SHANNON BROWNLEE AND BIBIANA BIELEKOVA

Ending the Reproducibility Crisis

Medical research that can't be replicated hinders discoveries. Could an artificial intelligence-powered tool change the incentives to benefit scientists, taxpayers, and patients?

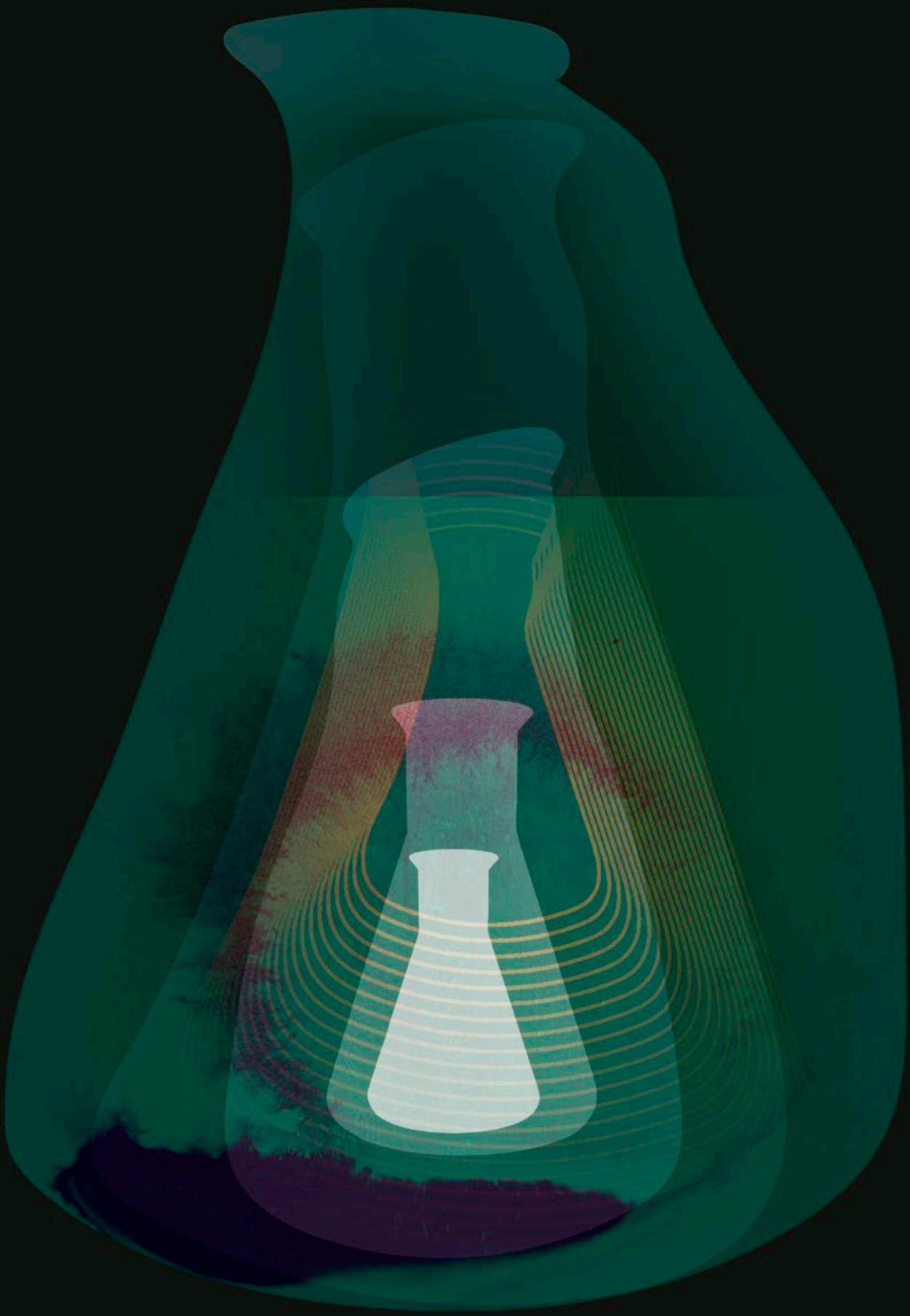
Biomedical science is capable of great feats, just one of which is the astonishingly rapid development of vaccines against the SARS-CoV-2 virus. Vaccine production generally takes 10–15 years, yet the first messenger RNA vaccines were rolling off the production line a mere 12 months after the novel virus was identified and its genome sequenced. This speed serves as a testament to the deep well of biomedical knowledge about mRNA and viruses that already existed. The basic mRNA science that was developed over the last decade—much of it funded by the federal government—paid off in spectacular fashion.

But not all of the billions of taxpayer dollars the United States invests each year in biomedical research produce such rapid gains for human health. One factor holding back the development of new treatments is a complex and long-standing problem: the widespread irreproducibility of biomedical research results. Many factors contribute to irreproducible results and addressing the problem will require strong leadership at the highest levels. In a paper published earlier in 2021 in the open peer review journal *F1000 Research,* we proposed a cost-effective, minimally intrusive solution for aligning the self-interest of researchers with the societal goal of maximizing research reproducibility and its value to human health. In this article, we briefly lay out some of the forces driving irreproducibility and the essential components of its solution.

As a physician-scientist, I (Bibi Bielekova) care for patients with multiple sclerosis (MS). I began my career in the late 1990s and soon observed the pervasiveness of poor experimental design in the field. In 2004, I coauthored a study in the neuroscience journal *Brain* based on a review of more than 200 papers on laboratory biomarkers for MS published between 1982 and 2002. Only a few of these published papers fulfilled widely accepted criteria of sound study design. In most, the sample size was too small and lacked appropriate controls. In many, the researchers used inappropriate statistical tests or failed to randomize their samples. I was not surprised that few produced replicable results.

At the time, I well understood that reproducibility is truth. Without reproducibility, we cannot make significant progress in human health; the search for MS biomarkers provides a case in point. Having biomarkers that reliably measure different biological processes that kill brain cells and cause disability—markers such as levels of chemicals that could indicate various functions of immune cells and cells in the brain—would lead to a better understanding of MS, its nature and progression,

Illustration by Shonagh Rae



and could revolutionize its treatment. As a young scientist, it was deeply discouraging to realize that research on MS biomarkers was of such poor quality that I could not determine if any of them might be useful for finding new treatments.

Since 2004, the problem of irreproducible findings has been documented in every area of preclinical and clinical research, on afflictions including Alzheimer's, depression, cancer, and stroke. Indeed, of the approximately 1.5 million papers published in biomedical journals each year, researchers have estimated that at least half are so poorly designed, conducted, analyzed, or reported that the results cannot be replicated and therefore cannot be trusted. The phenomenon is so widely recognized, it now has a name: the reproducibility crisis.

As a writer and health policy expert with a long-standing interest in research integrity and the ethics of clinical studies, I (Shannon Brownlee) first became interested in reproducibility while writing a review of Richard Harris's 2017 book, *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*, in which he describes the reproducibility crisis in biomedical research. Since I'm also a former scientist and senior vice president of a health care think tank, I read the book with shock followed by outrage. I knew there were many sources of waste in our health care system, but the failure on the part of the research enterprise to ensure the validity of publicly funded studies, especially those involving patients, seemed particularly egregious.

In 2020, we began collaborating, drawing on Bielekova's decades of research at the National Institutes of Health (NIH) and Brownlee's knowledge of health care policy.

We view the reproducibility crisis as a fundamental challenge to the biomedical research endeavor, one that nothing short of an integrated overhaul of the system can fix. And it is not merely an academic issue; irreproducibility has real-world consequences both for patients' health and for the relationship between researchers and the society that funds them. Altering the way research output is disseminated and evaluated, using transparent, data-driven measures of methodological rigor and social value, could transform the system's incentives, change the way scientists work, and produce public health benefits that are almost beyond imagining today.

Uncovering a long-brewing crisis

The scale of the reproducibility crisis surfaced in 2009, with a letter in *The Lancet* coauthored by Paul Glasziou, then the director of Oxford University's Centre for Evidence-Based Medicine and now a professor of evidence-based practice at Bond University in Queensland, Australia, and Iain Chalmers, the founder of the James Lind Library and cofounder of the Cochrane Collaboration. Glasziou and Chalmers estimated that as much as 85% of biomedical research funding was being wasted on studies with shoddy methodology and reporting. Though it got little attention at the time, the letter is now widely cited.

In 2012, *Nature* published a paper by C. Glenn Begley, a physician-scientist who was head of the oncology division at the biotech company Amgen. Begley had long been quietly frustrated by the low rate at which seemingly promising basic research translated into viable treatments for cancer. He and his team had tried to replicate 53 "landmark" preclinical cancer studies, going to extraordinary lengths to follow the authors' methods. For 20 of the papers, Amgen scientists traveled to the original labs to watch the experiment being redone. Only this time they required the researchers be blinded to which group of animals or cells was being subjected to the experimental treatment and which was the control. Of the 53 original studies, only 6 could be replicated—even by the original investigators themselves. Some of the 47 irreproducible studies had spawned entire fields, with hundreds of secondary publications.

Unlike Glasziou and Chalmers's estimate—and others, such as Bielekova's, that had documented the problem of irreproducibility but flown under the radar—Begley's paper was greeted by scientists with outrage and ridicule. By then, researchers had begun to acknowledge the low rate of return on preclinical cancer studies, but they routinely chalked it up to the complexity of the disease: "Cancer is hard." Begley was pointing not at the disease,

Altering the way research output is disseminated and evaluated, using transparent, data-driven measures of methodological rigor and social value, could transform the system's incentives, change the way scientists work, and produce public health benefits that are almost beyond imagining today.

but at the researchers themselves. Colleagues stood up at conferences and told him his paper would damage the cause of science and decrease research funding. He received hate mail and threats. Many subsequent studies, however, have backed up Begley's observations.

And the problem is not unique to biomedicine. In 2015, *Science* published a paper by a team of researchers from the Open Science Collaboration describing the results of an international effort to replicate 100 psychology studies considered part of the core knowledge for understanding personality, relationships, learning, and memory. After defining multiple criteria to determine whether a replication could be considered successful, the team tried to reproduce 100 original findings. Across the criteria, they succeeded less than half the time.

By 2016, however, what had seemed outrageous would be widely acknowledged. That year, *Nature* surveyed 1,576 scientists, finding that more than 70% reported that they had had trouble replicating experiments published by others. More than 50% reported that they sometimes could not repeat their own results. And 90% agreed that science was in the midst of a reproducibility crisis.

Incentives that encourage bad science

We see this crisis as a natural outgrowth of incentives in the current research environment—incentives that do not merely permit bad science but can actively encourage it. Hiring decisions, promotions, tenure, professional stature, and, for many scientists, even salaries depend first and foremost on bringing in grants and publishing papers—rather than producing validated and reproducible results. Yet it is valid results, not publications, that are essential to creating the knowledge that makes new and better treatments possible.

Funders, including NIH, do not reward research institutions for the rate at which their scientists publish replicable studies—and do not penalize them for irreproducible research. Therefore, neither institutions nor researchers are incentivized to focus on ensuring replicability. As Brian Nosek, a professor of psychology at the University of Virginia who was the corresponding author on the *Science* paper, described it: "Fundamentally I'm rewarded for publishing, not for getting it right. I don't get rewarded for demonstrating the validity of somebody else's work. I don't get rewarded for producing ambiguous or negative results. The reward system is misaligned with the reality of how science works."

At the heart of this reward system is a publication metric called the "impact factor," which is calculated by dividing the number of times a journal's papers have been cited by the number of articles the journal published over a two-year period. The impact factor has taken on a life of its own, signaling to funders, hiring and promotion committees, other scientists in the field, and the press that a published finding is important.

The impact factor has become a kind of shorthand adopted by every corner of science. Considering impact factors and citation counts allows peer reviewers of grant proposals and journal submissions to rate the quality of the science without taking the time to delve into the study's methods and its likely validity. The quantity of a scientist's publications and their impact factor also holds enormous sway in academia. A 2018 study of 92 elite universities around the world found that when hiring and promoting faculty members, 95% of these universities consider the number of papers a researcher had published and 28% look at impact factor. Research reproducibility and its value to health are rarely if ever listed by academic institutions as criteria for hiring, promotion, or tenure.

Despite the shortcomings inherent in publication metrics, big institutions have little appetite for moving away from them, according to Jeffrey Flier, the former dean of Harvard University's Faculty of Medicine. "The incentives are not aligned for the dean of Harvard Medical School to be writing or talking about the problem of irreproducibility," said Flier. "[Deans] are supposed to describe the work at their institutions as transformative."

The resulting legacy of irreproducible studies profoundly influences every generation of young researchers. For example, an early-career scientist visiting Bielekova's lab had access to a set of tissue samples from patients with MS in his home country and was intending to do a series of experiments based on the findings of an article published in a high impact journal that had accrued hundreds of citations. This high-impact paper, however, had been unequivocally discredited by two independent validation studies of much higher technical quality but with far fewer citations. Unfortunately, the literature, and search engines such as PubMed, don't give researchers any clues about such dead ends.

And because incentives for researchers are all about publishing positive results, the negative ones—upon which scientific progress also depends—rarely see print. A researcher in Bielekova's lab conducted an elegant, novel set of experiments with results that initially looked spectacular. Most researchers

would have rushed to publish their results at this stage. However, the experiments used so-called omics technologies—genomics, transcriptomics, proteomics, and metabolomics—which measure large numbers of markers, such as proteins or genes, in a single tissue sample. The more complex the analyses researchers apply, the greater the chance they will get impressive-looking results that are not reproducible. The machine-learning algorithms used in these analyses are so powerful that when apparently useful biomarkers are put together in a complex mathematical way, the process generates results than can seem to differentiate between sick patients and healthy ones. The only way to determine whether the results represent a fluke or a truly useful clinical tool is to apply the same analysis in a blinded way to a new set of patients. For this reason, Bielekova's lab has a policy that all results from omics technologies must be independently validated before they are published.

The researcher in her lab could not validate his results. Though he was urged to write them up and publish so that other researchers would not go down the same blind alley, he did not. Why? Because it would have required time and effort that would not be rewarded by a prestigious journal placement, or possibly any publication at all. He would rather devote his time to new experiments, which might be more likely to lead to a publication that would provide a career advantage.

This fierce competition for publications profoundly shapes the outlook of early-career researchers, who face a brutal path to securing scarce academic positions. "The stakes are very high," says Maryann Feldman, a leading scholar of innovation policy at the University of North Carolina at Chapel Hill and the former director of the Science of Science and Innovation Policy program at the National Science Foundation. "It's a pressure cooker with a lot of people jockeying for [few] positions."

A flurry of disconnected reforms

To date, there have been few effective reforms to this multifactorial problem, never mind a shift in the incentives governing the way research is conducted. Given the number of players in the system and its complexity, it should also be unsurprising that most of the existing solutions to the reproducibility crisis are focused on isolated aspects of it. For example, last year a group of researchers, journal editors, and grantmakers issued a set of guidelines called the Hong Kong Principles. The guidelines address, among other things, the criteria used for hiring, promotion, and tenure decisions. Thus far, 19 universities and other research institutions have endorsed the principles, but they are entirely voluntary.

Another voluntary movement was sparked by a meeting convened by NIH in 2014, leading several journal editors to endorse abolishing word limits on the methods sections of papers so researchers can include all the necessary details that others need to replicate the experiments. Some journal editors have gone further, initiating open peer review, where the peer reviewers and their statements are publicly available, and encouraging authors to make all their data available in supplemental materials.

These are worthwhile efforts, but of all the players in the system, funders could have the greatest influence over the behavior of scientists and research institutions, by insisting that grantees' research be designed and conducted in ways most likely to produce valid results. More than two dozen funders from around the world, including NIH, have joined a consortium, called the Ensuring Value in Research (EViR) Funders' Forum, to develop a set of standards that could begin to move toward greater rigor. For instance, EViR supports requiring grantees to offer a formal appraisal of the methods of prior experiments and to publish in open-access journals such as *PLOS* or *eLife*, a new publication started by the Howard Hughes Medical Institute.

Although piecemeal reforms such as these could have an impact on reproducibility, they do not add up to the integrated overhaul we believe is necessary. Funders could shift the incentives in favor of sound methodology, finding ways to turn scientists into collaborators—rather than competitors—whose common goal is improving health. Funders should also provide academic institutions with reasons to evaluate scientists based on the quality of their work.

Could AI change the culture and incentives of biomedical research?

What is required is a change in the research culture to realign incentives toward productivity and innovation. We believe that technological advances in artificial intelligence can be leveraged to build a tool that objectively assesses the true value of individual academic publications for society.

We've named this tool the Biomedical Research Network (BRN) and envision a dynamic machine-learning platform that would assess the methodological rigor, reproducibility, and utility of studies. By changing the powerful incentives driving institutional and individual behavior, the BRN can help transform the scientific enterprise into a self-regulating system that increases the rate at which it produces scientific breakthroughs that have meaningful impacts for society.

Using natural language processing and machine-learning algorithms, the BRN would start by creating a "family tree" for related papers. This tree would show the relationships between individual experiments, as well as how they

contributed to concrete societal benefits such as patents, new drug applications, or successful health campaigns. The BRN would also analyze and grade the quality of the methods used in studies, integrating and comparing related studies to determine their reproducibility and, separately, their societal impact. These mechanisms could simultaneously give researchers insights into successful experiments while helping them avoid blind alleys created by published studies of poor quality, or those known to be plagued by problems such as faulty statistical analysis or contaminated cell lines.

Consider, for example, the Interleukin-17 (*IL-17*) gene. First sequenced in 1993, *IL-17* has led to multiple lines of research into its role in autoimmune disorders. One line has been highly productive, leading to patents of successful treatments for psoriasis. Other lines of research, however, have proved irreproducible. The BRN family tree would show the links between the original discovery of *IL-17* and subsequent successful studies that led to treatments, alerting researchers that they could rely on the first set of studies as a foundation for further research. Conversely, researchers would be able to avoid any irreproducible research. Scientists and institutions affiliated with the reproducible research would be recognized for that work, as would those who contributed to societally important lines of discovery. Importantly, this recognition would accrue over decades, as lines of productive research continue to grow and branch out, building upon earlier studies.

In this way, the BRN could effectively curate knowledge, using transparent, objective methodologies to do so. Once the BRN was available, early-career researchers and the public would know which branches of scientific investigation have proven reproducible, which have succeeded in providing health benefits and which, though they may have generated fame and followers, were eventually discredited. Academic hiring and promotion committees and grantmakers would also be able to use the network to judge which scientists consistently produce validated results. Funders could recognize the institutions whose scientists' work has already contributed to human health or holds the potential to do so. This system would serve as a strong incentive for individual scientists to avoid publishing poorly designed and executed studies, while also providing disincentives for institutions to reward sloppy science.

Through such objective measures and transparency, the BRN could also create a cascade of other positive changes that could significantly increase the speed of biomedical innovation. It could, for example, enable a dynamic publication system, dramatically decreasing the time researchers now waste writing separate papers for every new result in a series of related studies. The current system, with its emphasis on the number of papers published, rewards slicing and dicing results to maximize the number of publications—and hoarding data to extract as many papers as possible. A dynamic publication process would allow new results to be incorporated into existing papers that would rapidly be available online. The BRN would also encourage collaboration through the sharing of data and materials, as scientists who make such valuable tools available would also share the positive societal impact score when the use of their tools resulted in new knowledge or societal value.

The Biomedical Research Network can help transform the scientific enterprise into a self-regulating system that increases the rate at which it produces scientific breakthroughs that have meaningful impacts for society.

The BRN could further reshape the culture of science by reforming processes such as peer review. Currently, single blinded peer review empowers reviewers to determine the fate of a publication without accepting public responsibility, enabling and institutionalizing bias without effectively weeding out bad science. Open peer review has failed so far in part because reviewers are not rewarded for the time it takes to do a thorough review. The BRN, by contrast, could generate a numerical score for reviewers, rewarding them for constructive reviews that identify flawed experiments and offer actionable recommendations for improving the quality of the paper.

This score would provide both funders and research institutions with information that can allow them to reward high-quality reviewers. A publicly available BRN peer reviewer score would encourage reviewers to scrutinize the methods and results of a paper far more closely than they generally do today, and it would reward creative critical thinking. Critiques from scientists with high cumulative reviewer scores would impel authors to consider the reviewers' arguments carefully and perform additional experiments to ensure the validity of findings.

Finally, the BRN could give credit where credit is due, creating an atmosphere where teams of creative individuals can flourish. Successful research requires a diverse workforce with complementary skills. Current publication criteria, however, favor intellectual over manual or technical contributions. The BRN could help institutions credit all team members by measuring scientific contributions more fairly, and potentially increasing productivity by motivating and rewarding all members of the research team.

Tapping artificial intelligence to assess published research is already being done. Nosek and his team have funding from the Defense Advanced Research Projects Agency (DARPA) to train AI to read published social and behavioral science research, including the methods sections, and aggregate data relevant to evaluating the papers' claims. Nosek says the algorithms are showing some accuracy in their assessment of the reproducibility of such studies.

Nosek's organization, the Center for Open Science, is spearheading another disruptive project, called Registered Reports. More than 300 journals have pledged to send study proposals, rather than completed papers, out for peer review—before the research has even commenced. The reviewers focus on spotting methodological flaws. If the proposal passes muster, the journal agrees to publish the results regardless of whether they are positive or negative, provided the researchers follow the registered methodology. This proposal, and other initiatives aimed at increasing transparency and usefulness, could be readily integrated into the BRN.

Patients can't wait

The time has come for the biomedical research community to acknowledge the failures of the current system and think expansively about how to change it. One of us has been waiting 20 years for this moment. And though they may not realize it, MS patients have been waiting, too. Two decades in the life of a person with this disease can mean progression from slight disability to profound impairment. While the scientific community has only slowly begun to acknowledge its dysfunction, MS has continued its path through millions of lives worldwide. A handful of new drugs have emerged in the past 20 years, but in a cruel twist, they work only for young people at the outset of their disease, leaving patients older than 54 with practically no effective treatments.

Imagine, instead, 20 years during which grantmakers funded well-trained, collaborative researchers, who were rewarded for the quality of their research and the benefit it offered to patients. Such a scenario might or might not have resulted in better therapies for MS and a host of other intractable conditions, but it's hard to believe that given the same amount of funding, therapies would have progressed as slowly as they have.

The biomedical research system now needs a strong dose of coordinated leadership from within. The COVID-19 pandemic and development of multiple vaccines have shown that medical innovation can be both rapid and transformative, provided administrators are willing to dismantle barriers and promote creativity and collaboration. The validity of science is the responsibility of all scientists, and courageous leadership is what is needed now to refocus the entire enterprise on ensuring the public's trust and improving the lives of patients.

Shannon Brownlee *is a lecturer at the George Washington University School of Public Health and a former senior vice president of the Lown Institute, a health care think tank.* ***Bibiana Bielekova*** *is a physician-scientist and the chief of the Neuroimmunological Diseases Section in the Intramural Research Program (IRP) of the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health.* *The writing of this article was supported, in part, by the IRP of the NIAID.*

RECOMMENDED READING

Bibiana Bielekova and Shannon Brownlee, "The imperative to find the courage to redesign the biomedical research enterprise [version 1; peer review: 1 approved with reservations]," *F1000 Research* 10, no. 641 (2021).

Richard Harris, *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions* (New York, NY: Basic Books, 2017).

Malcolm R. Macleod, Aaron Lawson McLean, Aikaterini Kyriakopoulou, Stylianos Serghiou, Arno de Wilde, Nicki Sherratt, Theo Hirst, Rachel Hemblade, Zsanett Bahor, Cristina Nunes-Fonseca, Aparna Potluru, Andrew Thomson, Julija Baginskaite, Kieren Egan, Hanna Vesterinen, Gillian L. Currie, Leonid Churilov, David W. Howells, and Emily S Sena, "Risk of Bias in Reports of In Vivo Research: A Focus for Improvement," *PLOS Biology* 13, no. 10 (2015): e1002273.

David Moher, Paul Glasziou, Iain Chalmers, Mona Nasser, Patrick M. M. Bossuyt, Daniël A. Korevaar, Ian D. Graham, Philippe Ravaud, and Isabelle Boutron, "Increasing value and reducing waste in biomedical research: who's listening?" *Lancet* 387, no. 10027 (2016): 1573–1586.

Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science* 349, no. 6251 (2015).