2012

Evidence-Based Litigation Reform

Mark Spottswood
Florida State University College of Law

EVIDENCE-BASED LITIGATION REFORM

Mark Spottswood[*]

I. INTRODUCTION

When we seek to assess the impact of existing rules of legal procedure or propose improvements to them, we are faced with stubborn questions defying easy answers. Who will benefit from a new rule? Who might be harmed? Will the rule cost more, or add delay to the resolution of cases? And perhaps most importantly, will the legal system as a whole become more just or fair if the rule is adopted?

When trying to answer such questions, we encounter a poverty of useful data. We can rely on our intuitions and theoretic understandings to choose between possible rules, but history is littered with examples of well-intentioned rule reform that led to effects drastically different from what rule designers imagined.[1] For instance, when the enactors of the Federal Rules of Civil Procedure added summary judgment and fact discovery to the litigation toolkit, they failed to envision the future they were creating, in which discovery became the most time-intensive aspect of litigation practice and summary judgment disposed of more cases than trials did.[2] Their failure to anticipate these things is not unusual, and it should not be surprising. The litigation system is vastly complex, and how it will change in response to a new rule depends on obscure interactions between a new rule, the background matrix of existing rules and practices, the preferences

---

[*] Assistant Professor, Florida State University College of Law. I am very grateful to Ted Eisenberg, Marty Redish, Suzanna Sherry, Janet Alexander, Sergio Campos, Sandra Sperino, Robin Effron, Verity Winship, Corey Yung, Nancy Leong, Justin Pidot, Dan Markel, Reid Fontaine, Marshall Kapp, Colin Miller, David Levine, Kara Hatfield, Sam Wiseman, Hannah Wiseman, Shawn Bayern, Franita Tolson, Jake Linford, Jay Kesten, Tetyana Payosova, Curtis Bridgeman, Jim Rossi, Mark Seidenfeld, Manuel Usted, Karen Sandrik, Wayne Logan, Raja Raghunath, and Sarah Mirkin for their helpful comments and suggestions, as well as to additional commenters at the 2012 Junior Faculty Federal Courts Workshop, the 2012 Annual Meeting of the Law and Society Association, and the 2011 FSU College of Law Summer Enrichment Series.

[1] *See* Laurens Walker, *A Comprehensive Reform for Federal Civil Rulemaking*, 61 GEO. WASH. L. REV. 455, 484–89 (1993) (describing the intuitive, theory-driven approach that has dominated civil rule making); Thomas E. Willging, *Past and Potential Uses of Empirical Research in Civil Rulemaking*, 77 NOTRE DAME L. REV. 1121, 1121–22, 1197 (2002) (noting that until very recently, rule makers have rarely looked to empirical research to support their conclusions, and that even in recent times most of the studies they employed were of limited utility for answering questions about the causal impact of rules).

[2] *See generally* discussion *infra* Part III.

of the many players involved, and their strategic adaptations and counter-adaptations to new realities.[3]

A number of other disciplines face similar design challenges. It is also hard to predict how the human body will react to a new drug, how a change in a product's design or manufacture will impact its reliability or safety, or how well a new regulatory policy will achieve the goals of its designers. At times policymakers in these other fields have relied on their intuitive understanding of how a system works to make inferences about the results of a new intervention. Sometimes this may go very well, but it might also go disastrously wrong.

To take just one chilling example, note the long-standing use of bloodletting as a therapeutic tool in medicine.[4] Therapeutic exsanguination persisted for millennia before a few enterprising physicians decided to test the theory that it improved health outcomes.[5] It turned out, of course, that it was bad for most patients to be bled, but this was not obvious before the systematic collection of data.[6] Looking at individual case anecdotes in the absence of controlled experiments, it was simply impossible to tell whether practices like bloodletting were saving lives or not. The rise of the more effective medicine we now enjoy was driven by a turn to evidence-based evaluation of both disease theory and therapeutic effectiveness, and could not have occurred absent a willingness to conduct controlled experiments as a means of identifying which therapies work and which ones do not.[7] More recently, the evidence-based medicine movement has offered strong arguments that medical care could be further improved if doctors depended more on experimental verification of treatments and less on intuition.[8]

Unfortunately, given our present methods of evaluating procedural success, we have little cause to be confident that our existing legal rules

---

[3] *See* Ronald J. Allen, *Taming Complexity: Rationality, the Law of Evidence, and the Nature of the Legal System* 16–17 (Thirteenth Int'l Conference on Artificial Intelligence and Law Workshop on Artificial Intelligence & Evidential Inference, Working Paper No. 11-52, 2011), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1845817. Allen describes the litigation environment as a "complex adaptive system with emergent properties" that are unpredictable from the underlying structure, in which long-term equilibria develop out of innumerable interactive choices by individual decision makers. *Id.* at 17. Such chaotic systems tend to resist reliable prediction. *Id.*

[4] *See* Edward Shorter, *Primary Care*, *in* THE CAMBRIDGE HISTORY OF MEDICINE 103, 109 (Roy Porter ed., 2006).

[5] *See id.* at 108–10.

[6] *See id.*; *see also* M. WEATHERALL, IN SEARCH OF A CURE: A HISTORY OF PHARMACEUTICAL DISCOVERY 16 (1990).

[7] *See generally* discussion *infra* Part II.

[8] *Cf.* Jeffrey A. Claridge & Timothy C. Fabian, *History and Development of Evidence-Based Medicine*, 29 WORLD J. SURGERY 547, 547 (2005).

work any better than bleeding and purging a sick patient. Like medieval doctors, we watch the patient (that is, individual cases) with some care, but we rarely try to systematically measure the differential effects of competing rules on case outcomes. What is worse, even when we do try to measure the effects of procedural rules, our investigations routinely neglect to measure the impact of a rule on the accuracy of case outcomes when we attempt to evaluate its effectiveness. This problem is a grave one. Outcome accuracy—meaning a correspondence between the factual understandings that motivate legal decision makers and the historical facts that gave rise to litigation—is among the most important values that we try to optimize through procedural rules.[9] But because we have never tried to systematically measure the accuracy of case outcomes, our ability to estimate the accuracy of civil and criminal case outcomes is severely limited.

Some may object that outcome accuracy is something that cannot be measured effectively,[10] but this is not the case. In order to measure the accuracy of case outcomes in general, we must first parse the concept of accuracy into a measurable form, and then create "gold standard" methods for assessing it. I propose one potential measurement protocol in this Article that might be able to allow some systematic measurement of factual accuracy in procedural outcomes, while acknowledging that this solution would be costly to implement on a large scale. In brief, this protocol entails obtaining a record of what facts motivate those who are responsible for producing legal outcomes, and then comparing those beliefs with the results of a more detailed, in-depth investigation into the factual background of a case. This protocol could be implemented on a relatively small scale if researchers are interested in particular questions regarding procedural validity, but it could also be scaled up as a basis for inter-systemic comparisons.

Although it might seem that such investigations would be second best to the outcome of a trial on the merits, this objection turns out to be less weighty than it seems, because the vast majority of cases are either settled or decided on a pretrial motion.[11] We have good reason to worry that in

---

[9] *See* Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181, 183–85 (2004).

[10] *Cf.* Steven B. Duke et al., *A Picture's Worth a Thousand Words: Conversational Versus Eyewitness Testimony in Criminal Convictions*, 44 AM. CRIM. L. REV. 1, 3–4 (2007) (noting, for one example, that most think of DNA evidence as a "gold standard" means of assessing verdict accuracy, but that it is available and dispositive in only a small percentage of criminal cases).

[11] *See* Herbert M. Kritzer, *Adjudication to Settlement: Shading in the Gray*, 70 JUDICATURE 161, 161–64 (1986).

many such cases, key decision makers know only a fraction of the information that a jury trial might produce.[12] Nor is it fatal to an evidence-based litigation reform project that many cases involve fundamentally ambiguous records or "he said, she said" credibility disputes.[13] A proper system for measuring outcome accuracy should be able to separately categorize those cases that the legal system viewed as close, and if legal decision makers and gold standard reference investigators agree that cases involve difficult ambiguities, then the legal decisions can be said to be just as accurate as when investigators and decision makers agree that a case has only one clearly correct outcome. But such an investigation would also help reveal when cases appear ambiguous only because insufficient investigation has been performed, or when cases are thought to have clear right answers only because complicating evidence has been ignored.

Unless and until we investigate the ways that our existing procedural devices affect outcome accuracy, we should have little confidence that any of our procedures are particularly effective ways of generating factually valid legal results at acceptable cost. Although a few scholars, following Charles Nesson, might wish to ignore evidence about our system's accuracy if it undermines public confidence in verdicts,[14] more will think that we should attend to both the system's accuracy and its legitimacy when designing rules.[15] Nor is the problem of inaccurate outcomes of merely theoretic interest. Inaccurate outcomes involve chilling social costs: some people go to jail for crimes they did not commit, some are forced to pay others for wrongs they did not cause, and some who deserve punishment or sanction evade it. Every so often, we get a brief window into the defects in our procedures. A prominent recent example, from the criminal procedure arena, was the rise of DNA evidence. Suddenly, a new forensic technique showed that many "ordinary" criminal convictions, which had seemed as reliable as most other case outcomes, were factually invalid.[16] But since we

---

[12] *See* Nora Freeman Engstrom, *Sunlight and Settlement Mills*, 86 N.Y.U. L. REV. 805, 816–17 (2011) (describing the practice norms of many settlement-focused law firms, in which there is little attorney-client contact or attorney investigation into case facts).

[13] *Cf.* Alex Stein, *An Essay on Uncertainty and Fact-Finding in Civil Litigation, with Special Reference to Contract Cases*, 48 U. TORONTO L.J. 299, 300 (1998) (explaining that fact-finding inherently involves unknowable uncertainty, due to the frequent situation of "evidential scarcity" that bedevils real-world litigation).

[14] Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1358–59 (1985).

[15] *See* discussion *infra* Part III.B.

[16] *See* BRANDON L. GARRETT, CONVICTING THE INNOCENT: WHERE CRIMINAL PROSECUTIONS GO WRONG 1–13 (2011); Richard A. Rosen, *Innocence and Death*, 82 N.C. L. REV. 61, 70 (2003) (noting

rarely get such insight into the validity of case outcomes, the real harms of inaccurate resolutions are mostly hidden from our sight.

In this Article, I attempt to show the need for investigations into the causal effects of procedural rules on the accuracy of case outcomes. Drawing on examples from the medical literature, I sketch out what would be involved in trying to measure the impact of various procedural rules on the accuracy of case outcomes. Although such research could, for the first time, enable us to identify which procedures succeed at resolving cases accurately and which do not, it would also be expensive and difficult. Nevertheless, if we value accurate litigation outcomes a great deal, we might wish to bear these burdens, as we already do in the name of protecting our health.

After imagining what a more evidence-based world of procedural design might require, I then turn to consider what we might learn about existing procedural debates and problems by comparing current practice to that idealized approach. One sobering implication is that the existing empirical literature on litigation rules may have less to teach us than we might hope. So long as we fail to measure accuracy, information about the variables we can track cannot provide a strong platform on which to base rule-design decisions, because we can never be sure that improvements in other procedural values are not coming at the expense of the system's accuracy.

At the same time, this discussion has important implications for those with the power to make new rules, even if they are unwilling to do the hard work needed to systematically test the accuracy effects of existing procedures. In particular, I urge that rules be made in a way that both allows continuing evaluation of comparative data and that incorporates as wide a spectrum of on-the-ground experience as possible. On the spectrum of possibilities, this means that the best rules will be those tested out first on a local level and then made broadly applicable through the formal rule making process. Conversely, rules established by constitutional mandate from high-level courts will tend to be under-informed and will stifle our ability to test the premises that prompted their announcement.

This Article will proceed in five parts. First, I will discuss the rise of an evidence-based research culture in Western medicine in order to highlight some of the lessons that legal reformers can learn both from its successes and from its continuing struggles. Second, I will make the case for why an evidence-based approach could be a valuable change to the way that

---

that "[u]ntil the moment when the DNA test results came back, almost none of these cases [in which DNA evidence revealed that a conviction was in error] would have been considered exceptional among criminal cases").

juridical procedures are designed, and why existing attempts at data collection fall short of being able to facilitate such an approach. In this discussion, I will draw primarily on examples in civil procedure, although similar observations could be applied to rule design in criminal procedure or evidence. Third, I will address the difficulties that such a movement will face, with particular emphasis on the importance of developing a gold standard baseline for assessing the accuracy of case outcomes. Fourth, I will discuss a research design that might address these problems and make systematic accuracy measurement an achievable possibility. Finally, I will discuss some of the smaller steps we could take towards evidence-based litigation reform.

II. THE BENEFITS OF EVIDENCE-DRIVEN DESIGN: THE MEDICAL EXAMPLE

Before discussing the problems I see in existing approaches to evaluating the impacts of rules of legal procedure, I will first take a detour into the history of medical innovation and practice, with a special focus on the gains that doctors have achieved by systematically measuring the effects of potential treatments on health outcomes. This discussion will have several payoffs for scholars of procedural reform. First, the world of medieval medicine illustrates the dark side of relying on intuition and theory alone when evaluating how successful a treatment is. In that time and place, doctors did an extraordinary amount of harm to patients, despite having good intentions, because it was impossible to assess whether a treatment was helping or hurting on a case-by-case basis. This story has disturbing implications regarding the extent to which modern legal procedures achieve high levels of case accuracy, an outcome that goes similarly untested in our modern world.

Next, I will describe how doctors were able to substantially improve the quality of patient care through a centuries-long effort to develop more accurate biological theories and to test the effect of possible treatments using controlled trials. This example offers a potential source of inspiration to legal reformers. Perhaps, if we were willing to invest resources on a similar scale with the goal of developing a measurably accurate procedural system, we could find similar success in improving the quality of justice.

Finally, I will discuss modern methods of medical testing, with special emphasis on the FDA-mandated system of pharmaceutical drug testing and the recent reform efforts of the evidence-based medicine movement. By surveying these modern approaches, we can find sources of guidance for how legal procedures could be similarly evaluated. In particular, the evidence-based medicine movement has offered insights into the best

means of evaluating the efficacy of diagnostic tests, and this model provides a worthy example for procedural reformers to emulate. We will also see, however, that implementing such procedures would require a willingness to invest sizeable amounts of time and money in improving the quality of justice, and that the approach can lose much of its value if it becomes politicized or captured by special interests.

A. Medicine Before the Rise of Systematic Effectiveness Testing

The phrase "evidence-based medicine" may be new,[17] but people have long been interested in curing disease and increasing physical health. Although there are a few intriguing early examples of medical experiments,[18] for a very long time medical practice doctors relied primarily on ancient authoritative texts and theoretical understandings of disease processes to choose treatments for the sick, supplemented by various forms of faith healing and folk medicine.[19] Throughout much of Western history, medicine was a mixed bag; it offered some surprisingly effective treatments for a few conditions,[20] but had little to offer for some of the most common medical challenges of the day, such as the often fatal process of childbirth, and favored the routine use of actively harmful procedures like bleeding and purging.[21]

It would be a mistake to think that doctors relied on such methods because they were uninterested in curing patients. Rather, they lacked the means by which to differentiate between good and bad treatments, and so tended to defer to ancient authorities to guide their decisions.[22] In the

---

[17] *See* Claridge & Fabian, *supra* note 8, at 547 (tracing the origin of the term to the mid-1990s).

[18] *See, e.g.*, *Daniel* 1:1-16 (relating a dietary experiment proposed by Daniel and implemented by his guards); Christian Gluud et al., *Commentary on the* Ben Cao Tu Jing (Atlas of Materia Medica), JAMES LIND LIB. (2003), http://www.jameslindlibrary.org/illustrating/articles/commentary-on-the-ben-cao-tu-jing-11th-century-atlas-of-mater.pdf. Also notable among early medical practitioners were the Empiricists, a sect of ancient Greek physicians who eschewed the fanciful anatomical theories of the day, preferring to prescribe treatments by analogy to prior successful interventions rather than based on theoretic understandings of human disease processes. *See* Vivian Nutton, *The Rise of Medicine*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 53.

[19] *See generally* Claridge & Fabian, *supra* note 8, at 547; Nutton, *supra* note 18, at 54–64 (relating how, for a long time after Galen, most medical writers devoted themselves primarily to collecting and restating past medical wisdom rather than challenging it).

[20] *See* Nutton, *supra* note 18, at 66. One historian lists medieval treatments for abdominal injuries, hernias, anal fistulae, bladder stones, and cataracts as particularly effective. *Id.*

[21] *Id.* at 68.

[22] Perhaps the most influential authority was Galen of Pergamum, a Roman physician and philosopher who went to great lengths to synthesize the medical literatures of antiquity with his own investigations into anatomy, which he based primarily on dissections of pigs and monkeys. Galen was

dominant paradigm, which dated back to Hippocrates, the body contained four "humors"—phlegm, red or yellow bile, black bile, and blood—and disease occurred when the balance between these humors was disturbed or when they were contaminated by toxins.[23]

Many popular medical treatments attempted to balance the humors or rid the body of toxins by draining it of fluids that were believed to be contaminated.[24] To accomplish this, physicians engaged in bloodletting, heaped blankets on feverish patients to make them sweat, gave them emetics to make them vomit, and purged their intestines with powerful laxatives.[25] Even worse, the compounds used to open patients' bowels were often highly toxic substances.[26] Some herbal remedies were also employed to help balance the humors, but such remedies were mixed together in complex formulations so that the body could "select[] whatever ingredient would correct the humoral imbalance."[27] In short, in order to improve the health of their patients, doctors systematically made them "anaemic through bloodletting, deplet[ed them] . . . of fluids and valuable electrolytes via the stool, and poison[ed them] . . . with compounds of such heavy metals as mercury and lead."[28]

It is easy to look back on such a calamity and think that it must have been obvious that such procedures were a terrible idea, but it was not so. In the world to which traditional medicine applied, many diseases were frequently fatal, with or without the addition of medical treatment.[29] In any individual case, it would be impossible to tell whether a death occurred *because of* a heroic treatment or *in spite of* it. Moreover, there were certainly many anecdotes that a doctor could point to in support of such theories. Thanks to our immune systems, people often recover from disease and other calamities without medical intervention, and with a strong enough constitution recovery could no doubt follow even a stringent regime of

---

forced to rely on animal studies due to the Roman ban on the dissection of human cadavers. The basic understanding of disease processes and therapeutic treatment he set out endured for millennia. *See id.* at 54–55; Roy Porter, *What Is Disease?*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 80; Shorter, *supra* note 4, at 103.

[23] *See generally* Shorter, *supra* note 4, at 103–04; Miles Weatherall, *Drug Treatment and the Rise of Pharmacology*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 214.

[24] *See* Shorter, *supra* note 4, at 104–05.

[25] *Id.*

[26] *Id.* at 108.

[27] WALTER SNEADER, DRUG DISCOVERY: A HISTORY 22 (2005).

[28] Shorter, *supra* note 4, at 109. Given the content of traditional "heroic medicine," it is perhaps unsurprising that some developed a dread of doctoring; indeed, Joseph Addison commented in 1711 that "when a nation abounds in physicians, it grows thin of people." *Id.* (quoting 1 THE SPECTATOR IN FOUR VOLUMES 64–65 (1945)).

[29] *See* L. Cilliers, *Where Were the Doctors When the Roman Empire Died?*, 26 ACTA THEOLOGICA SUPPLEMENTUM 62, 75 (2006).

bleeding, blistering, and purging.[30] And in some cases, these therapies may even have helped patients feel better, thanks to the power of the placebo effect.[31] A patient's belief that a particular cure will be beneficial can have therapeutic effects even where the treatment itself is biochemically useless, and many patients believed powerfully that traditional medicine could cure them.[32] Indeed, some patients attributed their healing to truly awful medical interventions.[33] So, despite its horrors, traditional medicine seemed plausible enough during the long stretch of history to which it applied, and continued to be demanded by patients even as doctors began to lose faith in it.[34] The example of medieval medicine shows us that well-meaning professionals can do great harm to the people they are trying to help by depending on conventional wisdom and theories to choose their treatments if those theories do not rest on a firm and tested foundation. This example should be sobering for those who would reform legal procedures, given that our existing procedural toolkit has been subject to only limited testing.

B. Sowing the Seeds of Doubt

Two early reformers, Paracelsus[35] and Vesalius,[36] took the first steps towards making medicine testable.[37] Although he had many other odd notions, Paracelsus pioneered the idea of using purified, specific compounds as remedies rather than the complex blends of herbs known as

---

[30] *See* DAVID L. SACKETT ET AL., EVIDENCE-BASED MEDICINE: HOW TO PRACTICE AND TEACH EBM 150–51 (2d ed. 2000) (noting that when observing a single patient's case, it can be difficult to distinguish true effects of treatment from natural healing processes or placebo effects).

[31] *See* Porter, *supra* note 22, at 83.

[32] *See* Shorter, *supra* note 4, at 104.

[33] *See id.* at 105. One particularly telling anecdote involves a German patient in the early nineteenth century who was "weak, losing weight, and unable to rise from bed." *Id.* The doctor sent along a mild placebo of sweet syrup, but ants colonized the vial when the messenger took a rest break on his way to the house. *Id.* The peasant later received the doctor looking quite recovered, and credited his recovery to the powerful vomiting that the "really tough medicine" had induced. *Id.*

[34] *Id.*

[35] *See* SNEADER, *supra* note 27, at 41. Born in Switzerland in the late fifteenth century, Paracelsus attended medical school but came to reject the ancient approaches favored by doctors of his day. *Id.* He famously burned Galenic books, seeking instead to develop his own version of medical science; he studied folk remedies, conducted chemical experiments, and invented a variety of chemical and mineral cures. *Id.*; Weatherall, *supra* note 23, at 213–14. Some of his ideas were heavily influenced by alchemy and astrology, and some of his cures consisted of toxic doses of heavy metals, but he also managed to propose the idea that living beings require air to live and to introduce the use of laudanum, an opioid tincture, for the treatment of pain. *See* SNEADER, *supra*, at 42; Weatherall, *supra*, at 213–14.

[36] *See generally* Walter Pagel & Pyarali Rattansi, *Vesalius and Paracelsus*, *in* MEDICAL HISTORY 309, 309–28 (F.N.L. Poynter ed., 1964) (discussing the life of Vesalius).

[37] *Id.* at 320.

"galenicals."[38] Ultimately, this would enable physicians to measure the impact of specific pharmaceutical compounds on patient health. Vesalius explored the details of human anatomy by means of extensive dissection of cadavers.[39] He catalogued more than two hundred errors in Galen's anatomical descriptions, including Galen's claims that venous blood originated in the liver and his poor understanding of the movement of blood through the human heart and lungs.[40] By both casting strong doubt on received authority regarding effective treatments and by proposing that "knowledge of the true anatomy . . . was only to be gained by dissection and close examination of the parts of the human body," Vesalius and Paracelsus planted the seeds for an empirical and experimental turn in medicine.[41]

Vesalius's work inspired a close attention to anatomical detail that enabled many advances, as anatomists across Europe elaborated much that had been mysterious about the shape and functioning of the circulatory, pulmonary, digestive, nervous, and reproductive systems.[42] A better understanding of the body's mechanics made it increasingly hard to make sense of its functioning in terms of humoral balance.[43] But to probe the body's deeper biological secrets, a broader set of tools was needed. Some researchers explored the realm of the very small, developing increasingly effective microscopes by which to elucidate the fine detail of living tissues, and eventually developing the cell model of biological systems and the roles of bacteria and parasites in causing diseases.[44]

Other important techniques and inventions enabled clinicians to expand their observations into the interior of the body. Doctors learned to listen to the inner workings of the body, first by percussing the chest to hear the special sounds of pulmonary disease, and later by employing the newly developed stethoscope to hear "the movement of blood, gas, and air within the limbs and major body cavities."[45] By the end of the 1800s, doctors were using X-ray technology to see what had previously been hidden behind the skin of their patients,[46] and using sensors to monitor the electrical activity

---

[38] *Compare* SNEADER, *supra* note 27, at 22, *with* Pagel & Rattansi, *supra* note 36, at 42.

[39] *See* Pagel & Rattansi, *supra* note 36, at 319.

[40] *Id.* at 318–19.

[41] *Id.* at 325.

[42] *See* SNEADER, *supra* note 27, at 74; Porter, *supra* note 22, at 138.

[43] *See* Porter, *supra* note 22, at 138–39.

[44] *See id.* at 141, 159–60. *See generally* GEORGE ROSEN, A HISTORY OF PUBLIC HEALTH 286–88 (1958).

[45] Shorter, *supra* note 4, at 113. *See generally* Porter, *supra* note 22, at 153.

[46] Stanley Joel Reiser, *The Science of Diagnosis: Diagnostic Technology*, *in* 2 COMPANION

of the heart.[47] These new disciplines, which incorporated knowledge drawn from physics and chemistry into the medical toolkit, enabled doctors to understand in far better detail the specific course of diseases in the body.

But observation and theorizing alone probably could not have swept aside the practices of ancient medicine. It was also necessary to show that the new theories could make a practical difference in treating patients. Unfortunately, for some time it was much easier to show what *did not* work than what did. As the mathematics of statistical analysis was coming of age during the 1700s, doctors began to track the results of larger numbers of cases to evaluate potential therapies instead of relying on single case histories as their unit of analysis.[48] Such methods soon showed the defects in common therapies. In 1835, Pierre Louis published his *Recherches sur les Effets de la Saigneé*, which reported his comparisons of matched pairs of patients with similar histories and conditions, some of whom received traditional therapies like bloodletting and emetics, while others did not.[49] He found that traditional therapies offered little to no therapeutic benefit for patients suffering from pneumonia or other diseases.[50] Increasingly, the old humoral ideas about treatment became increasingly unsupportable, as their theories could not be squared with internal anatomy and their treatment recommendations turned out to be ineffective when subjected to statistical analysis.

C. Developing Scientifically Valid Foundations for Medical Practice

Happily for us, the power of experimentation did more than just show us what not to do; when coupled with sustained inquiry and attempts to innovate new treatments, experimental methods allowed medicine to finally

---

ENCYCLOPEDIA OF THE HISTORY OF MEDICINE 826, 839–40 (W.F. Bynum & Roy Porter eds., 1993). The spread of this technology in particular was extraordinarily rapid. *Id.* at 840. One professor of surgery commented, just two years after the discovery of X-ray imaging, that "[p]roper surgery cannot be done in a certain variety of diseases without first using the X-ray." *Id.* at 840.

[47] Roy Porter, *Hospitals and Surgery*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 208.

[48] W.F. BYNUM, SCIENCE AND THE PRACTICE OF MEDICINE IN THE NINETEENTH CENTURY 42–44 (1994). In one early study, James Lind grouped together twelve scurvy patients in 1747 and compared the effects of six potential therapies, concluding that citrus fruit was the most effective in treating this disease, which was widespread among seamen. *Id.* at 42–43; Claridge & Fabian, *supra* note 8, at 549. Lind and others encouraged the systematic reporting of case histories, so that similar analyses could be conducted with respect to other diseases, and hospitals increasingly complied. BYNUM, *supra*, at 42–43.

[49] *See* PIERRE LOUIS, RECHERCHES SUR LES EFFETS DE LA SAIGNEÉ 7 (1835).

[50] BYNUM, *supra* note 48, at 43–44. Louis's decision to publish these results no doubt required a fair bit of bravery, given that his teacher, François Joseph Victor Broussais, was both a leader of the Paris school and a passionate advocate of "letting blood through leeching for virtually all diseases." *Id.* at 44; Harold J. Cook, *Physical Methods*, *in* 2 COMPANION ENCYCLOPEDIA OF THE HISTORY OF MEDICINE, *supra* note 46, at 950.

start accumulating treatments that were known to be effective. Even a summary of the major innovations in medicine would take far more space than a law article can afford, but a few highlights should help us recognize some of the ingredients that were crucial to the ultimate success of evidence-based medical treatment design.

Some medical innovations turned out to be beneficial even though they were based on faulty biological theories. One notable example is the asepsis revolution in surgery and wound care.[51] For a long time, surgery was an extraordinarily risky proposition; patients risked succumbing to shock from the great pain involved, and wounds almost inevitably became infected, frequently leading to fatal sepsis.[52] The identification and use of anesthetic agents helped to solve the first dilemma, while the second was much improved by attempts to improve hospital hygiene.[53] Interestingly, however, reforms aimed at improving the sanitary conditions of hospitals significantly predated the microbiological theories that justify antiseptic methods for preventing infection.[54] Rather, early reformers noted a connection between many diseases and the "filthy conditions and closed contaminated atmospheres" that prevailed not only in prisons and slums, but also within the hospitals of the period.[55] This fell in line with a popular disease theory of the time, which held that the miasmatic gases one encountered in squalid conditions caused many diseases.[56] Campaigns to make hospitals cleaner and better ventilated therefore predated the scientific understanding of bacterial infection by almost a hundred years.[57] Sometimes, in other words, one can stumble on efficacious ideas by accident.[58]

---

[51] ROSEN, *supra* note 44, at 315–16.

[52] *Id.* at 315.

[53] *See id.* at 317–19.

[54] *Id.* at 319.

[55] Dorothy Porter, *Public Health*, *in* 2 COMPANION ENCYCLOPEDIA OF THE HISTORY OF MEDICINE, *supra* note 46, at 1235–36; Porter, *supra* note 47, at 189.

[56] Porter, *supra* note 55, at 1236.

[57] *Compare* Porter, *supra* note 55, *with* ROSEN, *supra* note 44, at 319.

[58] A similar example can be found in smallpox inoculation, an effective immunological program that predated an understanding of basic immunological concepts by nearly a century. Early inoculation efforts took place in the mid-1700s, based on the observation that those who had previously survived the disease were resistant to new infections. Before long, a less dangerous inoculation was made available by Edward Jenner, who observed that milkmaids who had been exposed to cowpox also developed a resistance to smallpox. *See* Kenneth Kiple, *The History of Disease*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 33. The relationship between inoculation and the body's production of antibodies was not understood until the late 1800s. *See generally* Paul Weindling, *The Immunological Tradition*, *in* 1 COMPANION ENCYCLOPEDIA OF THE HISTORY OF MEDICINE, *supra* note 46, at 192–95.

More often, however, the identification of useful treatment methods will require both sustained development of underlying theories and the willingness to try many different means of achieving the desired result. The history of antibiotic development, surely one of the greatest triumphs of the evidence revolution in medicine, provides an instructive example. Detecting bacteria and understanding how they were transmitted and how they led to diseases required an extensive research program.[59] Important advances depended on Louis Pasteur's work on the chemistry of optically asymmetric organic compounds, which enabled him to detect the existence of biological activity in the fermenting of beer and wine.[60] This realization led to his subsequent studies of how such activity could be controlled and of how microbes can be transmitted through the air.[61] The ultimate detection of many microbes depended on the development of special chemical dyes that could be used to stain and then visualize them using newly developed microscopes.[62] Robert Koch's studies of infected mice were later critical to establishing the life cycle of bacteria in a living host.[63] Finally, the development of effective chemical antibiotics depended on a great deal of observational study of various conditions that inhibited bacterial growth, such as the proximity of certain fungi, as well as an exhaustive process of trial and error with a wide variety of chemical compounds.[64] All in all, it was more than eighty years from the date Pasteur first discovered anaerobic bacteria until generally effective antibiotic treatments began to be available on the mass market in the late 1930s.[65]

D. Evidence-Based Medicine in the Modern Age

Despite the advances described above, many medical challenges persist. Cancers, cardiac ailments, and other diseases of the long-lived and well-nourished remain to trouble the medical establishment, while the evolutionary process assures that some old diseases will occasionally become resistant to existing therapies and rise again to trouble us.[66] By

---

[59] *See generally* ROSEN, *supra* note 44, at 280–91.

[60] *Id.* at 305–06.

[61] *Id.* at 307.

[62] *Id.* at 312–13.

[63] *Id.* at 311–12.

[64] *See generally* SNEADER, *supra* note 27, at 287–313. S*ee also* WEATHERALL, *supra* note 6, at 149–54, 170–82.

[65] *See generally Chronology*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4, at 351–55.

[66] *See* OFFICE OF TECH. ASSESSMENT, U.S. CONG., OTA-H-629, IMPACTS OF ANTIBIOTIC-RESISTANT BACTERIA 1 (1995), *available at* http://govinfo.library.unt.edu/ota/Ota_1/DATA/1995/

observing the struggles that medicine continues to face as well as its obvious achievements, we can learn much about what a similar movement in the legal world can expect.

Several features of modern medical practice are particularly noteworthy. First, the search for new treatments involves inevitable tradeoffs between protecting the safety of existing patients and learning how to better treat future ones.[67] Second, many effective therapies would never have come into existence absent the very costly searches and testing conducted by for-profit entities with financial incentives to investigate many possible therapies.[68] Third, although attention to the minutiae of biological processes has brought many benefits, modern clinicians have recently made a strong case for experimental protocols that go beyond testing the effects of medicine on immediate biophysical markers by measuring medicine's impact on overall health as well.[69] This shift of attention from measuring surrogate outcomes to ultimate outcomes has allowed researchers to demonstrate that many popular treatments and diagnostic techniques actually do more harm than good.[70] Finally, even in the modern age, medical investigation remains subject to bias arising from financial interest and skewed reporting of trial outcomes.[71]

To begin with, let us consider the problem of identifying new effective therapies. The history already discussed should indicate that this problem is far from trivial; as already mentioned, many decades passed between the seminal work identifying bacteria as the source of many diseases and the identification of effective antibiotic treatments for many common infectious diseases. As a general matter, we cannot test or compare treatments until we have identified them as plausible candidates, and indeed the history of medical innovation is littered with useful inventions whose promise lay unrealized for some time.[72]

So a central challenge of maximizing medical effectiveness is the high cost of finding new candidate treatments. One of the reasons such costs can

---

9503.PDF.

[67] John Barbour, *The Century of Wonder Drugs/Antibiotics: Medical Savior for Millions/Advent of Sulfa and Penicillin Later Made Cardiac Surgery, Organ Transplants Possible*, HOUS. CHRON., Dec. 8, 1991, at A25, *available at* http://www.chron.com/CDA/archives/archive.mpl/1991_827186/the-century-of-wonder-drugs-antibiotics-medical-sa.html.

[68] *See* CONG. BUDGET OFFICE, RESEARCH AND DEVELOPMENT IN THE PHARMACEUTICAL INDUSTRY 2–3 (2006), *available at* http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/76xx/doc7615/10-02-drugr-d.pdf (noting that the cost of finding an innovative new drug in the modern regulatory environment exceeds $800 million per new molecular entity).

[69] *See generally* Richard A. Deyo, *Using Outcomes to Improve Quality of Research and Quality of Care*, *in* EVIDENCE-BASED CLINICAL PRACTICE: CONCEPTS AND APPROACHES 65–66 (John P. Geyman et al. eds., 2000).

[70] *See id.*

[71] *See id.*

[72] *See* Barbour, *supra* note 67.

be quite high is that the risks of employing the wrong treatments can be substantial.[73] This can lead us to demand high standards of evidence of safety and efficacy before we will tolerate the use of a treatment.[74] In modern medical history, this concern has manifested most notably in the area of pharmaceutical treatments.

Although pharmaceuticals represent some of our greatest modern medical innovations, they also have enormous power to do harm. Take the famous tragedy associated with a once popular sleeping tablet called thalidomide. Developed by the German firm Chemie Grünenthal in the 1950s, thalidomide showed enormous promise as a sleep aid,[75] producing a "deep, natural, all-night sleep without a hangover."[76] Unlike most other sleeping pills of the time, it had low acute toxicity, meaning that there was a large difference between the dosage that would induce sleep and the dosage that would induce death.[77] Once it was approved in many countries (although not yet in the United States), doctors began prescribing it for a wide range of conditions; it was used to treat "colds, coughs, influenza, nervousness, migraine and other headaches, . . . asthma," and, most tragically, nausea.[78] Unfortunately, thalidomide was a teratogen, capable of causing birth defects in developing fetuses if taken at very specific times during early pregnancy.[79] Since pregnant women frequently used it to treat their nausea at night, many thousands of deformed infants were born across Europe before the drug's teratogenic effect was understood.[80]

The case of thalidomide illustrates well the challenges involved in making sure a new treatment is safe and effective. The drug manufacturer had first tested thalidomide on mice, rats, guinea pigs, and rabbits, and it had observed no ill effects during those experiments.[81] But this sort of investigation proved unable to detect warning signs about the drug's

---

[73] *See generally* Jonathan V. O'Steen & Van O'Steen, *The FDA Defense: Vioxx and the Argument Against Federal Preemption of State Claims for Injuries Resulting from Defective Drugs*, 48 ARIZ. L. REV. 67 (2006).

[74] *See generally id.*

[75] SNEADER, *supra* note 27, at 367.

[76] Max Sherman & Steven Strauss, *Thalidomide: A Twenty-Five Year Perspective*, 41 FOOD DRUG COSM. L.J. 458, 460 (1986).

[77] Carrie L. Radomsky & Norman Levine, *Thalidomide*, 19 DERMATOLOGIC CLINICS 87, 87 (2001).

[78] Sherman & Strauss, *supra* note 76, at 460.

[79] *See* WEATHERALL, *supra* note 6, at 277.

[80] Sherman & Strauss, *supra* note 76, at 459. The lag between the first birth defects and the discovery of thalidomide as the cause was exacerbated by Grünenthal's attempts to conceal information regarding these birth defects from their potential customers. *See* SNEADER, *supra* note 27, at 367; WEATHERALL, *supra* note 6, at 276.

[81] SNEADER, *supra* note 27, at 367.

dangers.[82] Many drugs operate similarly in humans and other animals, but thalidomide is an exception to this rule.[83] It has no teratogenic effect on many non-human test subjects, so even if pregnant animals had been used as test subjects, the investigation might have missed the drug's dangers.[84]

The manufacturer also conducted human trials of the drug, observing no apparent ill effects during treatment.[85] Better-conducted tests in human subjects might have raised warning signs, but without advance knowledge of what the drug's dangers were, researchers were looking for the wrong things. The drug did cause a "tingling neuritis" in some patients, indicating possible damage to peripheral nerves,[86] which some astute observers (including FDA analysts) thought might indicate greater risks for gestating fetuses.[87] But such effects were noticed only late in the drug's history.[88] Researchers conducting early trials of the drug may have missed such signs because the primary dangers presented by most sedative compounds happen quickly in the form of overdoses.[89] As a result, early test subjects would not necessarily have been monitored for the span of time needed to observe the neuropathic side effect, which only occurred after "long-term use."[90] And although a controlled experiment involving pregnant women as subjects could have detected the teratogenicity, the narrow time window during which the drug could act as a teratogen during fetal development meant that a large number of pregnant women would have had to have been tested at varying phases of their pregnancies in order to create a high likelihood of detecting this dreadful side effect.[91] As this example illustrates, assuring the safety and efficacy of new drugs is a very hard thing to do. In order to catch all possible ways that a drug might be harmful, it must be tested for its effects on a variety of subpopulations, using realistic dosages and time-courses of treatment.

The FDA, which can count its refusal to approve thalidomide for sale in the United States as one of its great victories, now structures its approval

---

[82] *Id.*

[83] *Id.*

[84] *See* WEATHERALL, *supra* note 6, at 277.

[85] Sherman & Strauss, *supra* note 76, at 460.

[86] *See The Thalidomide Disaster*, TIME, Aug. 10, 1962.

[87] Sherman & Strauss, *supra* note 76, at 460–61; *see also* SNEADER, *supra* note 27, at 368.

[88] *See* Sherman & Strauss, *supra* note 76, at 460.

[89] *See id.* at 459.

[90] *Id.* at 461.

[91] *See* WEATHERALL, *supra* note 6, at 277. Disturbingly, it is unlikely that such a test would occur even today. Our existing FDA approval process, which was developed in part to prevent another thalidomide scandal, bars pregnant women from acting as test subjects during new-drug development. Sherman & Strauss, *supra* note 76, at 464.

process for new drugs in a manner that aims to forestall similar occurrences. First, before it will permit any drug trials in human subjects, the FDA demands the submission of a wealth of data regarding a drug's composition and how it is manufactured.[92] The agency also requires "[a]dequate information about pharmacological and toxicological studies of the drug involving laboratory animals or in vitro," of a quantity and quality sufficient to assure the agency that the drug is likely to be safe when tested on people.[93] Then, a company must conduct multiple rounds of human trials designed to show that the drug is acceptably safe, to demonstrate that it is therapeutically effective, and to quantify the nature and extent of its side effects.[94]

In deciding whether these trials have demonstrated that a drug is ready to be marketed, the FDA generally requires that the trials be designed in such a way as to credibly demonstrate the causal impact of the drug.[95] Thus, in order to meet the agency's standards, tests must generally compare the drug's effects with the effects of a placebo or of a competing therapy (if one exists), so that it can be made clear whether the drug improves on existing alternatives.[96] Beyond this, the agency also expects that a portion of the drug trials involve random assignment of patients to treatment and control groups to avoid problems of selection bias and to make the treatment groups and control groups as nearly identical as possible.[97] A selection bias would occur, for instance, if researchers consciously or unconsciously steered healthier patients towards the drug rather than a placebo, with the result that the treated group ended up healthier than the control group due to the selection effect rather than the treatment itself.[98]

---

[92] *See* 21 C.F.R. § 312.23 (2012).

[93] *Id.*

[94] *See* 21 C.F.R. § 312.21 (2012).

[95] *See* 21 C.F.R. § 314.126 (2012).

[96] *Id.*

[97] *Id.*; *see also* SACKETT ET AL., *supra* note 30, at 106–07.

[98] *See* JOSHUA D. ANGRIST & JORN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS: AN EMPIRICIST'S COMPANION 12–15 (2009); SACKETT ET AL., *supra* note 30, at 106–07. A classic example of the impact of selection bias was seen during World War II, when the British asked statistician Abraham Wald to recommend where they should add additional armor to their bombers. *See* John D. Cook, *Selection Bias and Bombers*, ENDEAVOUR (Jan. 21, 2008), http://www.johndcook.com/blog/2008/01/21/selection-bias-and-bombers/. Upon examining the planes in service, Wald recommended that additional plating be added only to those spots where he could observe no damage. *Id.* The reason, as he explained, was a selection effect: the sample provided consisted only of those planes that were shot in non-critical locations, given that all the critically damaged planes would have failed to return. *Id.* Thus, each bullet hole indicated not a location where planes were in danger of being shot, but rather a location where they could survive being shot. *Id.*

The agency also requires information showing which population subgroups the drug has been tested in.[99] This testing helps to make clear whether the drug behaves differently (and perhaps dangerously or inefficaciously) when given to some subpopulations.[100]

Finally, the agency advises researchers to devise procedures to "minimize bias on the part of the subjects, observers, and analysts of the data," and mentions "blinding" as one means by which to do so.[101] Blinding is a classic means by which to avoid a variety of problems in the measurement of causal effects.[102] Either test subjects alone can be blinded ("single-blinding"), or both subjects and researchers can be blinded ("double-blinding"), to whether the subjects involved are receiving a real treatment or only a placebo.[103] If subjects are kept ignorant, this can both preserve the ability of placebos to produce a placebo effect (thus assuring that real effects can be distinguished from placebo effects) and also can ensure that subjects do not change their behavior based on inclusion in the placebo group in a way that might jeopardize the overall results (such as by seeking alternative treatment during the pendency of the study).[104] Blinding researchers has many benefits as well. Researchers who know which patients are being treated and which are not may consciously or unconsciously communicate that information to patients, or may engage in wishful thinking when measuring results, thus enlarging the treatment's effect through biased measurement.[105] A randomly assigned, placebo-controlled, double-blinded study with a reasonably large sample size has enormous power to isolate real effects from false ones, and thus is commonly labeled as a "gold standard" for measuring clinical efficacy.[106] As such, it may provide valuable inspiration for better designed studies of the efficacy of legal procedures as well, as I shall discuss later.

---

[99] 21 C.F.R. § 314.50 (2012); *see also* FOOD & DRUG ADMIN., U.S. DEP'T OF HEALTH & HUM. SERVS., GUIDANCE FOR INDUSTRY: COLLECTION OF RACE AND ETHNICITY DATA IN CLINICAL TRIALS (2005), *available at* http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126396.pdf.

[100] *See* SACKETT ET AL., *supra* note 30, at 119 (noting that sociodemographic differences can occasionally make the results of prior trials inapplicable to a patient).

[101] 21 C.F.R. § 314.126 (2012).

[102] *See* Larry E. Miller & Morgan E. Stewart, *The Blind Leading the Blind: Use and Misuse of Blinding in Randomized Controlled Trials*, 32 CONTEMP. CLINICAL TRIALS 240, 240 (2011).

[103] David W. Peterson & John M. Conley, *Of Cherries, Fudge, and Onions: Science and Its Courtroom Perversions*, 64 LAW & CONTEMP. PROBS. 213, 217–18 (2001).

[104] *See id.*

[105] *See generally* Kenneth F. Schulz, *Blinding in Randomized Trials: Hiding Who Got What*, 359 LANCET 696 (2002).

[106] Ted J. Kaptchuk, *The Double-Blind, Randomized, Placebo-Controlled Trial: Gold Standard or Golden Calf?*, 54 J. CLINICAL EPIDEMIOLOGY 541, 541 (2001).

The FDA, therefore, goes to great lengths to ensure that pharmaceuticals are safe and effective before they can be marketed to the general public. All of this caution, unfortunately, has a downside. Just one out of every 1,000 tested compounds shows enough promise to make it through pre-clinical testing (the in-vitro and animal-based studies that precede human subject testing), and just one in four drugs tested in human subjects survives the long gauntlet of FDA-mandated clinical trials.[107] Companies can generally expect an eight to twelve year process of testing a product and awaiting the FDA's decision regarding its marketability,[108] and they must spend hundreds of millions of dollars in research and testing costs for each new drug they wish to develop and sell.[109] This means that although we can be confident that new drugs are an improvement over old ones, relatively few companies will be able to afford to engage in the extensive development process necessary to produce new drugs.[110] Indeed, the FDA has increasingly come under fire in recent years for imposing so many hurdles in the path of drug developers.[111] Some critics have successfully pressured it to reduce the amount of screening conducted for new drugs, even though such haste predictably results in more adverse drug reactions.[112]

Unfortunately, we face an inevitable trade-off between producing and testing as many new treatment ideas as possible and protecting patient safety. We could potentially find many more effective pharmaceuticals if we lowered the costs involved in trying out new drugs, but that would come at a human cost. Given the uncertainty involved in trying to predict what we might discover if we looked harder and what value such discoveries

---

[107] Michael A. Carrier, *Two Puzzles Resolved: Of the Schumpeter-Arrow Stalemate and Pharmaceutical Innovation Markets*, 39 IOWA L. REV. 393, 417 (2007).

[108] Mary K. Olson, *Regulatory Agency Discretion Among Competing Industries: Inside the FDA*, 11 J.L. ECON. & ORG. 379, 382 (1995) (describing a total time ranging from eight to twelve years for typical drugs).

[109] Joseph A. DiMasi et al., *The Price of Innovation: New Estimates of Drug Development Costs*, 22 J. HEALTH ECON. 151, 180 (2003) (estimating an average out-of-pocket cost of $403 million per new drug, with a total capitalized cost of $802 million).

[110] *See* Carrier, *supra* note 107, at 401 (noting that there are no longer any "garage inventors" producing new drugs and that the universe of potential drug innovators is now restricted to large, established entities).

[111] *See* Jeffrey S. Barkun et al., *Evaluation and Stages of Surgical Innovations*, 374 LANCET 1089, 1090 (2009) (noting that the FDA's mechanism has slowed drug development); Carrier, *supra* note 107, at 401, 417; Richard A. Epstein, *The Pharmaceutical Industry at Risk: How Excessive Government Regulation Stifles Innovation*, 82 CLINICAL PHARMACOLOGY & THERAPEUTICS 131, 131–32 (2007).

[112] *See* Barkun et al., *supra* note 111, at 1090 (noting that the FDA's mechanism has slowed drug development).

could have for human health, such trade-offs will necessarily resist quantified analysis.

It is interesting, nonetheless, to compare pharmaceutical development with other areas of medicine that the FDA does not regulate as heavily. New surgical procedures, for instance, do not fall within the FDA's purview and are rarely subject to placebo-controlled trials on random subjects.[113] This difference cuts both ways: it is much easier to innovate new surgical techniques than it is to develop new drugs, but we have much less information regarding the typical safety of surgical interventions than we do of pharmacological ones.[114] The downsides, when we become aware of them, can be large. A recent, rare example of a blinded, placebo controlled, randomly assigned surgical trial addressed one of the most common types of orthopedic surgery: arthroscopic lavage of arthritic knees.[115] When the investigators compared the operation with one in which patients received only skin incisions and anesthesia, but no surgery, it turned out that this operation, which was taking place more than 650,000 times per year, was no more effective than the placebo.[116] Numerous patients, therefore, had been undertaking the risks of surgery for no real therapeutic benefit.[117] Whether more surgeries should be tested to the extent that most drugs are remains a question that deeply divides the medical profession,[118] showing that the evidence-based revolution in medicine is still very much a work in progress.

---

[113] *See* Patrick L. Ergina et al., *Challenges in Evaluating Surgical Innovation*, 374 LANCET 1097, 1097 (2009) (noting that RCTs have been rare in surgery since the 1970s).

[114] *See generally* Barkun et al., *supra* note 111.

[115] J. Bruce Moseley et al., *A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee*, 347 NEW ENG. J. MED. 81, 81, 85 (2002).

[116] *Id.* at 81, 84, 87.

[117] *See generally id.* Very little research of this kind is done, and when it is done it is quite controversial. *See* Ergina et al., *supra* note 113, at 1097. Some feel that even if prior consent is obtained, it is ethically unacceptable to subject patients to sham surgeries for the sake of increasing medical knowledge. *See* Sam Horng & Franklin G. Miller, *Is Placebo Surgery Unethical?*, 347 NEW ENG. J. MED. 137, 137 (2002). Of course, there is a flip side to such a concern; if many widespread surgeries turn out to be therapeutically ineffective, a vastly larger group of people may be experiencing a different form of "sham" therapy without giving meaningful consent to this different risk. It is also hard to do such studies well; one important factor is the skill of the surgeon, and compiling a representative grouping for each study might prove difficult. This, indeed, is the best criticism of the Mosley study; it tested very carefully the effect of surgery by a single surgeon. Such a design cannot guarantee, of course, that the surgeon's performance is representative of his peers.

[118] *See* Ergina et al., *supra* note 113, at 1097; Horng & Miller, *supra* note 117, at 137–39.

E. The Special Challenge of Validating Diagnostic Tests

Between the extremes of high cost, high quality pharmaceutical design, and easy but potentially untrustworthy surgical innovation lies another target of the modern evidence-based medicine movement: diagnostic and screening tests, which are intended to detect whether patients harbor various hard-to-detect diseases.[119] Over the last twenty years, reformers have increasingly argued that the decision to engage in diagnostic tests should depend on the existence of evidence that those tests actually benefit patients.[120] Ideally, a diagnostic test would be risk-free, would be inexpensive to perform, and would have both perfect *sensitivity*, meaning that it generates no false positive results, as well as perfect *specificity*, meaning that it generates no false negatives.[121] In the real world, this will rarely be the case. Often the most accurate tests will be expensive or dangerous to a patient, and no test can ever be *perfectly* accurate.

There will therefore always be a need among doctors for new tests that offer reasonably accurate results in a quick, cheap, and safe manner, as well as for experiments that can sort the useful tests from the bad. Moreover, state-of-the-art methods of validating diagnostic tests have a unique relevance for designers of the rules governing the litigation process. The litigation system itself can be conceived of as a special type of diagnostic test, designed to detect when a set of facts has occurred that gives rise to legal consequences. Its methods and results may be different, but the goal of ensuring high sensitivity and specificity is the same.

Special research designs are needed to validate diagnostic tests. First, the general method for evaluating a test relies on comparing it with an existing "gold standard" of diagnosis.[122] Such gold standards are those existing tests that maximize specificity and sensitivity to the greatest extent possible.[123] Of course, both ethical and scientific constraints will make any gold standard chosen for research purposes a compromise. Doctors cannot dissect the brains of healthy individuals to validate a test for brain cancer, and even if they did, the possibility of misclassification of results can never

---

[119] *See* Holger J. Schünemann et al., *GRADE: Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies*, 336 BRIT. MED. J. 1106, 1106 (2008).

[120] *See, e.g.*, M. Carrington Reid et al., *Use of Methodological Standards in Diagnostic Test Research Getting Better but Still Not Good*, 274 JAMA 645, 645–50 (1995).

[121] Joann G. Elmore & Edward J. Boyko, *Assessing Accuracy of Diagnostic and Screening Tests*, *in* EVIDENCE-BASED CLINICAL PRACTICE: CONCEPTS AND APPROACHES, *supra* note 69, at 83, 85.

[122] *Id.* at 85–87; *see also* SHARON E. STRAUS ET AL., EVIDENCE-BASED MEDICINE: HOW TO PRACTICE AND TEACH EBM 71–72 (3d ed. 2005).

[123] Elmore & Boyko, *supra* note 121, at 85.

be entirely ruled out.[124] Still, the idea of a gold standard provides an ideal towards which research designs can aspire; the more accurate the basis of comparison, the more confidence we can have that a study of a diagnostic test accurately represents its true performance.[125]

Second, due to concerns regarding measurement bias, blinding is still a valuable principle in tests evaluating diagnostic procedures.[126] Many test results involve significant amounts of interpretation, and the validity of an assessment of diagnostic validity may be compromised if researchers are influenced in their interpretation of that ambiguity by the result of another test.[127] Thus, for instance, a radiologist interpreting an ambiguous chest X-ray might be more likely to conclude that it shows evidence of a tumor if she knows that a lung biopsy has indicated that cancer is likely.[128] To avoid this problem, it is ideal to have different individuals perform the test being assessed and the gold standard to which it is compared, and to keep them unaware of the other's conclusions.[129]

Third, if one wishes to get an accurate estimate of a test's validity, it is important to design the coverage of the test cases to resemble how the test will be used in the real world. Ideally, test subjects would be drawn randomly from the pool of people to whom the test might be applied under real world conditions, and then each would receive both the test under study and its gold standard comparison.[130] If researchers pick and choose the cases that will be studied, bias can result.

Finally, the best studies of screening and diagnostic tests will go beyond just measuring the tests' *accuracy*; they will also assess the extent to which the information gained tends to come at an unacceptable risk to the patient's safety and well-being.[131] Just knowing the rates at which tests produce false positives and negatives is not enough. A good study will track the real world consequences of such misestimates on patient health and well-being. Some false positives, for instance, may lead to painful or dangerous biopsies, while others will be less costly. Some false negatives

---

[124] *See* STRAUS ET AL., *supra* note 122, at 72–73 (noting that some reference standards, like angiography, might be unacceptably invasive or risky to perform on healthy patients).

[125] *See* Elmore & Boyko, supra note 121, at 87.

[126] *See id.*; STRAUS ET AL., *supra* note 122, at 72.

[127] *See* STRAUS ET AL., *supra* note 122, at 72 (noting that some seemingly "hard" measurement standards involve a great deal of discretion, and that blinding the reference standard assessment is a good means of avoiding bias).

[128] Elmore & Boyko, *supra* note 121, at 87.

[129] *See id.*; STRAUS ET AL., *supra* note 122, at 72.

[130] Elmore & Boyko, *supra* note 121, at 87.

[131] *Id.* at 91–92.

may be fatal, while others will be correctible at little long-term risk to the patient. By comparing measures of long-term health when the test is or is not used, researchers can offer some guidance as to whether its accuracy comes at an unacceptable price. In particular, if a test increases knowledge but generally makes patients less well, it should be avoided.

This distinction between what helps in the short term and what produces good long-term outcomes is one that modern proponents of evidence-based medicine emphasize. In general, the reformers urge, the best evidence regarding the quality of medical treatments or tests tracks "ultimate outcomes of interest," like overall patient mortality, morbidity, or pain levels, rather than "surrogate outcomes" in the form of discrete physiologic phenomena.[132] This is a shift from the approach that long predominated, in which physicians focused their efforts on changing easily measureable surrogate outcomes.[133]

As an example, a physician interested in preventing future heart attacks in a patient with high cholesterol might wish to examine the efficacy of prescribing a medicine that lowers cholesterol. A study using a surrogate outcome might rely on blood cholesterol levels to demonstrate the drug's impact, while one focusing on ultimate outcomes might look at the effect of the drug on overall mortality.[134] In some cases, the difference would be quite significant; some drugs, for instance, will lower cholesterol while *increasing* overall mortality.[135] Such cures are truly worse than the disease, and provide an important cautionary tale about overreliance on theoretical understandings of complex systems. As one doctor explains, "Recognition is growing that physiologic, laboratory, and imaging outcomes are sometimes poorly associated with symptomatic, functional, and survival outcomes."[136] For this reason, there is much value in tracking not just those changes in biological processes that are expected to improve a patient's quality of life, but also more direct measures of quality of life such as survey responses.[137] By doing so, clinicians can notice when their theoretical assumptions no longer guide the way towards improving patient well-being.

---

[132] *See* Deyo, *supra* note 69, at 65–66.

[133] *Id.*

[134] *See id*. at 66.

[135] *See* M.F. Oliver et al., *W.H.O. Cooperative Trial on Primary Prevention of Ischæmic Heart Disease Using Clofibrate to Lower Serum Cholesterol: Mortality Follow-Up*, 316 LANCET 379, 379–85 (1980).

[136] Deyo, *supra* note 69, at 70–71.

[137] *Id.* at 66.

One last note of caution is in order regarding the value of evidence-driven design of medical treatments. Those who innovate medical treatments often stand to gain a great deal of profit from them,[138] and even when treatments are developed in an academic setting, there may be incentives to advance an individual researcher's career at the expense of promoting accurate knowledge regarding effective treatments.[139] These factors can lead to "publication bias," in which trials favoring the safety and efficacy of drugs are published while those that cast doubt on efficacy are suppressed.[140] The medical establishment has taken some steps to rein in the biasing effects of having trials funded by those who look to profit from selling a treatment through disclosure rules, but such rules have proven challenging to enforce.[141] What is more, even government oversight may not always be sufficient to incentivize adequate investigation into the potential risks of highly profitable treatments.[142] Thus, in addition to the challenges described above, the development of an evidence-based culture faces an additional roadblock in the form of profit incentives that may inhibit the willingness of manufacturers to investigate and publicize evidence that blockbuster drugs are dangerous or ineffective.

Still, despite these challenges, an evidence-based approach to medical testing and treatments has gone a long way towards improving the quality of medical care over the long term. What is more, the existence of a culture that regularly performs such testing provides us with an important basis for trusting the efficacy of medical treatments as a general matter. Still, as we have seen, the dialogue is still ongoing within the medical community regarding how stringent such testing should be. There are costs and dangers associated with either too much or too little testing before treatments are put into general use.

---

[138] *See* Janet Spitz & Mark Wickham, *Pharmaceutical High Profits: The Value of R&D, or Oligopolistic Rents?*, 71 AM. J. ECON. & SOC. 26, 27 (2012) (pointing out that pharmaceutical firms profit at a rate of "2.5 to 37 times the non-pharmaceutical industry average over time").

[139] *See* Anna B. Laakmann, *Collapsing the Distinction Between Experimentation and Treatment in the Regulation of New Drugs*, 62 ALA. L. REV. 305, 318–19 (2011).

[140] *See id.*

[141] *See id.* at 319.

[142] *See* Theodore Eisenberg & Martin T. Wells, *Statins and Adverse Cardiovascular Events in Moderate-Risk Females: A Statistical and Legal Analysis with Implications for FDA Preemption Claims*, 5 J. EMPIRICAL LEGAL STUD. 507, 546–47 (2008) (noting that the FDA has at times failed to require that studies be designed in a way that could detect risks to vulnerable groups, and that its funding is dwarfed by that of the industry it is charged with regulating).

III. WHY PROCEDURAL DESIGNERS NEED BETTER DATA

The lessons of evidence-based medicine reform may be good or bad news, depending on how we look at them. The turn towards systematic measurement of treatment efficacy, coupled with a detailed investigation into the causes of disease, has enabled medicine to grow vastly more effective and safe than it once was.[143] Since we do not, in general, subject most new procedural devices to similar scrutiny, this suggests that there is a great deal of available low-hanging fruit in the form of improvements to procedural design that could be obtained just by looking carefully at which existing devices work well and which do not.[144] At the same time, doctors are still wrestling to consistently apply evidence-based criteria to their decision making, and many complain that the costs of requiring careful investigations might retard the creativity needed to locate the next generation of novel therapies.[145] This suggests that the road towards systematically testing procedural efficacy may be quite challenging and costly if we wish to do it right.

Why, then, should we bear these burdens? This Section will illustrate some of the challenges that arise for those who would design rules of procedure or evidence *without* drawing on data from controlled procedural trials, and also why such trials are hard to do well. It turns out that intuition and anecdotal experience are likely to be poor guides towards procedural success.[146] What is more, most existing attempts to assess procedural efficacy using statistical analysis of observational data fall short because their methods are unlikely to yield trustworthy conclusions about the causal impact of different rules.[147] We might solve many of these problems by running controlled experiments that compare the impacts of different

---

[143] *See generally Chronology*, *in* THE CAMBRIDGE HISTORY OF MEDICINE, *supra* note 4.

[144] A number of scholars have urged such a turn in procedural design. *See, e.g.*, Edward H. Cooper, *Simplified Rules of Federal Procedure?*, 100 MICH. L. REV. 1794, 1799 (2002); D. James Greiner & Cassandra Wolos Pattanayak, *Randomized Evaluation in Legal Assistance: What Difference Does Representation (Offer and Actual Use) Make?*, 121 YALE L.J. 2118, 2193–95 (2012); Richard A. Posner, *The Summary Jury Trial and Other Methods of Alternative Dispute Resolution: Some Cautionary Observations*, 53 U. CHI. L. REV. 366, 373–77 (1986); Carl Tobias, *More Proposals to Simplify Modern Federal Procedure*, 38 GA. L. REV. 1323, 1324–25 (2004); Laurens Walker, *Perfecting Federal Civil Rules: A Proposal for Restricted Field Experiments*, 51 LAW & CONTEMP. PROBS. 67, 67–68 (1988); Willging, *supra* note 1, at 1197, 1201–04; *see also* Michael Abramowicz et al., *Randomizing Law*, 159 U. PA. L. REV. 929, 931–34 (2011).

[145] *See* Sharon E. Straus & Finlay A. McAlister, *Evidence-Based Medicine: A Commentary on Common Criticisms*, 163 CAN. MED. ASS'N J. 837, 838–40 (2000).

[146] Posner, *supra* note 144, at 367.

[147] Walker, *supra* note 144, at 72.

procedural and evidential rule regimes,[148] but the vast and interconnected nature of judicial institutions will make such experiments hard to do and will mean that some important questions may lie beyond our ability to measure.

When deciding whether to adopt a new rule of procedure or evidence, rule makers typically seek to optimize a few variables.[149] They would like to make rules that achieve good outcomes, that do not create excessive costs for the litigants or the justice system, and that do not protract litigation for no benefit.[150] To achieve meaningful procedural improvements, they have to evaluate a complicated question: Would the litigation system, over time, strike a better balance among these values via the implementation of a new rule, or would it be better to continue using an existing one? This question is partly normative, given that there will always be value-driven disagreements regarding the appropriate balance between the differing values. No one can claim to have an objective basis, for instance, on which to decide how much we should spend to achieve a given increment of accuracy. Rather, the appropriate balance will always depend on the comparative importance one attaches to each factor. But there is also an objective component to such an inquiry: In a given set of cases, two different rule regimes, if implemented, might produce a different balance of accuracy in results, perceived fairness, time-to-case-completion, overall cost, and cost distribution. So in deciding what procedures we should adopt for settling disputes, we combine two main inputs: a set of normative preferences regarding the ideal balance between factors like accuracy, perceived fairness, and costs, and a set of factual judgments regarding how the competing rule alternatives would impact those factors.

A. The Challenges of Designing New Procedural Rules

For a variety of reasons, anticipating the impact of a proposed rule on these values is very hard to do. First, each legal case will be different, and may respond differently to the new rule. Here there is a strong analogy with medical research: Just as each individual has a unique medical history and constellation of symptoms, so each case involves a new combination of facts, legal rules, parties, lawyers, and judicial staff. Just as we cannot

---

[148] *See generally id.* at 67–68.

[149] *See* FED. R. CIV. P. 1; Arthur R. Miller, *From* Conley *to* Twombly *to* Iqbal*: A Double Play on the Federal Rules of Civil Procedure*, 60 DUKE L.J. 1, 130 (2010).

[150] *See* FED. R. CIV. P. 1; Miller, *supra* note 149, at 130.

know how safe drugs are in general by seeing how they behave in a small group of patients, we cannot easily estimate the impact of a rule by seeing how it operates in a few cases.[151] Doing so risks the legal equivalent of the thalidomide crisis.[152] There, initial testing failed to include pregnant women, and thus failed to catch the drug's dangerous side effects for developing fetuses.[153] In legal contexts, we might expect similar results, because a rule applied to good effect in one sort of setting may have more negative impacts if applied in new arenas.

Historically, it is easy to identify occasions when procedural rule makers made this mistake. Perhaps the most notable recent instance occurred during the enactment of the new Federal Rules of Civil Procedure in 1937. Reformers such as Charles Clark desired to make litigation faster and more efficient.[154] They noted that a relatively new procedural device, the motion for summary judgment, had advanced these ends in English practice.[155] In this early version of summary judgment, a plaintiff seeking liquidated damages could obtain speedy justice by filing an affidavit showing that there were no factual disputes necessitating a jury trial.[156] If the defendant failed to file a counter-affidavit showing that a factual dispute in fact existed, a court could then allow the plaintiff to recover damages without going through the formalities of a jury trial.[157]

Seeing the success of summary judgment in this narrow context, the drafters urged that the new rules should incorporate a similar provision that could be used in all types of cases, whether or not liquidated damages were involved, and that could be sought by either plaintiffs or defendants.[158] The enactors of the new Federal Rules obliged.[159]

---

[151] *See* Michael J. Saks, *Do We Really Know Anything About the Behavior of the Tort Litigation System – And Why Not?*, 140 U. PA. L. REV. 1147, 1159–62 (1992) (noting the frequency with which law reformers rely on anecdotal case evidence, and the very low utility of such information for assessments of procedural efficacy).

[152] Sherman & Strauss, *supra* note 76, at 459–61.

[153] *Id.*

[154] *See* Charles E. Clark & Charles U. Samenow, *The Summary Judgment*, 38 YALE L.J. 423, 423 (1929).

[155] *Id.* at 423–24.

[156] *See id.*

[157] *See id.* at 429–31.

[158] *See* Steven B. Burbank, *Vanishing Trials and Summary Judgment in Federal Civil Cases: Drifting Toward Bethlehem or Gomorrah?*, 1 J. EMPIRICAL LEGAL STUD. 591, 594–600 (2004); Clark & Samenow, *supra* note 154, at 469–71; Edson R. Sunderland, *An Appraisal of English Procedure*, 24 MICH. L. REV. 109, 111–12 (1925).

[159] *See* FED. R. CIV. P. 56 advisory committee's note. *See generally* Kevin M. Clermont, *Litigation Realities Redux*, 84 NOTRE DAME L. REV. 1919, 1941 (2009) (describing the mechanics of the new summary judgment procedure incorporated into the rules).

Unfortunately, the birth of this rule provides a classic example of how a rule can behave differently than expected when used in new ways and in new types of cases. Although a broader version of summary judgment, similar to the new federal experiment, had been tried in Michigan, the drafters of the new rule did not collect any data regarding its performance before copying it on a much larger scale.[160] Rather, they relied on their intuitive understanding of the judicial process and their imaginations to predict its future performance. They expected that the rule would primarily be used by plaintiffs to "pierce assumed or fictitious defenses,"[161] but instead the rule is now primarily employed by defendants who claim that plaintiffs have not collected enough evidence to prevail at trial.[162] What is worse, many observers believe that summary judgment, used in this unforeseen way, multiplies legal costs rather than reducing them (although, as we shall see, this is a very hard thing to show empirically).[163]

Why did our reformers not see this coming? It is not because they were uninterested in collecting data—Clark, in particular, was a major proponent of amassing descriptive statistics on the operation of law in action in everyday cases.[164] Rather, it seems that when they imagined expanded use of the device, they failed to anticipate how differing circumstances might impact its workings. Plaintiffs' attorneys, especially those paid on

---

[160] Burbank, *supra* note 158, at 594–600.

[161] *Id.* at 602 (internal quotation marks omitted).

[162] *See* Joe S. Cecil et al., *A Quarter-Century of Summary Judgment Practice in Six Federal District Courts*, 4 J. EMPIRICAL LEGAL STUD. 861, 886–88 (2007) (counting 967 plaintiff motions and 2,526 defendant motions in their survey of six federal district courts, making defendants roughly 2.5 times as likely to seek summary judgment compared with plaintiffs).

[163] *See, e.g.*, John A. Bauman, *The Evolution of the Summary Judgment Procedure: An Essay Commemorating the Centennial Anniversary of Keating's Act*, 31 IND. L.J. 329, 352–53 (1956) (arguing that, because the procedure appeared to be frequently employed but rarely resulted in the dismissal of cases, it might be adding delay to litigation); John Bronsteen, *Against Summary Judgment*, 75 GEO. WASH. L. REV. 522, 535–38 (2007) (urging that summary judgment is inefficient because trials are not that costly and because parties would likely choose settlement over trial in most cases if summary judgment was not an available option); D. Brock Hornby, *Summary Judgment Without Illusions*, 13 GREEN BAG 273, 274–75 (2010) (arguing that summary judgment motions fail to save either time or money); Diane P. Wood, *Summary Judgment and the Law of Unintended Consequences*, 36 OKLA. CITY U. L. REV. 231, 243–45, 249 (2011) (noting that a great deal of discovery activity is now focused solely on litigating the summary judgment motion, and questioning whether it actually saves money "from a systemic point of view"). *But cf.* Clermont, *supra* note 159, at 1946–51 (noting that time-to-completion in federal cases has been remarkably stable over the last several decades, and arguing that supply-side effects dynamically enforce such stability, as changes in delay tend to be balanced out by changing demand for litigation).

[164] *See generally* JOHN HENRY SCHLEGEL, AMERICAN LEGAL REALISM AND EMPIRICAL SOCIAL SCIENCE 82–98 (1995). Clark had previously directed a number of empirical projects and had urged that such data "may be used to illustrate and to test the efficacy of our rules of procedure." *Id.* at 85.

contingency, have little incentive to employ stalling tactics, but defendants and their attorneys can sometimes benefit by using summary judgment as a means of delaying cases and raising costs for plaintiffs.[165] Particularly in high-stakes cases, defendants will have much to win, and little to lose, by seeking summary judgment. So instead of being used in rare cases[166] as "a simple and quick way of disposing of routine . . . cases of debts or liquidated demands," as Clark expected,[167] litigants now routinely move for summary judgment in large and complicated cases where its efficiency-promoting value is less clear.[168]

Of course, none of this happened overnight, and there were many steps along the way that helped usher in the modern use of summary judgment. Some courts, including the Third Circuit, initially attempted to resist the movement towards defensive use of summary judgment by allowing plaintiffs to point to their pleadings as evidence of material fact disputes, but the Rules Committee amended the rule to make it clear that this was not appropriate.[169] The "basic purpose" of summary judgment, they explained, was to "assess the proof in order to see whether there is a genuine need for trial."[170] Two decades later, in the group of cases that have become known as the summary judgment "trilogy,"[171] the Supreme Court gave its blessing to the widespread use of summary judgment as a defensive device. *Celotex* was particularly important in this regard: noting its view that summary judgment should be viewed as an "integral part of the Federal Rules" rather than a "disfavored procedural shortcut," the Court held that defendants should be able to obtain summary judgment so long as they could show an absence of evidence favoring the plaintiff's claim; counter-evidence *disproving* the claim would not be necessary.[172] Perhaps these

---

[165] *See* Bronsteen, *supra* note 163, at 528–30 (describing the deterrent effect that frequent summary judgment motions can have on potential plaintiffs). The problem becomes more significant given that many defense attorneys bill by the hour; thus, delay tactics have extra benefits for them privately even if their clients gain little by waiting.

[166] *See* Stephen N. Subrin, *How Equity Conquered Common Law: The Federal Rules of Civil Procedure in Historical Perspective*, 135 U. PA. L. REV. 909, 980 (1987) (describing the view of Clark and others that summary judgment would be a rare event).

[167] *See* Burbank, *supra* note 158, at 602.

[168] *See* Bronsteen, *supra* note 163, at 522–27.

[169] *See* FED. R. CIV. P. 56 advisory committee's note.

[170] *Id.*; *see also* Adam N. Steinman, *The Irrepressible Myth of* Celotex*: Reconsidering Summary Judgment Burdens Twenty Years After the Trilogy*, 63 WASH. & LEE L. REV. 81, 91 (2006); Wood, *supra* note 163, at 241–42.

[171] Anderson v. Liberty Lobby, Inc., 477 U.S. 242 (1986); Celotex Corp. v. Catrett, 477 U.S. 317 (1986); Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574 (1986).

[172] *Celotex*, 477 U.S. at 325–27.

consolidating amendments and interpretations of the rule could be viewed, to a greater extent than the rule's original drafting, as effectively pursuing a goal of enabling widespread use of summary judgment by defendants. But the basic point remains: it was not long after summary judgment had been promulgated before it started to be used in a way that the original rule drafters never anticipated. And perhaps these subsequent ratifications and expansions would never have seemed like a worthy idea if not for the gradual spread of the defensive use of the device from non-existent, to an occasional rarity, into a method of case management that eclipsed trials on the merits. Thus, modern judges, lawyers, and litigants might be likened to the frog that, in the old tale, never realized that a pot of water in which he was swimming was slowing rising to a boil until it was too late; we have, by slow and steady pressure from the defense bar, become inured to a procedure that we might not have elected had we been offered a clear choice at the outset, at least in the absence of evidence that it saves either time or money in the aggregate. And given that other rules most likely undergo similar evolution over time, this might give us reason to doubt that our procedures, on average, strike an optimal balance between their costs and their benefits.

The remedy for such mismatches between expected and actual rule functioning might seem clear: give rules a limited trial run before implementing them on a wide scale.[173] Unfortunately, rules in practice may end up working very differently over the long term than they do during such trials. Summary judgment continues to provide an instructive example. The drafters of Rule 56 might have looked to Michigan's experience with a very similar rule in order to gauge its likely utility, but although this might have helped a bit, they still would have had a drastically limited picture of how it would operate over the long term in federal litigation. Part of the problem was that the rule designers sought to make many reforms to federal civil litigation simultaneously.[174] They thus did far more than just expand summary judgment; they also merged legal and equitable cases, enacted a

---

[173] Several commentators have made proposals along these lines. Some have urged a process of formal and rigorous experimentation with new rules. *See, e.g.*, Posner, *supra* note 144, at 366–68; Walker, *supra* note 144, at 67–68. Others have promoted the use of local informal "experiments" with rules even in the absence of randomized testing. *See, e.g.*, Cooper, *supra* note 144, at 1799; Tobias, *supra* note 144, at 1324–28. At one point in the early 1990s, the Civil Rules Advisory Committee briefly considered amending the rules to facilitate such informal experimentation, although the Standing Committee tabled this idea. Tobias, *supra*, at 1324.

[174] *See* Subrin, *supra* note 166, at 922–25.

plaintiff-friendly notice-pleading regime, and made a broad array of discovery available to litigating parties.[175]

The adoption and expansion of civil discovery, in particular, critically changed the dynamics of the summary judgment device. Discovery and factual investigation, taken together, are now the most time-consuming tasks that litigators face,[176] and an increasingly large volume of this time and expense is devoted to preparing elaborate records for supporting and defending summary judgment motions.[177] When the drafters envisioned the costs of moving for summary judgment, they imagined an affidavit-driven practice, but courts and rule makers, desiring to make the process as fair and accurate as possible, have invited litigants to take extensive oral depositions, introduce interrogatory answers, and amass large quantities of requested documents.[178] Preparing this, and reviewing it, takes a great deal of time, and—given the rarity of trials—it is quite possible that the time and expense we save by granting summary judgment motions in some cases is outweighed by the time and expense we lose preparing for, reviewing, and deciding summary judgment motions.

B. The Challenges of Detecting Procedural Effects Using Observational Data

Summary judgment, then, provides an object lesson in the difficulty of designing new procedural rules: our instincts about how litigators will employ such rules may be misguided, and even observing how the rule has worked in practice may be unilluminating, given that future use may juxtapose one new rule with others and that the interaction among rules may be unpredictable. Indeed, even though modern scholars have gone to great lengths to gather and evaluate observational data,[179] we still know very little regarding the question we started with: Does the availability of summary judgment increase the overall efficiency of litigation?

The basic problem we face is that the answer to this question is counterfactual. No matter how many times we count how often such

---

[175] *See* FED. R. CIV. P. 2, 8, 26; Subrin, *supra* note 166, at 923.

[176] *See* David M. Trubek et al., *The Costs of Ordinary Litigation*, 31 UCLA L. REV. 72, 90–91 (1983).

[177] *See* Hornby, *supra* note 163, at 274 (noting that almost half of lawyers surveyed believe that discovery is used primarily as a means of preparing summary judgment records, and only secondarily as a trial preparation device).

[178] *Id.* at 247–75.

[179] *See, e.g.*, Burbank, *supra* note 158, at 592–94; Cecil et al., *supra* note 162, at 874–81.

motions are brought, how often they succeed, or how expensive they are, we will not be able to determine how the same cases would have proceeded if summary judgment was not available. Nor will it do to compare cases in which such motions are brought with those in which they are not. Lawyers do not decide whether or not to make a summary judgment motion at random, so comparing the two groups of the cases will draw on biased samples and produce data that is unilluminating.[180] And we cannot compare cases decided before and after the new rule was implemented (even if we had easy access to good data from that period, which we do not), because we will not be able to tell whether any changes were wrought by the addition of summary judgment or by one of the many other changes to federal procedure that the new rules introduced.[181] A more promising answer would be to compare similar cases brought in two jurisdictions that were as similar as possible except that one lacked summary judgment. Locating such a perfect comparison, however, may prove frustrating, given the wide adoption of summary judgment in United States jurisdictions and the numerous dissimilarities one would encounter if one tried to draw on foreign examples.

These problems can be seen more clearly if we shift to an example that has sparked much more recent debate: the impact of the Supreme Court's recent pleading cases, *Bell Atlantic Corp. v. Twombly*[182] and *Ashcroft v. Iqbal*.[183] Courts had long interpreted Federal Rule of Civil Procedure 8 to require relatively modest "notice" pleading, but in these two decisions the Supreme Court instructed that plaintiffs must also provide sufficient detail to make their claims "plausible."[184] Most scholars agree that this new wording raised the bar by which lower courts evaluate the sufficiency of civil complaints, which in theory should make it harder for plaintiffs to press their claims.[185] But not all agree,[186] and rule makers are waiting to see

---

[180] *See* ANGRIST & PISCHKE, *supra* note 98, at 12–15 (describing the biasing impact of selection effects).

[181] *See* E. Allan Lind et al., *Methods for Empirical Evaluation of Innovations in the Justice System*, *in* EXPERIMENTATION IN THE LAW: REPORT OF THE FEDERAL JUDICIAL CENTER ADVISORY COMMITTEE ON EXPERIMENTATION IN THE LAW 87, 97–99 (1981) (discussing some difficulties with non-randomized before-and-after research designs, including the difficulty of eliminating the confounding effect of changes in external circumstances other than those under investigation).

[182] Bell Atl. Corp. v. Twombly, 550 U.S. 544 (2007).

[183] Ashcroft v. Iqbal, 556 U.S. 662 (2009).

[184] *Iqbal*, 556 U.S. at 680; *Twombly*, 550 U.S. at 570; *see also* JOE S. CECIL ET AL., MOTIONS TO DISMISS FOR FAILURE TO STATE A CLAIM AFTER *IQBAL*: REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES 1 (2011).

[185] *See* Miller, *supra* note 149, at 15–20 & n.52 (collecting some of the voluminous literature).

[186] *See, e.g.*, Martin H. Redish & Lee Epstein, Bell Atlantic v. Twombly *and the Future of Pleading*

evidence that these decisions are impacting dismissal practice before they consider whether the "new" rule should be returned to its former state.[187]

A number of studies have attempted to detect whether *Twombly* and *Iqbal* have made it harder for plaintiffs with valid claims to be heard by the courts.[188] Although these efforts are well-meaning and have involved substantial effort, they are instructive more for their shortcomings than their successes. In the first place, it is remarkably hard to get systematic data about what courts do. Most of the studies, in an attempt to save money and time, cull case data from legal search engines like Westlaw, even though most agree that this represents a biased sample of cases and almost surely overcounts grants of dismissals as compared with denials.[189] So far, one group of researchers has gotten fuller access to court data while conducting research for the Federal Judicial Center ("FJC"), and has used electronic docket data in order to get a broader picture of the grant rate for motions to dismiss.[190] Drawing on a large sample of cases, the FJC study authors compared case activity in 2006 (before *Twombly* was decided) and 2010 (a year after *Iqbal* was decided).[191] They reported that more motions to dismiss were filed in 2010 than in 2006, both on an absolute and on a percentage basis.[192] They also showed that courts more readily granted motions to dismiss "with leave to amend" in the later year, but that courts were not more likely to grant such motions "without leave to amend," which would indicate that the plaintiff was being finally thrown out of court.[193] They then explained that although the increase in "without leave to amend" dismissals was accompanied by a slight rise in case terminations soon after those orders, those increases were small enough that they could

---

in the Federal Courts: A Normative and Empirical Analysis 7 (Northwestern Univ. Sch. of Law Pub. Law and Legal Theory Series, Paper No. 10-16, 2008), *available at* http://ssrn.com/abstract=1581481.

[187] Lonny Hoffman, Twombly *and* Iqbal*'s Measure: An Assessment of the Federal Judicial Center's Study of Motions to Dismiss* 3–4 (Univ. of Hous. Law Found., Paper No. 1904134, 2011), *available at* http://ssrn.com/abstract=1904134.

[188] *See, e.g.*, CECIL ET AL., *supra* note 184, at 1–3; Redish & Epstein, *supra* note 186, at 8–9.

[189] *See* Cecil et al., *supra* note 162, at 869–72; Hoffman, *supra* note 187, at 8–9. The essential problem with such a sample is that courts are more likely to publish interesting or controversial decisions than routine ones, and the grant of a motion to dismiss is seen as a more significant step than a denial. Some, however, think that this reasoning may be overstated, and that the two data sources may be broadly comparable. *See* Patricia Hatamyar Moore, *An Updated Quantitative Study of* Iqbal*'s Impact on 12(b)(6) Motions*, 46 U. RICH. L. REV. 603 (2012) (arguing that the theoretical foundations for the expected difference are weak, and showing substantial similarity between the overall percentages of grants and denials in two samples).

[190] *See* CECIL ET AL., *supra* note 184, at 5.

[191] *Id.*

[192] *Id.* at 8–12; Joe S. Cecil, *Of Waves and Water: A Response to Comments on the FJC Study* 42 (Fed. Judicial Ctr., Draft, 2012), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2026103.

[193] *See* CECIL ET AL., *supra* note 184, at 14.

plausibly be the result of random variation between the two years.[194] Finally, they demonstrated that if one creates a statistical model that controls for a few factors[195] unrelated to the Supreme Court's decisions, the rise in "without leave to amend" grants disappears except in "financial instrument" cases, a category in which there has been a surge of litigation arising out of the economic downturn of 2008 and which plausibly might be dismissed at a higher rate due to the novel theories such cases often incorporate.[196]

But although this data is helpful in giving us some sense of how *Twombly* and *Iqbal* have affected civil case processing, it still leaves many of the most important questions unanswered. For one thing, the study is limited to looking at those claims that were actually filed, which means it cannot detect the degree to which plaintiffs with valid claims are deterred from filing lawsuits.[197] Given that one would expect rational plaintiffs to take into account the likelihood of losing a motion to dismiss in their calculations when deciding whether or not to sue, this means that this study may substantially undercount the harm that the new pleading rule does to some deserving plaintiffs.

On the flip side, of course, the study has no way of assessing the merit of dismissed claims.[198] Perhaps the increase in motions filed and granted is due to a rise in dubious filings over the last few years (perhaps in the recently swollen "financial instrument" category, which courts are able to screen out using dismissal orders). If that were the reason for the rise in dismissals, it would make a poor case for reform efforts. But it is equally possible that the new filings are in fact better than average, but that courts nevertheless abuse their authority under *Twombly* and *Iqbal* to disfavor novel claims. The basic problem is that, without a measure of the changing validity of complaints over time, any attempt to infer that *Twombly* and *Iqbal* have had either a beneficial or a negative impact based on grant rates alone misses the point. Even if other authors are correct that grant rates are increasing for reasons unrelated to changing filing patterns,[199] this could be

---

[194] *Id.* at 16.

[195] *See id.* at 5–6. The authors make no attempt to demonstrate that the factors they include as controls—the district the case was in, a broad category representing case type, and whether the complaint at issue had been previously amended—represent all plausible factors that might be responsible for a shift in dismissal rates between 2006 and 2010. *See id.*

[196] *Id.* at 21.

[197] *See id.* at 14, 26.

[198] *See id.* at 28.

[199] *See* Moore, *supra* note 189, at 637–38.

a sign that the new approach is succeeding, if there has been an uptick in frivolous filing. Conversely, even if dismissals are holding steady or even declining, this might be masking a real impact, if the average validity of complaints was rising.

Ultimately, however, there is a deeper problem with observational studies of this design: the dubiousness of inferring causal impact based on snapshots of litigation activity before and after a rule's enactment. In some areas, estimating causation based on before-and-after measurements can be a reasonably valid approach.[200] But when dealing with new rules of procedure or evidence, we should place low confidence in conclusions driven by such reasoning. Such research designs are subject to a dilemma. If we measure very soon before and after a new rule comes into effect, we may be reasonably confident that any before-and-after differences are due to the new rule.[201] But, as the authors of the FJC study explain, legal rules take time to percolate before arriving in a longer-term equilibrium form.[202] Lower court judges may disagree about how to apply a new rule, appellate courts may reshape the rule through interpretation, and lawyers may develop new tactics that change how the rule works in practice. The clearest example of why short-term measurement is a risky approach is summary judgment; for *decades* after the enactment of that device, it was used quite cautiously,[203] but over time it became increasingly attractive to federal judges facing rising docket pressures and perceiving an excess of frivolous litigation.[204] A study examining the few years after its enactment would have failed, therefore, to appreciate its likely long-term impact. Similarly, it may turn out that lawyers with good claims are almost always able to find the facts they need to satisfy a plausibility inquiry, but that it will take time for them to learn how. In that case, any initial impact of the new regime on meritorious plaintiffs may evaporate over the longer term, and a short-term study would badly overstate the new rules' demerits.

---

[200] *See generally* Lind et al., *supra* note 181, at 97–99.

[201] We can reasonably assume that most other variables within the legal system look fairly similar right before and right after a major change in law. Occasionally this is not the case, but using a small time horizon makes it easier to notice any other major confounding changes. It would *not* be safe to make such an assumption in situations where many changes were occurring at once, as is the case when a number of new rules come into existence simultaneously.

[202] *See* CECIL ET AL., *supra* note 184, at 2–3.

[203] *See, e.g.*, Arnstein v. Porter, 154 F.2d 464, 470, 475 (2d Cir. 1946).

[204] *See* Arthur R. Miller, *The Pretrial Rush to Judgment: Are the "Litigation Explosion," "Liability Crisis," and Efficiency Clichés Eroding Our Day in Court and Jury Trial Commitments?*, 78 N.Y.U. L. REV. 982, 1022–41 (2003).

Unfortunately, the other horn of the dilemma is equally troubling. If we lengthen our before-and-after period to try and wait for a real rule-application equilibrium to arise, it becomes increasingly likely that other changes in the legal system's operation will make it hard to isolate the causal effect of a rule change. Take the FJC study as an example: even though the authors wait only a year after the second of the two cases that propound the new pleading regime, they already report a significant rise in federal court caseloads and a shift in the types of cases being brought.[205] It may be possible to credibly control for such changes using statistical methods,[206] but such modeling can only correct for changes that researchers can anticipate and measure. Many other aspects of litigation relevant to motion-to-dismiss practice might be occurring, from changes in lawyering tactics or changes in the overall economic climate to changes in the makeup and ideology of the federal bench, any of which might affect pleading practice. The longer we wait, the surer we can be that we have not anticipated or measured all the relevant factors. In short, whether we engage in short-term or long-term before-and-after observation of rule impacts, we will always face significant unknowns regarding the true impact of new rules.[207]

It is, of course, theoretically possible that our intuition and ratiocination can succeed where our attempts at explicit measurement cannot. Maybe, in other words, we should not worry about defects in our ability to measure the impact of rules so long as the results of questionable research designs accord with what procedural scholars expect. After all, people were throwing and catching balls long before Isaac Newton provided a formalized understanding of gravity. Perhaps judges and lawyers, after enough immersion in the legal system, develop an intuitive understanding of litigation that rivals our in-built facility for predicting the movement of physical objects in space.[208] If that was the case, an evidence-based movement might actually do more harm than good by causing us to throw out our theoretic judgments based on (as we have seen) the questionably valid data we normally can gather regarding procedural questions.

Unfortunately, this theoretical possibility is almost certainly not a reality, no matter how much confidence individual lawyers and judges have

---

[205] *See* CECIL ET AL., *supra* note 184, at 7.

[206] The FJC authors, for instance, do control for changing case types. *See generally* CECIL ET AL., *supra* note 184.

[207] *See* Lind et al., *supra* note 181, at 97–99.

[208] *See* Robert G. Bone, *The Empirical Turn in Procedural Rule Making: Comment on Walker (1)*, 23 J. LEGAL STUD. 595, 599 (1994).

in their ability to say which new rules would work well and which poorly. For one thing, internal confidence is a poor guide to such accuracy. The ball thrower knows, with practice, that he can hit a target, but he knows this because regular, clear feedback is easy to get in such an environment. By contrast, it is hard to say for sure whether a new rule is improving legal practice or not; "success" is described as an optimal trade-off among competing values, and no one occupies an omniscient perspective from which all costs and benefits can be assessed. Our position is much more like the sixteenth century doctor than the baseball player. We know we want to resolve cases quickly, cheaply, and correctly, just as the doctor knew he wanted to make sick people better, but any one outcome could be the result of our new intervention or something else entirely. Just as it was hard to evaluate the success of bloodletting by evaluating cases one at a time, so it is hard to evaluate the effect of a rule providing for sanctions for discovery misconduct just by reading cases that employ it.

C. Building on the Medical Example: Using Randomized Controlled Trials of New Rules of Procedure

Perhaps, however, the medical analogy can do more than help us define the limits to our present knowledge. Can the recent rise of evidence-based medicine teach us more sophisticated ways to evaluate procedural success? Perhaps it can. One of the greatest tools in the rise of scientific medicine has been the randomized controlled trial.[209] Such trials represent a gold standard method of assessing causation because they isolate the impact of a single factor's presence or absence in a comparison between two groups that are as close to identical as is possible.[210] If we wish to have greater confidence that new rules actually provide a significant benefit over the status quo, randomizing rule application so that some litigants will be randomly assigned the new rule, while others the old one, has the potential to isolate the causal effect of the new rule. What is more, such a trial could be run for long enough that litigators and judges in a particular area develop enough experience using the new rule to provide an indication of what a "mature" version of such a rule-in-action will be.

This proposal has strong intuitive appeal, and a number of scholars have made proposals along these lines over the years.[211] On a few

---

[209] Abramowicz et al., *supra* note 144, at 933.

[210] *See* HANS ZEISEL ET AL., DELAY IN THE COURT 241–42 (1959); Abramowicz et al., *supra* note 144, at 935; Greiner & Pattanayak, *supra* note 144, at 2196–97.

[211] *See, e.g.*, ZEISEL ET AL., *supra* note 210, at 241–50; Posner, *supra* note 144, at 374–75; Tobias,

occasions, courts and scholars have dipped their toes into these waters, testing out new procedural rule proposals using randomized experiments.[212] Despite the promise of such an idea, however, we must be realistic about what would be involved in attempting to improve our litigation system through systematic testing of this nature. The history of medicine shows that although some effective therapies can be found by a brief bit of experimental testing,[213] there was a long period during which experimentation served primarily to show doctors what failed to help without providing much guidance as to what could do better.[214] Two transformations were necessary before we could develop high confidence in medical therapies (and in some areas of medicine, these two transformations may still be works in progress).

First, it was necessary to test many alternative models of how the body worked in order to develop an accurate understanding of the underlying biology of bodies and their diseases.[215] Here, legal scholars may start with an advantage, as we will not need to invent microscopes in order to develop plausible theories of "lawsuit physiology." Rather, once initial confidentiality challenges are overcome, the relevant players can be interviewed and a large amount of data about the internal workings of cases can be collected at comparatively low cost.

But the other transformation—conducting sufficient experimentation to locate the best possible rule configurations[216]—may prove more challenging. In order to find effective therapies, it is also necessary to try many things that will fail in order to identify those few things that reliably succeed.[217] Such an undertaking may test our mettle. First, we will have to develop a willingness to do far more procedural experimentation than has historically been the norm. Second, we will be challenged by the fact that we have far fewer courts than human beings. Trying to run multiple trials in the same court at one time may risk interaction effects between new rules

---

supra note 144, at 1324–25; Walker, *supra* note 144, at 67–68; Willging, *supra* note 1, at 1197, 1201–04.

[212] *See, e.g.*, Paul R.J. Connolly & Michael D. Planet, *Controlling the Caseflow–Kentucky Style: How to Speed Up Litigation Without Slowing Down Justice*, 21 JUDGES J. 9, 11 (1982) (describing a RCT involving competing approaches to case management); Valerie P. Hans et al., *The Arizona Jury Reform Permitting Civil Jury Trial Discussions: The Views of Trial Participants, Judges, and Jurors*, 32 U. MICH. J.L. REFORM 349, 365 (1999) (describing an RCT of a new rule permitting jurors to discuss their cases before trials were concluded); *see also* Willging, *supra* note 1, at 1131 n.41 (collecting a number of additional recent experiments).

[213] *See, e.g.*, BYNUM, *supra* note 48, at 42–43.

[214] *See id.* at 44.

[215] *See id.* at 43.

[216] *See id.*

[217] *See* Carrier, *supra* note 107, at 417 (noting the extraordinary number of potential drugs that must be tested before a single marketable treatment can be identified).

that we cannot reliably untangle through analysis. So there would be challenges involved in setting appropriate research priorities and allocating courts to different rule-trials.[218] In the end, it seems doubtful that we will identify many optimal procedural choices through the sort of sporadic, ad hoc experimentation that has been typical of the last few decades.

I will not belabor the value of randomized trials as compared with observational studies. Others have made that argument better than I can, and in any event it may be possible to approximate many of the benefits of randomized experiments through increasingly thoughtful design of observational studies. There is one final challenge involved in creating an evidence-based procedural reform movement that has no parallel in medicine, however, and which will inevitably plague attempts to make evidence relevant to procedural design whether we conduct experiments or count case outcomes.

As discussed above, doctors have sometimes disagreed about whether to measure objective immediate indicators of health or more ultimate, but subjective, qualitative information.[219] Procedural reform, however, will involve a much thornier challenge. Simply put, some of the outcome variables of greatest interest to procedural reformers go unmeasured in nearly all empirical investigations of procedural effects, even those observational or experimental studies that approach most closely a gold standard ideal. In particular, one of the key goals of procedure is to achieve "just" or "accurate" outcomes in individual cases, but almost no studies attempt to measure accuracy. The next few sections of this Article will discuss the severity of this problem and how it might be addressed through novel research designs.

---

[218] One important means of winnowing down the search space would be to first use quasi-experimental methods to learn of potential rule effects through analysis of the effects of rules that vary across existing jurisdictional lines or judicial districts. Such studies provide the advantage that they allow a researcher to analyze pre-existing data without waiting for a randomized assignment process to run its course, but they come at a necessary cost, in that the results of such quasi-experiments necessarily hinge on our ability to anticipate, measure, and control for potential confounding factors. *See generally* Elizabeth A. Stuart & Donald B. Rubin, *Best Practices in Quasi-Experimental Designs*: *Matching Methods for Causal Inference*, *in* BEST PRACTICES IN QUANTITATIVE METHODS 155–76 (Jason W. Osborne ed., 2008). Thus, such approaches might provide a valuable way of identifying some potentially useful rule choices, but would ideally be supplemented by a true procedural experiment before being relied on to make long-term procedural policy on a national scale. They could also prove valuable in measuring effects of procedural rules on pre-litigation conduct, a setting in which true random assignment is most likely unfeasible.

[219] *See* discussion *supra* Part II.

IV. THE IMPORTANCE OF MEASURING ACCURACY WHEN EVALUATING PROCEDURAL RULES

As we have discussed, rules of procedure are hard to design, both because rules behave unpredictably and because it is hard to detect the true impact of the rules governing the litigation system. Perhaps, then, rule makers should imitate those who promote evidence-based decision making in medicine, and seek to collect and use better data before arriving at decisions. Indeed, those who are attracted to such a solution might look optimistically to the rising tide of empirical analyses of legal policy questions. But before we confidently board the evidence-based train, it is critical that we pause and reflect on the ways that getting the right evidence relating to questions of litigation policy differs from what is typical in the medical world, or even what is typical for most questions of regulatory policy.

Unfortunately, those who would create the architectural rules of litigation face a special challenge: one of the key variables of interest when comparing rules is the degree to which they promote the accurate resolution of cases, but no easy metric exists external to the litigation system that allows for broad comparisons of the accuracy of case results. So if we want to get the benefits of evidence-based procedural design, neither the observation of existing cases, or even experiments where we compare differing rules, will give us the answers that we need. As a result, the costs of an evidence-based litigation movement are higher than they initially appear, and procedural designers should proceed with great caution before making decisions based on the sort of data that existing empirical scholarship can provide.

A. An Example: Assessing the Civil Gideon *Debate*

The need for accuracy measurement can best be illustrated by considering another example: the ongoing debate about the right to appointed counsel in civil cases. For most of Anglo-American legal history, it was commonplace for parties even in the most serious criminal cases to self-represent.[220] In serious criminal cases, in fact, defendants were actively *forbidden* counsel for hundreds of years.[221] After a series of scandals involving prominent convictions (and executions) obtained by

---

[220] *See* WILLIAM M. BEANEY, THE RIGHT TO COUNSEL IN AMERICAN COURTS 8–9 (1955).

[221] *Id.* at 9.

perjured evidence in the late seventeenth century, the tide began to turn, and it slowly became possible for those felony defendants who could afford attorneys to retain them as trial counsel.[222] But most criminal defendants are not rich, and so many went without legal advice or representation even in this new system.[223]

Only very recently has the tide shifted towards the widespread use of appointed counsel by the indigent.[224] In 1963, the Supreme Court held in *Gideon v. Wainright* that the Constitution required states to provide indigent felony defendants with state-funded counsel.[225] This decision is widely held up as a momentous improvement in the fairness and accuracy of American criminal prosecutions.[226] But although *Gideon* firmly entrenched a right for criminal defendants to receive counsel, there are numerous other areas of litigation, including most types of civil cases, where indigent parties do not have a right to be represented at state expense.[227] Those who oppose this state of affairs label themselves the "Civil *Gideon*" movement.[228]

So let us consider the debate that is joined between the Civil Gideonites and the defenders of the status quo. On the status quo side, there are some

---

[222] Powell v. Alabama, 287 U.S. 45, 60 (1932); BEANEY, *supra* note 220, at 10–11 (noting that England denied many accused felons the right to hire counsel until 1836).

[223] *See* BEANEY, *supra* note 220, at 12–33 (describing the slow movement towards full indigent representation in England and describing the similarly slow American progress from pre-Revolutionary days through the early twentieth century).

[224] As late as the early twentieth century, neither the federal government nor most states guaranteed the right to counsel to those defendants who could not afford an attorney. *See* Betts v. Brady, 316 U.S. 455, 468 & n.22 (1942) (collecting state authority); Johnson v. Zerbst, 304 U.S. 458, 463 (1938) (holding that appointed counsel is guaranteed under the right to counsel pursuant to the Sixth Amendment in federal courts); Benjamin H. Barton, *Against Civil* Gideon *(and for Pro Se Court Reform)*, 62 FLA. L. REV. 1227, 1236 (2010) (noting that *Zerbst* was the first case to recognize a right to appointed counsel at the federal level).

[225] Gideon v. Wainwright, 372 U.S. 335, 339, 342 (1963); *cf.* Argersinger v. Hamlin, 407 U.S. 25, 37 (1972) (extending the right to misdemeanor cases that result in jail time of any duration). The right to appointed counsel in federal prosecutions arose in *Johnson*, 304 U.S. at 463. Ultimately, *Gideon* nationalized a movement that was already underway at the state level; by 1963, most states provided for appointed counsel in all "serious" criminal cases. Yale Kamisar, *How Earl Warren's Twenty-Two Years in Law Enforcement Affected His Work as Chief Justice*, 3 OHIO ST. J. CRIM. L. 11, 20 (2005). *See generally* Yale Kamisar, *The Right to Counsel and the Fourteenth Amendment: A Dialogue on "The Most Pervasive Right" of an Accused*, 30 U. CHI. L. REV. 1 (1962).

[226] *See* David Strauss, *The Common Law Genius of the Warren Court*, 49 WM. & MARY L. REV. 845, 851 (2007) (describing the case as one of the two "most celebrated" criminal procedure decisions of the Warren Court).

[227] *See, e.g.*, Lassiter v. Dep't of Soc. Servs., 452 U.S. 18, 26–27 (1981). *See generally* Barton, *supra* note 224, at 1246 (concluding that the movement towards expansion of *Gideon* into civil contexts died with *Lassiter*).

[228] *See* Barton, *supra* note 224, at 1227–30.

reasons why we might be skeptical that *Gideon* has been beneficial either for defendants or for society in general. For one thing, the Court has been reluctant to extend its ruling to contexts not involving potential incarceration, suggesting some lingering uncertainty about the need for appointed lawyers in all cases.[229] For another, a right to appointed counsel is meaningful only to the extent that such counsel is well-funded and well-monitored; the modern history of indigent criminal defense, by contrast, is one of "grossly inadequate funding" in which courts are extremely deferential to decisions made by highly overburdened counsel.[230] Overburdened appointed attorneys have strong incentives to conduct little investigation into their cases and to pressure defendants to settle, and those incentives may lead the defendants with meritorious cases to be funneled into the same plea machine as the guilty ones.[231] Although this situation could in theory be resolved through better funding and better oversight of public defender agencies, in practice criminal defendants as a group lack the political capital to successfully advance a reform agenda. Still, the Civil Gideonites might reply that all of these reasons for doubt are merely suggestive and that the alternatives are likely to be worse. As the *Gideon* Court observed, "[e]ven the intelligent and educated layman" may "lack[] both the skill and knowledge adequately to prepare his defense, even though he have a perfect one."[232] So perhaps even a very poor appointed attorney is better than proceeding *pro se*.

So we have a pressing puzzle on our hands: Does appointing counsel help the indigent, and if so, does it make the system as a whole function better? There is a large empirical literature attempting to quantify the benefits and costs of appointed counsel, but sadly it turns out to be of little use. Advocates of Civil *Gideon* can point to a deluge of observational studies showing that represented litigants tend to fare better than those who represent themselves.[233] Despite this onslaught of case data, status quo defenders have a potent counterattack: observational studies tell us little about the true effects of representation because they are subject to severe selection effects.[234] The problem with comparing the outcomes that *pro se* and represented litigants achieve is that lawyers are not randomly assigned

---

[229] *Id.* at 1250–59.

[230] *Id.*

[231] *See id.* at 1253–54.

[232] Gideon v. Wainwright, 372 U.S. 335, 345 (1963) (quoting Powell v. Louisiana, 287 U.S. 45, 69 (1932)).

[233] *See* Greiner & Pattanayak, *supra* note 144, at 2127 n.154.

[234] *Id.* at 2183–84, 2197.

to indigent claimants in civil settings; rather, lawyers seek out cases with merit.[235] On the flip side, the indigent litigants who are most likely to seek out representation may constitute a "disproportionately worldy, future-looking, and risk-averse" subpopulation who may be more likely to prevail in litigation regardless of whether they have a lawyer's aid.[236]

Better data can be obtained by conducting randomized experiments, which can create control and treatment groups of indigent litigants who are as near identical as possible except with respect to the offer of representation.[237] But this inquiry has rarely been conducted, and when it has the results have been mixed, with a few studies showing positive impacts of representation and others finding that representation does not increase the odds of success but does slow down the process of litigation.[238]

One way to read such mixed results would be to conclude that in trying to decide when counsel should be appointed, detail is everything. Different case types and regions may raise or lower the value of having a lawyer in ways that are hard to predict in advance, although it is possible to concoct just-so stories to explain data once we have them. In short, we see again the problem of procedural complexity, and it seems like the most promising response is to run more experiments in a variety of settings, trying to see when appointing counsel is helpful and when it is counterproductive.[239] The only alternative is to trust our intuitions even when randomized experiments indicate that they are sometimes unreliable.

But there is a deeper problem with this debate, one that we cannot see unless we are willing to reframe the measures by which we evaluate the value of a right to counsel. Recall the discussion above regarding the modern medical movement towards outcome-based measurement in

---

[235] *See id.* at 2195.

[236] *Id.* at 2192.

[237] *See* ANGRIST & PISCHKE, *supra* note 98, at 15–22.

[238] *See* Greiner & Pattanayak, *supra* note 144, at 2118 (finding no association between the offer of legal assistance and success on the merits in administrative appeals of the unemployment assistance denials); D. James Greiner et al., *The Limits of Unbundled Legal Assistance: A Randomized Study in a Massachusetts District Court and Prospects for the Future*, 126 HARV. L. REV. (forthcoming 2012), *available* at http://ssrn.com/abstract=1948286 (finding that offers of full traditional legal assistance helped clients prevail in eviction disputes in Massachusetts district court).

[239] Of course, even this neglects the true level of legal complexity involved. A brief experiment regarding the effects of representation may not capture the value (or costs) that a system-wide guarantee would have over the long term, because many legal actors might change their behavior over time in response to the new guarantee in ways that are hard to predict. Likewise, the identity of the appointed representatives might be different in a study than would be the case once reforms are institutionalized. A civil "public defense" office may not litigate as effectively as volunteer attorneys whose performance is measured in a study—or perhaps they may develop specialized expertise and litigate more effectively.

validating the efficacy of medical procedures and diagnostic tests.[240] Often, a medical researcher has the choice between measuring something that is inexpensive to determine, like cholesterol levels or blood pressure, versus something that is harder to measure, like long-term changes in mortality or subjective assessments of overall patient well-being.[241] Unfortunately, choosing the easy-to-measure proxy can have serious penalties for the value of a research design: given that the body is a very complex environment, sometimes an intervention may seem to have a positive impact on a proxy outcome while also having a non-existent—or even harmful—effect on overall health.[242] Therefore, if one wishes to maximize the ability of medicine to improve patient health, it is best to validate tests and procedures by measuring their impacts on overall health rather than on temporary signs that may, or may not, signal true improvements in well-being.[243]

Returning to the Civil *Gideon* example, we can perceive a similar dichotomy between cheap-but-accessible and expensive-but-instructive measurements when we study the efficacy of procedural rules in litigation. Whether based on statistical analysis of ordinary cases or random assignment of representational offers, most existing studies of procedural impact have a common structure: they either observe or experimentally vary the availability of representation, and then they observe and record certain variables about case outcomes, trying to detect a link between the two.[244] In theory, well-designed studies that focus on case outcomes, without measuring accuracy, could help us answer the following questions: Does representation make a litigant more or less likely to win? Does it make the case take more or less time to resolve? Does it increase, or decrease, settlement rates? But they cannot possibly answer what may be the most important question: Does appointing counsel for indigent clients make it more or less likely that the ultimate outcome in the case is right?[245] After all, it is at least possible that appointing counsel tends to help more parties prevail, but that the increase comes primarily in the form of false positives rather than true positives, and that the overall accuracy of the system

---

[240] *See* discussion *supra* Part II.E.

[241] *See* Deyo, *supra* note 69, at 65–66.

[242] *See id. See generally* Oliver et al., *supra* note 135, at 379–85.

[243] *See* Oliver et al., *supra* note 135, at 385.

[244] *See* Greiner & Pattanayak, *supra* note 144, at 2118. *See generally* Greiner et al., *supra* note 238.

[245] Readers who worry that my use of the word "right" masks a great deal of complexity are correct and are urged to be patient. I will discuss the theoretic depths involved in trying to measure accuracy in outcomes *infra*.

suffers. Such a result would run strongly counter to the widely shared urge to make civil litigation outcomes reflect real world facts as accurately as we can,[246] subject to the constraints of cost and practicality. So long as we focus on what is relatively easy to measure, like who wins, but not on what matters more, like whether the case outcomes are more or less accurate, the real significance of these studies will be necessarily limited.

This concern—that without measuring variations in the accuracy of litigation results, existing studies tell us little about how well the appointment of indigent counsel furthers some of the justice system's key goals—is not unique to the right-to-counsel debate. For instance, the identical difficulty subverts scholarly attempts to assess the merits of the recent pleading revolution. Even at its best, this empirical literature is unable to get at the questions that are of real theoretical interest in the "*Twiqbal*" debate. Suppose, for instance, that despite its limits the FJC study correctly indicates that *Twombly* and *Iqbal* have caused judges to dismiss more cases for failure to state a claim.[247] That might be a very *good* thing, if most of the dismissed cases are in fact frivolous, or a very *bad* thing, if judges are using their newfound freedom to dismiss worthy cases brought by unpopular groups of plaintiffs.[248] One doubts, after all, that the Supreme Court granted certiorari in these cases hoping to have no effect on pleading practice. Rather, the aim was to screen out frivolous cases without affecting meritorious ones.[249] If the new rule truly achieves this, it should be subject to far less criticism than critics presently aim at it. At the same time, the situation might be inverted; frivolous plaintiffs may well be able to plead plausible cases most of the time, while some deserving plaintiffs may lack access to the facts they need. But unless we find a way to

---

[246] *See* Louis Kaplow & Steven Shavell, *Accuracy in the Determination of Liability*, 37 J.L. & ECON. 1, 1 (1994). *But cf.* 4 WILLIAM BLACKSTONE, COMMENTARIES *352 (urging that minimizing false convictions is more important than maximizing true convictions).

[247] *See* CECIL ET AL., *supra* note 184, at 21 (finding an increase in the rate at which motions to dismiss for failure to state a claim were granted in cases involving financial instruments); Hoffman, *supra* note 187, at 4–5.

[248] *See* Moore, *supra* note 189, at 654 (noting that case outcome data cannot tell us whether dismissals are occurring in the cases where they "should happen") (internal quotation marks omitted); Victor D. Quintanilla, *Beyond Common Sense: A Social Psychological Study of* Iqbal*'s Effect on Claims of Race Discrimination*, 17 MICH. J. RACE & L. 1, 54–55 (providing evidence that the post-*Iqbal* increase in the rate of reported dismissals is higher in race discrimination suits brought by black plaintiffs than it is in the larger population of suits examined by Moore).

[249] *See, e.g.*, Bell Atl. Corp. v. Twombly, 550 U.S. 544, 557–58 (2007) (expressing a worry that a plaintiff with "a largely groundless claim" could effectively "take up the time of a number of other people, with the right to do so representing an *in terrorem* increment of the settlement value," if such cases cannot be screened out at the pleading stage) (internal quotation marks omitted).

measure not just the difference in dismissal rates under the two regimes, but also the comparative validity of the claims that survive and the claims that are dismissed under each rule, we really are just guessing when we conclude that the extra dismissals are cause for concern.

B. How Worrying Is Our Failure to Measure Accuracy?

We presently lack any way to systematically assess whether a given change to a system of dispute resolution makes it function more or less effectively, because we have no external means to track whether a given change makes case outcomes more or less accurate on the margin. To see just how troubling this is, imagine if medical researchers could measure the cost of new drugs and some of the side effects they produce, but were barred by medical protocol from measuring whether their drugs make the patients healthier or sicker. Such an environment would give us little reason to trust the beneficence of drugs on the market, so by analogy we should harbor similar doubts with respect to current legal procedures. But some may wish to resist this analogy, and urge that we either have ways of determining accuracy without resort to systematic measurement, or that accuracy is not important enough to be worth measuring. Several arguments along these lines will be considered in turn.

Some might object that the existing appellate process already operates to test the accuracy of case outcomes and provides overall data regarding the accuracy of trial procedures. This response fails for a number of reasons. First, using data from appeals to measure the appropriateness of case resolutions suffers from serious selection bias: most cases filed are never appealed, even though every case filed is resolved in some fashion.[250] Some of the filters are legal rules restricting when an appeal can be taken, while others involve the incentives that parties have to avoid disturbing a status quo. One of the most common situations is when parties settle their disputes and thus lack either the legal standing or the desire to further litigate their claims.[251] But if we care about the accuracy of procedures, we must measure the degree to which settlements track the underlying merit of cases. What is more, this selection bias will exist even in many cases where an appeal was taken. Many potential grounds for appeal rely on procedural issues that do not touch the factual merits of the underlying claims.[252] In

---

[250] *See* Clermont, *supra* note 159, at 1972.

[251] *See, e.g.*, DeFunis v. Odegaard, 416 U.S. 312 (1974).

[252] *See* GARRETT, *supra* note 16, at 10–11 (observing that many wrongful convictions go undetected by an appellate process that focuses on detecting procedural, rather than factual, errors).

such cases, the result on appeal will tell us nothing about the factual validity of the pre-appeal outcome, and it may be very hard to separate out such procedural grounds of review from substantive ones when studying a large dataset of cases.

Second, appellate review ordinarily involves both formal and informal deference to the decision making of a lower court.[253] As a result, if we use appellate reversals as a measure of error, we will systematically undercount such errors. So if, for example, we are looking at a subset of cases that lead to jury verdicts, appellate data will not tell us how often such verdicts are incorrect, but instead only how often they were so incorrect as to be "unreasonable." This means that such data will systematically undercount factual errors at trial.[254]

Third, and perhaps most importantly, using appellate review as a measure of accuracy separates one piece of a complex system from the larger whole in a way that interferes with useful measurement. After all, if we are trying to measure the accuracy of legal outcomes, the outcome of most interest is what happens after all legal procedures have run their course. The appellate process is one aspect of the whole, and it may interact with whatever changes we are trying to assess. In the end, the error measurement we would get by using appellate reversal rates would be worthless: each error counted would be one that the system as a whole would make right, while none of those errors that are the true concern would be measured at all.

It seems, therefore, that we lack any systematic information on the litigation system's validity. Nevertheless, one can still ask, is that really a problem? For some, the realization that we just do not know how accurate our system is will be maddening, while others will shrug. So before turning to how we can go about measuring, it is worth considering whether it is worth the bother.[255]

There are several reasons why our lack of systematic data collection might seem to be less harmful than it initially appears. One sort of doubter might say that accuracy, all things considered, is not a weighty enough variable in the procedural calculus to be worth tracking, at least if the tracking must be costly. Surely, they would say, there is much more to legal decision making than getting the facts right, and indeed much of what

---

[253] *See, e.g.*, FED. R. CIV. P. 52(a)(6) (instructing appellate courts to defer to the factual findings of trial judges unless those findings are clearly erroneous).

[254] *See* GARRETT, *supra* note 16, at 10–11.

[255] *Cf.* Bone, *supra* note 208, at 599 (noting that scientific testing "is not the exclusive route to predictive confidence").

judges and juries do involves the exercise of wisdom and discretion rather than the reconstruction of historical events.[256]

There may well be a fairly broad ideological chasm between those who believe that the goal of accuracy in outcomes is the central benefit we seek in designing procedure[257] and others who think it is merely one virtue of trials among many, and perhaps not even the most important one. Nevertheless, I believe that the difficulty of measuring accuracy in legal outcomes should be of concern to most people interested in procedural design. For one thing, it will be rare indeed to find someone who goes beyond listing other procedural values to take the position that accuracy is harmful or even useless.[258] Rather, the more common intuition will be that we value accuracy as well as other things, such as results that do moral justice or procedures that feel fair to participants.[259] But once this much is granted, then the failure to measure one key goal among others still seems problematic. If we found out that medical researchers were failing to measure the effect of new drugs on health, it would be of little solace to discover that they were able to carefully measure side effects, even if we believed that minimizing side effects was very important.

Even if we think accuracy is less important than other values, this analogy still holds. We care most about curing disease, but we still think it valuable to collect data on the side effects and costs of new therapies. Moreover, our ability to reliably achieve some of these other values, like an emphasis on the ability of the system to allow individuals to vindicate wrongs done against them, may depend on its ability to regularly figure out the right facts.[260] This would mean that enhancing our ability to maximize accuracy also helps us further these other values. After all, few if any moral theories would permit us to determine who is blameworthy and who is not without reference to at least some facts about their past conduct or states of

---

[256] *See* ROBERT P. BURNS, A THEORY OF THE TRIAL 235–40 (1999) (urging that the structure of our trials enables the jury to integrate reconstructions of historical fact with practical and moral judgments to produce an amalgam justice "which is often immeasurably richer" than the outcome of a process focused solely on accuracy could be).

[257] Robert G. Bone, *Making Effective Rules: The Need for Procedure Theory*, 61 OKLA. L. REV. 319, 337 (2008) (expressing skepticism that the demands of participatory rights can ever rise above the needs of adjudicative accuracy).

[258] Indeed, I am aware of no author who has explicitly endorsed such a view.

[259] *See* BURNS, *supra* note 256, at 239; Solum, *supra* note 9, at 320–21.

[260] *See* Benjamin C. Zipursky, *Rights, Wrongs, and Recourse in the Law of Torts*, 51 VAND. L. REV. 1, 82–87 (1998) (describing the dignitary and social values that support permitting those who have been harmed to personally seek redress against their wrongdoers using state-supplied mechanisms, and noting that this value only extends to those cases where the defendant actually violated a right held by the plaintiff).

mind.[261] Therefore, most pluralists should be worried about the present failure to measure accuracy in adjudication.

Despite these responses, however, there is a deeper concern, which is that the concept of accurate case outcomes is too confused to be capable of measurement. Many, if not most, legal outcomes do not offer a clear distinction between factual and normative judgments.[262] A jury verdict, for example, may encode historical assumptions, future predictions, and normative judgments of blameworthiness into a finding of liability and a damages award.[263] Likewise, lawyers may rely on factual assumptions when negotiating civil settlements, but the ultimate agreement does not come with a narrative description of what happened that gave rise to the settlement.[264] And when judges dispose of a case on a procedural ground, they often deliberately avoid making any comment on its underlying factual merits.[265] So many case dispositions blur, or even omit entirely, an inquiry into what "really happened" in the past.

But even though many case dispositions may not articulate or rely upon a particular set of found facts, it is still meaningful to ask whether the relevant decision makers correctly understood the facts that gave rise to the dispute when making their decisions. Lawyers and judges must regularly make decisions that rely on an interplay of at least three key features: the facts as they understand them, the legal rules that oblige certain results, and their own preferences (to the extent they have discretion to give effect to them).[266] Even if a given outcome depends upon a discretionary choice or a

---

[261] Among ethical theories, the two dominant frameworks focus either on evaluating acts by reference to the consequences they produce or else by the "moral quality of the act . . . in itself." *See* R. George Wright, *Combating Civilian Casualties: Rules and Balancing in the Developing Law of War*, 38 WAKE FOREST L. REV. 129, 154–55 (2003) (contrasting consequentialist and deontological modes of ethical reasoning). Under either approach, one needs to describe conduct or its consequences in some detail before any reasonable assessment of its ethical propriety can be conducted.

[262] *See* BURNS, *supra* note 256, at 235.

[263] *See id.* at 235–40; *see also* FED. R. CIV. P. 49 (permitting but not requiring judges to obtain an explanation of jury verdicts by using special verdict forms); *cf.* FED. R. CIV. P. 52(a)(1) (requiring a reasoned explanation when judges, rather than juries, decide cases).

[264] *Cf.* FED. R. CRIM. P. 11(b)(3) (requiring a factual explanation when criminal defendants enter into plea bargains to resolve their disputes with the government).

[265] *See*, *e.g.*, 5B CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE & PROCEDURE § 1353 (3d ed. 2011) (noting that dismissals for improper service of process, like many other technical reasons for dismissing an action under Fed. R. Civ. P. 12(b), are not "on the merits" and are not intended to preclude future litigation in which the defect has been corrected).

[266] *See* Frank B. Cross, *Political Science and the New Legal Realism: A Case of Unfortunate Interdisciplinary Ignorance*, 92 NW. U. L. REV. 251, 309–10 (1997) (explaining that one must look to both ideological and legal perspectives in order to adequately understand judicial decision making, and that the operative facts of a case help to determine the direction of both political preferences and legal

legal rule, the choice or rule arises based on a factual foundation and might change if the facts were different.[267] So when a mistaken factual understanding leads to a different outcome than would have occurred were the true picture known to the decision makers, we can label that outcome inaccurate even if it does not come with an explicit explanation.

There is a second sense in which factual accuracy matters even for decisions that are not explicitly or solely factual. Take settlements as an example. We might care how well lawyers on average understand the facts before they settle, but we might also feel that even if some information is hidden from both sides, the process of negotiation might aggregate both sides' incomplete information into a single sum that in some sense incorporates more information than either party possesses. We must be careful here: there is no meaningful sense in which we can say that there is an amount that a case "should" settle for based solely on the historical facts that gave rise to it, so settlement amounts by themselves cannot be said to be accurate or inaccurate. Nevertheless, we might care how well settlement values correlate with certain important underlying case facts, such as the extent of injuries suffered by the plaintiff or the degree to which a physician acted contrary to ordinary treatment protocols. Although the settlement amounts cannot be said to be trying to "reconstruct" such data, they do arise from a process that we intend to respond to variances in those data. So if settlement amounts are failing to reflect variations in the underlying facts in the way we would desire, that information is relevant to the design of the overall system.

An example may make this point more concrete. Several studies, most published in medical journals, have sought to assess how well the malpractice litigation system functions as a means of separating negligent medical errors from other causes of patient morbidity.[268] One of the most recent and most thorough investigations was conducted by David Studdert and his co-authors in 2006.[269] This group of researchers analyzed a large sample of closed malpractice claims, using independent physicians to analyze the insurers' case files in order to determine, to the best of their ability, which cases involved injured patients and how many of those

---

analysis).

[267] *See id.* at 310 (noting that "political decision making takes account of facts as well").

[268] *See, e.g.*, Troyen A. Brennan et al., *Relation Between Negligent Adverse Events and the Outcomes of Medical-Malpractice Litigation*, 335 NEW ENG. J. MED. 1963 (1996); Frederick W. Cheney et al., *Standard of Care and Anesthesia Liability*, 261 JAMA 1599 (1989); Henry S. Farber & Michelle J. White, *A Comparison of Formal and Informal Dispute Resolution in Medical Malpractice*, 23 J. LEGAL STUD. 777 (1994); David M. Studdert et al., *Claims, Errors, and Compensation Payments in Medical Malpractice Litigation*, 354 NEW ENG. J. MED. 2024 (2006).

[269] Studdert et al., *supra* note 268, at 2024.

injuries were due to "medical errors."[270] They then compared these judgments with the cases' outcomes, in an attempt to see how often the tort system produced results similar to what the experts would have viewed as appropriate.[271] This allowed them to produce some interesting comparisons: 73% of the claims they studied appeared to be either true positives, in which compensation was provided to a claimant who had been injured by medical errors, or true negatives, in which compensation was not provided to a claimant who was either uninjured or whose injuries were not attributable to medical mistakes.[272] Of the remaining quarter of claims, just over 10% involved false positives, in the form of payment given to those who had not suffered an injury due to error, while a slightly larger fraction of 16% involved false negatives, in the form of payments denied to those who seemed deserving.[273]

Without belaboring the pluses and minuses of this particular investigation,[274] the approach employed by it and other similar malpractice studies is quite instructive. By comparing the underlying facts of litigation with its results, the authors are able to obtain insight on a question that is rarely explored. That is, we can get some sense of how successful the tort system is at figuring out what happened in the past. And although this study (and the others like it that I have been able to find) attempted only to provide a descriptive picture of how the current litigation system was functioning, one could combine this novel approach to measuring the validity of litigation outcomes with an experimental variation of procedural rules to gain a truly powerful tool for separating useful procedural innovations from harmful ones. So, when we would prefer that certain facts correlate substantially with litigation results, one way of assessing how accurate the legal system is at compensating the right claims is to measure

---

[270] *Id.*

[271] *Id.* at 2026–29.

[272] *Id.* at 2027–28.

[273] *Id.* at 2028. *But cf.* Cheney et al., *supra* note 268, at 1601 (finding a similar rate of correctly sorted claims, but with a much higher proportion of false positive awards, in a sample of 1,004 closed anesthesia malpractice claims).

[274] As the authors acknowledge, the reviewing physicians were not blind to the results of litigation, which increases the risk of biased coding. Studdert et al., *supra* note 268, at 2032. A more difficult concern relates to the decision to have only physicians do the coding, and to have them identify "medical errors" in the records. The first problem is that, as the authors acknowledge, satisfying the tort standard of care is not necessarily equivalent to following conventional medical protocol. *See id.* Perhaps more troublingly, having physicians identify which injuries were caused by medical mistakes creates a potential bias because doctors may be more likely to give other members of their profession the benefit of the doubt than would outsiders who were fully informed.

the relationship between the incidences of those facts with typical case results.

In the end, it seems that most policy makers and critics who wish to optimize procedural efficacy should care both about how well legal decision makers understand the cases they decide, and how well their decisions correlate with the objective facts that the law purports to respond to. So most of those who seek to maximize accuracy (along with other values) in a juridical system should be interested in finding ways to measure accuracy in a more systemic and sustained way.

There is, however, one procedural value that may cut strongly in the other direction. Those who privilege the system's legitimacy over all competing values may be hesitant to look too closely at the details of its performance, for fear of undermining the public's faith in our courts.[275] Many are properly concerned that legal procedures be viewed as legitimate both by those who use them as well as by the broader public who observe proceedings from a distance.[276] There are several reasons for this concern. For one thing, people more readily comply with legal rules concerning primary conduct when they view legal institutions in a positive light.[277] And for some, it may be more important for disputes "to be settled than [for them] to be settled right,"[278] if the alternative is viewed as private self-help and a gradual descent into social chaos. Indeed, if we are frank with ourselves, we must admit that exposing the large amount of ignorance we possess about the utility of our legal procedures might be a poor public relations move for the litigation system.

For some, this argument may be so powerful as to make any proposal for increased outcome measurement a non-starter, but I suspect that most will have a different intuition upon reflection. Caring so much about legitimacy that we ignore accuracy considerations should leave a very bad taste in our mouths.[279] There would be a strong element of deceit at play if policy makers and critics know that the justice system is not being careful to ensure it is getting it right as often as possible but avert their eyes from

---

[275] *See, e.g.*, Tanner v. United States, 483 U.S. 107, 120 (1987) (worrying that the jury system might not survive "efforts to perfect it" that involved post-hoc public scrutiny of the deliberation process).

[276] *See* Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1358–59 (1985).

[277] *See* TOM R. TYLER, WHY PEOPLE OBEY THE LAW 45–50 (2006).

[278] Burnet v. Coronado Oil & Gas Co., 285 U.S. 393, 406–07, 410 (1932) (Brandeis, J., dissenting).

[279] *See* DAVID M. ESTLUND, DEMOCRATIC AUTHORITY: A PHILOSOPHICAL FRAMEWORK 95 (2008) (urging that procedures that treat participants "equally" would nevertheless fail to treat them with appropriate respect if they ignored the relative epistemic merit of their contributions).

the problem for fear of the public's reaction.[280] It was one thing for medical science to treat without measuring when few realized that measurement could help; it would be another for modern medicine to eschew a practice that has such broad benefits for public health. Perhaps one guidepost that might caution us to avoid such a course is this: if we fully informed the public we were trying to protect of the reasons why we might wish to avoid putting our adjudicative procedures to the test, there are a number of reasons they might accept (such as a concern for their privacy or an outlandish cost). It would be very surprising, however, if they would accept the strong paternalism present in the decision to subject them to the whims of a legal system that we could have, but have not, tested for accuracy and fairness.

C. The Risks of Trusting an Untested Procedural System

Assuming that we agree that accuracy is important and that we could measure it if we were willing to try, is there any further reason to object to a call for investigations along these lines? Perhaps there is. Many readers may feel that this is all much ado about nothing; even if we do not track accuracy systematically, they might say, our procedures are so well-designed that there is little to worry about.[281] Note that this can be true even if we grant the point made in a previous section that we are bad at predicting how rules will affect the overall system.[282] Perhaps the social pressures that constrain our legal system gradually nudge it towards good equilibria (including high rates of accuracy) by aggregating the decisions of many different decision makers even though none of those individuals understand the whole.[283] Our failure to measure accuracy means we cannot use data to support such a notion, but if we had strong intuitions that the

---

[280] *See* Robert J. MacCoun, *Voice, Control, and Belonging: The Double-Edged Sword of Procedural Fairness*, 1 ANN. REV. L. SOC. SCI. 171, 193 (2005); Austin Sarat, *Authority, Anxiety, and Procedural Justice: Moving from Scientific Detachment to Critical Engagement*, 27 LAW & SOC'Y REV. 647, 656–59 (1993).

[281] *See, e.g.*, Herrera v. Collins, 506 U.S. 390, 420 (1993) (O'Connor, J., concurring) ("Our society has a high degree of confidence in its criminal trials, in no small part because the Constitution offers unparalleled protections against convicting the innocent."); United States v. Garsson, 291 F. 646, 649 (S.D.N.Y. 1923) (Hand, J.) (calling the idea of wrongful convictions in our legal system an "unreal dream").

[282] *See* discussion *supra* Parts II, III.

[283] *See, e.g.*, Paul H. Rubin, *Why Is the Common Law Efficient?*, 6 J. LEGAL STUD. 51, 61 (1977). *But see* Allen, *supra* note 3, at 16–18 (describing our legal system as a "system of interconnected nodes that looks somewhat like a neural network," but doubting that this network operates to optimize any consistent set of variables about legal rules or litigation results).

legal system was self-optimizing, that might be enough to make the costs of the necessary data gathering too much to bear.

There are two possible responses to such nonchalance. The first is to note that even if our system is doing pretty well, we might still reasonably wish to either improve its accuracy further, or else attempt to maintain its accuracy while decreasing the cost or delay associated with obtaining resolution of legal claims. Even among those who have the intuition that the American litigation system functions adequately, I doubt there are many who feel that it is beyond any possible improvement. But so long as we are unable to measure the accuracy of changes in procedure, we cannot tell when we are cutting costs at no detriment to justice and when we are saving money by sacrificing quality. In other words, if we must balance multiple factors in designing procedure, we cannot truly say that any change is an improvement (or even a neutral trade-off) unless we are measuring each of those factors. And given the extremely complex nature of the system itself, and the sometimes counterintuitive results of interventions,[284] we should be very reluctant to trust mere intuition that any given cost-cutting measure comes at no cost to our system's accuracy.

But we can, and should, go further in responding to those who see no need to measure how often we get it right or wrong: concluding too readily that our confidence in our system is predictive of its accuracy stands in stark contrast to the historical record of legal fact-finding. History shows that people have often placed their faith and trust in very inaccurate methods of finding facts.[285] And in the present day, the recent spate of DNA-based exonerations[286] serves to underline the point: so long as we fail to measure the quality of our system's outputs, we have little reason to trust that it is functioning in an optimal way.

Start with the historical record: the past provides numerous examples of people who, in good faith, relied on dispute resolution systems that we would find impossible to trust. For instance, consider the trial by ordeal. Like us, the residents of early medieval England had disputes and a need to resolve them, but when faced with two disputing parties they did not engage in a process of sifting through evidence to reconstruct historical facts to decide who should prevail.[287] Rather, they relied on God to decide whose version of events was true and whose was false, and they forced God to

---

[284] *See* discussion *supra* Part II.

[285] *See* discussion *supra* Part II.A.

[286] *See* GARRETT, *supra* note 16, at 1–13 (describing some of the common features of the first 250 DNA exonerations).

[287] LEONARD W. LEVY, THE PALLADIUM OF JUSTICE: ORIGINS OF TRIAL BY JURY 5 (1999).

show his hand by means of such procedures as the ordeal.[288] In the ordeal, one party to a dispute would be put to a physical test in which "he called upon God to witness his innocence by putting a miraculous sign upon his body," such as healing a burn wound without sepsis, causing him to sink rather than float upon water, or swallow a large morsel of dry food without choking.[289] Anyone who thinks that common sense alone is a reasonable guide towards designing fair procedures must face the difficult fact that common sense beliefs may be systematically wrong,[290] and the ordeal example shows that such widespread error is not only possible, but that it can dominate a legal system for centuries without widespread criticism.[291]

Some readers may balk at the first example because it involved explicitly religious assumptions about proof. If so, consider a practice that lasted for centuries and that was still in widespread use during the first century of American life: the bar on testimony by interested witnesses.[292] This rule totally precluded parties and other interested witnesses from giving testimony under oath.[293] From the 1500s until the mid-nineteenth century, the firm policy of the judiciary was that a long list of potential witnesses—including the parties in a civil case, criminal defendants, anyone with a financial interest in the case's outcome, anyone previously convicted of a felony, and atheists—were so likely to lie on the stand that it was better to dismiss cases entirely for lack of evidence than hear what they had to say.[294] In many cases, the result was that one party had no evidence to offer

---

[288] *Id.* at 5–6; S.F.C. MILSOM, HISTORICAL FOUNDATIONS OF THE COMMON LAW 359 (1st ed. 1969).

[289] LEVY, *supra* note 287, at 5.

[290] Nor is it only in the distant past that common sense can be so misleading. For a present-day example, consider the widespread assumption that observing witnesses testify increases the likelihood of accurate deception detection, which is probably false. *See* Mark Spottswood, *Live Hearings and Paper Trials*, 38 FLA. ST. U. L. REV. 827, 835–51 (2011).

[291] *See* LEVY, *supra* note 287, at 5. A parallel, and equally disturbing, example can be found in continental practice. In the Roman canon law tradition, tortured confessions were frequently the basis for criminal convictions. *See* Stephan Landsman, *The Rise of the Contentious Spirit: Adversary Procedure in Eighteenth Century England*, 75 CORNELL L. REV. 497, 594 (1990). This practice evolved as a work-around for the strict formal rule that had required at least two witnesses to testify against an accused before any conviction could be obtained. *Id.* The reliability of a confession obtained under torture is, of course, highly suspect.

[292] James Oldham, *Truth-Telling in the Eighteenth-Century English Courtroom*, 12 LAW & HIST. REV. 95, 107–09 (1994).

[293] *Id.*

[294] *Id.* at 102, 107–09; *see also* 3 WILLIAM BLACKSTONE, COMMENTARIES *363–64, *370. Blackstone lists "treason, felony, perjury, . . . conspiracy," "outlaw[ry]," "excommunicat[ion]," "attaint of false verdict," and "forgery" among the reasons that a person would have been too infamous to serve as a juror, and then notes that such persons were equally disallowed to testify as witnesses. *Id.* Perversely, the victim in a criminal case, who also acted as prosecutor, did not fall within the ambit of this rule. Oldham, *supra* note 292, at 107; *cf.* JEREMY BENTHAM, 5 RATIONALE OF JUDICIAL EVIDENCE

at all to press a claim or defend against it, even if that party would have resisted the temptation to testify deceptively. As a result, many valid claims were likely deterred.[295] Once again, we see that a great multitude of judges and lawyers placed their trust in procedures that most today would view as deeply inaccurate and unfair.

Finally, for those who think we have surely risen above the ignorance and errors of our past, consider a recent example: the string of DNA-evidence exonerations of wrongfully convicted felons.[296] On one level, the introduction of DNA evidence might seem like a basis for increased trust in our legal system. Indeed, the strength of a properly conducted DNA match provides our modern "gold standard" for placing a criminal defendant at the scene of a crime. But the very accuracy of this device has enabled it to expose ugly flaws in our system of justice: numerous cases that seemed quite ordinary at the time they were decided turned out to have placed innocent people in prison.[297] The distressing fact that the DNA exoneration cases illuminate is that most wrongfully decided cases look almost exactly like accurately decided cases. The defect in such cases does not lie in the facts known to actors in the legal system, but rather in the facts they failed to discover or bring to light.[298] Nor can we comfort ourselves much with the fact that the exonerations are relatively few when placed against the vast number of criminal convictions. Unfortunately, only a small proportion of decided cases involve untested, but still existing, DNA evidence that can implicate or exonerate a convicted defendant.[299] Because of this, we cannot form even a crude estimate of the overall accuracy of our criminal systems by counting these exonerations. But when placed alongside the historical record of trust in highly dubious procedural devices, the number of DNA-based exonerations of run-of-the-mill convictions should make us very hesitant to place blind trust in the accuracy of our system.

---

56 (1827) (noting the absurd extent to which these interest-based incompetency rules were enforced, based upon the apparent assumption that there was "no Englishman" who "would not perjure himself" for a farthing).

[295] *See generally* John Fabian Witt, *Making the Fifth: The Constitutionalization of American Self-Incrimination Doctrine, 1791-1903*, 77 TEX. L. REV. 825, 860–66 (1999) (tracing some of the many complaints that led to the doctrine's belated decline).

[296] *See generally* GARRETT, *supra* note 16, at 1–13.

[297] *See id.* (describing some 250 DNA-based exonerations).

[298] *See* Rosen, *supra* note 16, at 73 (noting that "[u]ntil the moment when the DNA test results came back, almost none of these cases [in which DNA evidence revealed that a conviction was in error] would have been considered exceptional among criminal cases").

[299] *See* GARRETT, *supra* note 16, at 12 ("If DNA is a 'truth machine,' it tells us only about a sliver of very serious convictions, most for rape, chiefly from the 1980s.").

V. ADDRESSING THE PROBLEM: PROCEDURAL EXPERIMENTS THAT MEASURE ACCURACY

In the preceding sections, we have seen that systematic measurement has revolutionized the practice of medicine, although there is still a long way to go before the evidence-based medicine movement can be said to have prevailed fully.[300] Rule makers face great challenges when they design or amend rules of procedure and evidence because the litigation environment is a complex system in which it is often difficult to link procedural cause with substantive effect.[301] For this reason, an evidence-based litigation movement might be very attractive, but reformers will not be able to meaningfully assess the impact of differing rules or rule applications on the accuracy of case outcomes unless they are willing to go beyond measuring who wins because simple outcomes make a poor proxy for the real variables of procedural interest. In particular, if studies fail to measure accuracy and focus instead on other variables like litigant satisfaction or cost, we will have little basis for confidence that improvements in these other variables are not coming at the cost of the system's ability to accurately divine the factual merits of litigants' cases. Until this problem is addressed, the value of both existing rules and proposed reforms to them must always remain uncertain for any person who thinks that accurate outcomes matter. In this section, I will discuss the challenges inherent in measuring the accuracy of a procedural system and sketch one possible means for accomplishing this difficult task.

The evidence-based medicine movement can provide us some inspiration as we try to develop methods of determining the impact of rules on the accuracy of legal outcomes. One important lesson we can learn from doctors is the value of attending both to easily measurable surrogate outcomes and also more subjective ultimate outcomes of deeper theoretic interest, depending on the specific goals of an investigation.[302] In medicine, doctors sometimes track immediate biophysical signs, such as biopsy results or cholesterol levels, either as indicators of broader patient health or as interesting in their own right.[303] Depending on the question of interest, procedural investigators could similarly compare legal outcomes with individual details of underlying cases. As discussed above, a number of

---

[300] *See* discussion *supra* Part II.

[301] *See generally* Allen, *supra* note 3.

[302] *See* Deyo, *supra* note 69, at 70–71.

[303] *See* STRAUS ET AL., *supra* note 122, at 72–73.

investigations into the accuracy of medical malpractice litigation outcomes have proceeded in this way, comparing the outcomes of claims with expert determinations of whether medical errors occurred.[304] These studies have given unique insight into this realm of litigation, enabling us to get rough estimates as to how often settlements are given to both deserving and undeserving claimants, as well as how often culpable doctors avoid legal judgments against them.[305]

One lesson from the evidence-based medicine movement, however, is that such intermediate measurement can lead to systematic errors, given that factors like blood pressure are imperfect predictors of overall health.[306] The modern trend in evidence-based medicine is towards looking at more ultimate indicators of health, like changes in overall mortality or morbidity on longer time scales, rather than immediate biophysical signs.[307]

Unfortunately, the factual accuracy of case outcomes is harder to measure than whether a patient lives or dies. It may be valuable to know how brute facts like who prevails or how much they win correlate with various underlying case facts, but such correlations will ultimately allow only a blurry estimate of the deeper idea of adjudicative accuracy. As discussed above,[308] one of the most critical quantities of interest for procedural design would be a measure of the correspondence between the factual understanding that motivated the legal result in a case and the actual set of historical facts that gave rise to the litigation. Such a comparison would isolate the component of the decision that is primarily about the facts that are being disputed from the legal doctrines, personal preferences, and moral judgments that help decision makers translate facts into outcomes. And while qualitative descriptions of similarities and differences would be interesting, more useful for the purposes of measuring the impact of possible rules on overall accuracy would be a single accuracy score that can be compared between many different types of case outcomes and subject matters under dispute. Thus, the product of accuracy measurement, in an ideal world, would be a single value that incorporates both the number of factual disagreements between the motivating and actual facts, as well as

---

[304] *See, e.g.*, Brennan et al., *supra* note 268, at 1963–64; Cheney et al., *supra* note 268, at 1599; Farber & White, *supra* note 268, at 795; Studdert et al., *supra* note 268, at 2024.

[305] *See, e.g.*, Studdert et al, *supra* note 268, at 2028 (finding that, in about a quarter of malpractice cases, the availability of a settlement did not correlate with whether medical error had occurred).

[306] *See* discussion *supra* Part II.D.

[307] *See* Deyo, *supra* note 69, at 70–71.

[308] *See* discussion *supra* Part III.

some indication of the relative importance of those facts to the decision maker.[309]

In the search for an objective method for evaluating procedural success, we are quickly confronted with the reality that the very concept of accuracy depends on both the subjective motivations of a legal decision maker and the qualitative judgment of a reviewer or group of reviewers. Despite what some readers may think, however, this is not fatal to the overall project of measuring adjudicative accuracy. Measuring overall patient health or pain levels involves a great deal of subjectivity, but tracking such data is extremely important if you wish to make real people healthier rather than just optimize various test results.[310] And in some cases, medical measurement properly incorporates even greater amounts of subjectivity. Rates of psychiatric disease, for example, are grounded in individual clinical comparisons of a patient's mental and social functioning with a range of behavior that is defined to be aberrant.[311] So long as the target of measurement is a quantity of important theoretic interest and the measurement process has an acceptable rate of reliability, subjectivity in coding measurements need not be fatal to the goals of accuracy tracking.[312]

A second challenge looms: Where do we get data concerning the motivating factual understanding and the historic factual reality? For some litigation outcomes, the motivating factual understanding may be fairly transparent. Judicial decisions, for instance, will often be supported by a written elaboration of both the facts as understood by the court and the reasons for the overall decision.[313] So long as there are sufficiently strong norms of sincerity and candor in the articulation of these facts and reasons,[314] such writings might provide an acceptable source of data as to the motivating factual understanding underlying a given outcome. But many other methods of resolving cases leave no paper trail. Settlements are often kept secret and are usually unaccompanied by an explanation of the

---

[309] In such a system, a "10 out of 10" might indicate that the original decision maker and the evaluator agreed on all of the facts that were important to the result, while a "0 out of 10" would indicate total disagreement. Most results, of course, would be intermediate values that represent partial, rather than total, agreement.

[310] *See* Joanmarie Ilaria Davoli, *Still Stuck in the Cuckoo's Nest: Why Do Courts Continue to Rely on Antiquated Mental Illness Research?*, 69 TENN. L. REV. 987, 1028–31 (2002).

[311] *See id.* (describing the behavioral qualities that modern-day psychiatrists use to delimit mental illness).

[312] *See* Deyo, *supra* note 69, at 66.

[313] *See, e.g.*, FED R. CIV. P. 52(a).

[314] *See generally* Micah Schwartzman, *Judicial Sincerity*, 94 VA. L. REV. 987 (2008) (defining and discussing the norms of judicial sincerity and candor).

lawyers' view of the case, in part due to practice conventions and in part due to professional confidentiality obligations.[315] Jury verdicts (in the absence of special verdict forms) are shrouded in mystery.[316] And sometimes, decisions may be the product of negotiation among parties with differing factual understandings, so that no one mind can be said to possess the full set of factual understandings that "motivate" the overall result.

These obstacles make it harder to implement systematic measurement of case outcome accuracy, but they do not make it impossible. Parties could be asked to voluntarily permit their lawyers to describe and disclose their view of cases as part of research programs, and jurors could similarly be asked to articulate their views of the case after a trial is concluded. If voluntary rates of participation were high enough, acceptably valid and reliable data might be collected. But herein lurks a problem; so long as disclosure is voluntary, individual research efforts might always be frustrated by holdouts who refuse to disclose relevant information, and we will never be able to know with confidence whether the data of those who voluntarily disclose is representative of the set of people who refuse. A more promising approach—although a harder one to realize—would be to amend existing procedures and professional obligations to require such disclosures as part of the duties of lawyers and jurors, while creating a corresponding duty of confidentiality on the part of any researchers who gain access to such information.

So the motivational-facts component of the accuracy measure could be systematically acquired at modest cost, provided that we are either able to establish a norm of widespread voluntary explanations by decision makers or else establish mandatory duties to the same effect. What, then, of the other half of the puzzle, the historical facts? Here we have a harder challenge, as we lack an existing mechanism, external to the legal process, capable of providing a gold standard picture of what events truly occurred that gave rise to the lawsuit.

---

[315] *See* Scott A. Moss, *Illuminating Secrecy: A New Economic Analysis of Confidential Settlements*, 105 MICH. L. REV. 867, 869–70 (2007) (noting that most settlements are confidential and that courts almost never order disclosure of confidential settlement terms). The obstacle of lawyer confidentiality obligations has an interesting parallel in the history of medicine. One of the great obstacles to the introduction of scientific methods to medical practice was the longstanding opposition of European religious organizations to the dissection of cadavers. The relaxation of this norm was a crucial ingredient in early attempts to put traditional medical theories to the test. *See* discussion *supra* Part I.

[316] *See* 9B CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE & PROCEDURE § 2505 (3d ed. 2011) (noting that the use of special verdict forms has never been widespread in federal civil cases, although it may be on the rise); Anne Bowen Poulin, *The Jury: The Criminal Justice System's Different Voice*, 62 U. CIN. L. REV. 1377, 1420 (1994) (noting the rarity of special verdict forms in criminal cases).

The malpractice-litigation studies I discussed above may provide us with the kernel of an accuracy measurement design that could be implemented more broadly.[317] Researchers in several of those studies retained physicians to review insurance files and medical records in order to determine whether they indicated adverse events, whether those adverse events were caused by medical negligence, and also the degree to which the patient suffered a disability.[318] This approach has some notable strengths: It employs subject matter experts with some expertise in the relevant domain, and it shields them from knowledge regarding how the legal system had resolved the claims, lessening the risk that they will be biased toward confirming its outcomes. And it guides the reference-standard evaluators by focusing their attention on variables of particular legal interest. Unfortunately, with the doctors' expertise may come a different form of bias: because they identify with other members of their profession, doctors may be less likely to attribute injuries to medical causes or find medical behavior to be negligent.[319] Also, having them code for a legal conclusion (negligence) meant that the data they produced combines normative impulses with factual assessment in a way that obstructs an attempt to focus on the factual validity of the results. That is, it is hard to say, on the basis of these studies, whether the "errors" involve misunderstandings of medical information or disagreements about what sort of errors deserve compensation. Finally, a focus on medical records or insurance files as the sole source of independent data may itself bias the results, especially if doctors, nurses, and insurers are reluctant to record some types of bad behavior. By addressing these deficiencies, we might approach a form of assessment we could justifiably claim as a gold standard for purposes of measuring the impact of procedures on juridical accuracy.

First, if the reference-standard evaluators are aware of the outcomes of the cases they are reviewing, then there is a risk that they will either be biased toward confirming those results or towards labeling them as erroneous.[320] For this reason, a gold standard reference evaluation should have full access to case files but be as blind as possible to case outcomes.

Second, we should try to provide our evaluators with as much independent data as possible so that they do not merely replicate the

---

[317] *See* discussion *supra* Part IV.B.

[318] *See, e.g.*, Brennan et al., *supra* note 268, at 1964.

[319] *See* Studdert et al., *supra* note 268, at 2031–32.

[320] *See* STRAUS ET AL., *supra* note 122, at 72 (noting that some seemingly "hard" measurement standards involve a great deal of discretion, and that blinding the reference standard assessment is a good means of avoiding bias); Studdert et al., *supra* note 268, at 2032.

mistakes of the system they are studying. Crucially, this means that case records alone may be an insufficient source of information, especially for cases resolved early on with little discovery or investigation. Ideally, evaluators should know all that the parties and lawyers know. In order to approach this ideal, evaluators should be given access to full accounts of the facts underlying the dispute from all the parties, lawyers, and witnesses, with issues of confidentiality and privilege waived for the purposes of investigation.

Clearly, for such a system to be effective, parties would have to have sufficient assurance that their unvarnished and candid accounts would not subsequently be used against them in court. A critical part of any such assessment process is that the information obtained is kept fully private and is never used to alter the results reached by the legal process. As a result, we must keep secret the errors detected by such methods on an individual basis so that we can reliably report the accuracy of the system as a whole.

There is also a difficult tension present between making evaluations blind to actual outcomes and making them fully informed. The more deeply evaluators probe the facts underlying a dispute, the greater the risk that a party or other individual involved in a case will deliberately or inadvertently reveal information regarding how the case actually turned out. Hopefully, such concerns could be minimized if all involved understood both the value of keeping such information from being disclosed, and also that the evaluators' judgment could have no effect on the outcome that had actually been achieved.

Third, the best means of measuring accuracy will be one that separates out those components of decision making that are factual in nature from those that are discretionary or normative.[321] Once our evaluators have combed through the statements of parties, witnesses, lawyers, and relevant documents, they can then prepare their own account of what most likely

---

[321] The conflation of normative and factual accuracy is a problem that is endemic to much of the studies that have actually attempted to measure accuracy. Of course, it may also be useful to compare a decision maker's understanding of relevant law (or even relevant norms) with some sort of reference standard on these questions. Indeed, many legal questions may have answers clear enough that it may be reasonable to refer to some outcomes as legally "accurate" and others as "inaccurate." *See* Ronald J. Allen & Michael S. Pardo, *The Myth of the Law-Fact Distinction*, 97 NW. U. L. REV. 1769, 1790–97 (2003) (explaining that questions of law are, at bottom, questions of fact and so may have right, wrong, and indeterminate answers to the same extent that other factual questions do); Lawrence B. Solum, *On the Indeterminacy Crisis: Critiquing Critical Dogma*, 54 U. CHI. L. REV. 462, 472 (1987) (arguing that there are many legal questions for which some answers are clearly correct and others are clearly incorrect). Some studies along these lines might be useful; for instance, it might be instructive to see how well attorneys engaging in early settlements understand the law that would govern their cases.

occurred in the past giving rise to the dispute. Because this process is aimed at assessing the legal system's performance—rather than producing binary outcomes between innocence and guilt or liability and non-liability—evaluators should not express certainty where none exists. Rather, an appropriate procedure would be for the evaluator to create a narrative in which she indicated both those facts that were clear and those that were subject to significant uncertainty. Once the narrative statement was complete, the evaluator could be given access to the actual decision in the case (if that decision involves enough factual components to be analyzable for accuracy) or to an account prepared by a decision maker of his reasons for reaching a particular result. The evaluator could then score that decision on a scale that allowed her to distinguish between those resolutions that were strongly supported by the underlying facts, those that were made on an ambiguous record, and those that ran strongly against the most likely version of past events. Alternatively, if the research question involves comparing some non-factual aspect of a decision (like a damages amount) with the underlying facts of cases, the evaluator's account can be used as a source for identifying which cases involve the relevant facts and which do not.

Establishing such a system would require thoughtful choices about how we could best ensure that evaluators work hard to evaluate cases closely and avoid bias to the extent that is humanly achievable. One important issue is deciding whom to use as evaluators. There are many plausible options, such as using lawyers retained on a special-master model, subject-matter experts (as in many of the malpractice litigation studies), retired judges, or even intelligent non-lawyers in something approximating a jury model.

Whoever is chosen, the research design would need to incorporate periodic checking of the evaluators' reliability. There are several tests that could be conducted to maintain confidence in our evaluators. Those in charge could assign multiple coders to evaluate the same cases, in order to test measurement reliability (or how likely it is that the measurement system would code the same case the same way in repeated encounters).[322] The researchers could also personally review a sampling of their results and the supporting evidence to get some sense of the validity of the measurements by seeing how closely the evaluative summaries track the supporting materials. And as a final validity check, test cases for which

---

[322] *See, e.g.*, James C. Phillips & Edward L. Carter, *Gender and U.S. Supreme Court Oral Argument on the Roberts Court: An Empirical Examination*, 41 RUTGERS L.J. 613, 625–26 (2010) (employing this common procedure in order to ensure adequate reliability in the measuring of subjective variables).

historical facts are reliably known could be inserted into the measurement stream from time to time to give a further basis for coding validity analysis. Indeed, a comparison of how reliable and valid the work of different evaluators is would be useful not just in monitoring their performance and incentivizing careful work, but also in making long-run decisions about who to turn to and how much to pay for the work. It may turn out that educated laypeople do as well or better than lawyers, in which case we might find that a great deal of evaluation can be obtained at relatively acceptable cost. It might also turn out that retired judges or subject matter experts are so superior that review by anyone else cannot plausibly be considered a gold standard against which to test litigation results. Absent experience with such systems, further speculation as to the best choice for reference-standard evaluators would be of little use.

Some may wish to object to any reference-standard evaluators we might select on the grounds that, regardless of the individual merits of those evaluators, they will operate at a disadvantage to viewers of a live oral trial, with its special rituals of cross-examination and its opportunity to observe witness demeanor while testifying as an aid to making credibility calls. This concern, however, would be weaker than it initially appears. First, most civil or criminal cases are not resolved through a trial on the merits, but rather through a settlement or a pre-trial procedural dismissal.[323] Thus, even if we worry that an inquisitorial-style reference evaluation lacks some reliability features associated with trials, it could still be made more accurate than a modal case resolution, given its greater access to case information and its lack of distortions due to cost constraints or disparities in litigation resources between parties. Second, contrary to received wisdom, live trials may not offer systematic accuracy advantages over paper-based fact-finding. Although live trials offer some benefits in terms of clarifying and simplifying complex case information, they also come at a cost: they may decrease the accuracy of detecting insincere or mistaken testimony, and they may also provide extra sources of appearance-driven bias that skew case evaluations.[324] Thus, there are good reasons to think that a well-funded and well-motivated investigation by an intelligent individual with complete access to both the facts in a case record and to concealed confidential information would usually meet, and sometimes exceed, the typical accuracy of the existing litigation system.

---

[323] *See* Spottswood, *supra* note 290, at 828.

[324] *See id.* at 835–51.

If we were willing to take the steps needed to establish such a system, there is a good chance that it would substantially increase our ability to acquire useful answers to a wide variety of procedural questions. For instance, we could gain useful insight into long-running debates on the best method of judicial selection by comparing the overall accuracy rates achieved by competing methods, controlling as far as we can for other confounding differences.[325] Alternatively, we could compare different areas of the law or different procedural regimes and see which do a better job at understanding the facts of cases. Likewise, we could compare the accuracy of resolutions at different stages in the lives of cases, and learn if pretrial decisions broadly correspond with the results in factually similar cases that go to trial, or if settlement amounts vary appropriately depending on the factual strength of the underlying case. Finally, we could use such measurement as a basis for truly evidence-based design of judicial procedures by conducting either randomized experiments or well-designed observational studies that assess the impacts of different procedural rule alternatives on the overall accuracy of the factual understandings that produce case outcomes, combined with assessments of the perceived fairness, cost, and time-to-decision of those procedures.[326] With such data in hand, we could make informed procedural policy decisions when we try and balance among these variables rather than relying on our (often wrong) intuitions about the ways that different rule regimes will play out in practice.

Consider a concrete example already discussed at some length above: the right-to-counsel debate.[327] Armed with tools described above, we could answer the question that is most important in these debates, but which even the most sophisticated, randomized experiments have been unable to probe. Imagine that we have identified an area in which appointed counsel does seem to improve their client's chances of prevailing, perhaps by means of a randomized experiment. This might be good or bad news: perhaps the lawyers are helping people with bad claims or defenses confuse judges and win what they are not entitled to through perjury and fancy lawyering.[328]

---

[325] *Cf.* CHRIS W. BONNEAU & MELINDA GANN HALL, IN DEFENSE OF JUDICIAL ELECTIONS 128–39 (2009); PHILIP L. DUBOIS, FROM BALLOT TO BENCH: JUDICIAL ELECTIONS AND THE QUEST FOR ACCOUNTABILITY 27–28 (1980); Mary L. Volcansek, *Judicial Elections and American Exceptionalism: A Comparative Perspective*, 60 DEPAUL L. REV. 805, 817–19 (2011).

[326] *See* Posner, *supra* note 144, at 374–77; Walker, *supra* note 144, at 67–68.

[327] *See* discussion *supra* Part III.

[328] *Cf.* Albert W. Alschuler, *How to Win the Trial of the Century: The Ethics of Lord Brougham and the O.J. Simpson Defense Team*, 29 MCGEORGE L. REV. 291, 299–317 (1998) (describing the tactics

Conversely, perhaps appointed lawyers are helping people with good claims but low advocacy skills obtain results that they deserve.[329] An assessment of the validity of the factual understandings of the relevant decision makers could provide a useful partial answer, telling us whether the lawyers were improving success by increasing rates of confusion or by better educating judges. If a random experiment was combined with a measure of the degree to which decision makers' justifications for their decisions corresponded with the results of gold standard independent evaluations, we might be able to get a better grasp on whether, in a given type of case, appointing counsel provides a general social benefit or merely a narrow parochial benefit for their clients.

Similarly, such an approach could shed valuable light on the *Twombly*/*Iqbal* debate. Recall that the question of true theoretical interest is whether the heightened pleading standard propounded by the Supreme Court successfully weeds out frivolous claims before discovery while allowing most non-frivolous claims to proceed.[330] An accuracy-measurement design could be useful, in the first instance, by providing descriptive statistics about how well judges do at understanding the underlying facts of cases at the motion-to-dismiss phase of a case. This alone could be valuable information, but we could learn much more if we conducted a controlled experimental trial comparing different pleading-review approaches. By randomly assigning some cases to a *Twombly*/*Iqbal* standard and some to the pre-existing regime, tracking them to completion, and comparing the overall accuracy of all results, we could learn whether the new standard raises or lowers the accuracy of the outcomes in cases it applies to. This might not tell us all we would wish to know about these cases; even if the new regime tends to be more accurate on average, it might be objectionable if, for instance, it tended to deter some claims by meritorious plaintiffs from being filed.[331] But knowing whether or not the

---

that enabled the O.J. Simpson defense team to obtain an acquittal).

[329] *See* Powell v. Alabama, 287 U.S. 45, 69 (1932).

[330] *See* discussion *supra* Part III.B.

[331] It might be possible to detect such an effect by tracking incidents of injury and seeing how many result in a filed complaint under either set of rules. *See, e.g.*, A. Russell Localio et al., *Relation Between Malpractice Claims and Adverse Events Due to Negligence*, 325 NEW ENG. J. MED. 245, 248 (1991) (finding that out of 280 incidents involving medical negligence—as defined by a medical review panel—only 8 patients filed malpractice claims). But conducting a true experiment along these lines would most likely involve difficult trade-offs. On the one hand, there is a need to implement experimental and control procedures on a geographically widespread and long-term basis in order to allow enough information about the consequences of rules to percolate down into the decision making processes of potential plaintiffs. *See* Abramowicz et al., *supra* note 144, at 978. On the other, there is a need to avoid confounding effects from other causes that might influence accuracy over long time scales, as well as the effects of forum-shopping by plaintiffs who have been geographically "assigned" to one condition but can pragmatically elect to use a different one. And of course, realistically detecting

Supreme Court's basic intuition regarding the value of plausibility screening is correct would be enormously helpful in deciding whether its new standard should be preserved or uprooted.

Having suggested that such an institution might be possible and very useful, I must emphasize that it cannot be a panacea for all juridical ills. For one thing, data collection of this sort, that works to generate simple measures that can compare widely different types of cases, can be very useful for testing hypotheses, but it will not do the hard work of generating them. Designing good research on the causal effects of procedure will require not just powerful means of comparing outcomes, but also a steady increase in our understandings of how procedural systems function at a theoretic level. To go back to the medical parallel, we would not do nearly as well if we had an extensive drug testing process but no knowledge of the internal composition of the human body. Indeed, historically a great deal of the early progress in making medicine a more scientific enterprise arose not from widespread experimentation comparing therapies and controls, but rather from the new availability of cadavers for dissection and an increased attention to nuanced evaluation of patient symptoms.[332] In the litigation realm, this translates into a need for a combined approach, where theories are developed based on close and sophisticated observation of cases, preferably with maximal access to private information, as well as through controlled experiments in artificial scenarios, and then tested in larger, real-world samples with appropriate controls for confounding variables and measurement of effects. If we are to achieve long-run optimization of procedural systems, theory and empirical analysis will have to walk hand-in-hand.

Ultimately we will need more than just promising theories and a design that enables us to test them; we will also need *money*. Perhaps the reason that a system along these lines has never been tried is that conducting an investigation into the accuracy of competing procedural regimes requires resources well beyond what juridical policy scholars normally spend. Indeed, the very nature of crafting a gold standard investigation contemplates that some case measurements will cost more than was spent on the original litigation.[333] If the judicial process has compiled a less-than-complete record, the reference-standard evaluator will need to expend more resources independently investigating the dispute.

---

all the sources of injury that *might* lead to a lawsuit and keeping track of whether they do, in fact, produce one is a task of near-Sisyphean magnitude. As a result, it may be extremely difficult to meaningfully measure the impact of differing rules on potential plaintiffs' willingness to sue.

[332] *See* discussion *supra* Part II.B.

[333] This is most likely to be the case in low-cost administrative settings or when evaluating case outcomes that arise through very early settlements or pleading dismissals.

The cost barrier is probably the biggest obstacle towards conducting this sort of research, but it might be surmountable. Note that in some settings, such as the approval of new pharmaceutical drugs, we are willing to bear quite high costs to be sure that our interventions are better than the status quo. As discussed above, our society expends hundreds of millions of dollars in evaluation costs for each new pharmaceutical drug we allow to be marketed.[334] We are willing, in other words, to spend lots of money to ensure that medical treatments are effective at improving our health, perhaps in part due to the high costs of treating serious diseases ineffectively.

Perhaps reformers could similarly mobilize political decision makers to invest in accuracy tracking by relying on the dramatic string of DNA-based exonerations of serious felony convictions.[335] Indeed, one special point of leverage that reformers might employ is that policy makers will likely expect the results of such studies to confirm their pre-existing views regarding good procedural policy.[336] Thus, even though many studies may reach conclusions that undermine the preferred policies of legislators or rule makers, they may be willing to support them because, *ex ante*, most of them will expect such investigations to provide support.

Moreover, we need not, and should not, collect so much data about all of the cases in our court systems at any one time. One way to keep costs down is to focus on those adjudicative settings or questions that are particularly worrying, and do targeted studies that are large enough to obtain relevant data, but not so large that the costs become overwhelming.[337] For instance, if we are very concerned (as some are)[338] about the quality of the decision making in immigration courts, we could embark on a research project comparing the factual judgments of Immigration Judges ("IJs") with those of independent evaluators given access to case data, the applicants, and appropriate country information.

---

[334] *See* DiMasi et al., *supra* note 109, at 180 (estimating an average out-of-pocket cost of $403 million per new drug brought to new market, with a total capitalized cost of $802 million). Although drug companies will often be able to recapture these costs by bringing valuable new products to market, the legal system does not operate as a for-profit enterprise and thus it is likely that similar research would require significant governmental or private funding sources.

[335] *See generally* GARRETT, *supra* note 16, at 1–13.

[336] *See* Abramowicz et al., *supra* note 144, at 985.

[337] *See id.* at 962.

[338] *See, e.g.*, Benslimane v. Gonzales, 430 F.3d 828, 829 (7th Cir. 2005) (Posner, J.) ("In the year ending on the date of the argument, different panels of this court reversed the Board of Immigration Appeals in whole or part in a staggering 40 percent of . . . petitions . . . that were resolved on the merits.").

This could be used both to get descriptive statistics on how well IJs are doing in general on this measure, and also to test the efficacy of specific reforms via either random assignment, if possible and politically feasible, or a second-best solution using some sort of statistical control technique and observational data.

Similarly, if we are particularly concerned about the impact of discovery costs on civil litigation and think that some new procedure might improve on the status quo, we could conduct a targeted experiment comparing the new and old rules in a few select jurisdictions, evaluating the impact of the new rule not just on legal costs but also on legal accuracy. Thus, we would be able to distinguish reforms that cut costs or delay in a beneficial way from those that trim away cost by dispensing with justice. Over the long term, such investigations might save far more social costs than they generate.

Of course, if we found such an approach helpful with respect to discrete questions, we might want to implement a broader system of monitoring that would enable many different research questions to be answered over time, and enable inter-system and inter-temporal comparisons of many kinds. Coupled with a regulatory regime that required randomized experimental trials of new rules and publication of data regarding the differential impacts of those proposals on the cost, time-to-completion, perceived fairness, and factual accuracy of case resolutions, we might truly be able to embark on a project of evidence-based procedure reform. For that to work, however, we would need a means of keeping costs to a manageable level.

One method of keeping costs down, while still generating data that may allow reasonably accurate inferences regarding an overall system, is randomly selecting a small percentage of cases in the system to be measured. For instance, we almost certainly could not find money in the federal budget to pay for systematic accuracy measurement over all federal civil cases, but it might be feasible to randomly choose 3,000 (out of a total of around 300,000) cases to be evaluated and coded each year.[339] That might still allow large enough sub-samples of case subject matters and resolution-types to permit a great deal of useful data gathering, and if finer-grained analysis was needed, cases could be gathered over a multi-year time frame into a larger sample.

---

[339] *See* Posner, *supra* note 144, at 375; U.S. COURTS, CASELOAD STATISTICS 2011 TABLE C (2011), *available at* http://www.uscourts.gov/Viewer.aspx?doc=/uscourts/Statistics/FederalJudicialCaseloadStatistics/2011/tables/C00Mar11.pdf (indicating that 285,603 civil cases were terminated by the federal courts in 2010, and that 324,190 were terminated in 2011).

Finally, we must also be realistic about what even a well-funded system of measurement along these lines is capable of achieving. Even given an elaborate accuracy-measurement system, there are some theories that will be beyond our capability to test. One central problem is that litigation rules do not only affect litigation conduct; they also affect out-of-court behavior and a party's choice of forum in which to file claims. So if one effect of legal rules is to deter filing of claims or shunt them into a different court system, a system that tracks factors like the cost, time-to-resolution, and accuracy of legal dispositions within that system will miss such an effect. The results could be very misleading data; if a new rule makes certain meritorious claims very hard to prove, it might lower the rate at which factually supportable claims are brought successfully even while the subset of claims actually filed stays the same.[340] This is not a failure of such a research design—no tool can answer any possible question—but it is an important limitation. Absent protocols that can follow large quantities of out-of-court behavior as well as the litigation process itself,[341] we will be limited to exploring questions that ask what effects procedural rules have on litigation behavior alone, which may not always be what rule designers would wish to know in a perfect informational world.[342]

So an evidence-based movement in procedural design probably is achievable, but would require shifts in confidentiality norms, new reporting obligations for many legal actors, and a willingness to absorb significant new costs. Should reformers band together to push for such large changes to the way we evaluate procedural success and failure? In the end, it depends on a number of factors and admits of no easy answers. For those who, like me, think we have little warrant for believing that our system is fairly accurate on average and think that accuracy is one of the most important qualities that a litigation system can possess, the proposal may be

---

[340] *See* Localio et al., *supra* note 331, at 247–48.

[341] *See* discussion *supra* Part V.

[342] One potential means of filling this gap might involve careful, cross-jurisdictional matching studies designed to closely approximate a controlled experiment. Such a design would involve two key elements. First, it would be necessary to have a way of sampling the frequency of law violations in a particular context. *See e.g.*, Brennan et al., *supra* note 268, at 1964 (sampling from medical case files to detect base rates of adverse medical events caused by negligence, and comparing such base rates with rates of filed claims). Second, variations in claiming rates due to variations in procedural or evidentiary rules might be detected through careful regression or matching analyses that aim to isolate the effect of the procedures from potential confounding covariance. Such analysis will necessarily leave more uncertainty on the table than a true experiment, but it is probably the only viable means of obtaining data on these questions, given the implausibility of randomly assigning individuals to differing rule regimes long before they have reason to sue or defend against a claim.

attractive. Others may doubt either of those two propositions, and so will be unwilling to bear the large transitional costs involved in creating a truly evidence-based litigation-design movement.

VI. SMALL STEPS TOWARDS AN ACCURACY MEASUREMENT CULTURE

Given that the large transformation I describe is hard to envision and that many will be reluctant to bear its costs, some readers may wonder whether the discussion above has any relevance for real-world litigation research and practice. Luckily, a focus on what a strong evidence-based litigation reform movement might involve also provides clues as to ways we can modestly improve existing methods of assessing, designing, and implementing procedural rules.

First, whether or not a large-scale program of validity evaluation for legal outcomes is implemented, procedural analysts can benefit from thinking about the gulf between those things that are typically measured in empirical studies about litigation rules and the factors of deepest relevance for evaluating rule optimality. One implication of this discussion is that we should be very cautious in drawing normative conclusions about the desirability of procedural options based on experiments and statistical data that only measure part of the relevant values. So long as studies are confined to describing who benefits from rules, how long it takes for cases to be resolved, and how much parties spend in the process, we may learn much about which interest groups stand to benefit from differing rule regimes, but we will know little about the broader social desirability of such regimes. Choosing to advocate for reforms that benefit some classes of litigants more than others without any assessment of the factual merits of their cases amounts to either political rhetoric or fairly shallow policy analysis. Some plaintiffs who lose deserve to win, others who win deserve to lose, and any policy reform suggestion that does not treat the two distinctly has little to recommend it.

To be sure, the factual-validity measurement protocol I describe above could never fully capture "who deserves to win," because in some cases the answer to that question will require normative and political judgment that lies beyond the limits of empirical analysis. But often in litigation, there would be broad agreement on a social level about who deserves to win and who to lose if only we could identify which claims were factually supportable and which were not. Until legal empirical research can find ways of tackling such questions, we must treat empirically derived recommendations in favor of particular procedural regimes with a large grain of salt.

Luckily, for those who are interested in conducting empirical research and who are sensitive to the critique above, there will be an abundance of research questions that can be delved into at relatively modest cost using the methods I have described. Tracking data that integrates cost, fairness, and validity data for all types of resolutions in the civil justice system will be hard, but smaller, more manageable projects might focus on simpler settings, such as administrative court systems or small arbitrations. In such arenas, the variety of legal issues and methods of case resolution will likely be smaller, as will the amount of relevant information needed for independent analysis of the underlying facts.

For questions of pressing interest about the formal court system, grant funding might enable localized research that incorporates validity assessment into a study design, although in the absence of broader reforms such studies will face challenges gaining voluntary access to sufficient data given existing confidentiality obligations. And some such projects may be made even more feasible if the research question of interest does not require a type of analysis that is sensitive to all of the different ways that a case might be resolved. For instance, imagine a rule changing the manner in which evidence is presented to juries. Assessing the true overall effect of such a rule would require a design that could capture the ways that earlier litigation behavior changes in response to it, but a critic might still find it useful to demonstrate that jury verdicts employing the new rule are, on average, less factually accurate than verdicts employing the status-quo control.[343] Such studies might profitably be conducted using relatively small samples of cases, and could therefore be much more affordable than the type of systematic monitoring I sketched above.

Judges can also aid in the development of more informed procedural design by treating the goal of making data available as among the considerations that guide their choices. For instance, even though existing randomized procedural trials have important limits, they still tell us much more than the alternative of poorly controlled observational studies. When

---

[343] Such an investigation may work better for certain types of evidence rules than others. One can imagine, for instance, that expert evaluations of the validity of verdicts could tell us a great deal about the comparative usefulness of two different types of scientific-evidence rules. By contrast, it is hard to see how we could justifiably investigate the utility of prejudice-based exclusions just by comparing the results of a system employing such methods to a fully informed reference-standard evaluation. A critic of such a study could justifiably worry that, if prejudice reduces verdict accuracy, fully informed investigators are just as likely to suffer from it as juries. *See* Chris Guthrie et al., *Inside the Judicial Mind*, 86 CORNELL L. REV. 777, 780–84 (2001) (presenting evidence that professional judges, just like jurors, fall prey to cognitive errors and biases).

acting as local rule makers, judges can aid in the production of data by adopting new rules in full only after testing them in a randomized fashion while encouraging litigants to consent to measurement efforts.[344]

Moreover, judges can also help researchers accumulate data merely by staying out of the way. One significant problem in assessing the current landscape of procedural rules in America is that many procedural practices have been constitutionalized.[345] As I discussed above, the literature on Civil *Gideon* shows that it is harder than one might initially expect to conclude that providing counsel to the indigent is a benefit in all types of cases.[346] Yet, because of *Gideon* itself, we may never be able to know whether the right to appointed *criminal* counsel is more clearly beneficial than the civil versions that have been studied experimentally. Grounding a procedural rule in the federal Constitution makes it much harder to know whether that rule is actually worthwhile. That does not mean that there will never be cases where the benefits of a rule are so clear that no one can see a need for future data,[347] but it does mean that the standard of evidence we demand before adopting new constitutional procedural rules should be particularly high.

A few further points may be useful to both judges and rule makers. One of the things that makes it very hard to tell how procedural rules impact adjudicative accuracy is the problem of hidden factual information. Judges, and those who read the opinions of judges to learn about the system's functioning, get only a limited window into the true spectrum of information about cases, especially the vast majority that are resolved during the pretrial process. Lawyers and parties probably know much more, including facts that were never disclosed during the litigation. Because of this, practicing litigators may have a more finely tuned sense for how often the process reaches a result that rests on dubious factual assumptions than judges or outside observers do. This may mean that those methods of making rules that draw more heavily on the input of practicing lawyers may have advantages over those that rely mostly on the work of judges. Lawyers, of course, may be biased in favor of rules that may aid

---

[344] *See* Tobias, *supra* note 144, at 1324–25 (urging a reform to the Federal Rules of Civil Procedure to make such experimentation easier).

[345] *See e.g.*, Gideon v. Wainwright, 372 U.S. 335, 344–45 (1963).

[346] *See* Barton, *supra* note 224, at 1232–33.

[347] *Cf.* Abramowicz et al., *supra* note 144, at 973 (advising that "[w]e should not allow randomized tests of parachutes because we already have strong evidence that they are effective" and that similar principles apply whenever we have no need to collect data in order to choose intelligently between policies).

client bases with whom they have a special relationship; any such input, therefore, would need to balance the input of differing segments of the bar. This observation, therefore, is one reason to favor procedural rules generated by the federal rule making process, which does incorporate lawyer input from a wide segment of practice areas, over rules that are purely judge-made.

Lastly, this discussion generates one final caution for rule makers. Some scholars have recently suggested that our government institutions should develop systematic policies requiring that new rules be experimentally tested before they can be implemented.[348] In theory, such an approach could have many benefits. However, the value of an experimental protocol will always depend on the usefulness of the questions it can ask. It would be a grave mistake, therefore, to choose procedural rules by running experiments that show their impacts on factors like cost with no means of validating outcome quality. Until we attempt to measure the impact of our procedures on the accuracy of case outcomes, experiments can tell us who will benefit from new rules and how much those new rules will cost, but not whether the changes are improving or worsening the quality of the justice provided by our institutions. Data of that sort may do more harm than good in the procedural policy making arena.

VII. CONCLUSION

Even after centuries, the medical profession still wrestles to systematically condition its policies and choices on evidence. Nevertheless, few can deny that the combination of careful investigation and theorizing about disease in the human body, coupled with an increased willingness to demand that interventions be experimentally validated before they are implemented, has resulted in an astonishing increase in the effectiveness of medical treatments compared with what prevailed a few hundred years ago.

If those who shape litigation environments wish to imitate the successes of medicine, they will operate with some advantages: litigation is a social process, not a biochemical one, and the ease of observing its operations means that it will be easier to devise plausible theories about how it works than it was to develop useful biomedical ideas. But when it comes to testing those theories, lawyers operate at a disadvantage: the health of a human body can be measured in numerous ways, some of which cost very

---

[348] *See id.* at 1005.

little, but we will face large challenges if we commit ourselves to measuring the accuracy of legal results.

The approach I sketched out in this Article, which attempts to separate out the factual accuracy component of case outcomes and systematically compare it with an independent, gold standard reference evaluation, would be costly to implement but potentially very powerful. It could answer many questions about the efficacy and optimality of competing procedural rules that are presently inaccessible to either observational or experimental testing. Nevertheless, developing a culture willing to provide the rule structure and financial support necessary for large scale accuracy testing may be beyond our reach, especially if most lawyers and law scholars believe that they can trust the procedural systems that we currently employ without testing their validity.

Continuing to gamble on the optimality of existing procedures seems untenable given the large downsides of being wrong, which include the human misery of the wrongfully convicted, the financial and emotional harms of imposing inappropriate liability on innocent parties, the injustice suffered by those who deserve social benefits but are denied them, and the social waste perpetrated by those who successfully claim benefits they should not receive. It is a near certainty that all of these forms of injustice happen nearly every day in our society, but they are mostly hidden from our view, in part because the public credibility of those who dispute a result blessed by the legal system is extraordinarily low. Whether we can develop an evidence-based litigation reform movement depends largely on how much we are willing to pay to reduce such uncertain, but potentially grave, harms.