The federal government currently publishes 196,284 searchable databases online, a number of which include information about private parties that is negative or unflattering in some way. Federal agencies increasingly publish adverse data not just to inform the public or promote transparency, but to pursue regulatory ends—to change the underlying behavior being reported. Such “regulation by database” has become a preferred method of regulation in recent years, despite scant attention from policymakers, courts, or scholars on its appropriate uses and safeguards.

This Article evaluates the aspirations and burdens of regulation by database. Based on case studies of six important data sets (published by the CFPB, CPSC, EPA, ...
FEC, FDA, and Medicare), the Article proposes what I call “Good Government Data Practices” to ensure that databases are reliable, useful, and fair. More optimal data disclosures require careful design choices that consider both data inputs and outputs, including how to gather and process data, how to characterize them, and how to present them. The article envisions a decidedly modern role for government agencies as data “stewards” rather than as mere publishers or repositories.

Agency databases have proliferated on the belief that markets, regulation, and even democracy all require transparency, that sunlight is the best disinfectant. But as transparency has moved online—becoming more pervasive, more powerful, and more burdened with regulatory dimensions—we also must recognize that sunlight can blind or even burn. It is in this spirit that I call for policymakers to embrace the government’s role as a data “steward,” a sentinel that helps maximize the quality of data inputs and outputs via tailored procedures. The more reliable government data are, the more they can enlighten us and perhaps even deter unwanted behavior.

INTRODUCTION .............................................................................. 3
I. FROM OPEN GOVERNMENT TO OPEN INDUSTRY .................. 10
   A. Publishing “Secret Laws” ............................................ 11
   B. The APA and Notice ................................................... 12
   C. FOIA and Transparency by Request ......................... 14
   D. The Internet Era ......................................................... 15
II. DATA AND ITS ASPIRATIONS ................................................ 21
   A. Market Ideals .............................................................. 21
   B. Regulatory Ideals ........................................................ 23
   C. Democratic Ideals ......................................................... 27
III. DATA AND ITS DISCONTENTS ........................................... 28
IV. AGENCY DATABASES ............................................................ 37
   A. The FEC’s Campaign Finance Data ............................... 37
   B. The EPA’s Toxic Release Data .................................... 43
   C. The CFPB’s Consumer Complaint Data ...................... 47
   D. The CPSC’s Product Safety Data ................................. 52
   E. Medicare’s Quality Data ............................................. 55
   F. The FDA’s Adverse Event Data ................................. 64
V. DESIGNING FOR OPTIMAL DISCLOSURE .............................. 69
PUBLICITY is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman.

— Louis D. Brandeis

Perhaps we should blame it on Brandeis. As is so often the case, the perfect turn of phrase often takes on a life of its own, rendering more difficult the likelihood of careful and balanced analysis of the topic to which the phrase, like a barnacle, has become attached.

— Frederick Schauer

INTRODUCTION

The federal government publishes tens of thousands of searchable online databases from hundreds of sources. The site Data.gov includes 196,284 unique data sets—roughly 150,000 from the federal government, with the remainder from sub-

federal and nongovernmental sources. Some of these data sets include information about private parties that is negative, unflattering, or adverse in some way. For example, users can search for consumer products that may have caused injuries; for drugs that may have caused side effects; for lenders that may have treated customers unfairly; for hospitals with higher-than-average mortality rates; for airlines that lose the most luggage; for lobbyists that contribute to federal candidates; or for nearby facilities that discharge toxic chemicals.

Federal agencies often publish these data not just to inform the public or promote government transparency, but also to pursue “regulatory” aims—to influence the underlying behavior being reported. “Regulation by revelation” is not at all new, of course, with a lineage stretching back decades. By now, disclosure has been so frequently used as a tool to discourage certain conduct that it is easy to take for granted. Though it is impossible to document all regulatory frameworks that rely on disclosure, even a partial list shows how ubiquitous it has become. We now rely on disclosure to regulate food nutrition, fuel economy, hospital quality, mortgages, securities, and sex


8. See discussion infra Section IV.E.


Disclosure has even become a preferred method of regulation internationally.\textsuperscript{15} Although policymakers have relied on disclosure-based regulation for decades,\textsuperscript{16} it has evolved from peculiarity to regularity as the cost of disclosure online decreases and as public demand increases. But disclosure also has evolved in other important ways. For example, sometimes the real goal of disclosure is to persuade rather than inform.\textsuperscript{17} After all, is the Surgeon General’s Warning on tobacco products meant to tell consumers something they do not already know? Or is it a suggestion not to smoke? Today, disclosure-based regulation is much less static (aimed narrowly at helping potential users of the information make better decisions) and more dynamic (aimed more broadly at trying to influence the disclosers’ underlying behavior).\textsuperscript{18} Frequently, the real party being targeted by mandatory disclosure is not the consumer, but the discloser,\textsuperscript{19} under the Brandeisian logic that shining a light on undesired behavior will deter it, or at least make it more costly.
Disclosure, then, remains in bloom. And this bloom is reflected in a relatively new species of disclosure—the searchable online database. In Part I below, I detail how agency databases derive from decades of federal policies promoting government transparency, particularly recent policies pushing the government to publish more information online. Online publication reached a crescendo with the Obama administration, which published an Open Government Memorandum on the President’s first day in office and then promoted scores of other transparency projects, including the websites FOIA.gov and Data.gov. It is possible, if not likely, that data transparency by federal agencies will experience a diminuendo under the Trump administration, if not a more aggressive, weaponized use of disclosure aimed at particular parties.

Nevertheless, decades of groundwork has enabled not only a swell of government data initiatives, but also innovative nongovernmental uses of these data. Thousands of government data sets beget perhaps thousands more third-party websites, mobile applications, and other informational products that rely on government data. Perhaps the best example is the federal government’s decision in the 1980s to publish Global Positioning System (GPS) data for civilian use, which made possible the recent wave of applications that incorporate geospatial location data, such as navigation, restaurant, and social media apps. Indeed, when President Obama announced his “Open Data Policy,” he remarked that “[t]his kind of

---

22. For example, when Donald Trump took to Twitter to criticize the contract price for Boeing to build a new Air Force One, Boeing stocks immediately dropped. See Phillip Bump, Did Donald Trump Tank Boeing’s Stock Because He Was Mad about a News Article?, WASH. POST (Dec. 6, 2016), https://www.washingtonpost.com/news/the-fix/wp/2016/12/06/did-donald-trump-tank-boeings-stock-because-he-was-mad-about-a-news-article/ [http://perma.cc/L9QM-6KQN].
innovation and ingenuity has the potential to transform the way we do almost everything.”

Lest readers discount this as hyperbole, disclosure policies are often burdened by their ambitions—being justified as promoting “autonomy, dignity, civility, community, citizenship, economic growth, and a variety of other virtues.” Part II examines these aspirations in light of the emerging, somewhat sobering evidence.

Given the policy justifications, then, many rightly wonder who could possibly oppose providing more information to the public. But as ubiquitous as disclosure has become, criticisms have emerged from scholars who doubt that it is “an unalloyed good.” I consider the burdens of “regulation by database” in detail in Part III, including problems with fairness, accuracy, and efficacy. But for introductory purposes, Schauer captures the skepticism well:

Secrecy, privacy, anonymity, and confidentiality also have their virtues, and we can all understand why transparency is a far more desirable attribute for sunroom windows than it is for bathroom doors. At times, it seems that transparency is a prime example of the old adage that where you stand depends on where you sit.

But policymakers rarely question whether the burdens of disclosure outweigh its purported benefits. One goal of this

25. Ben-Shahar & Schneider, supra note 17, at 734.
26. Fung et al., supra note 15, at xiii. Similarly, Fenster observes that “transparency appears to provide such a remarkable array of benefits that no right-thinking politician, administrator, policy wonk, or academic could be against it.” Mark Fenster, The Opacity of Transparency, 91 IOWA L. REV. 885, 888–89 (2006).
28. Schauer, supra note 9, at 1342.
29. Ben-Shahar & Schneider, supra note 17, at 683 (“[L]awmakers rarely inquire into the effectiveness or burden of disclosure.”).
Article is to help correct this asymmetry with regard to an increasingly important species of disclosure—the searchable federal database. In Part IV, I evaluate prominent databases published by six different agencies:

**Agency Databases Evaluated**

<table>
<thead>
<tr>
<th>Agency</th>
<th>Database</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFPB</td>
<td>Consumer Complaint Database</td>
<td>Complaints regarding financial product or service companies.</td>
</tr>
<tr>
<td>CPSC</td>
<td>SaferProducts.gov</td>
<td>“Reports of harm” for consumer products.</td>
</tr>
<tr>
<td>CMS</td>
<td>Hospital Compare, Physician Compare, etc.</td>
<td>Medicare quality of care metrics (such as mortality rates), with corresponding star ratings.</td>
</tr>
<tr>
<td>EPA</td>
<td>Toxics Release Inventory (TRI)</td>
<td>Production and release of roughly 650 dangerous chemicals by facility name, address.</td>
</tr>
<tr>
<td>FDA</td>
<td>FDA Adverse Event Reporting System (FAERS), Manufacturer and User Facility Device Experience (MAUDE)</td>
<td>Adverse events associated with drugs (FAERS) and medical devices (MAUDE) reported to the FDA by manufacturers, health providers, consumers.</td>
</tr>
<tr>
<td>FEC</td>
<td>Campaign Finance Disclosure Portal</td>
<td>Mandatory reports made by federal candidates, parties, committees, donors, lobbyists, others.</td>
</tr>
</tbody>
</table>

Part IV considers which of these databases succeed in producing reliable, usable data, and which have been able to influence the underlying behavior being tracked. I also consider how these databases suffer from different types of flaws, including incomplete or inaccurate data, unfriendly user interfaces, or databases whose costs of collection, maintenance, and presentation likely outweigh their utility.

Part V offers thoughts on how policymakers can design databases for more optimal disclosure, focusing on the inputs
and outputs of published data, including procedural safeguards to help ensure the quality and reliability of data. The recommendations include pre- and post-publication procedures for adjudicating “contested” data, evaluating administrative law problems that might arise when databases are implicated in enforcement actions, offering ideas on how to characterize and present the data fairly and accurately, and drawing lessons for agencies considering whether to publish “raw” or “polished” data, as well as “big” or “small” data.

Together, these recommendations envision a decidedly modern role for the government as a “data steward” rather than merely as a data source or publisher. For example, there are smart ways that the government can help gather and generate more data—making data “bigger.” But for some information, it might be preferable for the government to distill the data and make it more user-friendly—making data “smaller.” Either way, federal agencies can help ensure that data collection practices are fair, that data that purport to be accurate and objective meet those standards, and that the sources and any important context or limitations for the data are communicated clearly to users. In particular, policymakers should resist the notion that more data is always better data. Quality matters more than quantity. Part V thus builds on my recent work for the Administrative Conference of the United States (ACUS), which recently adopted a narrower set of recommendations on consumer complaint databases. We might refer to these recommendations collectively as “Good Government Data Practices.”

As the Article begins, consider the stakes. Successful data policies can help ensure healthy markets, empower consumers, inform citizens, and even influence the underlying conduct of disclosers. Failed data policies, conversely, can produce incomplete or “gerrymandered” information, create a false

---

31. Id. at 1627–28.
32. Id.
sense of security, waste resources, undermine public trust, and even put lives at risk. To design databases for optimal disclosure, we must also appreciate that disclosure is an exercise of power, that “for one person or institution to have information about another is for the former to have power over the latter.” Similarly, “transparency” can be recharacterized as “adverse transparency” if the information is unflattering or harmful to the subject in some way. How, then, can policymakers exercise this power responsibly and fairly? If we value disclosure and accept it as a baseline, how do we best manage it? As federal data is used more and more to achieve regulatory ends, both the means and ends require more purposeful policies.

I. FROM OPEN GOVERNMENT TO OPEN INDUSTRY

Recent efforts to shine light on the activities of the regulated derive from very old efforts to shine light on the activities of regulators themselves. For decades, perhaps even centuries, citizens have pushed the U.S. government to be more open and transparent. Indeed, the long arc toward government transparency is a defining hallmark of American administrative law, sitting comfortably “among the pantheon of great political virtues.” But today’s transparency emerged only after decades of reforms, arriving roughly in four waves. The 1930s brought efforts to publish so-called “secret laws” generated by agencies during the New Deal. The 1940s brought the Administrative Procedure Act’s mandates to give regulated parties advanced notice of agency actions. The 1960s and 1970s

34. Fung et al., supra note 15, at 172.
35. Schauer, supra note 9, at 1347.
37. Article I, section 5 of the Constitution requires each chamber of Congress to “keep a Journal of its Proceedings, and from time to time to publish the same,” which some view as a deliberate departure from the secrecy practiced by the British Parliament. U.S. Const. art. I § 5; James J. Brudney, Canon Shortfalls and the Virtues of Political Branch Interpretive Assets, 98 CAL. L. REV. 1199, 1218 (2010).
39. Fenster, supra note 26, at 888.
introduced FOIA and the era of transparency by request. And the 1990s and 2000s introduced mandates for agencies to post information on the internet, establishing important agency norms of online publication.

A. Publishing “Secret Laws”

The New Deal birthed a generation of new executive agencies and corresponding regulations. But agencies published their regulations at will, if at all.40 One of the earliest efforts to address the lack of transparency among federal agencies was the Federal Register Act of 1935 (the 1935 Act),41 which created the now-familiar daily gazette of executive documents. Before the 1935 Act, executive branch agencies “would each publish their own regulations in various separate publications, be they gazettes, bulletins, rulings, digests, pamphlets, notices, codes, certificates, orders, and the like.”42

The Act was motivated in part by the famous “hot oil” case, Panama Refining.43 The “hot oil” law was part of the National Industrial Recovery Act, the flagship New Deal bill passed during the Great Depression.44 The National Industrial Recovery Act authorized President Roosevelt, via the Secretary of the Interior, to limit oil production and stabilize prices during the discovery of vast new oil fields in Texas.45 During litigation over the new authority, “the government was embarrassed to admit that a reexamination of the relevant documents (which were not publicly available) had revealed that the Secretary had inadvertently revoked the relevant regulation before the lawsuit had been filed.”46 Just weeks before oral argument in the case, Erwin Griswold published a

42. McKinney, supra note 40, at 10.
law review article, *Government in Ignorance of the Law*, arguing for a Federal-Register-like system to publish executive branch laws. The “furor” over the case reflected longstanding and “widespread dissatisfaction with the unsystematic manner in which executive orders, agency regulations, and similar materials were being made available to the public.” As the federal government swelled with new agencies and new regulations, frustration reached even high-level government officials, who found it difficult, if not impossible, to locate what became known as “secret laws.” Before 1935, agencies did not even have to publish the regulations they imposed.

B. The APA and Notice

Publication requirements were further embedded in American law by the Administrative Procedure Act of 1946 (APA). The APA was a response, in part, to complaints from industry that administrative agencies during the New Deal

---

47. Erwin N. Griswold, *Government in Ignorance of the Law—A Plea for Better Publication of Executive Legislation*, 48 HARV. L. REV. 198 (1934). Griswold, the future U.S. Solicitor General and Dean of Harvard Law School, most likely anticipated the *Panama Refining* decision. He was an attorney at the Solicitor General’s office until 1934 during the briefing of the case. His article was published the same month (December 1934) as oral argument in the case. STRAUSS ET AL., supra note 46, at 446 n.4.


49. See *Cervase v. Office of the Fed. Register*, 580 F.2d 1166, 1169 (3d Cir. 1978) (“The basic object of this statutory reform was to eliminate secret law.”).


were too opaque and insular, particularly towards the private interests most affected by regulation. The original APA, in section 3, required agencies to publish important materials in the Federal Register, and in fact prohibited agencies from enforcing rules not published there. New APA procedures creating “notice and comment” rulemaking have since become a hallmark of citizen participation in government. Today, the APA requires agencies to publish a wide variety of information in the Federal Register, including basic information about their organizational structure, procedures, and substantive rules.

However, many eventually came to view APA section 3 more as a tool to withhold information than disclose it. And even between the Federal Register Act and the APA, a significant portion of agency documents—guidance, opinions, and other important “soft law” adopted by agencies—were not available in the Federal Register. Just as before 1935, such documents were accessible only on a haphazard basis.

Moreover, the APA’s disclosure provisions were largely designed to give notice to those whose legal rights were directly affected by the agency action, which were “almost invariably businesses.” Indeed, notice-and-comment rulemaking procedures were premised on the view that regulated parties, rather than the public at large, should be given notice and an opportunity to comment on proposed rules. As Bill Funk observes, the APA addressed participation in rulemaking to “interested persons,” meaning those with a “direct and palpable interest,” and required that public records be made available to persons “properly and directly concerned,” rather than the

53. Funk, supra note 38, at 172–73, 178.
54. APA § 3, 60 Stat. at 238.
57. STRAUSS ET AL., supra note 46, at 451.
58. Id.
59. Funk, supra note 38, at 173.
60. Id.; ATTORNEY GEN.’S COMM. ON ADMIN. PROCEDURE, DEP’T OF JUSTICE, FINAL REPORT OF THE ATTORNEY GENERAL’S COMMITTEE ON ADMINISTRATIVE PROCEDURE 2 (1941).
61. Funk, supra note 38, at 174.
general public. Thus, the APA’s disclosure provisions were aimed to inform regulated parties, not shine a light on them, as became the focus decades later.

C. FOIA and Transparency by Request

If the APA is viewed as a reaction to agency hostility towards regulated businesses, then the 1960s and 1970s could be viewed as a movement toward recognizing the public interest on equal footing. The modern open government movement really began in 1967, when Congress passed the Freedom of Information Act (FOIA), requiring agencies to index and make public vast amounts of materials not published in the Federal Register. Like the APA, FOIA was motivated in part by the desire to ensure an informed citizenry. President Lyndon Johnson, who signed the bill reluctantly and only under pressure from the press corps, noted that FOIA “springs from one of our most essential principles: a democracy works best when the people have all the information that the security of the nation will permit.” Half a century later, modern scholars still acknowledge FOIA’s importance to our democratic government.


66. Vladeck, supra note 64, at 1798.


68. Vladeck, supra note 64 (noting that FOIA “embodies the ideal that information is the lifeblood of democracy”).
FOIA created a “strong presumption in favor of disclosure.” In fact, it reversed the burden in the original APA that opened access to government records only if the requester could demonstrate a compelling need.

FOIA, of course, has been criticized for falling short of its lofty goals of pursuing democracy through transparency and accountability. A major complaint is that FOIA produces transparency only by request. It imposes few affirmative disclosure obligations on agencies, and relies on a complex framework that often requires relatively sophisticated private intermediaries with sufficient “time, money, and expertise” to “press a recalcitrant administration for disclosure.” Scholars have also criticized FOIA for being too malleable, particularly under administrations that construe its disclosure requirements narrowly and exemptions broadly (the George W. Bush administration is frequently cited). As such, FOIA envisions passive disclosure by agencies rather than active disclosure or publication of the information they hold. Moreover, the basic premise of FOIA’s “request-and-wait-for-a-response approach” is seen as obsolete in the Internet era.

D. The Internet Era

In the 1990s, as the internet came of age, a series of laws pushed the government to use it. For example, the Paperwork

---

72. Vladeck, supra note 64, at 1789.
74. Scholars often point to the George W. Bush administration on this point. President Bush instructed federal agencies to deny FOIA requests when they could invoke a “sound legal basis.” Memorandum from John Ashcroft, U.S. Att’y Gen., to the Heads of All Fed. Dep’ts & Agencies (Oct. 12, 2001); Vladeck, supra note 64, at 1790; Shkabatur, supra note 55, at 89.
76. Vladeck, supra note 64, at 1792–93.
Reduction Act of 1995,77 the Electronic Freedom of Information Act of 1996 (sometimes called “E-FOIA”),78 and the Government Paperwork Elimination Act of 199879 directed federal agencies to use the internet to publish more information online and to “improve the productivity, efficiency, and effectiveness of Federal programs.”80 In particular, E-FOIA required agencies to publish online their final opinions and orders, as well as records likely to be requested,81 thus spawning the surge in agency online “reading rooms.”82

During this time, regulators began to rely more on mandatory reporting to inform agency actions, including both rulemaking and enforcement. Somewhat quickly, internet technologies reduced the costs of gathering and disseminating such information online, which increased both public demand and expectations for agency records. Thus, federal agencies began to publish copious amounts of information not just about their own activities, but about regulated parties as well.83

In the 2000s, as federal agencies built sprawling websites,84 a new generation of laws pushed for even more online disclosure. For example, the E-Government Act of 2002 required federal agencies to post more information online and make it more accessible through improved organization.85 The Act also created the Office of Electronic Government within the Office of Management and Budget (OMB), headed by a Chief Information Officer that would coordinate with agencies through a Council.86 As with prior laws, the stated goal of the E-Government Act was to encourage the federal government to enhance public access to information and government services,

84. Id.; Cortez, supra note 75, at 1393.
86. Id.
this time using internet technologies. In 2007, the Open Government Act addressed various frustrations with FOIA, requiring new agency procedures and new public liaison offices designed to address persistent agency delays in responding to FOIA requests.

Likewise, during the 2000s, Congress also passed laws requiring more transparency in federal spending, directing the OMB to publish online the details of federal grants, loans, and contracts. Today, the public can search the federal website USAspending.gov to view entities that have received federal money, or more narrowly focused sites like Recovery.gov to see how the federal government has spent money from the economic stimulus package of 2009. The former includes a searchable database with the name and location of the entity receiving federal money, the amount received, the type of transaction, the funding body, the purpose of the funding, and other information. As Vladeck notes, the site “was able to piggyback on the work of OMB Watch, a nonprofit watchdog organization that with foundation support had already constructed a comprehensive, searchable database that is also

87. Id. § 101(a), 116 Stat. at 2902 (defining “electronic government”).
available free of charge to the public.”93 This work reflects a marked shift from focusing on information about the government to information about private parties.

In 2001, Congress passed an important but less frequently discussed law, the Information Quality Act (IQA), sometimes referred to as the Data Quality Act (DQA).94 The Act required the OMB to issue government-wide guidelines for “ensuring and maximizing the quality, objectivity, utility, and integrity of information . . . disseminated by the government.”95 It also required the OMB to “establish administrative mechanisms allowing affected persons to seek and obtain correction of information maintained and disseminated by the agency” that does not meet those standards.96 In 2002, the OMB finalized guidelines implementing the Act,97 followed by agencies issuing their own guidelines.98

These IQA guidelines would seem well suited to regulate the quality of information posted in agency databases. Yet the Act’s application to databases is highly unclear, and probably varies by database. The broad wording of the IQA states that the OMB guidelines should apply to agency “dissemination of public information, regardless of the form or format.”99 And the OMB guidelines define “information” as “any communication or representation of knowledge such as facts or data, in any

---

93. Vladeck, supra note 64, at 1829–30 (citing About, FEDSPENDING.ORG, http://www.fedspending.org/aboutthissite.php (last visited August 1, 2017) [https://perma.cc/QPC2-7ZT5]). Note that OMB Watch later changed its name to the Center for Effective Government, but ended operations in March 2016, transferring most of its resources to the Project on Government Oversight (POGO). See CTR. FOR EFFECTIVE GOV’T, http://www.foreffectivegov.org (last visited June 5, 2017) [https://perma.cc/7GF7-LGSD].


medium or form,”100 including “information that an agency disseminates from a web page.”101 However, the OMB excludes from coverage “opinions, where the agency’s presentation makes it clear that what is being offered is someone’s opinion rather than fact or the agency’s views.”102 The guidelines also exempt “adjudicative processes.”103 These exemptions might thus exclude important agency databases, such as the CFPB’s Consumer Complaint Database.104

All this is prelude, however, to the Obama administration and its efforts toward open government. On his first full day in office, President Obama published the Open Government Memorandum,105 as well as a memorandum on FOIA.106 Although various internet-driven transparency initiatives emerged during the Clinton and Bush administrations, the two Obama documents were viewed as a gesture toward openness and a turn from the secrecy that characterized the Bush administration.107

In December 2009, the OMB published the Open Government Directive,108 following on President Obama’s Open Government Memorandum, urging agencies to “take prompt steps to expand access to information by making it available online in open formats.”109 The directive required each agency, within 45 days, to “identify and publish online in an open format at least three high-value data sets . . . on Data.gov” that

101. Id. at 8460.
103. Id.
104. See discussion infra Section IV.C.
107. STRAUSS ET AL., supra note 46, at 441. Of course, even the Obama administration has been criticized for rejecting more transparency in matters of terrorism and national security. Jeff Kahn, Terrorist Watchlists, in CAMBRIDGE HANDBOOK OF SURVEILLANCE LAW (David Gray & Stephen E. Henderson eds., Cambridge Univ. Press, 2017).
109. Id. at 2.
had previously not been available.\textsuperscript{110} Within sixty days, each agency was to create an Open Government webpage. Today, eighty different federal agencies and subagencies have posted 196,284 datasets on Data.gov.\textsuperscript{111}

In 2011, the Justice Department created FOIA.gov, a website that publicizes data on how agencies have performed their FOIA duties.\textsuperscript{112} The searchable online database displays the number of FOIA requests received by each agency, the disposition of those requests, and the current backlog. Ironically, like other mandatory disclosure regimes, the Justice Department is using “naming and shaming” to encourage agencies to be more responsive to FOIA requests.\textsuperscript{113} Still, scholars question how effectively “naming and shaming” is at convincing under-performing agencies to increase their responsiveness to FOIA requests.\textsuperscript{114}

As this history shows, federal transparency efforts gradually evolved from general right-to-know laws like FOIA, aimed at increasing transparency in the government itself, to disclosure of information held by the government regarding the activities of corporations and regulated entities.\textsuperscript{115} Thus, the current gestalt that pursues transparency from industry owes much to earlier right-to-know efforts that pursued transparency from government.\textsuperscript{116} Moreover, as with so many other things, information technology has enabled the use of databases and disclosure as a regulatory tool. Internet technologies are being used to mine the data of countless industries and activities, post them in the public domain, and make them accessible through searching, sorting, and other data-sifting tools.

\textsuperscript{110} Id.
\textsuperscript{111} Federal Agency Participation, supra note 3.
\textsuperscript{113} Shkabatur, supra note 55, at 100.
\textsuperscript{114} Id.
\textsuperscript{115} See also FUNG ET AL., supra note 15, at xii–xiii. They call this “targeted transparency.” Id.
\textsuperscript{116} Id. at 28. Of course, one could support more transparency from government without supporting more transparency from industry, and vice versa. But the two trends seem to be part of the same historical arc towards more public reporting and openness by regulatory agencies.
II. DATA AND ITS ASPIRATIONS

Agency databases have become ubiquitous in part because they appeal to so many of our intuitions about how government, markets, and regulation should work. So much so, in fact, that disclosure skeptics criticize the optimists for too often posing it as a panacea—a Swiss Army policy “intended to promote autonomy, dignity, civility, community, citizenship, economic growth, and a variety of other virtues.”

Contemporary scholarship, of course, has focused on the many ways in which internet technologies have facilitated communication between the government and the public. But after years of scholarly praise of the internet’s role in facilitating transparency, accountability, and democracy, inevitable critiques have emerged. Still, disclosure is frequently offered as a tool that can achieve market, regulatory, and democratic ideals.

A. Market Ideals

A frequently invoked rationale for regulatory disclosures is that disclosure can improve consumer decision-making, facilitate markets, and “protect the naïve from the sophisticated.” Various disclosure regimes, at their heart, try to resolve the famous “lemons problem” framed by George Akerlof, who argued that in markets with information asymmetries between buyers and sellers, sellers may have an incentive to sell inferior products or services, which can

---

117. Ben-Shahar & Schneider, supra note 17, at 734.
120. Indeed, Sage evaluated these types of justifications almost twenty years ago, characterizing them somewhat differently as four separate rationales (competition, agency, performance, and democratic). See Sage, supra note 5, at 1710–11. He also observes, however, that disclosure rationales can often be contradictory or at least in tension. Id.
121. Ben-Shahar & Schneider, supra note 17, at 649.
undermine the market.\textsuperscript{122} By requiring disclosure, the
government can correct these asymmetries and facilitate
efficient markets. Schauer calls this “transparency as
efficiency”—the idea that freely available information “is
precisely what makes markets operate effectively.”\textsuperscript{123}

At their best, agency databases can inspire a “race to the
top” by encouraging firms to compete based on their published
activities. One of the original aspirations for the CFPB’s
Consumer Complaint Database was to encourage companies to
use the data to publicize how well they respond to consumer
complaints compared to competitors.\textsuperscript{124} The Bureau points to
this phenomenon in the airline industry, where airlines use
data by the Department of Transportation and FAA to market
their low rates of passenger complaints compared to
competitors, and where third party airline ratings systems
make use of the same government data.\textsuperscript{125} The Bureau
concludes that after the data is made public, “The marketplace
of ideas then does the rest.”\textsuperscript{126}

Disclosure thus satisfies both our free-market intuitions\textsuperscript{127}
and the policymaker’s urge to do something. In the law that
created the CFPB, Congress repeatedly asserted that the
Bureau would publish information that helped consumers
make more informed choices about financial products and
services\textsuperscript{128}—a refrain repeated frequently by the Bureau in its


\textsuperscript{123} Schauer, \textit{supra} note 9, at 1350.

\textsuperscript{124} Disclosure of Certain Credit Card Complaint Data, 76 Fed. Reg. 76,628, 76,630 n.9 (Dec. 8, 2011).

\textsuperscript{125} Id. at 76,631.


\textsuperscript{127} Ben-Shahar & Schneider, \textit{supra} note 17, at 681.

Consumer advocates, in fact, have encouraged the Bureau’s database efforts by arguing that “disclosure is one of the best tools government agencies can use.” A related consumer-centered ideal served by disclosure is autonomy. As Ben-Shahar and Schneider argue, mandated disclosure is alluring because “[i]t supposes that people make better decisions for themselves than anyone can make for them and that people are entitled to freedom in making decisions.” Countless mandatory disclosure laws rely on this logic. The CFPB’s complaint data, the CPSC’s product safety data, and many other data sets are predicated on consumers using the data to vote with their wallets, avoiding substandard performers.

Finally, corporations and industry groups often “urge greater transparency as an alternative to allegedly more heavy-handed regulation.” Of course, many scholars embrace this view as well. As Archon Fung and colleagues emphasize, the “ingeniousness” of regulation via disclosure “lies in its mobilization of individual choice, market forces, and participatory democracy through relatively light-handed government action.” Regulation by disclosure thus appeals across both political and ideological spectra.

B. Regulatory Ideals

A second aspiration of disclosure is to achieve regulatory ends—using publication to preempt or at least deter undesired behavior. Corporate and securities law, for example, rely heavily on disclosure of company holdings and transactions, with the idea that corporations whose dealings are transparent...
and publicly accessible will think twice before acting in ways that harm investors.\textsuperscript{137} Similarly, requiring hospitals to publish mortality rates is really a device to encourage hospitals to reduce mortality rates.\textsuperscript{138} Again, the same logic motivates many disclosure regimes, with a long lineage. In 1796, Jeremy Bentham observed that “the more strictly we are watched, the better we behave.”\textsuperscript{139}

Agencies also frequently use databases to publish compliance and enforcement data.\textsuperscript{140} Scholars have called for agencies to actively publish enforcement records that are available under FOIA but must be requested.\textsuperscript{141} For example, David Vladeck argues that Congress should require the OMB to compile enforcement records in a searchable database to “permit the public to track repeat-offender corporations in the same way the public can now track grants and contracts given to the same corporate recipients.”\textsuperscript{142} For years, a nonprofit based at Syracuse University, called the Transactional Records Access Clearinghouse (TRAC), has published enforcement data gathered via FOIA from the Bureau of Alcohol, Tobacco, and Firearms (ATF), the Department of Homeland Security (DHS), the Department of Justice (DOJ), the Drug Enforcement Administration (DEA), the Federal Bureau of Investigation (FBI), and the Internal Revenue Service (IRS).\textsuperscript{143}


\textsuperscript{138} Michael B. Rothberg et al., \textit{Choosing the Best Hospital: The Limitations of Public Quality Reporting}, 27 HEALTH AFF. 1680 (2008); Schauer, supra note 9, at 1348.


\textsuperscript{140} See discussion \textit{infra} Part IV, regarding several examples from the EPA, FEC, FDA, etc.

\textsuperscript{141} Vladeck, supra note 64, at 1830–31.

\textsuperscript{142} Id. at 1830.

\textsuperscript{143} \textit{Id.}; TRAC: About Us, TRANSACTIONAL RECORDS ACCESS CLEARINGHOUSE, http://trac.syr.edu/aboutTRACgeneral.html (last visited July 15, 2017) [https://perma.cc/L9NF-VZ4R]. The FTC maintains a massive database of complaints against companies, though it is nonpublic and is available only to enforcement agencies, such as the FTC, Department of Justice, and participating state and local agencies. Consumer Sentinel Network, FED. TRADE COMM’N, https://www.ftc.gov/enforcement/consumer-sentinel-network (last visited July 15,
Enforcement data may also help counter underenforcement by agencies, what Matthew Stephenson calls “agency slack.”\textsuperscript{144} Scholars note widespread underenforcement by a variety of agencies in a variety of contexts.\textsuperscript{145} Even when agencies do pursue regulatory violations, they often fail to enforce them.\textsuperscript{146} Underenforcement may derive from several sources—insufficient agency resources, ideology, anti-regulatory pressures, political oversight, inertia, or agency self-interest.\textsuperscript{147} Regardless of the contributors, making compliance and enforcement data public might inspire agencies to reach more optimal levels of enforcement—or even inspire companies to reach more optimal levels of compliance. Observers also suspect the converse. When the U.S. Department of Agriculture (USDA) “abruptly removed inspection reports, warning letters, and other documents on nearly 8000 animal facilities that the agency regulates” from its website in February 2017, critics worried that the move would shield violators from journalists and animal rights groups.\textsuperscript{148}

If publication alone does not encourage compliance, perhaps use of the data by third-party intermediaries can. In justifying its Consumer Complaint Database, the CFPB pointed to companies offering commercial intelligence products based on data from the FDA’s drug and device adverse events databases (FAERS and MAUDE).\textsuperscript{149} The CFPB itself notes that

\textsuperscript{144} Matthew C. Stephenson, Public Regulation of Private Enforcement: The Case for Expanding the Role of Administrative Agencies, 91 VA. L. REV. 93, 110 (2005).


\textsuperscript{147} Stephenson, supra note 144.

\textsuperscript{148} Meredith Wadman, Courts Ponder How Public Animal Reports Must Be, 356 SCIENCE 790, 790 (2017).

third-party users, like the consulting firm Deloitte and the U.S. News & World Report, are relying on the Bureau’s data to publish findings and recommendations.\footnote{150} Moreover, there is always the lingering fear that shareholders, plaintiffs’ lawyers, media, bloggers, or other enforcement agencies will use published data against companies.

An emerging potential use of government data, and one encouraged by the government itself, is “crowdsourcing.” Crowdsourcing is a method of soliciting answers, ideas, resources, or services from a large network of people, typically online.\footnote{151} Technologists have envisioned the government serving as a “platform” for innovation by providing data that inspires outside parties to create innovative uses for the data.\footnote{152} Government agencies cannot predict how their data sets might be used by the public, but the act of publishing data in raw, open, and machine-readable format allows the public to generate innovative and perhaps more enlightening uses of the data.\footnote{153}

The Obama administration pursued several crowdsourcing initiatives, published on websites like Challenge.gov, which features various prize competitions sponsored by over eighty federal agencies.\footnote{154} The site claims that the government has awarded “[m]ore than $250 million in prize money” since 2010.\footnote{155} The idea is that “U.S. federal agencies invite the public’s help to solve perplexing mission-centric problems.”\footnote{156} Indeed, even the administration’s original Open Government Directive offered contests and prizes that incentivized the
public to “tinker” with the data released. Thus, some of the most provocative uses of agency data may just be emerging.

C. Democratic Ideals

A third and more lofty justification for publishing government data is to enhance government accountability. Transparency is often assumed to be its precondition. Leading thinkers—such as John Milton, John Stuart Mill, James Madison, Oliver Wendell Holmes (father of the marketplace of ideas), and Louis Brandeis—have long drawn an explicit link between transparency, accountability, and democracy. Centuries of writing are filled with paens to the virtues of transparency. Today, the modern open source movement, which owes much to these forebears and helped seed the open government movement, holds that information is a necessary precondition for truth and progress. Seen in this way, agency databases are a logical vehicle for both technological and democratic progress.

161. See, e.g., Letter from James Madison to W.T. Barry (Aug. 4, 1822), in 9 THE WRITINGS OF JAMES MADISON 103 (Gaillard Hunt ed., 1910) (“A popular Government, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy; or perhaps both.” As Fung et al., supra note 15, at 24, discuss, this language is carved on the Library of Congress building); Abrams, 250 U.S. at 630 (Holmes, J., dissenting) (“[T]he best test of truth is the power of [an assertion] to get itself accepted in the competition of the market . . . .”); Stanley Ingber, The Marketplace of Ideas: A Legitimizing Myth, 1984 DUKE L.J. 1, 2–3 (1984).
162. Schauer, supra note 9, at 1350. However, despite the fanfare, some scholars question whether recent transparency and open government initiatives actually “strengthen public accountability.” See, e.g., Shkabatur, supra note 55, at 81. Shkabatur argues that current transparency policies, driven by new technologies, give agencies too much control over which data are published, “prioritizes quantity over quality of disclosures,” and reinforces older barriers to accessing information. Id.
Government databases can also serve an expressive function by acting as a conduit for consumers to air their grievances. Presidents Kennedy and Nixon both promulgated a Consumer Bill of Rights to pose the government as an intermediary or a tribunal through which consumer complaints could be given a voice. Such a role can increase the public’s confidence in government, providing “a positive point of contact” between agencies and citizens, and promoting the idea of “government as a positive force in society.” The CFPB’s Consumer Complaint Database essentially serves this role today.

In short, there are compelling reasons why data disclosure appeals to policymakers so much. Disclosure seems consistent with free-market and autonomy principles, and seems to be an easy and effective intervention compared to more traditional regulation. Politically, regulation by disclosure is cast as a “path of least resistance for administrative agencies seeking to promote meaningful change.” Indeed, calls for disclosure and transparency are often justified, on a more fundamental plane, as furthering the pursuit of truth, knowledge, and societal progress.

III. DATA AND ITS DISCONTENTS

But does disclosure live up to its many promises? In this Part, I consider the shortcomings of database disclosures, before evaluating several prominent agency databases in Part IV. Part V then considers ways to design databases for more optimal, effective disclosure.


165. Porter, supra note 163, at 76.

166. Ben-Shahar & Schneider, supra note 17, at 681–82.

167. Sage, supra note 5, at 1772.

168. Schauer, supra note 9, at 1350.
First and foremost, agency disclosure of information about regulated parties can itself cause a variety of harms, ranging from concrete (a devaluation of stock price) to less tangible, reputational harms. Although scholarship on these harms is not voluminous, the harms are relatively well documented. For example, in 2008 the FDA and CDC mistakenly identified tomatoes as the source of a salmonella outbreak, costing the tomato industry an estimated $200 million in lost sales. And there are numerous instances in which a company’s stock value plummeted after an agency announcement criticized the company. As Vladeck notes, “[t]here is also force, as a general matter, to the argument that companies should not be subject to commercial harm simply because they are compelled to report their activities to the government.”

Questions about the value of so-called “naming and shaming” have crept into various disciplines. In the book Is Shame Necessary?, Jennifer Jacquet considers the virtues and flaws of modern naming and shaming, particularly how social media and other modern modes of communication might amplify shame effectively to change the behavior of corporations or even governments. Indeed, organizations like Wikileaks derive tremendous power and influence from disclosure—the kind of nongovernmental power normally reserved for mainstream media.

169. See, e.g., Ernest Gellhorn, Adverse Publicity by Administrative Agencies, 86 HARV. L. REV. 1380 (1973); Cortez, supra note 75; CORTEZ, supra note 143, at 9–12.


171. Denis G. Maki, Coming to Grips with Foodborne Infection—Peanut Butter, Peppers, and Nationwide Salmonella Outbreaks, 360 NEW ENG. J. MED. 949 (2009).

172. See Cortez, supra note 75 (discussing examples); see also CORTEZ, supra note 143, at 9–12.

173. Vladeck, supra note 64, at 1793 (noting, however, that secrecy should give way to publication, for example, when necessary to avoid death or serious injury).


175. JENNIFER JACQUET, IS SHAME NECESSARY: NEW USES FOR AN OLD TOOL (2015).

Less sanguine views argue that “shaming is the very antithesis of the law,” particularly when wielded by the government. For example, Donald Trump’s Twitter posts, the focus of significant handwringing, demonstrate the unfair destructive power of adverse publicity. In December 2016, Boeing stock took a quick plunge after President-elect Trump took to Twitter to criticize the cost of Boeing’s contract to build a new Air Force One. After the episode, some investors and market analysts began to monitor Trump’s tweets for potential market-moving proclamations.

A second objection to the use of disclosure as a regulatory tool is that it may be ineffective. There is growing scholarly skepticism that openness necessarily leads to knowledge or that more information necessarily produces better decisions. As David Vladeck observes, “there is now a significant and growing dissonance between the promises made by our federal right-to-know laws and their performance.” And as Ben-Shahar and Schneider emphasize in their magisterial article, The Failure of Mandated Disclosure, disclosure regimes often fail completely in meeting their goals, and in fact can have unintended consequences that hinder them. Although their bearish views are challenged, the virtues of disclosure no longer remain uncontested.

A third objection to agency disclosure in general, and to databases in particular, is that they are not always complete or accurate. To be most useful, data that purport to present truthful, objective information must meet those standards.

178. Bump, supra note 22.
180. Schauer, supra note 9, at 1351.
181. Vladeck, supra note 64, at 1792.
182. Ben-Shahar & Schneider, supra note 17.
Otherwise, they risk succumbing to a problem known well to computer and software engineers: “Garbage in, garbage out.”\textsuperscript{184} Flawed inputs produce flawed outputs, and inaccurate databases will be unreliable. Even well-known advocates of regulation by information, such as Cass Sunstein, acknowledge the dangers of regulating based on flawed data.\textsuperscript{185} Litigation challenging inaccurate reports on the CPSC’s SaferProducts.gov database demonstrates that agencies sometimes post inaccurate data, and sometimes that data lingers online for years before it is corrected or retracted.\textsuperscript{186} The FEC’s campaign finance data is widely acknowledged to be incomplete.\textsuperscript{187} The FDA’s medical device database, which tracks problems with devices and is subject to mandatory reporting requirements, is undermined dramatically by under-reporting.\textsuperscript{188} Even relatively noncontroversial databases like USAspending.gov and Recovery.gov have been found to have widespread inaccuracies.\textsuperscript{189} Both the OMB and GAO have acknowledged that data on the sites has been inaccurate, untimely, or incomplete.\textsuperscript{190} An independent review by the Sunlight Foundation found that the sites had “over 1.2 trillion dollars’ worth of misreported spending in 2009 alone.”\textsuperscript{191}

\begin{flushright}
\textsuperscript{187} See discussion infra Section IV.A.
\textsuperscript{189} See, e.g., Shkabatur, supra note 55, at 103.
\textsuperscript{191} Transparency Through Technology: Evaluating Federal Open-Government Initiatives: Hearing Before the Subcomm. on Tech., Info. Policy, Intergovernmental Relations, and Procurement Reform of the H. Comm. on Oversight and Gov’t Reform, 112th Cong. 3 (2011) (statement of Ellen Miller, Executive Director, Sunlight Foundation); Shkabatur, supra note 55, at 103.
\end{flushright}
Inaccurate or incomplete data sets can be even more problematic if agencies process them with software algorithms or artificial intelligence to identify regulatory violators or single out firms for further investigation. The use of such algorithms for regulatory enforcement purposes raises a host of novel questions about agency delegations, justifications, and reasoning. Moreover, increased skepticism of the accuracy and objectivity of algorithmic decision making warrants further investigation, particularly as regulators rely on these methods more.

A related criticism of agency databases is that they often present data without appropriate context. Jennifer Shkabatur notes that it is hard for a lay person browsing USAspending.gov, for example, to evaluate whether a $20 million contract between the Department of Commerce and Industrial Economics for “continued support for the Deepwater Horizon oil spill” is money well-spent, or whether an $817 million contract between the Department of Defense and Lockheed Martin for “incremental funding” is wasteful. But that is the only information provided. Thus, she argues, “even if data is timely and reliable,” when stripped of context, it may not always be particularly meaningful or useful. Providing appropriate context is one way for agencies to act as data

192. For example, the FTC uses data from its nonpublic Sentinel database to look for patterns of complaints against a potential enforcement target. See CORTEZ, supra note 143, at 55–56. For a series of posts on the use of automated decision-making by regulatory agencies, see Mariano-Florentino Cuéllar, Artificial Intelligence and the Administrative State, REG. REVIEW (Dec. 19–22, 2016), https://www.theregview.org/2016/12/artificial-intelligence-and-the-administrative-state/ [https://perma.cc/WS9U-USFN].


195. See, e.g., Shkabatur, supra note 55, at 104 (citing reports and concerns that federal spending databases, like USAspending.gov and Recovery.gov, provide voluminous data on spending, but do not provide criteria or justifications for those spending decisions and whether they square with congressional directives).

196. Id. at 104–05 (noting that these are two real-life examples in USAspending.gov).

197. Id. at 105.
“stewards,” rather than to serve as mere publishers or repositories.

Scholars also note that agencies exercise significant discretion to decide what data to disclose and the scope of that data, which can skew user perceptions.\textsuperscript{198} Industry repeatedly made such arguments as the CFPB built its Consumer Complaint Database. Commenters objected that the database would necessarily include only self-selected complaints that were nonrandom and thus not representative of the consumer population.\textsuperscript{199} Industry commenters also objected that the complaints lacked context and that users might overlook the data’s limitations, despite disclaimers by the Bureau.\textsuperscript{200}

Another problem with agency disclosure is the volume of it, ironically, and the risk of drowning the public with information. Data.gov currently hosts over 195,000 data sets,\textsuperscript{201} some of which include millions of records or data points. But more information is not always better information—although many disclosure regimes assume otherwise.\textsuperscript{202} Today, federal agencies post so much information online that many scholars wonder who is served by these data dumps.\textsuperscript{203} The conventional wisdom is that few benefit from an “undifferentiated mass of information” posted online, as the “cost of sifting through it would overwhelm its value.”\textsuperscript{204} Genuinely useful information is often buried.\textsuperscript{205} As Paredes notes, sunlight can be the best disinfectant, “[b]ut sunlight can also be blinding.”\textsuperscript{206}

So how does salient information stand out? Evaluations of mandatory disclosures regimes find that even the most well-known are ignored: “Next to the warning label on cigarette packs, \textit{Miranda} is the most widely ignored piece of official advice in our society.”\textsuperscript{207} Gradually, because mandated

\begin{small}
\begin{enumerate}
\item[198.] Id. at 117.
\item[199.] \textit{See}, e.g., Disclosure of Certain Credit Card Complaint Data, 77 Fed. Reg. 37,558, 37,561 (June 22, 2012).
\item[200.] See id. at 37,562.
\item[202.] See Ben-Shahar \& Schneider, supra note 17, at 650.
\item[203.] Shkabatur, supra note 55, at 118.
\item[204.] Vladeck, supra note 64, at 1832.
\item[205.] Ben-Shahar \& Schneider, supra note 17, at 737.
\item[206.] Paredes, supra note 10, at 419.
\item[207.] Richard A. Leo, \textit{Questioning the Relevance of Miranda in the Twenty-First Century}, 99 MICH. L. REV. 1000, 1012–13 (quoting Patrick Malone, “You Have the
\end{enumerate}
\end{small}
disclosures are so attractive to policymakers, they tend to accumulate over time, which only compounds the problem—“disclosures are added, never removed.” As Ben-Shahar and Schneider observe, consumers “encounter too many disclosures to digest most of them.” Such concerns lend credence to the idea that sophisticated intermediaries will continue to serve a valuable channeling and interpretive role. Indeed, the idea behind the massive data dumps on Data.gov and other federal databases “is that nongovernmental intermediaries can step in and translate the raw data for the general public.”

A related shortcoming of disclosure is the complexity of the data. Mandated disclosure regimes have become ubiquitous in federal and state law, ensconced in statutes, ordinances, agency regulations, and common law. Sometimes, these sources of law demand “marvelously elaborate disclosures” that are difficult if not impossible for the intended beneficiaries (usually consumers) to understand. Consumers, of course, are not perfectly rational, but exercise “bounded rationality” due to various cognitive biases and distortions. Thus, many scholars have come to recognize that disclosure of objective information may not, in itself, generate optimal outcomes—rather, disclosure regimes “may need to aggregate, translate, simplify, or benchmark the facts.” Even if users do not understand the science or statistical techniques behind the data, the data may still be successful in improving the product or conduct targeted by the disclosure.

Given the volume and complexity of most data, intended beneficiaries often do not and cannot use it, particularly in the

---

Right to Remain Silent”: Miranda After Twenty Years, 55 AM. SCHOLAR 367, 368 (1986); Ben-Shahar & Schneider, supra note 17, at 678.

208. Ben-Shahar & Schneider, supra note 17, at 679.

209. Id. at 705.

210. Shkabatur, supra note 55, at 118.

211. Ben-Shahar & Schneider, supra note 17, at 650.

212. Id.


214. FUNG ET AL., supra note 15, at 34.

215. Craswell, supra note 18, at 361.
idealized way policymakers intend. In a variety of legal contexts, the targets for information disclosure “often do not read disclosed information, do not understand it when they read it, and do not use it even if they understand it.”

Examples abound. A troubling one is the extensive campus crime data reported by colleges and universities to the Department of Education under the Clery Act, which often goes unread. Nevertheless, the Department promises that the Clery Act “is intended to provide students and their families, as higher education consumers, with accurate, complete, and timely information about safety on campus so that they can make informed decisions.”

The current reality is that, for most databases, such aspirations outstrip reality.

Databases and other disclosure regimes can also be costly. Disclosure is often assumed to be simple and low-cost, particularly compared to more conventional regulation and enforcement. But successful disclosure systems often require “a distinctive and demanding architecture.” Any thoughtful disclosure regime must determine what information must be disclosed, by whom, to whom, the optimal format, and appropriate quality assurance and enforcement mechanisms.

Indeed, many assume that disclosure avoids many of the compliance and enforcement costs that attend to traditional command-and-control regulation. But, compliance with disclosure regimes must also be monitored and enforced—frequently with both civil and criminal penalties. For regimes with low rates of compliance, policymakers have often increased both penalties and enforcement efforts.

---

216. Ben-Shahar & Schneider, supra note 17, at 665.
219. OFFICE OF POSTSECONDARY EDUC., U.S. DEP’T OF EDUC., THE HANDBOOK FOR CAMPUS CRIME REPORTING 3 (2005); Ben-Shahar & Schneider, supra note 17, at 702–03.
220. Ben-Shahar & Schneider, supra note 17, at 735–37.
222. Id. at 39–46.
223. Ben-Shahar & Schneider, supra note 17, at 735–37.
224. FUNG ET AL., supra note 15, at 46 (examining civil and criminal penalties to enforce the disclosure mandates in campaign finance law and corporate law).
225. Id. at 45–46.
Moreover, the cost to disclosers can be significant. For example, a single new SEC requirement that companies file “current reports” of insider transactions was expected to generate 215,000 additional filings to the SEC annually, at an estimated cost of over $89 million per year.\(^{226}\) Another recent study found that U.S. physicians in just four common specialties spend $15.4 billion annually reporting under various quality measurement programs.\(^{227}\) Thus, regulation by disclosure can be costly.

Opportunity costs can also be significant. Relying on disclosure as a means to pursue regulatory ends may mean bypassing other, better means for achieving those ends.\(^{228}\) Traditional command-and-control regulation long ago lost its luster among policymakers and academics.\(^{229}\) But modern replacements like “new governance,” despite their many promises, can underwhelm. Despite the widespread use of disclosure, “it remains an open question whether transparency as regulation is better or worse, all things considered, than more direct forms of regulation.”\(^{230}\) One might reasonably wonder whether the time and personnel that agencies spend on disclosure would be better spent writing regulations and enforcing them.

Finally, there is some evidence that mandatory disclosure regimes can backfire. Consumers might be tempted to let their guards down when presented with mandated disclosures that give transactions a “veneer of legality.”\(^{231}\) There is also


\(^{228}\) See Ben-Shahar & Schneider, supra note 17, at 737–42.

\(^{229}\) Disclosure also responds to the Hayekian criticism that command-and-control regulation is insufficient for regulating a large, complex society. F.A. Hayek, The Use of Knowledge in Society, 35 AM. ECON. REV. 519, 524–25 (1945).

\(^{230}\) Schauer, supra note 9, at 1348.

\(^{231}\) Ben-Shahar & Schneider, supra note 17, at 740 (citing Lauren E. Willis, Decisionmaking and the Limits of Disclosure: The Problem of Predatory Lending: Price, 65 MD. L. REV. 707, 794–95 (2006) (using as an example the many disclosures required in consumer loans, which give a “veneer of legality and authority to the loan process”).
evidence that the party required to disclose information often interprets their compliance with the disclosure requirement as granting them license to act more harshly.\textsuperscript{232} Although there is much more research to be done, again, the virtues of disclosure are no longer uncontested.

IV. AGENCY DATABASES

Nascent skepticism with the use of disclosure as a regulatory tool has not stopped disclosure efforts from proliferating. Today, thousands of federal agency websites host hundreds of thousands of agency databases (as of February 2017, over 195,000).\textsuperscript{233} As such, writing about government data is difficult because databases have become so very common.\textsuperscript{234} A comprehensive survey of agency databases would “pointlessly burden” the audience, risking the same information overload often imposed by policymakers that rely on disclosure.\textsuperscript{235} Nevertheless, as becomes quickly obvious,\textsuperscript{236} agency databases are becoming the norm rather than the exception.

Below I survey six of the most salient agency efforts to post searchable data sets online. Of course, there are many more worth discussion—too many for a single article. The following represents a cross-section of databases that try to achieve regulatory goals, using online disclosure of behavior to try to affect that behavior.

A. The FEC’s Campaign Finance Data

For decades, federal campaign finance law has relied on disclosure not only to police limits on campaign contributions and spending, but also to pursue deeper goals of deterring

\textsuperscript{232} Id. at 739 (citing such evidence related to conflict-of-interest disclosures).

\textsuperscript{233} DATA.GOV, supra note 201 (“Search over 194,126 data sets” as of March 22, 2016). Note that as of April 1, 2016, 158,301 data sets were published by the federal government, with the rest published by sub-federal units of government, including a smattering from the private and nonprofit sectors. Data Catalog: Organization Types, DATA.GOV, http://catalog.data.gov/dataset#sec-organization_type (last visited July 14, 1017) [http://perma.cc/L5SY-5M9S] (See “Organization Types” in the left sidebar).

\textsuperscript{234} Ben-Shahar & Schneider, supra note 17, at 652.

\textsuperscript{235} Id.

\textsuperscript{236} CORTEZ, supra note 143, at app. E.
corruption and the appearance of it.\textsuperscript{237} Indeed, disclosure has been, perhaps, the one leg of the campaign finance law tripod to be spared by the Supreme Court.\textsuperscript{238} Unlike limits on campaign contributions and expenditures, disclosure requirements have endured repeated First Amendment challenges—from early cases like \textit{Burroughs}\textsuperscript{239} to more strident recent cases like \textit{Citizens United} and \textit{McCutcheon}.\textsuperscript{240} Of course, in the seminal case \textit{Buckley v. Valeo},\textsuperscript{241} the Court upheld various disclosure requirements introduced in the Federal Election Campaign Act of 1971 (FECA) and the FECA Amendments of 1974.\textsuperscript{242} In the ensuing decades, as courts invalidated various restrictions on campaign contributions and expenditures on First Amendment grounds, disclosure requirements endured. Thus, through attrition, disclosure has become the preferred choice for regulating money in politics.\textsuperscript{243} Indeed, recent reform proposals would address lingering problems with campaign finance regulation by using even more disclosures, for example, by trying to triangulate FEC data with data from other agencies like the SEC and IRS.\textsuperscript{244}

Today, the FEC maintains several searchable online databases on its website, which are centralized on the FEC's Campaign Finance Disclosure Portal.\textsuperscript{245} Users can search FEC data based on reports required of federal candidates, parties,

\begin{itemize}
\item \textsuperscript{237} \textit{Buckley v. Valeo}, 424 U.S. 1, 66–68 (1976).
\item \textsuperscript{238} The other two legs being limits on campaign contributions and limits on expenditures.
\item \textsuperscript{239} \textit{Burroughs v. United States}, 290 U.S. 534 (1934) (upholding an indictment under the Federal Corrupt Practices Act).
\item \textsuperscript{241} \textit{Buckley}, 424 U.S. at 66–68.
\item \textsuperscript{244} \textit{See, e.g.}, Lucian A. Bebchuk & Robert J. Jackson, Jr., \textit{Shining Light on Corporate Political Spending}, 101 Geo. L.J. 923, 931, 935–36 (2013).
\end{itemize}
committees, donors, and lobbyists, among others. The data are generally searchable by name, date, and location, and are presented in list, map, and chart form, making them more accessible to lay users. Thus, for example, a user can learn how much money a federal candidate in her district has raised and spent, or how much a certain political action committee (PAC) has dedicated to electioneering, or find detailed information about independent expenditures or bundled contributions. The FEC’s Disclosure Data Catalog publishes these datasets in downloadable .CSV, .XML, or .XSD formats, thus making them more useful for sophisticated users.

Despite their broad scope, FEC databases pay special attention, of course, to the activities of candidates and committees. Users can search for federal candidates and political committees by name and view on a single page all reports filed by that person or committee, including a multi-year summary of the money they raised and spent.

Like many other agencies, the FEC also publishes online searchable databases of enforcement records. Its Enforcement Query System is a searchable depository of FEC enforcement documents, including complaints, responses, settlements, and other relevant documents. The system includes, for example, a searchable and sortable list of parties required to pay
administrative fines for violating reporting requirements. Users can view the name of the party fined, the type of report filed late (or not at all), the amount of the fine, and any related candidate information.

Congress has recognized that data held by the FEC can be more meaningful when combined with data held by other institutions, including Congress itself. For example, the Lobbying Disclosure Act of 1995 requires “lobbyists” to register with the Clerk of the House and the Secretary of the Senate and disclose their lobbying activities, including who they lobbied and on what issues, bills, or other government action. The Act requires House and Senate officials to use “computerized systems” with “coding” and “cross-indexing” to “maximize public access to materials filed.” Congress also requires these reports to be available over the internet. In 2007, frustrated with the slow rate of publication online, Congress amended the law to require publication online in a searchable, sortable, and downloadable format. The 2007 amendments also linked lobbying information with campaign contribution data reported to the FEC, so that users can better track financial ties between lobbyists and public officials. Thus, the amendments demonstrate how data regimes can be dynamic rather than static.

But, as with many other disclosure-based regulatory schemes, researchers question how effectively campaign finance disclosures have achieved their stated goals of preventing corruption and the appearance of it. Despite

254. Id.
256. Id. § 1605.
257. Id. § 1605(a)(4),(9).
relatively thoughtful presentation and formatting of the data, the data themselves are not very reliable. Compliance with FEC reporting requirements is spotty, filings are often selective and incomplete (the FEC does not require filings to be complete to be accepted), FEC enforcement is limited in several important ways, data collection is not always standardized, and thus the data is often unreliable. To note just one example, although the FEC tracks campaign contributions, it does not track individual contributors well because there is no unique identifier assigned to them. Thus, the data “gives the illusion of transparency, but functions instead to obscure the most pertinent financial constituencies in a sea of data.” These design failures mean that the intended users—voters, intermediaries, and regulators—are relying on flawed data, even if it is presented in multiple formats.

Moreover, scholars have questioned at length whether voluminous campaign finance data succeeds in achieving its stated goals and have identified a long list of preconditions necessary for the data to do so. For example, Malbin and Gais identify several requirements that align with the wisdom applied to other disclosure-based regimes—the disclosure must be accurate, usable, and accessible, both to the intermediaries who can synthesize it, and to voters who might cast votes based on it. The preconditions are substantially similar to the conditions prescribed by Fung, Graham, and Weil (which I discuss in more detail in Part V, infra). Numerous articles by campaign finance scholars suggest specific improvements to the FEC’s data practices, including how it collects, processes, and disseminates data, and enforces compliance. Simple fixes, like assigning unique identifiers to contributors and

262. Id. at 1484.
263. Id.
264. Id. at 1486–89.
266. Heerwig & Shaw, supra note 243, at 1475 (citing FUNG ET AL., supra note 15, at 11).
267. For just a small sampling, see, e.g., Heerwig & Shaw, supra note 243; Hasen, supra note 243; Richard Briffault, Updating Disclosure for the New Era of Independent Spending, 27 J.L. & POL. 683 (2012).
allowing the FEC to conduct random audits, might greatly improve the quality of the data.\footnote{268}{Heerwig & Shaw, supra note 243, at 1494–99.}

Finally, as with other data regimes, third-party intermediaries play an important role in translating and synthesizing the government’s campaign finance data. Groups like the Center for Responsive Politics (CRP), the Campaign Finance Institute (CFI), and the Sunlight Foundation rely on FEC data to provide more digestible information to the public.\footnote{269}{Id. at 1487–88.} Indeed, an old 2002 survey found that political journalists relied on the CRP website more than any other, with the FEC’s own site ranking third.\footnote{270}{Id. (citing ALBERT L. MAY, INST. FOR POLITICS, DEMOCRACY, & THE INTERNET, THE VIRTUAL TRAIL: POLITICAL JOURNALISM ON THE INTERNET 22 (2002), http://www.pewtrusts.org/~media/legacy/uploadedfiles/wwwpewtrustsorg/reports/the_practice_of_journalism/pponlinejournalistpdf.pdf [https://perma.cc/H3X2-M8JK]).}

Another group called MAPLight.org mashes up the publicly available voting records of members of Congress with campaign finance data, trying to find correlations.\footnote{271}{MAPLIGHT, http://www.maplight.org (last visited July 22, 2017) [https://perma.cc/5EFN-MCCA].} MapLight gathers data from not only the FEC, but also from the CRP, which runs OpenSecrets.org and FollowTheMoney.org (for California data).\footnote{272}{Find Contributions, MAPLIGHT, http://maplight.org/us-congress/contributions (last visited July 22, 2017) [https://perma.cc/7T2W-8S8Z].} However, because the FEC data is so flawed, scholars worry that it is not suitable for use by researchers and other informational intermediaries who might otherwise be able to extrapolate larger patterns or trends.\footnote{273}{Heerwig & Shaw, supra note 243, at 1490.} Thus, the FEC’s regime is characterized by noble sentiments but flawed data.

Therefore, although there is low confidence in the FEC’s data sets, there remains clear demand for such data and somewhat feasible fixes that are likely to appeal to many interested parties, such as voters, watchdogs, and the disclosers themselves. Improved data quality—using the methods recommended in Part V and the recommendations of earlier projects—should enable better, more frequent uses of the information and perhaps even affect disclosers’ behavior.
B. The EPA’s Toxic Release Data

Like campaign finance law, environmental law relies on public disclosure to regulate underlying conduct. Environmental scholars sometimes refer to this as “regulation by revelation”—leveraging the threat of public backlash to change the underlying behavior that leads to pollution. As the EPA’s former General Counsel observed, “[i]nformation . . . can be a supplement, sometimes even an alternative, to regulation. When broadly available, information can change behavior.”

This logic undergirds several well-known environmental statutes that are predicated on disclosure, such as the Clean Air Act; the Emergency Planning and Community Right-to-Know Act; the Federal Insecticide, Fungicide, and Rodenticide Act; the National Environmental Policy Act; the Safe Drinking Water Act; and the Toxic Substances Control Act. The commonality is that these statutes all “place affirmative duties on federal agencies to make information available to the public.” In combination, these laws “seem to provide a right of public access to virtually all environmental information in the hands of the federal government.”

In fact, perhaps the most well-known agency database dates back to 1986, when Congress passed the Emergency Planning and Community Right-to-Know Act, requiring the EPA to establish a national toxic chemical inventory, with the information “in a computer data base . . . accessible to any person.” The database became the Toxics Release Inventory (TRI) program.

The TRI program requires facilities to report their production and release of roughly 650 dangerous chemicals.

276. Vladeck, supra note 64, at 1788 (citations omitted).
277. Id.
278. Id.
280. 30 C.F.R. § 372.65 (2016); see also 42 U.S.C. § 11023.
The EPA first reported TRI data in 1989, and first published it online in 1998. Today’s version allows users to search for toxic release data by state, county, city, or ZIP code, and will generate a customized “factsheet” based on the query, listing all facilities in the geographic area and the quantity of chemicals they release. The data are presented in colorful chart, graph, and map forms. For each reporting facility, the EPA maintains a “Facility Profile Report” with more granular data regarding the amount of chemicals managed, released, or transferred. Still more data is available by link to each company’s full reports in the EPA’s Envirofacts database.

TRI has been widely hailed for having a “significant impact on firm-level emissions” and even inspiring several other disclosure-based regulatory efforts, both in the United States and overseas. The initial success of the TRI program even surprised the EPA and environmental groups who had toiled for years to regulate toxic pollution. Ten years after initiating the TRI program, the amount of pollution released had dropped by half. In fact, initial media interest in the program and the resulting threat of negative publicity appeared to have a powerful impact on companies—even before the first reports were required, executives of some companies promised to reduce their toxic outputs by as much as 90 percent. As one of the earliest programs of its kind, the TRI

288. Id.
289. Id.
has been applauded by many as the best example of regulation via disclosure.290

Perhaps inspired by its own success, the EPA now publishes hundreds of datasets online. The EPA website publishes so many datasets that it includes several landing pages that help users search for and navigate the data available.291 Some of the more well-known datasets after TRI include the EPA's Enforcement and Compliance History Online (ECHO) website, which allows users to search for recent and historical enforcement actions, including the last date of inspection.292 The EPA currently lists ninety-six datasets on its website,293 with 1,738 listed on Data.gov.294 The agency even hosts an online discussion forum for data developers.295

Despite the initial success of TRI, Fung, Graham, and Weil found that in comparison with seven other disclosure regimes, toxic release disclosure was only moderately successful, at best.296 They found, in particular, that toxic release data is not embedded in potential users' decision-making, as “[m]ost home buyers, renters, job seekers, consumers, and investors do not consider toxic pollution” when making decisions.297 The TRI data did succeed in better informing policymakers, such as Congress and the EPA itself.298 And many manufacturers quickly embedded the new data into their decision-making, recognizing the reputational and regulatory consequences.299

290. See, e.g., Esty, supra note 4; Pederson, supra note 274.
297. Id. at 85.
298. Id. at 86.
299. Id.
But researchers have gradually curbed their enthusiasm based on studies showing flaws in reporting (including inaccurate data) and longitudinal studies showing less impact on potential users’ and disclosers’ actual conduct.\textsuperscript{300} Scholars have long worried that EPA datasets are “patchy” and “unreliable.”\textsuperscript{301}

In the last decade, despite its success, the EPA has “drastically scaled back the information made public” under the TRI program.\textsuperscript{302} Prior to 2006, the EPA required facilities to report information regarding any chemical release over 500 pounds.\textsuperscript{303} But a 2006 rule increased the threshold to 5,000 pounds, provided the total annual release into the environment does not exceed 2,000 pounds.\textsuperscript{304} The GAO criticized the EPA’s rule as reducing the “quantity and detail of information” released to communities.\textsuperscript{305} Under the Trump administration, one can envision even more drastic changes to the TRI program.

But these stories have not deterred scholars and policymakers who still believe that data is the path to environmental regulation. For example, Daniel Esty argues that “[a]s data become easier to analyze and disseminate, and dramatically less costly to acquire and use, our capacity to identify and solve environmental problems will increase substantially.”\textsuperscript{306} Indeed, he predicts that information technologies will enable an “environmental revolution perhaps as important as that which launched the modern environmental movement” in the 1960s.\textsuperscript{307}

Of course, even data optimists believe that there are significant challenges in producing environmental data that are reliable and usable.\textsuperscript{308} But, as in other fields, environmental scholars see great promise in third-party watchdogs and data intermediaries translating voluminous

\textsuperscript{300} Id. (citing studies).
\textsuperscript{301} Esty, supra note 4, at 156.
\textsuperscript{302} Vladeck, supra note 64, at 1791.
\textsuperscript{304} Id. at 76,937.
\textsuperscript{305} JOHN B. STEPHENSON, U.S. GOV’T ACCOUNTABILITY OFFICE, ENVIRONMENTAL RIGHT-TO-KNOW: EPA’S RECENT RULE COULD REDUCE AVAILABILITY OF TOXIC CHEMICAL INFORMATION USED TO ASSESS ENVIRONMENTAL JUSTICE 16 (2007); Vladeck, supra note 64, at 1791 n.33.
\textsuperscript{306} Esty, supra note 4, at 119.
\textsuperscript{307} Id.
\textsuperscript{308} Id. at 171–74.
data into usable heuristics for consumers, perhaps creating a race to the top among regulated firms. Nevertheless, not even the most well-resourced and well-meaning intermediaries can cure flawed data. The EPA’s many data sets, then, suffer more for quality than for quantity.

C. The CFPB’s Consumer Complaint Data

The Consumer Financial Protection Bureau (CFPB) is the newest agency among those surveyed here, and its newness makes it interesting. Born of the 2008 financial crisis, the Bureau was “designed in a world of new technology.” In short, the CFPB “is a new agency operating under a new statute and is on the frontier of the open data trend.” The Bureau’s Office of Consumer Response operates a process for consumers to file complaints regarding financial products and services, which quite notably are published online by the CFPB in a massive, searchable Consumer Complaint Database, identifying companies by name.

The CFPB website allows consumers to file complaints for eleven categories of financial products, including mortgages, student loans, and credit cards. Complaints can specify the name of the company, the type of product or service at issue, the type of problem with it, and the consumer’s ZIP code, all of which the Bureau authenticates and sends to the company for a response. The published data is searchable and sortable,
and includes not only the company’s name, but also its response (if any) and whether the response was timely or further disputed by the customer.\footnote{315} Companies select their responses via a pull down menu that includes options such as “Closed with monetary relief,” “Closed with non-monetary relief,” “Incorrect company,” and “In progress.” Companies have a total of sixty days to respond, and late responses are tagged by the CFPB as “Past due” or “No response” if the delay exceeds thirty days.\footnote{317} Each complaint and response is published, but only if it meets numerous publication criteria.\footnote{318}

In 2014, the Bureau expanded the database (reversing its previous position) by posting narrative commentary by consumers.\footnote{319} It originally declined to do so, citing privacy and the risk of disclosing consumers’ personal information.\footnote{320} But after considering industry objections (including the potential harm to company reputations) and devising ways to scrub the information of personally identifiable information, the Bureau finalized its plan to include consumer narratives in the Complaint Database.\footnote{321} As a measure of symmetry, the Bureau proposed to allow companies to post their own narrative responses,\footnote{322} but companies preferred to respond with a preset list of “structured” responses, such as “Company acted appropriately,” “Factual dispute,” “Misunderstanding,” and “Opportunity for improvement.”\footnote{323} However, these responses

\footnotesize

\footnote{315. Cortez, supra note 143, at 62–63.}
\footnote{317. Id. at 24.}
\footnote{318. The Bureau will not publish complaints if they are missing critical information, have been referred to other regulators, are duplicative, would reveal trade secrets, are fraudulent, or identify the wrong company. Id. at 26; Final Policy Statement: Disclosure of Consumer Complaint Data, 78 Fed. Reg. 21,218, 21,225 (Apr. 10, 2013).}
\footnote{320. Disclosure of Consumer Complaint Data, 77 Fed. Reg. at 37,568.}
\footnote{323. COMPANY PORTAL MANUAL, supra note 316, at 29.
are optional; companies need not select one for publication.\textsuperscript{324} Today, then, the Complaint Database includes narrative descriptions of consumers’ problems (if they choose to narrate them), which can make the problems more concrete and compelling than displaying relatively sanitized data entries alone.\textsuperscript{325}

The intended users of the Complaint Database are consumers, researchers, the Bureau, other regulators, and even the subjects of the complaints themselves—companies.\textsuperscript{326} The Bureau and consumer groups emphasize that the primary intended beneficiaries are consumers, and publishing complaint data is a “public service” that can “empower” consumers and help them avoid “bad actors” in these markets.\textsuperscript{327} The Bureau itself emphasizes that “disclosure is one of the best tools government agencies can use.”\textsuperscript{328} Former Bureau Director Richard Cordray encouraged “the public, including consumers, the companies that serve them, analysts, data scientists, civic hackers, developers, policymakers, journalists, and academics, to analyze, augment, and build on the public database.”\textsuperscript{329} Bureau staff also hope that intermediaries develop mobile apps and other information products based on complaint data.\textsuperscript{330}

This latter aspiration—that intermediaries will use the government’s data—is being realized, at least modestly so far. Academics are publishing empirical analyses of the CFPB’s data.\textsuperscript{331} Public interest research groups (PIRGs) are producing reports of certain financial product categories, like credit cards


\textsuperscript{325} CORTEZ, supra note 143, at 65.


\textsuperscript{328} Id. at 21,220.


\textsuperscript{330} CORTEZ, supra note 143, at 66.

\textsuperscript{331} See, e.g., Ian Ayres et al., Skeletons in the Database: An Early Analysis of the CFPB’s Consumer Complaints, 19 FORDHAM J. CORP. & FIN. L. 343 (2014).
and debt collection, based on Bureau data. In 2015, the rankings-crazed *U.S. News & World Report* ranked credit cards by relying, in part, on data from the Consumer Complaint Database.

Similar to other agencies, the CFPB hopes that the act of data publication itself will encourage companies to improve their underlying behavior. Bureau staff report that complaints have inspired some companies to address potential problems of their own, such as long customer service phone trees. Indeed, one of the Bureau’s original aspirations was that published complaints would encourage companies to compete in a race to the top, based on how they handled customer service and customer complaints. Management consulting firms now advise companies to “turn what they hear from the CFPB’s consumer complaint database into a business advantage.” Bureau staff also report that some companies have tied executive compensation to how well the company has responded to published complaints. In short, the Bureau sees its role somewhat modestly as a publisher—providing a window into a dialogue between companies and their customers. Once complaints are published, “[t]he marketplace of ideas then does the rest.”

Of course, the Consumer Complaint Database has not been without controversy, particularly to the firms identified in it. Firms and industry groups filed scores of public comments objecting to Bureau proposals to publish complaint data. The objections fall into seven general categories: (i) it is unfair to


334. CORSZ, supra note 143, at 66.

335. Id.


338. CORSZ, supra note 143, at 66.

339. Id. at 66–67.

publish complaints that are not verified by the Bureau; (ii) the complaints are self-selecting and thus are non-random and non-representative of customer experiences; (iii) the data lack context and might appear to be endorsed by the Bureau; (iv) the data are susceptible to manipulation and fraud; (v) the companies will suffer reputational harm in the media and might draw the attention of plaintiffs’ lawyers; (vi) the database is overinclusive because it includes complaints that are not necessarily legal or regulatory violations; and (vii) the Bureau lacks statutory authority to publish complaint data online.\footnote{CORTEZ, supra note 143, at 67–70.} The Bureau responded at length to these objections in the Federal Register, showing a basic sensitivity to industry concerns, though disagreeing with industry conclusions that the Bureau should not publish the data online.\footnote{Id. at 67–71.}

Moreover, the Bureau has fielded industry complaints about the database in several formats, including in public comments filed during notice and comment periods, in letters to the Director, and in complaints to the Bureau’s Ombudsman.\footnote{Id. at 75.} The Federal Reserve’s Office of Inspector General, which has oversight responsibility for the CFPB, has audited the database “to assess the effectiveness of the [CFPB’s] controls over the accuracy and completeness of its public-facing Consumer Complaint Database.”\footnote{FED. RESERVE OFFICE OF INSPECTOR GEN., 2015-FMIC-C-016, OPPORTUNITIES EXIST TO ENHANCE MANAGEMENT CONTROLS OVER THE CFPB’S CONSUMER COMPLAINT DATABASE, AUDIT REPORT 1 (2015).} Still, the database remains a target for industry, and a bill proposed in the 115th Congress would remove it from the public domain,\footnote{Financial CHOICE Act of 2017, H.R. 10, 115th Cong. § 725 (2017–18). H.R. 10 passed the House on June 8, 2017.} sparking an outcry from consumer watchdogs and public interest groups.\footnote{See, e.g., Ed Mierzwinski, Financial Choice Act: A Cruel Choice for the CFPB & Consumers, U.S. PIRG (Apr. 24, 2017), http://www.uspirg.org/blogs/eds-blog/usp/financial-choice-act-cruel-choice-cfpb-consumers} 

Thus, as a new database being run by a new agency, the Consumer Complaint Database continues to be refined and new uses (and objections) continue to emerge. Although the database seems well-designed to defuse potential criticisms, it
is still susceptible to being undermined by an unsympathetic Trump administration and a Republican-led Congress, per the pending bill. Nevertheless, the CFPB’s efforts can serve as a model for other agencies considering publishing consumer reports or complaints.

D. The CPSC’s Product Safety Data

The Consumer Product Safety Commission’s database, SaferProducts.gov, enjoys more clear statutory authority and attention from Congress than most agency databases. Since the 1970s, Congress has required the CPSC to “protect the public against unreasonable risks of injury” and “assist consumers in evaluating the comparative safety of consumer products.” But in 2008, Congress passed the Consumer Product Safety Improvement Act, requiring the Commission to create a searchable online database of product safety incidents. The law required the database to include “reports of harm relating to the use of consumer products,” including reports from consumers, physicians, state and local governments, and others. Each report must describe the product or substance at issue, identify the manufacturer or labeler, and describe the harm reported. Supporters hailed the effort to “empower consumers,” “expedite recall disclosure,” and “enhance a family’s right to know about dangerous and defective products on the market.” The CPSC launched SaferProducts.gov in March 2011. Today, SaferProducts.gov includes a searchable online database of thousands of “reports of harm” related to identified

347. H.R. 10.
products. Like the CFPB’s Consumer Complaint Database, the site has a portal for users to report incidents and a portal for companies to respond to them. The database is keyword-searchable with advanced search options that include product name, company or brand name, and the product model. The advanced search also allows users to search for incidents by date, location, the “Victim’s Age,” and by “Injury Information,” including reports of death.

Interestingly, the Act also requires the CPSC to “provide clear and conspicuous notice to users of the database that the Commission does not guarantee the accuracy, completeness, or adequacy of the contents of the database.” As such, SaferProducts.gov includes a disclaimer that tracks this language almost verbatim.

An innovation in SaferProducts.gov that might be emulated by other agency databases is that the CPSC allows manufacturers to comment on reports and object to inaccuracies. This feature derives from the Act itself, which dictates that the CPSC “shall” provide manufacturers and labelers an opportunity to comment on incident reports and request that such comments be included in the report posted.


358. Id.


361. CORTEZ, supra note 143, at 20–21.
Likewise, the CPSC must consider objections that any information in a report is “materially inaccurate,” which the CPSC defines as information “that is false or misleading, and which is so substantial and important as to affect a reasonable consumer’s decision-making about the product.” Congressional attention to the validity of CPSC reports is not an accident. A series of inaccurate product safety warnings inspired Congress to amend the Consumer Product Safety Act in 1981 to improve procedural safeguards for Commission announcements.

The procedures for SaferProducts.gov are already being tested. In a recent case, Company Doe v. Tenenbaum, a company anonymously challenged an inaccurate product safety report in the database. The court found that the CPSC database “bears the Government’s stamp of approval through its publication on an official website that, by its terms, is a repository of reports regarding ‘unsafe products.’” The court sustained the company’s challenge, somewhat remarkably, finding that the CPSC posting “materially inaccurate” information on SaferProducts.gov was not only “final agency action” under the Administrative Procedure Act (APA), but also violated the CPSC’s own regulations. In the vast majority of similar cases, courts find that negative statements or other adverse public disclosures by agencies are not “final” under the APA and thus cannot support a cause of action against the government. The CPSC litigation thus could provide a roadmap for other litigants similarly aggrieved.

Although SaferProducts.gov’s pre- and post-publication procedures might serve as a useful model for other agency databases, challenges certainly remain. For instance, it is not

---

363. Id. § 2055a(c)(4).
367. Id. at 597.
368. CORTEZ, supra note 143, at app. C (surveying similar cases and finding few successful challenges).
370. Cortez, supra note 75, at 1443–44.
clear how often consumers use the data to make purchasing decisions (perhaps the data is more useful to distributors and others in the supply chain), nor is it clear whether publishing the data inspires manufacturers to improve product safety apart from other requirements. And presentation remains critical. A GAO review found that some users were confused about the purpose of SaferProducts.gov, viewing it as a site featuring safe rather than unsafe products. The “upbeat” name of the database might have contributed to such misperceptions. The GAO also found that although the CPSC had used various methods to inform consumers about SaferProducts.gov (which at the time generated at least 100,000 page visits a month), including use of social media, the agency could do more to publicize the resource.

Perhaps third-party informational products could leverage the data more effectively—for example, by creating mobile phone apps that can scan products in the aisle and display a brief product safety profile generated by data from SaferProducts.gov. Or, large online retailers like Amazon.com might find novel uses for such data, triangulating it with user reviews. Like other agencies, the CPSC’s open government website allows users to download raw data files, with voluminous links to the agency’s open data plans. Third-party intermediaries might be particularly helpful at translating “big” product safety data into comparative ratings accessible to ordinary consumers.

E. Medicare’s Quality Data

The Centers for Medicare and Medicaid Services (CMS) administers Medicare, our federal health insurance program for the elderly and chronically disabled. Because Medicare pays

---

373. CPSC: AWARENESS, USE, AND USEFULNESS, supra note 371, at 12–14.
thousands of nongovernmental physicians, hospitals, and other types of providers for care, beneficiaries often must choose among multiple local providers—sometimes scores or even hundreds of such providers. But for a Medicare beneficiary that needs cardiac surgery, for example, choosing a specific surgeon in a specific hospital can be daunting.\textsuperscript{375}

To facilitate such decisions, CMS operates five searchable databases that compare Medicare providers—Hospital Compare,\textsuperscript{376} Physician Compare,\textsuperscript{377} Nursing Home Compare,\textsuperscript{378} Home Health Compare,\textsuperscript{379} and Dialysis Facility Compare.\textsuperscript{380} The search functions all work in roughly the same way. Users can search for providers by city, state, or ZIP code, and then view a list of results within a twenty-five mile radius, each sortable by different criteria. For example, the Nursing Home and Dialysis Facility databases allow users to sort facilities based on how they rate on a five-star scale.\textsuperscript{381} Hospitals are sortable by distance, by whether they offer emergency services, and by “hospital type.”\textsuperscript{382} Home Health facilities are listed by the type of care offered, such as physical therapy and occupational therapy.\textsuperscript{383} Physician searches are more complicated (and are probably the least useful of the five),

\begin{itemize}
\item For in-depth critiques of the consumer-driven health policy and the heroic assumptions it often makes about patients’ capacity to understand and make important medical and spending decisions, see \textsc{Timothy Stoltzfus Jost}, \textit{Health Care at Risk: A Critique of the Consumer-Driven Movement} (2007) and \textsc{Mark A. Hall} & \textsc{Carl E. Schneider}, \textit{Patients as Consumers: Courts, Contracts, and the New Medical Marketplace}, 106 \textit{Mich. L. Rev.} 643 (2008).
\item \textsc{Hospital Compare}, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/hospitalcompare/search.html (last visited Aug. 6, 2017) [https://perma.cc/SG2Q-RVBY].
\item \textsc{Physician Compare}, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/physiciancompare/search.html (last visited Aug. 6, 2017) [https://perma.cc/E7DN-6P59].
\item \textsc{Nursing Home Compare}, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/nursinghomecompare/search.html (last visited Aug. 6, 2017) [https://perma.cc/E7DN-6P59].
\item \textsc{Home Health Compare}, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/homehealthcompare/search.html (last visited Aug. 6, 2017) [https://perma.cc/N6Q7-K3BQ].
\item \textsc{Dialysis Facility Compare}, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/dialysisfacilitycompare/#search (last visited Aug. 6, 2017) [https://perma.cc/R278-XGKV].
\item See \textsc{Nursing Home Compare}, supra note 378; \textsc{Dialysis Facility Compare}, supra note 380.
\item See \textsc{Hospital Compare}, supra note 376.
\item See \textsc{Home Health Compare}, supra note 379.
\end{itemize}
requiring the user to also search for a physician’s name, specialty, or medical condition to help narrow the results.\textsuperscript{384}

The five databases offer quite distinct data on “quality.” Hospital Compare includes data on over 4,000 hospitals nationwide,\textsuperscript{385} allowing users to compare up to three hospitals at a time, using six categories: “Survey of Patients’ Experiences,” “Timely & Effective Care,” “Complications, Readmissions & Deaths,” “Use of Medical Imaging,” and “Payment & Value of Care.” Each tab, moreover, includes several subcategories of information. For example, “Timely & Effective Care” is divided into ten subcategories, such as “Heart Attack Care” and “Stroke Care.”\textsuperscript{386} Results are compared to state and national averages as reference points.

The data displayed, however, are less satisfying than the categories might suggest. A frustrating proportion of data for hospitals is listed as “Not Available,” with numbered footnotes explaining why.\textsuperscript{387} The tab titled “Payment & Value of Care” might tantalize health policy wonks, but unfortunately, comparative data often is not displayed directly (e.g., “Get Results for This Hospital” is displayed when searching for Medicare spending per beneficiary) or meaningfully (e.g., “No Different than the National Average Payment” is frequently displayed). Also, “Value of Care” metrics tend to display charts full of “No Different than the National Rate” and “No Different than the National Average Payment.”

An afternoon on Hospital Compare leaves one with the impression that there are probably too many categories and subcategories of data for the average person to make meaningful comparisons, particularly when much of the data is “Not Available.” Thus, although CMS touts Hospital Compare as “an important tool for individuals to use in making decisions.

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{384} Physician Compare: About the Data, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.medicare.gov/physiciancompare/staticpages/data/aboutthedata.html (last visited Aug. 6, 2017) [https://perma.cc/6NDV-4TG7].
\item \textsuperscript{385} What is Hospital Compare?, MEDICARE.GOV, http://www.medicare.gov/hospitalcompare/About/What-Is-HOS.html (last visited Aug. 6, 2017) [https://perma.cc/ZFG3-DC4C].
\item \textsuperscript{386} The ten categories are Heart Attack Care, Heart Failure Care, Pneumonia Care, Surgical Care, Emergency Department Care, Preventative Care, Children’s Asthma Care, Stroke Care, Blood Clot Prevention & Treatment, and Pregnancy & Delivery Care.
\item \textsuperscript{387} Hospital Compare, Footnotes, MEDICARE.GOV, https://www.medicare.gov/hospitalcompare/Data/Footnotes.html (last visited Aug. 6, 2017) [https://perma.cc/P4VF-HZ7D].
\end{itemize}
\end{footnotesize}
about health care options,” it is probably more successful at meeting CMS’s other aspiration as “a way to encourage accountability of hospitals for the care they provide to patients.” Still, given how incomplete the data is, it is questionable whether it meets even this goal.

Physician quality data is even more limited, perhaps reflecting physicians’ longstanding concerns over reputation and liability. Physician Compare includes only directory-type information, such as the name and location of the physician, hospital admitting privileges, and information about physician’s education and board certifications. But there are few quality metrics. Physician Compare does feature a Physician Quality Reporting System, but it merely asks physicians to report whether they follow certain best practices. Physicians who report to CMS are then given performance scores. Unfortunately, the system is voluntary, and only some Physician Compare profiles include their performance scores. Although there are over 200 reportable quality measures, few are listed on Physician Compare.

389. For a discussion of how transparency often challenges professional reputation while privacy reinforces it, including a discussion of corporate reputation, see William M. Sage, Reputation, Malpractice Liability, and Medical Error, in ACCOUNTABILITY: PATIENT SAFETY AND POLICY REFORM 159 (Virginia A. Sharpe, ed. 2004).
391. Id.
392. The site itself explains how the data is limited:
At this time, not all health care professionals and group practices have quality measure performance scores on their Physician Compare profile page. Some health care professionals and group practices are committed to providing high quality care, but do not have quality measures. There are many reasons why health care professionals and group practices do not have quality measures for review. There are more than 200 quality measures that can be reported to CMS through multiple reporting methods. However, only certain measures reported through some of the reporting methods are currently available. Over time, more quality measures will be added to Physician Compare and more health care professionals and group practices will have measures available.

Id.
Still, the star ratings available on databases like Nursing Home Compare and Dialysis Facility Compare are easily accessible across a number of facilities and probably do help users searching for facilities nearby. Of course, Medicare quality data has many potential uses for many potential audiences, ranging from patients looking to choose the best surgeon or hospital, to providers evaluating their own performance, to policymakers seeking to understand broader trends in care.\footnote{Madison, supra note 30, at 1625.}

These databases had humble beginnings. In 2005, CMS first published ten different quality measures for hospitals across the United States, sprouting from a partnership between CMS and the Hospital Quality Alliance.\footnote{Id. at 1626; Hospital Compare, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinitiatives/hospitalcompare.html (last visited July 22, 2017) [https://perma.cc/PY8M-5H2K]. The lineage of quality reporting in health also stretches back to precursors, like patient-centered nursing home standards in the wake of scandals involving nursing home deaths, passed under the Omnibus Budget Reconciliation Act of 1987, Pub. L. No. 100-203.} In subsequent years, CMS has continued to add data from a variety of sources, including patient experience ratings, mortality rates for certain conditions, and hospital readmission rates, among many others.\footnote{Madison, supra note 30, at 1626 (citing Hospital Compare, supra note 394).} Today’s Compare databases are an amalgam of data from a variety of sources, gradually added like ornaments to a Christmas tree.\footnote{As with other agencies, the raw data files at a separate site. See DATA.MEDICARE.GOV, http://data.medicare.gov (last visited Aug. 6, 2017) [https://perma.cc/9TVL-DJNM].}

Perhaps the richest potential source of data is Medicare claims data. CMS processes over a billion Medicare claims each year,\footnote{Medicare Fee-for-Service, Payment Accuracy, PAYMENTACCURACY.GOV, https://paymentaccuracy.gov/program/medicare-fee-for-service/ (last visited Aug. 21, 2017) [https://perma.cc/985Q-LGJP]; Madison, supra note 30, at 1607.} with each claim including multiple data points, including whom Medicare is paying and for what. Medicare claims data have long been used by academics, government researchers, and providers themselves to better understand the U.S. health care system.\footnote{Madison, supra note 30, at 1609–10.} The growth of “big data” in health
care, in fact, roughly parallels and depends on the growth and accessibility of Medicare claims data.\textsuperscript{399}

Although Medicare has released various data to the public for years, CMS was long restricted from releasing physician claims data by court order.\textsuperscript{400} In 2014, a year after the order was lifted, CMS released claims data for over 880,000 providers.\textsuperscript{401} The trend has been followed by states—sixteen of which have created all-payer claims databases that compile claims data from almost all payers in the state, including public and private insurers.\textsuperscript{402} However, in 2016 the Supreme Court threw a wrench in these efforts when it interpreted federal law as preempting state laws mandating reporting from self-insured employer plans.\textsuperscript{403} As is common for many data regimes, agencies often take two steps forward, then one step back.

Traditionally, claims data has meant merely payment data, without regard to the quality of care being provided. In 2003, Congress amended Medicare’s payment formulas to encourage hospitals to report quality data,\textsuperscript{404} and today over 1,300 hospitals report data about infection rates, mortality rates, and other quality indicators in order to boost their Medicare reimbursements,\textsuperscript{405} all of which feed into the Hospital Compare site. Medicare extended these incentives from hospitals to physicians in 2006 and, by 2017 will require it.\textsuperscript{406} Yet, realizing that not everyone is covered by Medicare,

\textsuperscript{399} Id. at 1610.
\textsuperscript{406} Madison, supra note 30, at 1613; CTES. FOR MEDICARE & MEDICAID SERVS., PHYSICIAN QUALITY REPORTING SYSTEM (PQRS) OVERVIEW 2–3 (Aug. 8,
Congress required CMS to release broad swaths of Medicare claims data regarding hospital care, physician care, prescription drugs, and other goods and services to enable private entities to add Medicare data to other data, on the condition that such entities generate publicly accessible quality ratings.\(^{407}\) Thus, various data sources are being combined in novel ways.

Nevertheless, scholars have long questioned the utility of performance data and report cards in the health industry,\(^{408}\) and disclosure more generally has long been a point of interest to health law scholars.\(^{409}\) Patients report that they seldom rely on publicly available data or even more comprehensible report cards of physicians, hospitals, or procedures.\(^{410}\) A 2012 survey found that only 15 percent of patients reviewed online quality rankings or reviews when choosing doctors or hospitals, with the most frequent users being the most educated middle-aged users.\(^{411}\) Patients frequently are not aware of the information, do not understand it, or do not use it.\(^{412}\) Indeed, of the numerous disclosure regimes analyzed by Fung, Graham, and Weil, they found patient safety disclosures to be among the least effective, due to the complexity of the information, cognitive biases, the likelihood the data would be

---


\(^{409}\) See, e.g., Sage, supra note 5.

\(^{410}\) Id.


\(^{412}\) Ben-Shahar & Schneider, supra note 17, at 672, 711–12 (citing studies).
misinterpreted, and the risk of strategic behavior by providers.\textsuperscript{413}

As a result, CMS has tried to “shrink” the voluminous and varied data by translating them into star ratings, using a five-star scale.\textsuperscript{414} Five stars represents facilities that are “much above average,” four stars are “above average,” three stars are “average,” and so on.\textsuperscript{415} Users can access the underlying data on the same site, including charts comparing each facility to the state and national averages.

More problematic is that studies reveal that those being measured—hospitals, physicians, and other providers—can respond in perverse ways to protect their ratings. Providers have been known to avoid sicker or more complicated patients for fear of compromising their scores on outcomes measurements.\textsuperscript{416} For example, a study of cardiac surgery report cards in Pennsylvania found that cardiac surgeons responded to the new disclosure requirement by becoming more reluctant to operate on sicker patients.\textsuperscript{417} The data is decidedly mixed, however. For every finding that public reporting of mortality rates reduced the rate of mortality, there are reports of selection bias by surgeons avoiding more severe, complex cases.\textsuperscript{418}

Moreover, despite the prevalence of doctor and hospital ratings and report cards, it is not clear whether consumers really want them: “Most consumers do not believe clinical quality varies significantly across doctors, hence the low consumer demand for clinical quality report cards.”\textsuperscript{419} One

\textsuperscript{413} Fung et al., supra note 15, at 76–77 (analyzing disclosure requirements in Pennsylvania and New York).
\textsuperscript{414} See, e.g., Nursing Home Compare, supra note 378; Madison, supra note 30, at 1627.
\textsuperscript{415} Five-Star Quality Rating, Nursing Home Compare, Medicare.gov, https://www.medicare.gov/nursinghomecompare/About/HowWeCalculate.html (last visited Aug. 6, 2017) [https://perma.cc/GP9L-QT83].
\textsuperscript{416} See, e.g., David Dranove et al., Is More Information Better? The Effects of “Report Cards” on Health Care Providers, 111 J. Pol. Econ. 555 (2003); Sage, supra note 5, at 1793.
\textsuperscript{417} Schneider & Epstein, supra note 408.
\textsuperscript{418} Fung et al., supra note 15, at 89 (citing studies).
study found that less than one percent of patients knew how their hospital or surgeon was rated under mandated ratings systems.\textsuperscript{420} Thus, ratings might be better in theory than in practice, at least for now.

CMS’s ratings have also generated litigation. A nursing home in Illinois sued Department of Health and Human Services (HHS) and CMS for mistakenly calculating its star rating on Nursing Home Compare, giving the facility two stars out of five rather than four.\textsuperscript{421} The mistaken star rating was published on the CMS website, which did not correct it for almost two years.\textsuperscript{422} The nursing home argued that HHS and the Illinois Department of Public Health, which conducted the underlying inspections, had violated its procedural due process rights under the Fifth and Fourteenth Amendments.\textsuperscript{423} The district court found that although the nursing home’s low star rating probably did affect its reputation, “reputational harm does not require due process protection.”\textsuperscript{424} The court agreed with the government that although a “mistaken rating could have caused some potential patients to look elsewhere for their care,” it did not rise to a property interest.\textsuperscript{425} To qualify, the nursing home would have to show that the reputational harm also included some sort of “change in legal status,” as required by the “stigma-plus” test.\textsuperscript{426} But the nursing home did not present evidence that there was any such change in legal status—such as a ban on referrals to the facility, a change in licensing status or reimbursement status, or some other tangible harm.\textsuperscript{427} Thus, the court called the mistake (and the nearly two-year delay in fixing it) “unfortunate,” but not something rising to a liberty or property interest protected by due process.\textsuperscript{428}

\textsuperscript{420} Epstein, \textit{supra} note 408, at 1694.
\textsuperscript{422} Id.
\textsuperscript{423} Id. at 1011–12.
\textsuperscript{424} Id. at 1012.
\textsuperscript{425} Id. at 1013.
\textsuperscript{426} Id. at 1014.
\textsuperscript{427} Id. at 1014–15.
\textsuperscript{428} Id. at 1018. For a survey of federal opinions between 1974–2014 in which a private party challenged an agency announcement, identifying cases that invoked due process arguments, see CORTEZ, \textit{supra} note 143, at app. C.
Still, quality ratings and other disclosure-based regulation remains all the rage in health policy, with contemporary proposals littered with patient surveys, outcomes data, star ratings or rankings, and of course federal databases like those mentioned above that combine many of these data points. Again, scholars have long been aware of the benefits, burdens, and limitations of using disclosure as a regulatory tool in health care. Yet, it is possible, if not probable, that efforts to replace the Affordable Care Act will rely heavily on disclosure to facilitate market-based reforms.

F. The FDA’s Adverse Event Data

The U.S. Food and Drug Administration (FDA) maintains several online databases that track problems with the products and companies it regulates. For example, the agency publishes several enforcement databases that allow users to search for FDA inspections, warning letters, recalls, and enforcement reports. The FDA also maintains a database of good news—agency product approvals. But perhaps the most

429. In April 2016, for example, Health Affairs dedicated an issue to “Patients’ and Consumers’ Use of Evidence,” with several articles focused on recent data-reporting initiatives targeted at patient use. See Patients’ and Consumers’ Use of Evidence, 35 HEALTH AFF. 1 (2016).
431. See, e.g., Sage, supra note 5.
well-known FDA databases are those that track adverse events associated with pharmaceuticals and medical devices.

The FDA Adverse Event Reporting System (FAERS) includes a database of medication errors and adverse drug events reported to the agency.\textsuperscript{438} Adverse event reporting dates back at least thirty-five years, and perhaps even longer, when the agency received reports by paper.\textsuperscript{439} Manufacturer reports are required by regulation,\textsuperscript{440} but reports by health care professionals and consumers are only voluntary.\textsuperscript{441}

Evolving from previous iterations,\textsuperscript{442} today’s database remains primarily a tool for the FDA to monitor safety problems, rather than a tool aimed for use by the general public. In fact, the FAERS “database” is not really searchable to most users. The FAERS website includes aggregate statistics, as well as links to raw data files that include individual case reports.\textsuperscript{443} But the raw data files are published only in quarterly increments,\textsuperscript{444} and are not amenable to simple searches, as the agency notes (“A simple search of FAERS data cannot be performed with these files by persons who are not familiar with creation of relational databases”).\textsuperscript{445}


\textsuperscript{439} Postmarketing Safety Reports for Human Drug and Biological Products; Electronic Submission Requirements, 79 Fed. Reg. 33,072, 33,073 (June 10, 2014). Indeed, one source notes that the FDA has maintained an adverse event database in some form since 1967. CORTÉZ, supra note 143, at 44 (citing David Gortler, Adverse Event Databases (AERS Database), http://50.63.91.31/FDA-adverse-event-database.html (last visited Aug. 6, 2017) [https://perma.cc/6CQG-7MMP]).

\textsuperscript{440} 21 C.F.R. §§ 310.305 (adverse event reports for drugs marketed without approved new drug approval applications), 312.32 (investigational drug safety reports), 314.80 (postmarketing reporting).

\textsuperscript{441} Questions and Answers on FDA’s Adverse Event Reporting System (FAERS), U.S. FOOD \& DRUG ADMIN., https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/ (last updated May 5, 2016) [https://perma.cc/2C5Q-R3SB] [hereinafter FAERS].

\textsuperscript{442} In September 2012, FAERS replaced the previous Adverse Event Reporting System (AERS), which itself derives from an earlier reboot of the FDA’s pharmacovigilance system. See 74 Fed. Reg. 42,184, 42,185 (Aug. 21, 2009); 63 Fed. Reg. 65,000, 65,030 (Nov. 24, 1998).

\textsuperscript{443} FAERS, supra note 441.


\textsuperscript{445} FAERS, supra note 441.
The FDA also instructs potential users to request individual case safety reports by submitting a FOIA request.\textsuperscript{446} Individual reports are accessible, however, by searching the FDA’s MedWatch website, which aggregates “clinically important safety information” for “human medical products,” including drugs, devices, and biologics.\textsuperscript{447}

The device counterpart to FAERS is MAUDE (Manufacturer and User Facility Device Experience), a database of device adverse events,\textsuperscript{448} which the agency began collecting in 1984.\textsuperscript{449} Today, MAUDE includes both mandatory and voluntary adverse event reports, with downloadable data files, including reports dating back to the 1990s and an online searchable database covering the last ten years.\textsuperscript{450} The MAUDE database allows users to search for medical devices that may have malfunctioned or caused death or serious injury.\textsuperscript{451} Users can search by a pull down menu of product problems, by the class of product, or by manufacturer, model, or brand name.\textsuperscript{452} Like FAERS, MAUDE is designed to help the FDA monitor emerging product safety problems, but unlike FAERS, the centralized search function makes the data more accessible to lay users.

Like other agencies, the FDA is trying to publish FAERS, MAUDE, and other databases in more user-friendly formats on its Open FDA site.\textsuperscript{453} The site, launched in 2014, includes

\begin{footnotesize}
\begin{enumerate}
\item \textsuperscript{446} Id.
\item \textsuperscript{448} \textit{MAUDE - Manufacturer and User Facility Device Experience}, U.S. FOOD & DRUG ADMIN., https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm (last updated June 30, 2017) [https://perma.cc/GJ8S-LG88] [hereinafter MAUDE].
\item \textsuperscript{449} Medical Device Reporting, 49 Fed. Reg. 36,326 (Sept. 14, 1984).
\item \textsuperscript{450} \textit{Mandatory Reporting Requirements: Manufacturers, Importers, and Device User Facilities}, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/ReportingAdverseEvents/ucm2005737.htm (last updated Nov. 7, 2016) [https://perma.cc/Y4BF-49CG].
\item \textsuperscript{451} Id.
\item \textsuperscript{452} MAUDE, supra note 448.
\item \textsuperscript{453} CORTEZ, supra note 143, at 44; U.S. Food & Drug Admin., \textit{Open-source APIs and a Developer Community for FDA Data}, OPENFDA, https://open.fda.gov
\end{enumerate}
\end{footnotesize}
separate pages for food products, drugs, and devices, with all three including enforcement reports, and the drug and device pages, including adverse event databases. Open FDA publishes both individual reports and larger trend analyses. For example, as of March 2016, the site included almost 5.9 million records in its adverse drug event database dating back to 2004. Open FDA includes extensive data tools and downloadable raw data files, obviously directed at third-party users.

Also like other agencies, the FDA is beginning to incorporate multiple data sources to pursue regulatory goals, in this case uncovering trends with medical product safety. In 2007, Congress required HHS and the FDA to coordinate with non-FDA sources, including “public, academic, and private entities” to “link and analyze safety data from multiple sources,” with an idea of uncovering emerging product safety risks. The goal was to include at least 100 million patients in the dataset by 2012. Called the FDA Sentinel Initiative, the effort has been designed to monitor product safety across different data sources, including data from Medicare, the Veterans Health Administration, and large private health insurers.

But again, like other databases, the FDA’s adverse event databases are not always complete or accurate. A 2011 study found widespread errors and incomplete reports filed in MedWatch, including more than 25 percent of reports using inaccurate product names. More than most agencies, then,
the FDA includes prominent disclaimers of the accuracy and reliability of its data.\textsuperscript{460} For example, the MAUDE database includes the following disclaimer:

Although [Medical Device Reports] are a valuable source of information, this passive surveillance system has limitations, including the potential submission of incomplete, inaccurate, untimely, unverified, or biased data. In addition, the incidence or prevalence of an event cannot be determined from this reporting system alone due to potential under-reporting of events and lack of information about frequency of device use.

\* \* \*

Confirming whether a device actually caused a specific event can be difficult based solely on information provided in a given report. Establishing a cause-and-effect relationship is especially difficult if circumstances surrounding the event have not been verified or if the device in question has not been directly evaluated.\textsuperscript{461}

Similarly, FAERS emphasizes that the “data does have limitations”:

First, there is no certainty that the reported event (adverse event or medication error) was actually due to the product. FDA does not require that a causal relationship between a product and event be proven, and reports do not always contain enough detail to properly evaluate an event. Further, FDA does not receive reports for every adverse event or medication error that occurs with a product. Many factors can influence whether or not an event will be reported, such as the time a product has been marketed and publicity about an event. Therefore, FAERS data cannot be used to calculate the incidence of an adverse event or medication error in the U.S. population.\textsuperscript{462}

The Open FDA site also confronts users, via pop-up window, with a note of caution that “[t]his API is not for

\textsuperscript{460} Corz, supra note 143, at 44.
\textsuperscript{461} MAUDE, supra note 448.
\textsuperscript{462} FAERS, supra note 441.
clinical or production use. While we make every effort to ensure that data is accurate, you should assume all results are unvalidated." Should policymakers settle for data of such quality? Are incomplete and unrepresentative data better than none? If so, how best can scarce resources be deployed to improve the accuracy, reliability, and usefulness of the FDA’s data?

There remains great hope that modern tools like agency databases can improve drug safety, even if the data are imperfect, by combining FDA data with other data sources, like Medicare claims, for example. Thus, an alliance between the FDA and CMS to combine their data might help cure some defects in the FDA’s adverse event reporting systems. Still, policymakers might counter widespread underreporting by experimenting with automated monitoring systems relying on digital technologies. For example, digital pill trackers and other mobile monitoring devices might send automatic problem reports to the FDA, subject to pre- and post-publication safeguards described above.

V. DESIGNING FOR OPTIMAL DISCLOSURE

How can policymakers design databases that realize their many aspirations while minimizing their shortcomings and burdens? At core, how can policymakers ensure the quality and reliability of agency data, so that users trust the data being published? My first prescription is modesty. As Richard Craswell cautions, “people who expect disclosure laws to solve

465. Id.
almost every problem—quickly, easily, and with very little cost—are doomed to have their expectations crushed.”

The truth is that regulation by database requires just as many difficult design and implementation choices as any other form of regulation. The early successes of the CFPB’s Consumer Complaint Database and the CPSC’s SaferProducts.gov site, for example, are owed to several canny decisions by Congress and the agencies to ensure the quality of the data posted. It is a mistake, then, to assume that creating a public database is necessarily less difficult and more cost-effective than traditional regulation.

Another important consideration is that agency databases vary widely in their purposes, scope, design, sources, and presentation. Although it is neither possible nor worthwhile to prescribe universal rules of thumb here, I try to highlight emerging best practices from the databases I have evaluated to date, in the hopes that these discussions will be useful to policymakers. Again, these recommendations build on, and in many ways exceed, those recently promulgated by ACUS.

In short, the most successful disclosure regimes will carefully consider both data inputs (how data will be collected and from whom) and outputs (how the data will be published and presented).

I thus offer a series of recommendations that address both dimensions, proposing a decidedly modern role for the government as a “data steward.” For example, agencies must consider how to gather and process the data, including carefully choosing sources and crafting adequate pre- and post-publication procedures to ensure data quality. Agencies also should not underestimate the costs required to generate and maintain their databases, as well as the potential overlap with agency enforcement and adjudication procedures. Finally, agencies should think carefully about potential audiences and users in deciding how to characterize and present the data.

---

468. Craswell, supra note 18, at 379.
469. Id.
470. See discussion supra Sections IV.C & IV.D.
A. Gathering and Processing the Data

Policymakers should think carefully, and ideally in advance, about data inputs. Whose data will populate the data sets? How will it be gathered? To what extent will the agency try to verify, validate, or otherwise authenticate the data? And how will the agency handle contested data?

As a threshold matter, it is particularly important that agencies identify reliable data sources. The irony here is that the internet not only facilitates many of the disclosures discussed in this Article, but also (by virtue of soliciting data from a variety of sources) raises problems with reliability. 473 Thus, government agencies can play an important role by ensuring that published data is credible. 474 Indeed, because the very fact of publication by a government agency often signals credibility, it is incumbent on agencies to ensure the accuracy of data that purports to be objective. 475 Notice-and-comment procedures might be useful in soliciting feedback on what sources are reliable and why.

As Kristin Madison argues, the federal government is more than just a repository for data—it is also a “data steward” responsible for actively managing the data it holds, helping to ensure its integrity. 476 The CFPB, for example, does not verify that consumer complaints are “accurate” (in other words, that the conduct alleged in the complaint actually occurred), but does help “authenticate” that each complaint is made by an actual customer of the company, giving the company ample opportunity to identify false or fraudulent complaints. 477 Likewise, when Congress authorized SaferProducts.gov, it required the CPSC to consider objections that the information is “materially inaccurate.” 478 Again, the CFPB and CPSC demonstrate two frameworks for active data stewardship.

In short, agencies that purport to publish accurate and objective data should adopt procedures to these ends.

---

473. See, e.g., Fung et al., supra note 15, at 165 (“Those who contribute information can do so without identifying themselves or their sponsoring organizations, or taking responsibility for what they are saying.”).
474. Heerwig & Shaw, supra note 243, at 1476.
475. Cortez, supra note 143, at 44.
476. Madison, supra note 30, at 1607–08.
emphasized in Part II, because “transparency” is frequently invoked to support disclosure, it is important to remember that “transparency” allows objects to be “seen without distortion.”\(^\text{479}\) To the extent feasible, then, government databases should try to achieve genuine transparency rather than “translucency.”\(^\text{480}\) Indeed, some scholars note that, inevitably, disclosure regimes created through the political process are forged by conflict and compromise, and thus generate only partial or imperfect transparency.\(^\text{481}\) But scholars also find that the more successful disclosure regimes tend to increase the accuracy and quality of the information they publish over time.\(^\text{482}\) Databases should be dynamic rather than static—constant works in progress. Although initial design choices are important, agencies should not hesitate to tinker with data collection procedures if flaws in the data become apparent, as in the case of the EPA’s Toxic Release Inventory.

1. Pre-Publication Procedures

Data sets that purport to publish accurate, objective information should be buttressed, ideally, with both pre- and post-publication procedures that allow parties to request that any information not meeting these standards be corrected or retracted. Again, some of the newer agency databases provide parties with pre-publication procedures to comment on, challenge, or request corrections and retractions of information before publication.\(^\text{483}\) For example, by statute the CPSC must give companies whose products are reported to SaferProducts.gov the opportunity to comment on any Reports of Harm.\(^\text{484}\) The CPSC must consider objections that the information is “materially inaccurate,”\(^\text{485}\) and the Commission publishes these procedures in the C.F.R.\(^\text{486}\) There are clear

\(^{479}\) Schauer, supra note 9, at 1343 (referring to the Oxford English Dictionary’s definition of “transparency”).

\(^{480}\) Id. at 1345. The Schauer distinction between transparency and translucency is subject to the objection, of course, that it assumes that the state could ever be completely “transparent,” providing undistorted access to itself.

\(^{481}\) FUNG ET AL., supra note 15, at xii.

\(^{482}\) Id. at 109.

\(^{483}\) CORTEZ, supra note 143, at 94–95.


\(^{485}\) Id. § 2055a(c)(4).

\(^{486}\) 16 C.F.R. § 1102.26.
timelines for parties to object to alleged inaccuracies and for the CPSC to resolve disputes before publication. Likewise, the CFPB authenticates that complaints are coming from actual customers of the company. Bureau procedures allow companies to use an online company portal to verify a commercial relationship with the customer and post the company’s response. The Bureau also makes clear that each complaint, before being published in the database, must meet several publication criteria. Of all the databases discussed in Part IV, the CPSC and CFPB procedures serve as the best models for pre-publication quality control.

2. Post-Publication Procedures for Contested Data

Nevertheless, errors in published data sets are probably inevitable, no matter how robust the pre-publication procedures may be. As a result, policymakers should also consider post-publication procedures as a backstop to help ensure the quality and reliability of data. Scholars have long recognized that such procedures can be an important safety valve for parties named in agency publications, as legal recourse is generally not available.

Here, the controversial Information or Data Quality Act might be of help. The Act required the OMB to publish government-wide guidelines for “ensuring and maximizing the quality, objectivity, utility, and integrity of information . . . disseminated by the government.” The Act applies broadly to “[f]ederal agency dissemination of public information, regardless of the form or format.” It also directed the OMB to

---

489. COMPANY PORTAL MANUAL, supra note 316.
490. CORTEZ, supra note 143, at 63–64.
491. Gellhorn, supra note 169; CORTEZ, supra note 143, at app. C.
492. The Act has been criticized as an attempt to thwart regulators in the guise of data quality. See, e.g., Levy & Johns, supra note 27, at 2.
494. 67 Fed. Reg. 8452, 8460 (Feb. 22, 2002). OMB guidelines define “information” as “any communication or representation of knowledge such as facts
establish procedures that allow “affected persons to seek and obtain correction of information maintained and disseminated by the agency.”

Per the OMB’s guidelines, dozens of federal agencies have published their own such guidelines and post-publication procedures for correcting or retracting information. Although these procedures would seem to have clear application to agency databases, the OMB guidelines include two important exemptions. First, they exclude from the IQA’s coverage “opinions, where the agency’s presentation makes it clear that what is being offered is someone’s opinion rather than fact or the agency’s views.”

Second, they exclude “adjudicative processes.”

Thus, both exemptions could be read as excluding, for example, the CFPB’s Consumer Complaint Database, which might be fairly characterized as including “opinions” or even “adjudicative processes.”

Nevertheless, the IQA and resulting agency guidelines articulate both substantive and procedural values that agencies should observe. To ensure the quality and reliability of government-published information, there should be a safety valve that allows the subjects identified to request correction or retraction by the agency.

A 2015 study by the GAO found eighty-seven publicly-reported requests for corrections or retractions under the IQA sent to thirty agencies between 2010 and 2014. Although the agencies denied fifty-nine out of the eighty-seven requests, the agencies made full corrections in eleven cases and partial

or data, in any medium or form,” including “information that an agency disseminates from a web page.” Id.


498. Id. at 8454.


corrections in fifteen cases (two cases were unresolved as of the date of the report).\textsuperscript{501} For example, HHS updated information on several CDC websites regarding bicycle helmet safety data in response to a request made under CDC, HHS, and OMB data quality guidelines.\textsuperscript{502} Moreover, in fifteen of the fifty-nine cases denying the IQA request, the agency used alternative procedural mechanisms, usually systems that predated the IQA.\textsuperscript{503} In some cases, the agency engaged in long substantive exchanges on the accuracy, presentation, and usability of data.\textsuperscript{504}

Post-publication procedures might also reside outside, rather than inside, the agency. My review for ACUS, for example, considered whether independent bodies like the OMB, ombudsmen, or inspectors general might play a role in superintending disputes over agency data.\textsuperscript{505}

First, the OMB already exerts both centripetal and centrifugal pressures on agency data collection and publication practices. In addition to the OMB’s IQA guidelines, the Paperwork Reduction Act requires the OMB to preapprove significant information gathering efforts by agencies.\textsuperscript{506} The agency must explain to the OMB why it needs the information, why it has “practical utility,” and why it is relevant to the agency’s regulatory functions.\textsuperscript{507} Thus, the OMB can play a useful standardizing role. But it is not well suited to resolving disputes between agencies and regulated parties, and agencies may bristle at having to endure further layers of OMB review.

Second, many agencies maintain an Office of the Ombudsman or its equivalent, which can field complaints about data published by agencies. For example, the CFPB’s Ombudsman has heard complaints about the Consumer
Complaint Database,\textsuperscript{508} and the FDA’s many ombudsmen have fielded complaints under the IQA.\textsuperscript{509} Ombuds can serve important customer service functions with regulated parties. As such, the use of ombuds in federal agencies has increased in recent years, as have calls for standards regarding their independence, duties, and information-providing roles.\textsuperscript{510}

A third option is review by agencies’ inspectors general. Inspectors general (IGs) are independent officers, directed by law to detect and prevent fraud, waste, and abuse in federal agencies.\textsuperscript{511} They also maintain, by design, crucial independence from agency heads, and thus can serve as an independent arbiter.\textsuperscript{512} They can also function as an avenue for fielding industry complaints.\textsuperscript{513} The Federal Reserve’s Office of Inspector General is, in fact, auditing the CFPB’s Consumer Complaint Database “to assess the effectiveness of the CFPB’s controls over the accuracy and completeness of the public complaint database.”\textsuperscript{514}

Finally, chief information officers (CIOs) within agencies might play an important role in not only answering important questions regarding database design, but also in participating in pre- and post-publication procedures described above. They are most likely to be informed of other agencies’ experiences and able to critically evaluate whether those models might translate well to their own data regimes.

\textsuperscript{508} CONSUMER FIN. PROT. BUREAU, OMBUDSMAN’S OFFICE, ANNUAL REPORT TO THE DIRECTOR 5, 6 (2014).
\textsuperscript{512} See, e.g., Neal Kumar Katyal, Internal Separation of Powers: Checking Today’s Most Dangerous Branch from Within, 115 YALE L.J. 2314, 2347 (2006).
\textsuperscript{513} CORTEZ, supra note 143, at 101.
3. Considering Costs

Procedural safeguards can be essential for ensuring data quality, though they are not without cost. Too often government agencies try to achieve disclosure on the cheap.\footnote{515}{I am indebted to Frank Pasquale for raising many of the issues in this paragraph.} Data collection and processing requires not just automation, but also human labor. Unfortunately, such labor is dismissed as the task of “data janitors” who receive inadequate compensation.\footnote{516}{See, e.g., Lilly Irani, Justice for "Data Janitors," (Jan. 15, 2015), http://www.publicbooks.org/nonfiction/justice-for-data-janitors [https://perma.cc/YK9H-6AYU].} The lack of sufficient infrastructure to ensure data quality can generate “big bad data”—data that are voluminous but of low quality. Thus, meaningful data often require meaningful investment to create a sufficient information infrastructure.

Moreover, posting the data online can be costly. Running an agency website is not a simple proposition. Web masters for federal websites must comply with at least two dozen different regulatory systems “[r]anging from privacy and usability to FOIA compliance to the demands of the Paperwork Reduction Act.”\footnote{517}{Sharona Hoffman & Andy Podgurski, Big Bad Data: Law, Public Health, and Biomedical Databases, 41 J. L. MED. & ETHICS 56 (2013).} Although each separate requirement may stand on its own logic, together they can limit how agencies present data, and generally favor standardization above experimentation.\footnote{518}{David Robinson et al., Government Data and the Invisible Hand, 11 YALE J.L. & TECH. 160, 162 (2009).} As Robinson and colleagues observe, “[a]s long as the government has a special role in the presentation and formatting of raw government data, certain desirable limits on what the government can do become undesirable limits on how the data can be presented or handled.”\footnote{519}{Id. at 163–65.} In this vein, nongovernmental intermediaries have proven useful in rendering government data more accessible and usable,\footnote{520}{Id. at 165.} as I discuss below in Section V.C.

Finally, all this assumes that policymakers have already made the threshold decision to publish the data and accept
responsibility for data stewardship. But because data stewardship can be resource intensive, policymakers should think more critically in advance about which data regimes warrant the government’s scarce stewardship resources. Which databases might be especially useful to consumers, or particularly effective at inducing optimal behavior from regulated entities? And of existing databases, which low-quality data sets might be worth salvaging? The FEC’s and FDA’s databases might underwhelm in several ways, but few would argue that more accurate, comprehensive campaign finance data or drug safety data would not be worthwhile. Thus, if stewardship is worth pursuing, limited resources dictate that it must be targeted stewardship, focusing on the data sets that are most likely to achieve the twin goals of being useful to target audiences and changing behavior.522

4. Administrative Law Dimensions

Most interesting, from an administrative law perspective, is that database publication procedures might be viewed as a unique species of agency adjudication and regulatory enforcement. Seen this way, database publication procedures fall along a continuum ranging from very minimal verification to more searching pre- and post-publication adjudication of contested data. No database regimes currently approach the full panoply of procedural safeguards that attach to more formal administrative adjudications, pursuant to statutory and due process requirements.523 But like traditional agency adjudications, the amount of “procedure” appropriate for database disputes will depend on what information each database includes, the regulatory goals of publishing the information, the statutory scheme in which it sits, and the cost and value of “getting it right” versus “making it public.”

Another possibility is that legal sanctions begin to attach more formally and more forcefully to database reporting such that companies reporting inaccurate or incomplete data will face fines, penalties, and other measurable burdens in addition to any reputational damage. Could the Medicare program, for example, condition reimbursement on accurate data

522. I credit Nick Bagley for seeding the ideas in this paragraph.
523. Bill Sage inspired many of the thoughts in this paragraph.
reporting? Could plaintiffs or prosecutors use the federal false statements statute or the False Claims Act to punish material inaccuracies or misleading data reporting by regulated firms? Both laws are broad and powerful, and are deployed in increasingly creative ways. Their use in database reporting cases could raise novel statutory and due process questions about the procedural safeguards agencies have selected for specific databases. Moreover, such actions could be undermined by how agencies themselves “characterize” the data, including any disclaimers about the accuracy, reliability, or objectivity of the data.

B. Characterizing the Data

Because not all data will be objective in nature, verifiable, or even worth the cost of verifying, it is equally important that agencies accurately characterize the data they present, listing the sources and any important context or limitations for the data.

1. Identifying Sources

Databases should be labeled and characterized accurately, much as we expect of product labeling by industry. This is particularly so because government agencies are one of the most trusted sources of information, and the information they publish carries the imprimatur of the federal government. Thus, agency databases should clearly identify the sources of their data. Agency databases should also indicate whether the data are contested, and detail steps the agency takes (or, more importantly, does not take) to resolve such contests.

---

524. For a discussion of this possibility, see Frank Pasquale, Grand Bargains for Big Data: The Emerging Law of Health Information, 72 MD. L. REV. 682 (2013).
528. CORTEZ, supra note 143, at 94.
529. Id. at 96.
Data sets that do not purport to be accurate or objective might require special precautions. Federal databases can be populated with data from a variety of sources—including consumers, regulated parties, or the agency itself—and each might require different quality controls and presentations.\textsuperscript{530} For example, the FDA’s adverse event databases are populated by reports from manufacturers and users that a product may have been “associated” with an adverse event, without any firm claims as to causation.\textsuperscript{531} Similarly, being listed in the CFPB’s Consumer Complaint Database does not mean that a company has committed any legal violation; many complaints are simply “vague expressions of being wronged.”\textsuperscript{532} Just like the FDA does not verify whether a product caused a specific adverse event, the CFPB does not verify that a company even engaged in the conduct alleged in the consumer complaint. Doctors subject to “report cards” also lament that death and complication rates are presented without being normalized for treating riskier patient populations.\textsuperscript{533} Should such “data” even be published? For better or worse, routinely they are.

2. Explaining Context, Limitations

The solution, perhaps, is for agencies to represent the data accurately, which often means explaining the context and any limitations of the data. Several agencies already endeavor to do so. For example, in the FDA’s medical device database, the agency notes that its “surveillance system has limitations, including the potential submission of incomplete, inaccurate, untimely, unverified, or biased data.”\textsuperscript{534} The FDA’s adverse drug event database also notes that “there is no certainty that the reported event . . . was actually due to the product.”\textsuperscript{535} Likewise, the CFPB disclaims that “[w]e don’t verify all the facts alleged in these complaints but we take steps to confirm a

\textsuperscript{530} Id.
\textsuperscript{531} See discussion supra Section IV.F.
\textsuperscript{532} CORTEZ, supra note 143, at 69–70; Porter, supra note 163, at 78.
\textsuperscript{533} Sandeep Jauhar, Opinion, Giving Doctors Grades, N.Y. TIMES (July 22, 2015), http://www.nytimes.com/2015/07/22/opinion/giving-doctors-grades.html [perma.cc/WH3T-Q6NJ]. Although death and complication rate data are usually normalized through various methodologies, physicians frequently object that such methodologies are inadequate.
\textsuperscript{534} MAUDE, supra note 448.
\textsuperscript{535} FAERS, supra note 441.
commercial relationship between the consumer and the company.”\textsuperscript{536} Congress requires the CPSC’s SaferProducts.gov database to “provide clear and conspicuous notice to users of the database that the Commission does not guarantee the accuracy, completeness, or adequacy of the contents of the database.”\textsuperscript{537} Although one court called the CPSC’s language “boilerplate” that “would not interest an ordinary consumer,”\textsuperscript{538} providing appropriate context and disclosing the limitations of data is relatively easy and helps answer several criticisms of disclosure noted in Part III.

A “reliable” database may depend not only on publishing accurate data (if that is what it purports to do), but also on publishing relatively complete and representative data. Industry commenters, for example, objected that the CFPB’s database of self-selected consumer complaints would necessarily be incomplete, nonrandom, and thus nonrepresentative of company performance and consumer experiences.\textsuperscript{539} The CFPB responded that the data are not portrayed as such, and promised to “inform consumers and any other public database users that the data reflect only the . . . complaints that consumers submit to the Bureau.”\textsuperscript{540}

Data selection or filtering criteria might thus generate published data that are technically accurate but misleading as a whole. Transparency initiatives often fail when “transparency is either not sufficiently mandatory or not applicable to categories of information that meaningfully contribute to public accountability.”\textsuperscript{541} Agencies with discretion to disclose information may tend to disclose “information that makes the administration look public spirited, effective, and efficient, but withhold information to the contrary.”\textsuperscript{542} The data chosen for publication may not paint a complete or representative picture. Thus, as Shkabatur argues, the answer for incomplete transparency may be even more
transparency.\textsuperscript{543} She finds support among scholars who argue that Congress should place affirmative disclosure duties on agencies, shifting away from the “passive” disclosure required by laws like FOIA that have been rendered as anachronisms in the Internet era.\textsuperscript{544} If it is neither possible nor cost-effective to publish comprehensive or representative data, the agency should provide adequate context for what is being published and explain why the dataset is incomplete.

C. Presenting the Data

Policymakers should also consider the “outputs” of agency databases—how the data will be published, presented, and used. Thoughtful designs will evaluate the optimal format, size, and scope of the database, as well as the target audiences and their potential uses.

1. Raw or Polished?

First, data can be published in raw or relatively polished formats, with gradations in between. Should agencies rely on massive raw data dumps targeted at more sophisticated users? Or should they package, stylize, and distill the data for lay users? The trend, as noted above, is to publish data sets in both more polished, packaged formats and in raw, open data formats. The former requires agencies to think carefully about how to convey the information, and in what packaging—which inevitably raises costs and includes normative judgment calls that might draw into question how objective or neutral the presentation is. But the latter (publishing raw data) is a relatively recent phenomenon.

Historically, agencies have been reluctant to publish information in open, raw, machine-readable data formats (particularly information requested via FOIA). For example, 2011 congressional testimony revealed that “[m]ost requests for correspondence and other documents are fulfilled by printing them, redacting, then re-scanning into unsearchable

\textsuperscript{543} See generally Shkabatur, supra note 55.

\textsuperscript{544} See, e.g., Vladeck, supra note 64, at 1828–29.
images.”\textsuperscript{545} Yet, as far back as 2004, the OMB encouraged agencies to “provide all data in an open, industry standard format permitting users to aggregate, disaggregate, or otherwise manipulate and analyze the data to meet their needs.”\textsuperscript{546} And scholars continue to argue for agencies to publish data online in open, structured, and machine-readable formats such as XML, consistent with the Open Government Working Group’s recommendations.\textsuperscript{547} Thus, there is a clear trend toward publishing in raw, open formats.

At the same time, some scholars argue that publishing data in raw, “naked” formats can itself serve as a barrier to access for nonprogrammers and others who are not able to understand or use such data.\textsuperscript{548} Raw government datasets might require, ironically, technically sophisticated intermediaries to decipher.\textsuperscript{549} Thus, open government efforts that encourage agencies to present data in a raw, naked, and “neutral” way may erect separate barriers to accessing and understanding the information.

A related idea is that government transparency can exist on different planes, from “relative” to “absolute.” Relative transparency occurs when someone—usually the government or a data intermediary—“relates” the data from one reporting entity to another for easy comparison. For example, star ratings, grades, and other distilling criteria essentially grade reporting entities on a curve. To wit, most users would understand that a hospital receiving one out of five stars on overall quality underperforms most other hospitals. But it is


\textsuperscript{547} See, e.g., Robinson et al., supra note 518, at 167 (arguing that original data should be posted in documents in XML formats with unique and permanent addresses); The Annotated 8 Principles of Open Government Data, OPEN GOV’T WORKING GRP., http://opengovdata.org/index.php/OpenDataPrinciples (last visited Oct. 8, 2017) [https://perma.cc/5YEA-HM6] (recommending that data be complete, primary, timely accessible, machine-readable, non-discriminatory, non-proprietary, and license-free).

\textsuperscript{548} See, e.g., Shkabatur, supra note 55, at 112.

\textsuperscript{549} Id.
much more difficult for users to understand what a two percent complication rate associated with cardiac surgeries performed at a specific facility should signal, as an absolute number. Of course, even “relative” transparency may not be particularly useful. If HospitalCompare.gov lists a hospital’s mortality rates as “[n]o different than the national average,” that might signal to users that they should not worry about that factor when selecting a hospital. But it could also mean that the national average is equally disappointing for everyone. Moreover, “relative” transparency is only realized after gathering “absolute” data points. But who should take on the task of turning absolute, raw data into relative, packaged comparisons?550

Some argue that the government should focus its energies less on presenting packaged information and more on publishing “reusable data.”551 The idea, inspired by the engineering principle that separates data from interaction, is that agencies should worry less about designing user-friendly websites, and more about releasing raw data for nongovernmental users.552 Robinson and colleagues argue that the latter will be better able to experiment with how to present the data effectively, whether it be with advanced search functionalities, automated content analysis, indexing among multiple sources, and various data visualization tools.553 They call this new role for agencies an “invisible hand,” enabling a “marketplace of engineering ideas.”554 Some users will also value being able to access “genuine” data that is not mediated, framed, or translated by an intermediary (including, or even particularly, by the government).555

550. I credit Kristin Madison with raising the notion of “relative” versus “absolute” transparency, which parallels some of the considerations when deciding whether to prioritize publishing raw versus polished data sets. Larry Lessig also touches on this dynamic when comparing campaign finance data (absolute) with EPA fuel economy stickers on new cars (relative). See LESSIG, supra note 260, at 257–58.
551. Robinson et al., supra note 518, at 160.
552. Id. at 161.
553. Id. at 161, 169.
554. Id. at 161.
555. Id. at 174. Fenster discusses how the process of the government communicating information necessarily involves imperfect judgment calls about what information to disclose and how. Fenster, supra note 26, at 926–27.
2. Big Data or Small Data?

Disclosure enthusiasts often assume that more is better. But recently, scholars have begun to acknowledge that it is more important that information be accessible and usable, rather than simply available.556 Perhaps the relevant question, then, is not what policymakers think users need to know, but what users want to know.557 Disclosure policies that consider what information users want, and what they can comprehend, tend to be more successful over time.558

Moreover, perhaps consumers do not necessarily need more data, but more advice.559 The opposite of making data “bigger,” of course, is making data “smaller”—usually by simplifying, tailoring, and targeting the information to make it easier to process.560 Thus, rating systems and other information made available at the point of purchase could be particularly useful for consumers.561 Mere publication on a government website might not be particularly useful, unless intermediaries make the data available where and when it can be used. Such information is more likely to become embedded in the decisions targeted by the disclosure.562 Thus, databases that allow users to simplify complex information, or that allow experts to easily convert it to actionable advice (such as a ratings system or a reliable heuristic), will be more successful in achieving regulatory goals.563 Restaurant hygiene grades, for example, are more embedded in the decision of where to eat than complex and voluminous patient safety disclosures are in the decision of where to seek medical care.564

Despite the current fascination with “big data,” many also appeal for simplification. Agencies are thinking more carefully today about ensuring the “utility” of data, perhaps owing in

556. See, e.g., Schauer, supra note 9, at 1344; Fenster, supra note 26, at 942.
557. Ben-Shahar & Schneider, supra note 17, at 746.
558. FUNG ET AL., supra note 15, at 11.
559. Ben-Shahar & Schneider, supra note 17, at 746.
560. Madison, supra note 30, at 1621–22 (noting astutely that recent laws like the Affordable Care Act include language like “patient-centered” and “patient engagement,” which can be code for making big data smaller).
561. FUNG ET AL., supra note 15, at 57.
562. Id. at 65–74.
563. Id. at 57.
564. Id. at 65.
Moreover, notions of “utility” continue to evolve. For example, the Nutrition Facts label on food products has had some modest success—consumers report that they are increasingly aware of nutrition labeling and make decisions based on it. Still, as Ben-Shahar and Schneider note, even with nutrition labeling, which they call “the simplest and most understandable case of daily disclosures,” studies still find high levels of consumer confusion that largely correlate with low consumer literacy and numeracy. How much should these findings deter agency disclosure efforts?

In general, ratings systems that communicate data that has been simplified and “translated” for lay users seem to enjoy moderate success. For example, there is evidence that simple letter grades for restaurant sanitation (from “A” to “C”) have led to cleaner restaurants in Los Angeles County. Restaurants with high letter grades posted in their store windows saw an increase in revenues, and conversely, restaurants with the lowest “C” grades saw a decrease. More tellingly, prominent disclosure of these grades encouraged restaurants to improve their sanitation practices, which correlated with a significant local drop in hospitalizations related to food-borne illnesses. Thus, simple, comprehensible, and easily accessible ratings not only allowed

566. Ben-Shahar & Schneider, supra note 17, at 675.
567. Id. at 675–76 (citing Gary Jones & Miles Richardson, An Objective Examination of Consumer Perception of Nutrition Information Based on Healthiness Ratings and Eye Movements, 10 PUB. HEALTH NUTRITION 238 (2007); Gill Cowburn & Lynn Stockley, Consumer Understanding and Use of Nutrition Labeling: A Systematic Review, 8 PUB. HEALTH NUTRITION 21, 23 (2005); Russell L. Rothman et al., Patient Understanding of Food Labels: The Role of Literacy and Numeracy, 31 AM. J. PREVENTATIVE MED. 391, 391 (2006)).
568. FUNG ET AL., supra note 15, at 743.
570. FUNG ET AL., supra note 15, at 50 (citing studies).
571. Id. Note, however, that studies of other local restaurant grading efforts have found them to be less successful. See, e.g., Daniel E. Ho, Fudging the Nudge: Information Disclosure and Restaurant Grading, 122 YALE L.J. 574 (2012) (examining similar efforts in multiple cities).
consumers to vote with their wallets, but also encouraged restaurants to compete based on cleanliness—undoubtedly the underlying motivation of the letter grade system.  

For disclosure policies to succeed on multiple levels, then, they must affect not only the decision-making of consumers and regulatory beneficiaries, but also the decision-making of the discloser—the regulated party. Thus, effective disclosure systems become “doubly embedded.” The way disclosure policies affect discloser behavior is intuitive—by affecting their profits, market share, and reputation. Disclosers may change their behavior, in fact, simply in anticipation that releasing information may affect one of these three things.

Thus, agencies may choose two very different courses: massive raw data dumps intended for sophisticated intermediaries, or highly distilled presentations intended for lay users. The correct choice, if one must be made, depends very much on the data and what the agency hopes to achieve by publishing it. 

On one hand, simplified ratings or grades are able to distill dozens or even hundreds of different complex criteria into a single understandable metric, like restaurant hygiene grades, hospital star ratings, or five-star crash safety ratings, which are based on complex engineering standards and test results. Ratings and grades also combat the problem of overdisclosure. Scholars that have evaluated the effectiveness of mandatory disclosure regimes sometimes observe that parties can “overdisclose” information to try to “overwhelm and distract” the intended audience. Ratings can ameliorate this problem.

On the other hand, sometimes efforts to make the complex more understandable fail, as evidenced by the vague five-color scheme for communicating the threat of a terrorist attack. Unlike dirty restaurants or unsafe cars, it is hard for most people to understand the significance of the terror threat

572. FUNG ET AL., supra note 15, at 50–51.  
573. Id. at 65–74.  
574. Id. at 65.  
575. Id. at 66.  
576. Id. at 59–61.  
577. Ben-Shahar & Schneider, supra note 17, at 700; Willis, supra note 231, at 790.  
578. FUNG ET AL., supra note 15, at 61.
changing from yellow to orange, and more importantly, how to act on that signal.\footnote{Id.} Thus, not all data is so easily distilled.

Given these considerations, should agencies design databases to be accessible to the lay public, or to be used by more sophisticated information intermediaries? An ideal answer is “both,” of course. To maximize accessibility, the data should be available in multiple formats, as many agencies now recognize, and as Data.gov demonstrates. If “both” is not a feasible option, the agency must decide whether “big data” or “small data” better achieve regulatory ends, including the relative costs of both approaches.

3. Intermediaries and Collaborative Data

Designing databases for use by third-party information intermediaries is compelling for several reasons.\footnote{For an early discussion of the use of data intermediaries, see Sage, supra note 5, at 1737–41.} Publicly-minded watchdogs like Pro Publica, the Sunlight Foundation, and the Project on Government Oversight (POGO) can serve a translational role, sifting large amounts of data into more understandable bits.\footnote{See, e.g., Shkabatur, supra note 55, at 118.} Although these organizations focus on government transparency, they also can (and do in fact) help extract and translate information about regulated parties.\footnote{See, e.g., Charles Ornstein et al., Dollars for Docs, PRO PUBLICA, https://projects.propublica.org/doedollars/ (last visited June 6, 2017) [https://perma.cc/W7UC-SSBF]. The Dollars for Docs project at Pro Publica posts a searchable database of physicians that have received money from pharmaceutical and device firms, using data reported under federal law, including the Physician Payment Sunshine Act. CMS released the data, but Pro Publica gathers it in a single searchable database with rankings and analysis. See Charles Ornstein & Ryann Grochowski Jones, About the Dollars for Docs Data, PRO PUBLICA (July 1, 2015), https://www.propublica.org/article/about-the-dollars-for-docs-data [https://perma.cc/6TYP-SLHQ].} Even complex datasets that are not translated by agencies into ratings, grades, or other digestible metrics can be translated by thoughtful intermediaries. For example, various consumer groups have tried to translate toxic release data into more user-friendly websites.\footnote{FUNG ET AL., supra note 15, at 62; The Pollution Information Site, SCORECARD, http://scorecard.goodguide.com (last visited July 16, 2017) [http://perma.cc/B7B8-7375]; The Right-to-Know Network, HOUS. CHRON., http://www.rtknet.org (last visited July 16, 2017) [http://perma.cc/VLSP-QJBT].} Thus, even raw data sets can be
repurposed for lay users like consumers and other regulatory beneficiaries. Indeed, Fung, Graham, and Weil found that the most successful disclosure regimes “featured strong groups representing information users, offered benefits to at least some information disclosers, and provided comprehensible content.”

However, translating voluminous, complex government data requires not only minimum technical and programming expertise, but also a basic understanding of the agency and its regulatory framework (and perhaps also an understanding of the regulated industry). The number of organizations that can fit comfortably into such a Venn diagram might be quite small. Indeed, even proponents of publishing raw government data concede that it is not immediately accessible to most lay users. And some doubt that these organizations derive their value from information supplied by the government, rather than from their own surveys and information-collecting activities.

Still, the fact that there are individuals like Joshua Tauberer (who created Govtrack.us in his spare time) and Carl Malamud (who painstakingly made SEC data available online) demonstrate that the barriers are far from insurmountable. Moreover, these extraordinary individual efforts inspired the government to publish the data in open formats.

Unsurprisingly, agencies are also discovering that data is becoming a more collaborative endeavor. There is optimism that once raw data is published, the private, public, and nonprofit sectors will make the data more accessible and useful to their constituents. Fung and colleagues note what happened with the EPA’s Toxic Release Inventory, as consumer groups like Scorecard and RTK refined the data and made it more user friendly, while the Chemical Manufacturers Association launched its own site emphasizing not only

584. FUNG ET AL., supra note 15, at xiv.
585. Robinson et al., supra note 518, at 173.
586. Ben-Shahar & Schneider, supra note 17, at 731–32.
588. Id. at 166, 171.
589. Id. at 171.
companies’ improving safety data, but also the number of jobs they created and taxes they paid by ZIP code.\textsuperscript{591}

As such, we might be experiencing a major shift in the government’s informational role from controller to facilitator.\textsuperscript{592} Indeed, modern agencies may be best suited to facilitating rather than controlling information—that is, gathering and publishing data, ensuring its quality, and then enabling the private and nonprofit sectors to maximize its uses.\textsuperscript{593} Data users might also become contributors, as in the case of consumers who report food poisoning from restaurants and thus supplement relatively infrequent restaurant inspections.\textsuperscript{594} Just as the CFPB endeavors, the government can serve as an aggregator of disparate data sources.

Craswell calls this “government-aided disclosures (GADs),” in which the government creates a baseline for disclosure, but allows companies to use the information dynamically or go beyond the baseline in some way.\textsuperscript{595} Such disclosures are mandated by government but are also integrated by disclosers because the information is useful to consumers or users.\textsuperscript{596}

Another factor that improves success is whether the information varies between disclosers, such that disclosers have an incentive to race to the top.\textsuperscript{597} Cigarette brands have little incentive to highlight the Surgeon General’s mandatory warnings, because the same preset warnings rotate among all products regardless of manufacturer, but they may have more incentive to reduce the tar and nicotine content of their products, which varies from brand to brand.\textsuperscript{598} Thus, data can serve as an important differentiator between competitors, and they might spend their own resources publicizing differences.\textsuperscript{599} Again, if databases aspire to affect underlying behavior and achieve regulatory ends, this is one way to do so.

In some notable instances, the federal government devotes remarkable resources to encourage users to collect and deploy

\textsuperscript{591} Id. at 158–60.

\textsuperscript{592} Indeed, Sage long ago called attention to the government’s role as a data facilitator. Sage, supra note 5, at 1712.

\textsuperscript{593} See, e.g., FUNG ET AL., supra note 15, at 166.

\textsuperscript{594} Id.

\textsuperscript{595} Craswell, supra note 18, at 369.

\textsuperscript{596} Id. at 369–70.

\textsuperscript{597} Id. at 371.

\textsuperscript{598} Id.

\textsuperscript{599} Id. at 370–71.
certain data, as in the case of electronic health records (EHRs). Through various pieces of legislation,\textsuperscript{600} Congress not only established standards for collecting and using electronic health records, but also devoted billions in incentives—an average of more than $40,000 per physician.\textsuperscript{601}

Extending the principle even further, the federal site HealthData.gov aggregates over 2,000 unique datasets from agencies like the CDC, CMS, FDA, and numerous state and local governments.\textsuperscript{602} The goal is to put open, machine-readable data in the hands of programmers, entrepreneurs, journalists, providers, scientists, consumers, and other policymakers who might, in turn, help improve health care in the United States.\textsuperscript{603} Thus, given the current fascination with “big data,” it helps to remember that countless government agencies (and Congress) are helping to make data “bigger,” consonant with their traditional goal of providing public goods.\textsuperscript{604}

The next generation of disclosure will thus be more collaborative, in the sense that various sectors will both contribute to and use the data.\textsuperscript{605} Indeed, there seems to be wide agreement that the government should not have a monopoly on generating data,\textsuperscript{606} but can play an important centripetal role in compiling data and helping assure their quality.

CONCLUSION

Agency databases have proliferated on the belief that markets, regulation, and democracy all thrive on transparency—that sunlight is the best disinfectant. Ideally,


\textsuperscript{603} About, HEALTHDATA.GOV, http://www.healthdata.gov/content/about (last visited July 15, 2017) [http://perma.cc/B629-5EXX].

\textsuperscript{604} Madison, supra note 30, at 1620–21; Sage, supra note 5, at 1771 (discussing the public good nature of information in the health industry).

\textsuperscript{605} Fung et al., supra note 15, at 153.

\textsuperscript{606} See, e.g., Esty, supra note 4, at 199–200.
shining a light on things like campaign contributions, pollution, and hospital outcomes will encourage more optimal behavior—or at least deter the worst of it. An added benefit is that “regulation by database” avoids the costs and formalities of traditional regulation.

But as transparency has moved online—becoming more pervasive, more powerful, and more burdened with regulatory dimensions—we also must recognize that sunlight can also blind or even burn. The case studies demonstrate how problems with accuracy, fairness, and efficacy can undermine even well-established, well-meaning data regimes. These problems can be avoided if agencies act less like passive publishers or repositories for data, and more like data stewards actively tending to a valuable (and dynamic) public good.

Policymakers must embrace the government’s role as a data steward, a sentinel that helps maximize the quality and reliability of data inputs and outputs via administrative safeguards. Thinking carefully about publication procedures, how to balance the interests of both subjects and users, how to present the data accurately and fairly, and how to maximize uses by audiences of varying sophistication can be just as resource-intensive as traditional regulation. But these steps are necessary for data to achieve “regulatory” ends. The more reliable government data are, the more they can enlighten us and deter unwanted behavior.