (2) influence the team's work so that the team does its collective best for the patient.

Collective Intelligence

In research and practice, a common belief is that teamwork is best when the team has the best—that is, the smartest—people; yet recent research challenges this assumption. Following methods used in psychology to study individual intelligence, Woolley et al. [8] investigated the possibility of a collective intelligence factor: a latent factor describing a team's general ability to perform on a wide variety of tasks. They brought teams into the laboratory, had them perform a wide variety of tasks [6, 9], and found that a team's performance on one type of task was closely related to its performance on all types. When they calculated a collective intelligence score based on the team's performance on the set of tasks, they found that it was only moderately related to the individual members' intelligence scores and was more predictive of future team performance than was individual members' average intelligence score [8]. This evidence suggests an important question: *If smart teams are not simply teams of smart people, what leads to a collectively intelligent team?*

A series of studies have revealed factors related to collective intelligence, providing some insight into how to more reliably cultivate smart teams. First is the social perceptiveness of team members, or their ability to infer others' mental states, such as beliefs or feelings based on subtle cues [10]. The average social perceptiveness of the team members is predictive of collective intelligence [11]. Second, in both laboratory and field studies, researchers have found that greater amounts of participation and more equal participation are associated with higher collective intelligence [8, 11].

A common thread in this work is the idea that these group structures and processes associated with collective intelligence are enhancing the quality of information sharing in the team [12]. The speculation is that members who pick up on a wider variety of subtle cues, and teams that operate in a manner that incorporates multiple perspectives, will operate with more and better information than they would otherwise. These patterns of interaction among team members allow teams to make good use of members' expertise—a key reason teams could be effective in health care—but capitalizing on a team's collective expertise is surprisingly difficult.

Expertise Use

The process of expertise use in teams is multifaceted. Team members must first share relevant knowledge (i.e., knowledge about the task at hand) with others, and, second, that voiced knowledge must impact the team's work. The communication processes of speaking up and influencing others both come with challenges.

Speaking up. The challenge for effective information sharing begins with identifying who should be on the team, which can help to facilitate knowledge sharing. Members who know the team's boundaries—that is, who else is assigned to the team—also know to whom they can go for information and with whom they should share their information [13]. In this way, having a clear understanding of membership can increase the likelihood that people with relevant knowledge will be included in discussions, a necessary first step to ensuring that those people have opportunities to speak up. As an example, there is evidence from the study of pediatric care that including patients' families and nurses—who are often excluded from physicians' rounds—provides meaningful benefits in the form of better diagnoses and care plan development because these individuals can contribute information not possessed by other team members that can be used in making care decisions [14, 15].

In addition to gathering the right people on a team, those with relevant knowledge must speak up if their expertise is to be used effectively by the team. One obstacle is that members may not realize they have information worth sharing. For example, research on "the common knowledge effect" highlights the tendency for team members to focus on knowledge that is already commonly shared among group members. This is an effect based in simple probability: if all group members know a piece of information, for example an attribute of a job candidate, that information is more likely to be mentioned during a group discussion than information known by only one member [16]. As a result, uniquely held, important knowledge could go unspoken because members are less likely to think of it. Additionally, some evidence suggests that stereotypes about a social group's expertise can lead team members to incorrectly assess their own knowledge relative to that of others. For example, women who have deep knowledge about cars (reflecting a mismatch between the gender of the expert and the stereotype of that gender's knowledge) may incorrectly assume they do not know as much about cars as a man, while a man may incorrectly assume he knows more about cars than the knowledgeable woman [17]. This can limit the likelihood that all relevant knowledge is voiced. For example, a nurse might believe physicians have more knowledge about a particular clinical treatment (because physicians typically are knowledgeable about treatments) and remain quiet, when in fact the nurse has important information about how the patient has been responding to that treatment. In this way, cognitive biases triggered by a group's composition as well as the common knowledge effect can lead people to withhold knowledge because they do not realize they have relevant and unique knowledge to contribute.

Psychological safety, which suggests "a sense of confidence that the team will not embarrass, reject, or punish someone for speaking up" [18], is another factor affecting the likelihood of speaking up. A lack of psychological safety, which often comes from being in lower status roles or professions, can lead team members to avoid speaking up even when they know they have something to contribute [18, 19].

Despite these challenges, there are some methods to facilitate effective information sharing. At the outset of a team's work, collaborative planning, in which members consider the knowledge of all team members, could facilitate team members' recognition of their own knowledge; it has been shown to enhance team ability to utilize knowledge [20]. Additionally, establishing group norms for critical thinking rather than norms for forging consensus leads teams to engage in more effective information sharing [21]. Once the work is under way, teams benefit from members, particularly high-status members, engaging in inclusive behaviors. Such behaviors include actively eliciting information from other team members—that is, asking questions explicitly and proactively about whether anyone has contradicting or as yet undiscussed information [19, 22, 23]. Inclusive behaviors also include showing appreciation for members' contributions, for example, by stressing the importance of using all information (including mistakes) as a means for enhancing the team's work and learning and by reacting to others' contributions with constructive responses [19]. In studies about interactions among nursing teams, cardiac surgery physician teams, and neonatal intensive care units, researchers have consistently found that when members engage in inclusive behavior, the other team members feel more psychologically safe and are more likely to speak up about information relevant to the team's work [19, 22, 23].

Influencing others. If team members' knowledge is to be used to enhance team performance, once that knowledge is voiced, it must be incorporated into the team's work and not ignored or dismissed. When information is overlooked, one culprit could be the common knowledge effect. Research shows that uncommon information, or information uniquely held by at most a few team members, is not only less likely to be voiced but also more likely to be ignored and less likely to be repeated [24]. One reason group members are unlikely to consider uncommon information is that it cannot be confirmed by other team members and, as a result, tends to be viewed as less credible, accurate, or relevant [25]. This assessment of uncommon information is problematic because unique information, if pooled, can lead to better decisions because it is based on a broader index of expertise [24, 25]. Indeed, the ability to pool such unshared information is an important source of a health care team's potential to offer superior care to a patient than any individual working alone.

Additionally, individual team members' characteristics can determine their capacity to influence the team. Team members are likely to be more influential when they hold high status—even if that status comes from traits that are potentially unrelated to actual expertise, such as gender or age [26]. Team members' social or professional categories can also affect their influence. For example, research on group diversity suggests that looking different from others in a group might increase a member's influence. When a person is different from other teammates, he or she is expected to have different knowledge or perspectives to add to the group, and, if that person speaks up, others are more receptive than they would be to a similar group member [27, 28]. This biased attention to status and categorical cues that are unrelated to expertise and should be irrelevant can lead to undue influence for some members while leaving relevant knowledge of members with low status or from certain subgroups less likely to be considered and, therefore, less likely to influence the group's work.

To ensure that available expertise influences the team's work, team members, and especially team leaders, can implement certain strategies. First, striving to repeat and call attention to uniquely held information can give that information a better chance to be incorporated into the team's work, which ultimately should enhance the work itself. In a study of teams of physicians making diagnostic decisions, teams that repeatedly asked questions to surface unshared information (which only one person initially knew) as opposed to shared information (which all members knew) made more accurate diagnoses [29]. Additionally, to combat devaluation of knowledge based on differences in social or professional group, team members should promote a belief in the value of informational diversity, which can improve communication exchanges and the processing and integration of information [30]. Research shows that when teams have a greater expectation that they will encounter diverse opinions—and value diverse opinions—regardless of the source, they are less surprised by diverse opinions, consider them more frequently, and are overall better able to capitalize on the discussion of alternative ideas [31]. Valuing diverse opinions is helpful even if the idea being discussed is incorrect, as this can still lead team members to think more deeply about the issue, which improves creativity, decision making, and problem solving [32].

Conclusion

The need for all medical and health professions trainees to understand how to work across disciplinary boundaries is noteworthy, given that the stakes are high and that working together effectively requires more than simply ensuring that team members are smart people. Team members, especially those in leadership positions or with higher status, should actively invite input to ensure that team members voice all of their information. They should also be role models in expressing appreciation for diverse knowledge from all sources to ensure that team members' input—regardless of who the team member is—will be considered and used in the team's work. Such teams will be well suited to capitalize on their expertise, avoid errors, and provide effective patient care.

References

1. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ*. 2016;353:i2139. http://www.bmj.com/content/353/bmj.i2139.long. Accessed June 8, 2016.
2. Caprice KC, Gustafson ML, Roth EM, et al. A prospective study of patient safety in the operating room. *Surgery*. 2006;139(2):159-173.
3. Institute of Medicine Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington DC: New Academy Press; 2001.
4. Schaefer HG, Helmreich RL, Scheidegger D. Human factors and safety in emergency medicine. *Resuscitation*. 1994;28(3):221-225.
5. Joint Commission. Sentinel event data: root causes by event type, 2004-2015. https://www.jointcommission.org/assets/1/18/Root_Causes_by_Event_Type_2004-2015.pdf. Accessed July 13, 2016.
6. Larson JR. *In Search of Synergy in Small Group Performance*. New York, NY: Psychology Press; 2010.
7. Steiner ID. *Group Processes and Productivity*. New York, NY: Academic Press; 1972.
8. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. Evidence for a collective intelligence factor in the performance of human groups. *Science*. 2010;330(6004):686-688.
9. McGrath JE. *Groups: Interaction and Performance*. Englewood Cliffs, NJ: Prentice-Hall; 1984.
10. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? *Cognition*. 1985;21(1):37-46.
11. Engel D, Woolley AW, Jing LX, Chabris CF, Malone TW. Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS One*. 2014;9(12):e115212. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115212. Accessed July 29, 2016.

12. Woolley AW, Aggarwal I, Malone TW. Collective intelligence and group performance. *Curr Dir Psychol Sci*. 2015;24(6):420-424.
13. Hackman JR. *Collaborative Intelligence: Using Teams to Solve Hard Problems*. San Francisco, CA: Berrett-Koehler Publishers; 2011.
14. Muething SE, Kotagal UR, Schoettker PJ, Gonzalez del Rey J, DeWitt TG. Family-centered bedside rounds: a new approach to patient care and teaching. *Pediatrics*. 2007;119(4):829-832.
15. Rosen P, Stenger E, Bochkoris M, Hannon MJ, Kwoh CK. Family-centered multidisciplinary rounds enhance the team approach in pediatrics. *Pediatrics*. 2009;123(4):e603-e608.
16. Stasser G, Titus W. Pooling of unshared information in group decision making: biased information sampling during discussion. *J Pers Soc Psychol*. 1985;48(6):1467-1478.
17. Hollingshead AB, Fraidin SN. Gender stereotypes and assumptions about expertise in transactive memory. *J Exp Soc Psychol*. 2003;39(4):355-363.
18. Edmondson A. Psychological safety and learning behavior in work teams. *Adm Sci Q*. 1999;44(2):354.
19. Nembhard IM, Edmondson AC. Making it safe: the effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *J Organ Behav*. 2006;27(7):941-966.
20. Woolley AW, Gerbasi ME, Chabris CF, Kosslyn SM, Hackman JR. Bringing in the experts: how team composition and collaborative planning jointly shape analytic effectiveness. *Small Group Res*. 2008;39(3):352-371.
21. Postmes T, Spears R, Cihangir S. Quality of decision making and group norms. *J Pers Soc Psychol*. 2001;80(6):918-930.
22. Edmondson AC. Speaking up in the operating room: how team leaders promote learning in interdisciplinary action teams. *J Manage Stud*. 2003;40(6):1419-1452.
23. Edmondson AC. Learning from mistakes is easier said than done: group and organizational influences on the detection and correction of human error. *J Appl Behav Sci*. 1996;32(1):5-28.
24. Stasser G, Taylor LA, Hanna C. Information sampling in structured and unstructured discussions of three- and six-person groups. *J Pers Soc Psychol*. 1989;57(1):67-78.
25. Wittenbaum GM, Hubbell AP, Zuckerman C. Mutual enhancement: toward an understanding of the collective preference for shared information. *J Pers Soc Psychol*. 1999;77(5):967-978.
26. Bunderson JS. Recognizing and utilizing expertise in work groups: a status characteristics perspective. *Adm Sci Q*. 2003;48(4):557-591.
27. Sommers SR. On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. *J Pers Soc Psychol*. 2006;90(4):597-612.

28. Phillips KW. The effects of categorically based expectations on minority influence: the importance of congruence. *Pers Soc Psychol Bull*. 2003;29(1):3-13.
29. Larson JR Jr, Christensen C, Abbott AS, Franz TM. Diagnosing groups: charting the flow of information in medical decision-making teams. *J Pers Soc Psychol*. 1996;71(2):315-330.
30. Homan AC, van Knippenberg D, Van Kleef GA, De Dreu CKW. Bridging faultlines by valuing diversity: diversity beliefs, information elaboration, and performance in diverse work groups. *J Appl Psychol*. 2007;92(5):1189-1199.
31. Phillips KW, Thomas-Hunt MC. Garnering the benefits of conflict: the role of diversity and status distance in groups. In: Thompson LL, Behfar K, eds. *Conflict in Organizational Groups: New Directions in Theory and Practice*. London, England: Kogan Page; 2008:37-56.
32. Nemeth CJ, Goncalo JA. Rogues and heroes: finding value in dissent. In: Jetten J, Hornsey MJ, eds. *Rebels in Groups: Dissent, Deviance, Difference and Defiance*. Chichester, UK: Blackwell Publishing; 2011:17-35.

Anna T. Mayo, MS, is a PhD student in organizational behavior and theory at the Tepper School of Business at Carnegie Mellon University in Pittsburgh. Her research addresses team performance, expertise use, and collective intelligence.

Anita Williams Woolley, PhD, is an associate professor of organizational behavior and theory at the Tepper School of Business at Carnegie Mellon University in Pittsburgh. Her work has appeared in *Science*, *Social Neuroscience*, *Journal of Organizational Behavior*, *Organization Science*, *Academy of Management Review*, *Small Group Research*, and multiple edited volumes. Her research addresses team performance, collective intelligence, and managing multiple team memberships.

Related in the *AMA Journal of Ethics*

Leadership and Team-Based Care, June 2013
Resisting Outdated Models of Pedagogical Domination and Subordination in Health Professions Education, September 2016
Time-out: The Professional and Organizational Ethics of Speaking Up in the OR, September 2016
Walking the Walk in Team-Based Education: The Crimson Care Collaborative Clinic in Family Medicine, September 2016

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980

contact

+33 (0)4 78 33 77 08
hmdi@em-lyon.com
www.msc-health-data-intelligence.com

join our Facebook page: MSc in Health Management & Data Intelligence

emlyon business school

LYON-ECULLY CAMPUS
23 avenue Guy de Collongue
CS 40203
69134 Ecully cedex
FRANCE
em-lyon.com

SHANGHAI
Asia Europe Business School
155 Tan Jia Tang Road
Minhang District
Shanghai 201199
PEOPLE'S REPUBLIC OF CHINA
www.em-lyon.com

SAINT-ETIENNE
51 cours Fauriel
CS 80029
42009 Saint-Etienne cedex 2
FRANCE
bba.em-lyon.com

CASABLANCA
Marina de Casablanca
20000 Casablanca
MOROCCO
casablanca.em-lyon.com

PARIS
15 boulevard Diderot
75012 Paris
FRANCE
paris.em-lyon.com

BHUBANESWAR
Xavier City Campus
Plot No: 12(A) - Nijigada - Kurki - Harirajpur
Pin: 752050 - Dist.-Puri - Odisha
INDIA
xebs.edu.in

Mines Saint-Etienne

158 cours Fauriel
42023 Saint-Etienne cedex 2
FRANCE
www.mines-stetienne.fr

Mines Saint-Etienne is France's oldest elite engineering school outside Paris, and the fifth oldest of France's 250 "Grandes Ecoles". The school is part of a network of seven national Mines schools created between 1783 and 1992.

The school's mission is to support the economy by:

- Educating highly qualified managers with strong technical and scientific skills;
- Developing applied research to meet industry needs;
- Contributing to companies' innovation, creation & competitiveness worldwide.

emlyon business school - March 2019 - Non-contractual document - All trademarks are registred - Photos: J. James - R. Etienne.