# On the Benefits and Costs of Public Access to Data Used to Support Federal Policy Making

Randall Lutter and David Zorn

*September 2016*

MERCATUS WORKING PAPER

MERCATUS CENTER
George Mason University

3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

*Randall Lutter and David Zorn. "On the Benefits and Costs of Public Access to Data Used to Support Federal Policy Making." Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, September 2016.*

**Abstract**

Congress is considering two bills that would require the Environmental Protection Agency (EPA) to make publicly available all data from studies that it relies on as it develops regulations. The Congressional Budget Office estimates that it would cost $250 million per year for the EPA to comply with such a requirement. As an alternative to these bills, the Obama administration points to an Office of Science and Technology Policy directive requiring that agencies spending more than $100 million per year on research issue plans to maximize public access to federally funded data. We show that this directive has not been implemented by the EPA and that there is good reason to question the validity of scientific research when the data used to create it is not publicly available. Furthermore, there is good reason to believe that the CBO significantly overestimated the cost of the bills. We recommend that all regulatory agencies generally provide public access to the data they rely on to develop economically significant regulations.

*JEL* code: H11

Keywords: data access, scientific research, regulation, public access, regulatory reform, data quality, government transparency, regulatory best practice, federally funded research, secret science

**Author Affiliation and Contact Information**

Randall Lutter
Professor of Public Policy
Frank Batten School of Leadership and Public Policy, University of Virginia
Visiting Fellow
Resources for the Future
randall.lutter@virginia.edu

David Zorn
Adjunct Professor
Antonin Scalia Law School, George Mason University
davidjosephzorn@gmail.com

**On the Benefits and Costs of Public Access to Data Used to Support Federal Policy Making**

Randall Lutter and David Zorn

Over the past few decades, the quality of published scientific research has increasingly come into question.[1] Researchers seeking to verify independently the results of articles published in prestigious scientific journals have reported different results with surprising frequency.[2] In August 2015, for example, researchers investigating 100 published papers in psychology found that, while 97 percent of original studies reported statistically significant results, only 39 percent of efforts to reproduce estimates of these effects reported finding the original results.[3] Irreproducible results pose such a serious problem that there is a growing awareness that all interested parties need to do more to contribute to a lasting and effective solution.[4]

To protect reproducibility, many scientific journals, including *Science*, *Nature*, and *Environmental Science & Technology*, have adopted policies that require authors to provide access to supporting data, statistical models, and even lab specimens. The *American Economic Review* (*AER*) investigated the reproducibility of results of published papers after the implementation of stronger data access rules.[5] The *AER* researchers found that data posting requirements are quite effective at promoting reproducibility—in the sense that analysis of original data with identical methods generates the original results. Based on their review of data

[1] John P. A. Ioannidis, "Why Most Published Research Findings Are False," *PLOS Medicine* 2, no. 8 (2005): e124.
[2] Florian Prinz, Thomas Schlange, and Khusru Asadullah, "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?," *Nature Reviews Drug Discovery* 10, no. 9 (2011): 712.
[3] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 349, no. 6251 (2015): aac4716. See also Daniel T. Gilbert et al., "Comment on 'Estimating the Reproducibility of Psychological Science,'" *Science* 351, no. 6277 (2016): 1037, and Christopher J. Anderson et al., "Response to Comment on 'Estimating the Reproducibility of Psychological Science,'" *Science* 351, no. 6277 (2016): 1037.
[4] Francis S. Collins and Lawrence A. Tabak, "Policy: NIH Plans to Enhance Reproducibility," *Nature* 505, no. 7485 (2014): 612–13.
[5] Robert A. Moffitt, "Report of the Editor: *American Economic Review* (with Appendix by Philip J. Glandon)," *American Economic Review* 101, no. 3 (2011): 684–93.

and code placed in repositories for published papers, the *AER* researchers concluded that "all but two of the articles (95 percent) could be replicated with little or no help from the author(s)."[6] Researchers had earlier found that inadvertent errors in empirical economics research were "commonplace."[7] A recent paper in *Environmental Health Perspectives*, a journal of the National Institutes of Health, proposes guidance for judging the quality of risk assessments. The guidance includes the following as a criterion for the selection of literature to be used in a risk assessment: "Sufficient data for the critical studies and the models used in the assessment are available to interested external parties so as to enable them to replicate/verify the assessment outcomes and to judge the scientific credibility of the data/models."[8]

Recognizing the need to ensure both reliability of the scientific underpinnings of its policy decisions and public confidence in that reliability, the federal government took steps in 2002 to improve the quality of and access to information it uses in policy making. In its Information Quality Guidelines, the Office of Management and Budget (OMB) states, "If an agency is responsible for disseminating influential scientific, financial, or statistical information, agency guidelines shall include a high degree of transparency about data and methods to facilitate the reproducibility of such information by qualified third parties."[9] It elaborates that "making the data and methods publicly available will assist in determining whether analytic results are reproducible."[10] OMB defines reproducibility to mean the "information is capable of being substantially reproduced, subject to an acceptable degree of

---

[6] Ibid., 7.
[7] William G. Dewald, Jerry G. Thursby, and Richard G. Anderson, "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *American Economic Review* 76, no. 4 (1986): 587–603.
[8] Penelope A. Fenner-Crisp and Vicki L. Dellarco, "Key Elements for Judging the Quality of a Risk Assessment," *Environmental Health Perspectives*, forthcoming.
[9] OMB, "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies; Notice; Republication," 67 Fed. Reg. 8460 (February 22, 2002).
[10] Ibid.

imprecision."[11] OMB further explains that "'capable of being substantially reproduced' means that independent analysis of the original or supporting data using identical methods would generate similar analytic results, subject to an acceptable degree of imprecision or error."[12]

An OMB directive to federal agencies provides for public access under the Freedom of Information Act (FOIA) to federally funded research data related to published research findings used in developing federal regulations.[13] The directive covers federal grants to and agreements with institutions of higher education, hospitals, and other nonprofit organizations:

> In response to a Freedom of Information Act (FOIA) request for research data relating to published research findings produced under an award that were used by the Federal Government in developing an agency action that has the force and effect of law, the Federal awarding agency shall request, and the recipient shall provide, within a reasonable time, the research data so that they can be made available to the public through the procedures established under the FOIA.[14]

To promote scientific integrity, President Obama signed a memorandum on scientific integrity in March 2009,[15] and the Office of Science and Technology Policy (OSTP) issued an implementing memo on scientific integrity in December 2010[16] and one on increasing access to the results of federally funded scientific research in February 2013.[17]

Some members of Congress have sought additional action by introducing two bills—H.R. 1030[18] and S. 544[19]—that would require the Environmental Protection Agency (EPA) to make publicly available supporting data from any studies that it relies on in its policy making. The

---

[11] Ibid.
[12] Ibid.
[13] OMB, *Circular A-110*, amended September 30, 1999.
[14] Ibid.
[15] Barack Obama, "Memorandum for the Heads of Executive Departments and Agencies: Scientific Integrity," March 9, 2009.
[16] John P. Holdren, "Memorandum for the Heads of Executive Departments and Agencies: Scientific Integrity," December 17, 2010.
[17] John P. Holdren, "Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research," February 22, 2013.
[18] Secret Science Reform Act of 2015, H.R. 1030, 114th Cong. (2015).
[19] Secret Science Reform Act of 2015, S. 544, 114th Cong. (2015).

Obama administration has issued statements of administration policy on these bills, indicating that a veto is likely because the bills "would undermine EPA's ability to protect the health of Americans, would impose expensive new mandates on EPA, and could impose substantial litigation costs on the Federal government. It also could impede EPA's reliance on the best available science."[20]

Ultimately, public access to data affects not only the efficacy of public policies but also public trust in the federal government's actions. Distrust can prevent the timely adoption of effective solutions to policy problems. Increasing access to the research data used in developing federal regulations may promote public trust.

In this paper, we review current federal policies and procedures intended to ensure that scientific and technical research meets appropriate quality standards and we compare them with similar practices and procedures used by nonfederal institutions. We focus on access to data and computer code because we find that requirements for public access to data and code have become a best practice in nonfederal scientific institutions.[21]

The scientific experience can inform us about the likely success of new federal policies intended to improve the quality and accessibility of information because federal policies and institutions have analogs in the scientific world. Some organizations in the scientific community are adopting best practices to protect scientific integrity. Identifying and characterizing these practices and describing their possible use by federal government agencies should help inform us about how to promote access and reproducibility. Contrary to the findings of some earlier work,

[20] "Statement of Administration Policy, H.R. 1030, Secret Science Reform Act of 2015," March 3, 2014.
[21] This industry best practice has not yet been adopted by any of the top federal scientific journals, including *Environmental Health Perspectives, Emerging Infectious Diseases*, and *Journal of Rehabilitation Research and Development*. See Randall Lutter and David Zorn, "Reinforcing Reproducibility: What Role for the Federal Government?," *Regulation* 38, no. 4 (2015–2016): 15–16.

our analysis suggests that data and code access can be provided at a reasonable cost that the benefits of transparency and greater reproducibility will likely exceed.

We also summarize major initiatives that the federal government has undertaken to improve the quality and public accessibility of federal policy making (including initiatives of the Obama administration) and their limitations. We then describe evidence that many scientific research papers present results that are irreproducible (and thus unreliable), and we describe steps that high-quality scientific journals have taken to address the issue. We next assess the benefits and costs of implementing a policy of general access to the data and code used in developing economically significant federal regulations. Finally, we make recommendations for improving the policy-making process by requiring public accessibility to the data and code underlying research that federal agencies use to support policies.

**Federal Policies on Data Quality and Public Access**

Public debate over federal policies that limit public access to the data used in regulatory decisions dates to at least the 1970s. In 1970, the FDA recommended that doctors prescribe oral hypoglycemic drugs only for patients with adult-onset (Type 2) diabetes that could not be controlled by diet and only when the patients were not insulin dependent.[22] The recommendation was made on the basis of a federally funded study carried out by the University Group Diabetes Program (UGDP), which found that the oral hypoglycemic drug tolbutamide was associated with an increased death rate from cardiovascular disease among mildly diabetic patients.[23] The results of the UGDP study were immediately controversial. Some researchers raised questions about the

[22] "Status of Problem of Usage of Tolbutamide, Preliminary Statements: FDA Statement, Friday May 22, 1970," *Diabetes* 19, no. 6 (1970): 467.
[23] Dave R. Kelleher, "Applying the Freedom of Information Act in the Area of Federal Grant Law: Exploring an Unknown Entity," *Cleveland State Law Review* 27, no. 2 (1978): 294–311.

study's design. In addition, the unavailability of oral hypoglycemic drugs would significantly reduce the treatment options for many patients. And the FDA's actions would potentially expose physicians to malpractice lawsuits.

In 1974, the proponents of oral hypoglycemic drugs, who organized as the Committee on the Care of the Diabetic (CCD), criticized the UGDP study and, using the FOIA, began requesting the data underlying the UGDP study in order to replicate the results. In 1975, the FDA proposed restrictive changes to the labeling of oral hypoglycemic drugs largely on the basis of the results of the UGDP study. In 1977, Secretary of Health, Education, and Welfare Joseph Califano declared that phenformin, another oral hypoglycemic drug, was an imminent public health hazard and withdrew FDA approval of drug products that contained it. The FDA denied the CCD's FOIA request on the grounds that the study's data were not agency records subject to the FOIA because the data were maintained by the UGDP and not by an agency of the federal government. A case was brought in federal district court as *Forsham v. Califano* on whether the data should be subject to the FOIA. Forsham and members of the CCD seeking access to the data lost in US District Court and the US Court of Appeals. In 1980, the US Supreme Court ruled that, even though the UGDP study was federally funded, the UGDP data were not subject to the FOIA as long as a federal agency did not have physical possession of the data.[24]

Two decades later, a similar issue arose when the EPA issued the National Ambient Air Quality Standards for Particulate Matter largely on the basis of federally funded research, particularly the Harvard School of Public Health's Six Cities study and an American Cancer Society (ACS) study.[25] Challenges to the regulation included criticisms of the studies' design and

[24] Forsham v. Harris, 445 U.S. 169 (1980).
[25] EPA, "National Ambient Air Quality Standards for Particulate Matter; Final Rule," 62 Fed. Reg. 38652–760 (July 18, 1997).

analysis. Efforts to obtain access to the data underlying the studies failed. The EPA did not possess the data, so the FOIA requests for the data came up empty. Researchers for both studies then gave access to the data to a team of researchers selected by the Health Effects Institute, a nonprofit research institute jointly funded by the EPA and the automotive industry that specialized in the health effects of air pollution so that the researchers could attempt to replicate the studies. In 2000, the reanalysis team reported that it had found very few coding problems with the data used in either study and that it had been able to replicate the point estimates made by the studies' researchers.[26] The team also performed a number of sensitivity analyses, a few of which showed a reduction in the estimated effects of particulate matter on mortality.[27] The Harvard and ACS researchers refused to share the data more widely on the grounds that they had promised the study participants anonymity and that the data contained personally identifiable information.[28] Lack of access to data has continued to play a significant role in the policy debate over EPA's clean air rules. In response to a congressional subpoena seeking the data used in the Six Cities and ACS studies, the EPA stated in 2014 that it still did not possess sufficient data to replicate the results of the original studies, even after multiple interactions with the owners of the data.[29]

In 1998, partly in response to the difficulties in obtaining data from the Harvard and ACS researchers, Congress passed the Shelby Amendment as part of Public Law 105-277. The amendment directs OMB to revise *Circular A-110* ("Uniform Administrative Requirements for

[26] Health Effects Institute, "Reanalysis of the Harvard Six Cities Studies and the American Cancer Society Study of Particulate Air Pollution and Mortality," July 2000, ii.
[27] Ibid., ii–iii.
[28] Elaine Appleton Grant, "Prevailing Winds," *Harvard Public Health*, Fall (2012): 30–37.
[29] Gina McCarthy, letter to Lamar Smith, March 7, 2014, accessed January 20, 2016, http://science.house.gov/sites/republicans.science.house.gov/files/documents/EPA%20letter%20to%20Smith%20March%207%202014%20(2).pdf. Lack of transparency has been alleged against the National Marine Fisheries Service and agencies in the Department of the Interior at a May 19, 2016, hearing titled "Examining Deficiencies in Transparency at the Department of the Interior" before the House Committee on Natural Resources, Subcommittee on Energy and Mineral Resources. See the testimony of Peter Seidel and Kathleen Sgamma.

Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations") "to require Federal awarding agencies to ensure that all data produced under an award will be made available to the public through the procedures established under the Freedom of Information Act."[30] The Shelby Amendment effectively negated the US Supreme Court's ruling in *Forsham v. Harris*, at least with respect to research that might be funded later. After two rounds of public comment, OMB revised *Circular A-110* in 1999 to improve public access to federally funded data. Specifically, if federally funded research findings are published in a journal or "when an agency publicly and officially cites the research findings" in issuing a federal regulation, then in the event of a FOIA request for the research data, "the awarding agency shall request, and the recipient shall provide, within a reasonable time, the research data so that they can be made available to the public" through FOIA procedures.[31]

The Shelby Amendment and *Circular A-110* improve public access to federally funded research and data, but obstacles remain. One limitation is that *Circular A-110* does not apply to research by for-profit contractors. Moreover, it is dependent on the use of the FOIA to petition for access to a specific piece of research. The FOIA allows the public to request that federal agencies provide records that the government possesses or has funded (because of the Shelby Amendment). The law requires that agencies "respond" to a request within 20 business days (four calendar weeks), plus an additional 10 business days (for a total of six calendar weeks) if a request involves searching multiple sites (which will usually be the case with research funded through contracts or grants).[32] Agencies may respond within the timeframe by affirming that records relating to the request exist. The records may be delivered to the requester at a later time.

[30] Omnibus Appropriations Act for FY 1999, Pub. L. No. 105-277 (1998).
[31] OMB, *Circular A-110*.
[32] Openness Promotes Effectiveness in Our National Government Act of 2007, Pub. L. No. 110-175 (2007).

Relying on the FOIA process requires that a requester file a request with the funding agency; that the funding agency determine whether the data requested are subject to the FOIA; and, if so, that the agency then request the data from the researcher. The researcher then sends the data to the funding agency, and the funding agency reviews the data to ensure that no data are protected from public disclosure under established FOIA exceptions.

Agencies often respond to FOIA requests quite slowly, according to independent assessments of agency responsiveness. Using 2015 statistics reported by the Department of Justice, which oversees FOIA activities for the federal government, the Center for Effective Government scored how well agencies performed at processing FOIA requests.[33] Using the data that agencies reported in their annual FOIA reports for 2013, the center rated agency performance based on 16 factors most highly weighted toward the percentage of requests fully or partially granted, the percentage of requests responded to within 20 days, the average number of days to respond to requests, and the size of each agency's request backlog. Also, the Cause of Action Institute tested agency response times to FOIA requests in 2012.[34] Table 1 shows the results of both studies of FOIA responses by agency. Other studies by Bloomberg[35] and the FOIA Project[36] yielded similar results, except they showed that many agencies do not respond to requests even after 180 days.

Other administration initiatives have sought to promote public access to data. For example, a 2013 memorandum from the Office of Science and Technology Policy (OSTP) goes beyond *Circular A-110.* The OSTP states that federally supported data should be publicly accessible and directs executive branch agencies that spend more than $100 million a year

[33] Sean Moulton and Gavin Baker, "Making the Grade: Access to Information Scorecard 2015," Center for Effective Government, March 2015.
[34] Cause of Action Institute, "Grading the Government: A Look at How Federal Agencies Measure Up on FOIA Requests," Cause of Action Institute, 2013.
[35] Jim Snyder and Danielle Ivory, "Obama Cabinet Flunks Disclosure Test with 19 in 20 Ignoring Law," Bloomberg, September 27, 2012.
[36] The FOIA Project, "Agency FOIA Backlogs and Processing Times," accessed January 15, 2016.

funding research to develop and issue plans to maximize access by the general public to digitally formatted data created with federal funds.[37] Also, federally funded research data should be deposited in a repository for public access, according to the OSTP memo. The Obama administration touts the OSTP initiative as a reason that H.R. 1030 and S. 544, requiring the EPA to provide public access to research data used in policy making, are unnecessary.[38]

**Table 1. Agency Responsiveness to FOIA Requests**

| Agency | CEG FOIA score for processing requests (%)[(a)] | Average number of days to respond to Cause of Action Institute FOIA request[(b)] |
|---|---|---|
| Department of Agriculture | 94 | 84.1 |
| Department of Transportation | 63 | 125.0 |
| Department of Health and Human Services | 60 | 135.4 |
| Department of Defense | 55 | 155.0 |
| Environmental Protection Agency | 52 | 47.0 |
| Department of Veterans Affairs | 51 | 79.5 |
| Department of Homeland Security | 51 | 148.0 |
| Department of Education | Not covered by CEG study | 21.0 |
| Department of Energy | Not covered by CEG study | 101.0 |
| Department of the Interior | Not covered by CEG study | 147.0 |
| Department of Commerce | Not covered by CEG study | No response in 240 days |

Note: FOIA = Freedom of Information Act; CEG = Center for Effective Government.

Sources: (a) Sean Moulton and Gavin Baker, "Making the Grade: Access to Information Scorecard 2015," Center for Effective Government," March 2015; (b) Cause of Action Institute, "Grading the Government: A Look at How Federal Agencies Measure Up on FOIA Requests," Cause of Action Institute, 2013.

To evaluate implementation of the OSTP initiative, we collected the data policies that each agency posted on its website to comply with the OSTP memo. Table 2 shows the major funding agencies covered by the memo,[39] the policies of each agency on how data will be accessible, and when the agency stated that the policy would be effective.

[37] Holdren, "Memorandum: Increasing Access to Results."
[38] "Statement of Administration Policy."
[39] John P. Holdren, letter to House and Senate Appropriations Committees, March 24, 2014, https://www.whitehouse.gov/sites/default/files/microsites/ostp/OpenAccess_March-2014.pdf.

## Table 2. Policies of Federal Agencies for Publicly Posting Taxpayer-Funded Research Data

| Agency | Policy for posting data | Effective date |
|---|---|---|
| Agency for Healthcare Research and Quality[(a)] | Researchers are expected to share data at the time of publication of the main findings from the dataset. The agency promotes use of publicly accessible databases. | Oct. 2015 |
| Assistant secretary for preparedness and response, Department of Health and Human Services[(b)] | Researchers must publish digital scientific datasets in a recognized scientific data repository that is capable of long-term preservation of the data and open access to the public within 30 months from the creation of the dataset or on publication of a peer-reviewed article based on the dataset, whichever is sooner. | Oct. 2015 |
| Centers for Disease Control and Prevention[(c)] | Researchers should make data available no later than 30 months after completion of collection, but only on request and only to an agency-approved party for an agency-approved public health purpose. | Oct. 2015 |
| Department of Defense[(d)] | Researchers should store digitally formatted scientific datasets at the time of publication of research where the data are publicly accessible. | pending rulemaking |
| Department of Energy[(e)] | Researchers should propose appropriate plans to provide access to data. | Oct. 2015 |
| Department of Transportation[(f)] | Researchers must ensure that unclassified data are available for public download and analysis. | Dec. 2015 |
| Food and Drug Administration[(g)] | Researchers are expected to commit to sharing digital data underlying their research findings on publication of their findings in a peer-reviewed article. | Jan. 2016 |
| National Aeronautics and Space Administration[(h)] | Researchers whose work has appeared in peer-reviewed publications must provide a plan for making the research data that underlie their results and findings digitally accessible within a reasonable time period after publication. | Feb. 2015 |
| National Institute of Standards and Technology[(i)] | Researchers must provide a plan for storage and preservation of the data and for how data will be made available to the public. | Dec. 2014 |
| National Institutes of Health[(j)] | Researchers are expected to make data available at the time the study appears in a peer-reviewed publication. | Dec. 2015 |
| National Oceanic and Atmospheric Administration[(k)] | Researchers must make data available typically within two years of collection or when an article using the data is published if earlier than two years. Data must be publicly discoverable through the agency's data inventory and must be publicly accessible via online services in widely used machine-readable formats. | March 2016 |
| National Science Foundation[(l)] | Researchers should deposit at an appropriate repository all data resulting from the research funded by an award from the foundation, regardless of whether the data support a publication. | no earlier than Jan. 2017 |
| Smithsonian Institution[(m)] | Researchers must submit digital research data supporting publications via an electronic copy or link to such copy to Smithsonian-managed or -approved repositories within a negotiated period of time. | Oct. 2015 |
| Department of Agriculture[(n)] | Researchers will be required to make the digital data underlying the conclusions of peer-reviewed scientific research publications freely available in public repositories in machine-readable formats. | 2017 |

*continued on next page*

| Agency | Policy for posting data | Effective date |
|---|---|---|
| Department of Veterans Affairs[o] | Researchers will be required to share all digital data underlying their published results from all agency-funded research at least under controlled public access mechanisms where privacy, intellectual property, or other concerns preclude open public access. | Dec. 2015 |
| Department of Education | No posted plan approved by the Office of Science and Technology Policy. | |
| Department of Homeland Security | No posted plan approved by the Office of Science and Technology Policy. | |
| Department of the Interior | No posted plan approved by the Office of Science and Technology Policy. | |
| Environmental Protection Agency | No posted plan approved by the Office of Science and Technology Policy. | |
| Office of Director of National Intelligence | No posted plan approved by the Office of Science and Technology Policy. | |
| Agency for International Development | No posted plan approved by the Office of Science and Technology Policy. | |

Note: All the agency policies are lengthy and detailed, and all include exceptions where release of data would compromise personal privacy, confidentiality, intellectual property, or national security.

Sources: (a) Agency for Healthcare Research and Quality, "AHRQ Public Access to Federally Funded Research," February 2015, accessed June 29, 2016; (b) Office of the Assistant Secretary for Preparedness and Response, "Public Access to Federally Funded Research: Publications and Data," accessed July 29, 2016; (c) Centers for Disease Control and Prevention, "CDC Plan for Increasing Access to Scientific Publications and Digital Scientific Data Generated with CDC Funding," January 2015; (d) Department of Defense, "Plan to Establish Public Access to the Results of Federally Funded Research," February 2015; (e) Department of Energy, "Public Access Plan," July 24, 2014; (f) Department of Transportation, "U.S. Department of Transportation Public Access Plan: Increasing Access to Federally Funded Research Results," accessed June 29, 2016; (g) Food and Drug Administration, "Plan to Increase Access to Results of FDA-Funded Scientific Research," February 2015; (h) National Aeronautics and Space Administration, "NASA Plan: Increasing Access to the Results of Scientific Research (Digital Scientific Data and Peer-Reviewed Publications)," November 21, 2014; (i) National Institute of Standards and Technology, "Plan for Providing Public Access to the Results of Federally Funded Research," December 4, 2014; (j) National Institutes of Health, "Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research," February 2015; (k) NOAA (National Oceanic and Atmospheric Administration) Research Council, "NOAA Plan for Increasing Public Access to Research Results," February 2015; (l) National Science Foundation, "NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation," March 18, 2015; (m) Smithsonian Institution, "Plan for Increased Public Access to Results of Federally Funded Research," August 18, 2015; (n) US Department of Agriculture, "Implementation Plan to Increase Public Access to Results of USDA-Funded Scientific Research," November 7, 2014; (o) Department of Veterans Affairs, "Policy and Implementation Plan for Public Access to Scientific Publications and Digital Data from Research Funded by the Department of Veterans Affairs," July 23, 2015.

The OSTP initiative has not been effectively implemented. First, it would not accomplish the goals of H.R. 1030 and S. 544 because it covers only federally funded research and not other research that agencies rely on for policy making, and it is clear that public access to research data is not *required* by most agencies. Most only require commitments to share the data. In the next

section, we will see that such policies employed by scientific journals have proven to be ineffective at ensuring accessibility to research data. Also, more than three years after OSTP issued its directive, a number of major research funding agencies have failed to issue the necessary plans. The EPA, the Office of the Director of National Intelligence, US Agency for International Development, and the departments of Education, Homeland Security, and the Interior have not posted plans to comply with the OSTP initiative.[40]

Public access to data would be inadequate or not timely even for those agencies that have posted approved plans. Plans for the National Science Foundation and the Department of Agriculture are not scheduled to go into effect until sometime in 2017 and for the Department of Defense possibly later, depending on the length of its rulemaking process. Also, several agency plans only require data access for research once it has been published in a peer-reviewed journal, while others make provision for access to data not associated with a publication. Anyone wanting gain access to data funded by the Centers for Disease Control and Prevention must apply via a questionnaire that includes describing the requester's research qualifications and reasons for wanting the data.[41] If the requester's credentials and interest are deemed meritorious, the requester may still have to wait 30 months after final collection of the data.[42]

**Irreproducibility in Scientific Research and Policies to Enhance Reproducibility**

Access to the data necessary to replicate scientific studies is essential because the results of so many peer-reviewed scientific publications have proven to be impossible to reproduce. For

[40] John P. Holdren, letter to House and Senate Appropriations Committees, April 29, 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/public_access_report_to_congress_apr2016_final.pdf.
[41] CDC (Centers for Disease Control and Prevention), "CDC Plan for Increasing Access to Scientific Publications and Digital Scientific Data Generated with CDC Funding," January 2015.
[42] Ibid.

example, researchers at Amgen were able to replicate only 11 percent of 53 major cancer research papers published between 2001 and 2011.[43] Researchers at Bayer reported that they could reproduce the results reported in a set of drug research studies relevant to the company only 25 percent of the time.[44] Researchers reviewing articles in the fields of neuroscience, developmental biology, immunology, cell and molecular biology, and general biology showed that in 54 percent of papers the methods and materials were not identified well enough to permit replication.[45] A survey of psychologists found that researchers could successfully replicate results of only 49 percent of 257 peer-reviewed papers.[46]

The federal government recognizes the challenges that irreproducible scientific research poses for innovation, the greater scientific enterprise, science-based policy development, and the efficient allocation of research funding. In 2014, the OSTP and the National Economic Council issued a request for information on how the federal government can "identify policy opportunities to promote innovation and its economic benefits in the United States."[47] One of the questions was the following: "Given recent evidence of the irreproducibility of a surprising number of published scientific findings, how can the Federal Government leverage its role as a significant funder of scientific research to most effectively address the problem?"[48] One approach to this question is to review how the nonfederal scientific community seeks to ensure reproducibility specifically and research quality generally.

---

[43] C. Glenn Begley and Lee M. Ellis, "Drug Development: Raise Standards for Preclinical Cancer Research," *Nature* 483, no. 7391 (2012): 531–33.
[44] Prinz, Schlange, and Asadullah, "Believe It or Not."
[45] Nicole A. Vasilevsky et al., "On the Reproducibility of Science: Unique Identification of Research Resources in the Biomedical Literature," *PeerJ* 1 (2013): e148.
[46] Joshua K. Hartshorne and Adena Schachner, "Tracking Replicability as a Method of Post-publication Open Evaluation," *Frontiers in Computational Neuroscience* 6, no. 8 (2012): 1–13.
[47] Office of Science and Technology Policy and National Economic Council, "Strategy for American Innovation, Action: Notice of Request for Information," 79 Fed. Reg. 44064–68 (July 29, 2014).
[48] Ibid., 44066.

OMB uses peer review as a standard for research relevant for policy, stating that properly peer-reviewed articles deserve a rebuttable presumption of substantial reproducibility.[49] Peer review in academia is the process that journal editors use to judge the significance and originality of research papers submitted for publication. When journal editors send manuscripts to referees for peer review, they typically ask whether a manuscript properly reviews the existing literature, uses methods adequate to support its conclusions, and reaches conclusions that represent a meaningful contribution to the literature. Reviewers are rarely asked to verify the findings of studies they review, and they typically lack the incentives or resources to do so. [50] Thus, peer review does not address whether research findings are reproducible.[51] In 2002, a report of the National Research Council stated that "peer review alone does not detect fraud, validate factual findings . . . or substitute for the judgments of the scientific community as a whole."[52]

Redoing experiments for research or policy-making purposes may be prohibitively costly for studies that were conducted over a period of years or that required special access to research subjects. Replication using the original data is still important to ensure reliability. For this reason, a number of the most prominent scientific journals require that authors commit to data sharing.[53]

---

[49] Joshua B. Bolten, "Memorandum for the Heads of Departments and Agencies, Subject: Issuance of OMB's 'Final Information Quality Bulletin for Peer Review,'" December 16, 2004.
[50] Sara Schroter et al., "What Errors Do Peer Reviewers Detect, and Does Training Improve Their Ability to Detect Them?," *Journal of the Royal Society of Medicine* 101, no. 10 (2008): 507–14. Researchers intentionally inserted eight errors into a 600-word paper and sent the paper to 300 reviewers. None of the 300 reviewers noted more than five of the eight errors, and 20 percent of reviewers failed to note any of the eight errors. The median number of errors identified by reviewers was two.
[51] Tom Jefferson, Philip Alderson, Elizabeth Wager, and Frank Davidoff, "Effects of Editorial Peer Review: A Systematic Review," *Journal of the American Medical Association* 287, no. 21 (2002): 2784–86
[52] National Research Council, *Access to Research Data in the 21st Century: An Ongoing Dialogue among Interested Parties, Report of a Workshop* (Washington, DC: National Academy Press, 2002).
[53] Lutter and Zorn, "Reinforcing Reproducibility."

Even in cases in which journal policies require authors to commit to sharing data on request, authors rarely follow through on those promises.[54] This finding has been instrumental in persuading journal editors to require that data be placed in repositories or otherwise made publicly accessible as a condition for publication, rather than just requiring authors to commit to sharing data on request.[55] The use of public repositories for the archiving of data has become virtually universal in evolutionary biology.[56] Some research shows that journal requirements to archive data increase data availability 1,000-fold compared with journals with no policy at all, suggesting that requirements for data archiving are very important.[57]

Posting study data has proven to be effective at improving the reliability of research in economics. In empirical economics, a study of replication of well-regarded peer-reviewed research in a highly regarded journal suggested that inadvertent errors may be "commonplace rather than rare occurrences."[58] The *AER* subsequently adopted a policy "to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication." Further, the *AER* conducted an evaluation of its policy and reported in 2011 that about 80 percent of 39 sampled papers met the spirit of the data availability policy and that 95 percent were substantially reproducible. Independent efforts at replication of nine selected papers found no serious errors (with almost exact replication for five studies and "several small discrepancies . . . immaterial to the

[54] Alawi A. Alsheikh-Ali et al., "Public Availability of Published Research Data in High-Impact Journals," *PLOS ONE* 6, no. 9 (2011): e24357; Caroline J. Savage and Andrew J. Vickers, "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals," *PLOS ONE* 4, no. 9 (2009): e7078; and Dewald, Thursby, and Anderson, "Replication in Empirical Economics." Also, Feigenbaum and Levy show that researchers have professional incentives not to voluntarily share data or willingly assist in replication efforts. Susan Feigenbaum and David M. Levy, "The Market for (Ir)Reproducible Econometrics," *Social Epistemology* 7, no. 3 (1993): 215–32.
[55] Moffitt, "Report of the Editor."
[56] Bryan T. Drew et al., "Lost Branches on the Tree of Life," *PLOS Biology* 11, no. 9 (2013): e1001636.
[57] Timothy H. Vines et al., "Mandated Data Archiving Greatly Improves Access to Research Data," *FASEB Journal* 27, no. 4 (2013): 1304–8.
[58] Dewald, Thursby, and Anderson, "Replication in Empirical Economics," 587.

conclusions" for another four).[59] This result represents a marked improvement relative to the results of the original 1986 study of replication by Dewald, Thursby, and Anderson. The difference is attributable, at least in part, to the change in the *AER*'s policy of data accessibility. Although analytic methods underlying papers published in the *AER* are different from those used in other disciplines, the experience of the *AER* suggests that data accessibility improves the reliability of the results of published, peer-reviewed scientific papers. Economic methods are broadly similar to those used in other types of scientific research in that they involve complicated statistical analyses of large volumes of nonexperimental data.

Administrative measures taken to date by the federal government have not been adequate to provide timely access to the data and code necessary to assess the independent reproducibility of scientific findings used in federal regulations. Yet the experience of scientific journals suggests that such replication is important because published articles have been found to contain errors with surprising frequency. Thus, one might ask what the benefits and costs are of a policy change that would require agencies to make publicly available all the data and code underlying their regulatory decisions. We next turn to the two parts of this question, focusing on the requirements of H.R. 1030 and S. 544, if extended to the federal government.

**Costs of Greater Access to Data Relevant to Federal Rulemaking**

The cost of providing access to data has been one of the primary concerns about requiring access to data used by the federal government.[60] H.R. 1030 and S. 544 would require the EPA to ensure that the data and computer code underlying any scientific research that an agency relies on in a "risk, exposure, or hazard assessment, criteria document, standard, limitation,

[59] Moffitt, "Report of the Editor."
[60] A. A. Rosenberg et al., "Congress's Attacks on Science-Based Rules," *Science* 348, no. 6238 (2015): 964–66.

regulation, regulatory impact analysis, or guidance" is publicly available online. According to the Congressional Budget Office (CBO), "Based on information from EPA, CBO estimates that the agency would spend, on average, $10,000 per scientific study for activities to meet the bill's requirements. Specifically, such funding would cover the costs of obtaining all of the underlying data used in a study, reviewing the data to address any confidentiality concerns, formatting the data for public access, providing access to the computer codes and models used in the study's analysis, and providing descriptions and documentation on how to access the data. Such activities could entail correspondence and negotiations with study authors and publishers and computer processing services to construct and maintain databases to store study-related information."[61] On the basis of that number and an estimate that the EPA references about 25,000 scientific studies per year in its rulemaking, CBO estimates that it would cost the EPA about $250 million per year to comply with the requirements of the bills if they were enacted.[62]

We develop an alternative and more transparent estimate of the costs of complying with those bills using estimates that the EPA has already developed for existing requirements that certain firms submit data. The costly activities and services that need to be performed to provide data access can be divided into two categories—data collection and data accessibility. Data collection includes most of the activities listed by CBO:[63] correspond with researchers and publishers to obtain the data, review the data for confidentiality concerns, format the data for public access, publicly post the computer code and models used in each study's analysis, and provide descriptions and documentation on how to obtain the data. Data accessibility includes

[61] CBO, "Cost Estimate, S. 544, Secret Science Reform Act of 2015," June 5, 2015.
[62] Ibid., 3.
[63] Ibid., 2.

the last activity mentioned by CBO: provide "computer processing services to construct and maintain data bases to store study-related information."[64]

When federal agencies require that industries or individuals provide information to the government, under the provisions of the Paperwork Reduction Act they must estimate the amount of time needed to provide the information. The EPA's Health and Safety Data Reporting Rule (40 C.F.R. 716), most recently updated in 2012, requires the chemical industry to undertake activities similar to the data collection activities that the EPA would need to perform under H.R. 1030 and S. 544.[65] The rule requires manufacturers, processors, and distributors to identify any health and safety studies in their possession that relate to the health or environmental effects of certain chemical substances and mixtures, to copy and summarize the relevant studies, to make lists of studies that are currently in progress, and to review the studies for confidential business information.

The EPA's supporting statement for its information collection request under the Paperwork Reduction Act gives the number of hours that the EPA estimates for data collection activities. The following estimates are from the EPA's 2015 supporting statement for the Health and Safety Data Reporting Rule.[66] The EPA estimates that it would take chemical manufacturers and processors 3.0 hours to determine which of their locations might have relevant studies, plus 4.5 hours to search through the files at those locations for the relevant studies. Those activities should roughly correspond to the efforts that a federal employee would need to spend communicating with researchers and publishers to locate the data underlying a published study and to obtain the data for compliance with a data access policy.

---

[64] Ibid.
[65] EPA, "Health and Safety Data Reporting; Addition of Certain Chemicals," 77 Fed. Reg. 71561–67 (December 3, 2012).
[66] EPA, "Supporting Statement for a Request for OMB Review under the Paperwork Reduction Act," August 31, 2015.

The EPA estimates that it would take chemical manufacturers and processors 1.0 hour to review each study for confidential business information. That should closely correspond to the amount of time needed for a federal employee to review study data for confidentiality concerns in preparation for public disclosure under a data access policy.

The EPA estimates that it would take chemical manufacturers and processors 1.0 hour to photocopy all relevant studies for submission to the agency. Given modern technology, by the time research has been published, almost all relevant underlying data and computer code and models will be in electronic format, so photocopying will be unnecessary. However, formatting unformatted data for public access can take a significant amount of time. In the absence of better information on this point, we surmise that formatting unformatted data and making the analytic models and computer code used in EPA analyses available may, in some cases, take 10.0 hours per study.

The EPA estimates that it would take chemical manufacturers and processors 12.0 hours to make a robust summary of the studies they would submit to the agency. That should roughly correspond to the amount of time needed for a federal employee to provide descriptions and documentation on how to access the data. We note that the level of effort and education necessary to provide a robust summary of scientific research is significantly greater than that needed to write metadata descriptions of study data and instructions on how to make the data available for use by the public. Based on the EPA's estimates, we can presume that the data collection activities needed to make public the data underlying the studies that the EPA uses in its rulemaking would take 30.5 hours per study.

In the same information collection request, the EPA explains how it calculates the monetary cost of paperwork processing activities that the agency must perform for its Health and

Safety Data Reporting Rule. The EPA uses the basic hourly wage for a Grade 13, Step 5 federal employee and adds 60 percent to account for benefits and overhead (the nonwage costs of employee time).[67] The Office of Personnel Management's 2015 General Schedule Locality Pay Table for the Washington–Baltimore–Northern Virginia area lists that basic hourly wage as $49.32.[68] Adjustment for benefits and overhead brings the full labor cost to $78.91 per hour. At that rate, the 30.5 hours spent on data collection activities would cost $2,407 for each study relied on by the EPA.

Once the EPA collects and prepares the data for public posting, there will be a cost for storage and maintenance of the data for public accessibility. Researchers at Indiana University have estimated the cost of constructing and maintaining a scientific data repository large enough to contain the data for 64,340 scientific publications, with data files of 32 GB per publication.[69] The average annual number of new research publications supported by National Science Foundation funding is 64,340, and 32 GB is the average size of a dataset associated with such research.[70] The Indiana University researchers estimate the cost of providing storage, maintenance, and access to the data for each publication to be $151.[71]

Based on this information, we estimate the total cost to the EPA for data collection and public accessibility would be $2,558 per study, or about 26 percent of the $10,000 per study cost estimated by CBO. These cost estimates (both CBO's estimate and the one we present here) assume a baseline of no public access to the EPA data. We estimate, however, that $592 (or 23

[67] Ibid., 15.
[68] Office of Personnel Management, "Salary Table 2015-DCB, Incorporating the 1% General Schedule Increase and a Locality Payment of 24.22% for the Locality Pay Area of Washington-Baltimore-Northern Virginia, DC-MD-VA-WV-PA, Total Increase: 1%, Effective January 2015," accessed July 29, 2016.
[69] Beth Plale et al., "Repository of NSF-Funded Publications and Related Datasets: 'Back of Envelope' Cost Estimate for 15 Years," March 2013.
[70] Ibid.
[71] Ibid., 8.

percent) of the total cost is just for obtaining the data. To the extent that the agency uses the same scientific research in its decision making regarding multiple rules, the cost of making research data publicly accessible would be less than $2,000 per study in those cases, or less than 20 percent of the cost estimated by CBO. Finally, to the extent that study authors posted the necessary data when their studies were published, the costs would be lower still. Many journals require authors to post their supporting data as a condition of publication.

CBO's cost estimate of $250 million per year for the EPA to comply with H.R. 1030 and S. 544 depends not only on the cost per study but also on the number of studies that the EPA relies on per year. CBO estimates that the EPA uses an average of 25,000 studies per year, based on a midpoint of 12 to 50,000 studies referenced for two different regulations.[72] We can use information from Regulations.gov to make a more transparent estimate. During the 10 years between January 1, 2005, and December 31, 2014, Regulations.gov listed 177,772 documents as being placed in the EPA's dockets and categorized as "Supporting and Related Material." That category includes some scientific research, some documents summarizing many pieces of scientific research, and many other nonscience-related documents such as administrative documents produced by the agency. A reasonable estimate is that each supporting document represents a single piece of scientific research. In this case, the EPA would reference, on average, 18,000 pieces of scientific research each year.

Using any estimate of the number of pieces of research *referenced* by the EPA is, however, very likely to be an overestimate of the number of pieces of research that would be covered by the texts of H.R. 1030 and S. 544. Both bills refer to research "relied upon" by the agency. The bills do not define the phrase or clarify what research is included by the term, but it

[72] CBO, "Cost Estimate, S. 544."

is reasonable to interpret the phrase "relied upon" as more narrow than referenced. The agency may reference many pieces of research that are related to a rulemaking but that it does not truly rely on to influence or justify a provision of the rule. For example, all of the EPA's recent National Ambient Air Quality Standards rules present estimates of the costs to comply with Executive Order 12866, but these costs are irrelevant during judicial review.[73]

Assuming a cost of $2,558 per study, our estimate of the total annual cost for the EPA to obtain and post the data for the amount of scientific research that the agency has traditionally cited per year would be $46 million. In its estimate, CBO mentions that costs over time would decline; once data had been obtained and posted for a study, there would be no additional cost to relying on that study again. The same would be true of our estimate.

The EPA may find that it is unable to obtain the underlying data for many scientific studies. Researchers have shown that, even when authors say their data are available on request, a large percentage of authors do not provide data on request.[74] They do not respond; they respond after months of delay; or they respond without sharing their data. If this is the case with research that the EPA wants to rely on, the EPA's costs associated with such studies will only be $592 for attempting to obtain the data and a small additional amount for asking the authors repeatedly. Based on the studies that have attempted to obtain access to data from peer-reviewed studies, we estimate that after spending 7.5 hours attempting to obtain data from study authors, the EPA will receive data for only 20 percent of the requested studies.[75] In that case, we estimate that the full $2,558 cost per study will apply to only 3,600 studies per year (20 percent of 18,000)

[73] Whitman, Administrator of Environmental Protection Agency, et al. v. American Trucking Associations, Inc., et al., 531 U.S. 457 (2001).
[74] Timothy H. Vines et al., "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology* 24, no. 1 (2014): 94–97, and Youngseek Kim and Melissa Adler, "Social Scientists' Data Sharing Behaviors: Investigating the Roles of Individual Motivations, Institutional Pressures, and Data Repositories," *International Journal of Information Management* 35, no. 4 (2015): 408–18.
[75] Ibid.

and that the $592 cost of attempting to obtain the data will apply to 14,400 studies per year (80 percent of 18,000), for a total cost of less than $18 million.

Those who object to H.R. 1030 and S. 544 say that, when the EPA is not able to use scientific studies because supporting data are not available, it will "weaken the ability of science to inform federal rule-making."[76] Such a claim seems to ignore the fact that a large percentage of published studies are unreliable. Further, the willingness to make data available is related to the strength of the evidence and the quality of reporting of the statistical results.[77] So one may presume that regulatory policies are more likely to be based on valid scientific relationships where data are available.

**Benefits of Greater Access to Data**

Public access to the data underlying studies used by federal agencies in making significant public policies may lead to increases in the true net benefits of federal policies by helping to ensure that the policies are based on valid science and not on published studies with irreproducible results.

Available data let us calculate how large the increases in net benefits of regulations from improved reproducibility would need to be to exceed the costs of providing this greater reproducibility. To calculate this increase, we begin with a 2014 OMB report that states that the EPA's estimates of the annualized benefits of 34 major rules, finalized by the EPA between October 1, 2003, and September 30, 2013, were $165 billion to $850 billion.[78] This report also states that the estimated annualized costs for major rules issued during that decade were $38 billion

[76] Rosenberg et al., "Congress's Attacks."
[77] Jelte M. Wicherts, Marjan Bakker, and Dylan Molenaar, "Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results," *PLOS ONE* 6, no. 11 (2011): e26828, and Moffitt, "Report of the Editor."
[78] OMB, "2014 Report to Congress on the Benefits and Costs of Federal Regulations and Unfunded Mandates on State, Local, and Tribal Entities," June 15, 2015.

to $46 billion. Dividing the benefits and costs for the 10 years of rules suggests that the annual net benefits of these major rules are $12 billion to $81 billion, a range derived by subtracting the largest cost estimate from the smallest benefit estimate and the smallest cost estimate from the largest benefit estimate. Improvements in reproducibility can be thought of as increasing the net benefits of regulations because they would avoid situations in which costs or benefits are wrongly estimated to occur or in which regulatory costs are imposed without corresponding benefits. More specifically, we can calculate an increase in existing net benefits from greater reproducibility, which, if it occurred, would cover the costs of obtaining the data and making the data available.

To address fully the uncertainty in such a calculation, we consider both the range of uncertainty in annual net benefits of the EPA's rules and the uncertainty in the costs of providing accessibility to the data underlying those rules. As just discussed, the baseline annual net benefits could be either $12 billion or $81 billion, as in two rows of table 3. Similarly, the incremental cost of providing accessibility could be either of the two estimates presented in the last section ($18 million or $46 million) or CBO's estimate of $250 million. We represent these possibilities as three columns in table 3. The content of each of the six cells in the table represents how large the incremental improvement in annual net benefits from the EPA's rules would have to be for such improvements to outweigh the costs of achieving them. As shown, an improvement in net benefits of 0.02 to 2.08 percent would imply that the net benefits of requiring data access are positive.

These estimates are conservative insofar as they ignore the incremental net benefits of the 287 nonmajor final rules that the EPA issued during the 10-year period ending in September 2013.[79] The estimates also ignore other important benefits of transparency, public participation, and collaboration. Making the data publicly available to verify the findings of research that influences

[79] From a search of Regulations.gov, accessed January 19, 2016, http://www.regulations.gov/#!searchResults;rpp=25;po=0;dct=FR;a=EPA;dkt=R;pd=10%257C01%257C03-09%257C30%257C13;docst=Final+Rule.

policy-making may increase the level of trust in federal policies. It is also likely that providing all sides of controversial issues with access to relevant scientific data would serve to focus debates more on strengthening the relationship of policies to reproducible science.

**Table 3. Percentage Increases in Estimated Regulatory Net Benefits Needed to Equal Various Estimates of the Cost of Data Access**

| Annual net benefits of new EPA regulations | Range of annual costs of ensuring access to data used in EPA regulations, % | | |
|---|---|---|---|
| | $18 million (assuming 20% data availability) | $46 million (assuming 100% data availability) | $250 million (CBO estimate, assuming 100% data availability) |
| $12 billion | 0.15 | 0.38 | 2.08 |
| $81 billion | 0.02 | 0.06 | 0.31 |

Note: EPA = Environmental Protection Agency; CBO = Congressional Budget Office.

Source: Authors' calculations. See text for explanation.

Some research institutions assert that data availability enhances the scientific enterprise. As the National Institutes of Health explains, "Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined."[80] In the National Institutes of Health's estimation, data sharing is required to speed the implementation of efforts to improve public policies.[81] Beyond data sharing, providing public access to data in archives has important social benefits in

[80] National Institutes of Health, "NIH Announces Draft Statement on Sharing Research Data," March 1, 2002. For some illustrations of the benefits of widely shared information, see the discussion on crowdsourcing by Jerry Brito in "Hack, Mash, & Peer: Crowdsourcing Government Transparency," *Columbia Science and Technology Law Review* 9 (2008): 119.

[81] "Data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health." Quoted in National Institutes of Health, "Final NIH Statement on Sharing Research Data," February 26, 2003. A similar point is made in the editorial, "Sharing Data to Save Lives," *Nature Medicine* 21, no. 1235 (2015): nm.3991.

preserving the public stock of data, which otherwise is easily lost. As some evolutionary biologists explain, "Once the results of a study are published (if ever), the data on which those results are based are often stored unreliably, subject to loss by hard drive failure and (even more likely) by the researcher forgetting the specific details required to use the data. Moreover, most data are never available to the broader community, even after publication of the results; in most cases this unavailability is permanent due to the eventual death of the researchers involved. We are losing nearly all of this important legacy."[82] Of course, our estimates of the benefits of public access to data supporting federal regulatory decisions fall short of proving that the benefits outweigh the associated costs. They do show, however, the plausibility of such a claim.

**Policy Recommendations**

We show that, without public access to data, federal agencies are at risk of making policy decisions based on flawed information that can misdirect public and private resources. Moreover, public access to influential data is essential for agencies to maintain transparency and for the public to have a meaningful opportunity to participate in the regulatory process in an informed manner. A policy prescribing public access to data in studies that the EPA relies on for its rulemaking would likely offer net benefits with costs much smaller than those estimated by CBO for H.R. 1030 and S. 544. Still, those legislative proposals could be improved. We suggest several refinements to require public access for data used in federal rulemaking.

First, those legislative proposals should be broadened to cover all regulatory agencies. As we show here, papers in numerous scientific disciplines frequently contain irreproducible results,

[82] Mark D. Rausher et al., "Data Archiving," *Evolution* 64, no. 3 (2010): 603–4. Rausher et al. also note the value that data archives provide for reproducibility: "The availability of data for published studies also allows error-checking, making science more open, and letting us more rapidly reach accurate conclusions."

making every federal agency that uses such research results vulnerable to having irreproducible results inadvertently influence policy.

Second, H.R. 1030 and S. 544 should target regulations that are economically significant as defined by Executive Order 12866. According to a search of RegInfo.gov, executive branch agencies published 66 economically significant final regulations in 2015; according to a search of Regulations.gov, the federal government published 1,124 final regulations during 2015. Even though economically significant regulations represent a small percentage of the regulations published, OMB considers them to account for the "vast majority of costs and benefits of new Federal regulations."[83]

H.R. 1030 and S. 544 should be amended to define "relied upon" to clarify that those legislative proposals affect only research that an agency uses to support or define key dimensions of policy. Research that merely provides background information relating to a policy is not influential research that is "relied upon" by an agency.

We are not recommending that agencies use the data obtained to replicate the results of studies, although it would be a sensible approach to show reasonable due diligence in regard to the scientific basis for public policies. We are, however, recommending that agencies seek to obtain the data underlying the studies that they rely on and then post the data publicly (after adopting appropriate protections for confidential business information and human subject and patient privacy), so that interested parties can attempt to replicate the results of the studies.

The existence of personally identifiable information (PII) in research data need not be an insurmountable barrier to broader access. A 2007 OMB memorandum with the subject "Safeguarding against and Responding to the Breach of Personally Identifiable Information"

[83] OMB, "Draft Report to Congress on the Costs and Benefits of Federal Regulations," 62 Fed. Reg. 39366 (July 22, 1997).

recognizes that different data have different levels of impact with PII generally having moderate or high impact.[84] OMB should also instruct the agencies to maximize access to such data if they are used by a federal agency in rulemaking. Depending on the risks to privacy posed by the PII at issue, OMB should encourage agencies to select controls from a suite of measures that can be adopted to protect PII. The range of potential measures includes the following:

- requiring applications for access,
- imposing nondisclosure agreements,
- requiring online training for researchers on how to protect PII,
- implementing digital rights management technologies to prevent copying or redistribution of data,
- establishing physical controls on how data is stored,
- air-gapping computers used to access the data so that the data is never exposed to the Internet,
- restricting the printing of data,
- allowing access to data only at Federal Statistical Research Data Centers,
- allowing data to be used only for the purposes of replication, validation, and sensitivity evaluation,
- requiring background checks,
- requiring users to post performance bonds that will be forfeited if they inadvertently act to release PII,
- imposing civil or criminal penalties for the release of PII, and
- blacklisting violators from accessing PII in the future.

These special considerations for providing access to data containing PII cannot all be legitimately applied to all research data. Specifically, agencies cannot treat all data as though they contain equally sensitive PII.

[84] Clay Johnson III, "Memorandum for the Heads of Executive Departments and Agencies, Subject: Safeguarding Against and Responding to the Breach of Personally Identifiable Information," May 22, 2007. OMB has long recognized that agencies should continue to protect the confidentiality of data to the degree promised to research subjects in the consent forms that were approved by the Institutional Review Board for the research.

Many federal agencies already provide access to data containing PII under certain circumstances and already have guidelines for handling PII. Examples include Internal Revenue Service data that include confidential information on income and audits,[85] Agency for Healthcare Research and Quality Medical Expenditure Panel Survey data that include medical diagnosis, treatment, and billing information,[86] and Bureau of Labor Statistics National Longitudinal Survey of Youth data that include criminal records, intellectual achievement statistics, sexual activity, and substance use.[87]

The US Government Accountability Office has issued several reports in the last few years finding that federal agencies should protect PII better.[88] However, we could find no concerns the Government Accountability Office expressed in any of these reports about problems caused by providing access to data for research purposes.

When the PII in research data has the highest degree of sensitivity, so that the data are accessible only after an application process, we recommend that agencies significantly lengthen the standard 60-day comment period on proposed regulations in order to make allowances for the delays in accessing data.

In the event that authors do not supply their underlying data and an agency still believes that relying on the results of a study is warranted, the agency ought to explain why it has

[85] The following papers illustrate the use of such data: Jason DeBacker et al., "Once Bitten, Twice Shy? The Lasting Impact of IRS Audits on Individual Tax Reporting," March 25, 2015; Raj Chetty et al., "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility," *American Economic Review* 104, no. 5 (2014): 141–47.

[86] Agency for Healthcare Research and Quality, "Medical Expenditure Panel Survey Restricted Data Files Available at the Data Centers," October 8, 2009, accessed June 29, 2016.

[87] US Bureau of Labor Statistics, "National Longitudinal Survey of Youth 1997 Topical Guide to the Data," accessed June 29, 2016.

[88] See, for example, Government Accountability Office, "Information Security: IRS Needs to Further Improve Controls over Financial and Taxpayer Data," March 2016; Government Accountability Office, "Federal Information Security: Agencies Need to Correct Weaknesses and Fully Implement Security Programs," September 2015; Government Accountability Office, "Information Security: VA Needs to Address Identified Vulnerabilities," November 2014; and Government Accountability Office, "Information Security: Agency Responses to Breaches of Personally Identifiable Information Need to Be More Consistent," December 2013.

sufficient confidence to use the study. For example, the agency might note that other researchers have already reproduced the study results or that the data are available to third parties who sign nondisclosure agreements but that the data cannot be posted publicly.

Our recommendation is similar to one by the Administrative Conference of the United States. In 2013, this independent federal agency made recommendations regarding the use of science in administrative processes. Specifically, regarding policy making, the agency recommended that "agencies should seek to provide disclosure of data underlying scientific research, including both privately and federally funded research being considered by the agencies. Where practicable, such information should be disclosed in machine-readable format. Where such data are not subject to legal or other protections, and the data's owners nonetheless will not provide such access, agencies should note that fact and explain why they used the results if they chose to do so." Furthermore, "each agency should identify and make publicly available (on the agency website or some other widely available forum) references to the scientific literature, underlying data, models, and research results that it considered. . . . Consistent with the limitations in the Information Quality Act (IQA) guidelines . . . each agency should ensure that members of the public have access to the information necessary to reproduce or assess the agency's technical or scientific conclusions."[89]

We want to clarify that we are calling for access only to the data necessary to replicate a study. We are not calling for access to all raw research data, which are all the data collected in the course of a research study. The data needed to replicate a study will usually have been processed to standardize, format, and organize the information for analysis and distribution and to exclude some raw data (e.g., lab notes that are not relevant to the results of the study as

[89] Administrative Conference of the United States, "Science in the Administrative Process," June 14, 2013.

presented). This distinction is also made by journals such as *Science*[90] and *PLOSOne*,[91] which require the posting of all data necessary for replication as a condition of publication.

Our recommendation is more targeted than the requirements of H.R. 1030 and S. 544 in that it would initially require public access to data underlying a much smaller set of regulatory decisions—those that are economically significant. This targeting would greatly reduce the expected number of actions subject to mandatory public data access.

[90] "*Science*, Editorial Policies," accessed January 11, 2016, http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail.
[91] PLOS One, "Data Availability," accessed January 11, 2016, http://journals.plos.org/plosone/s/data-availability.