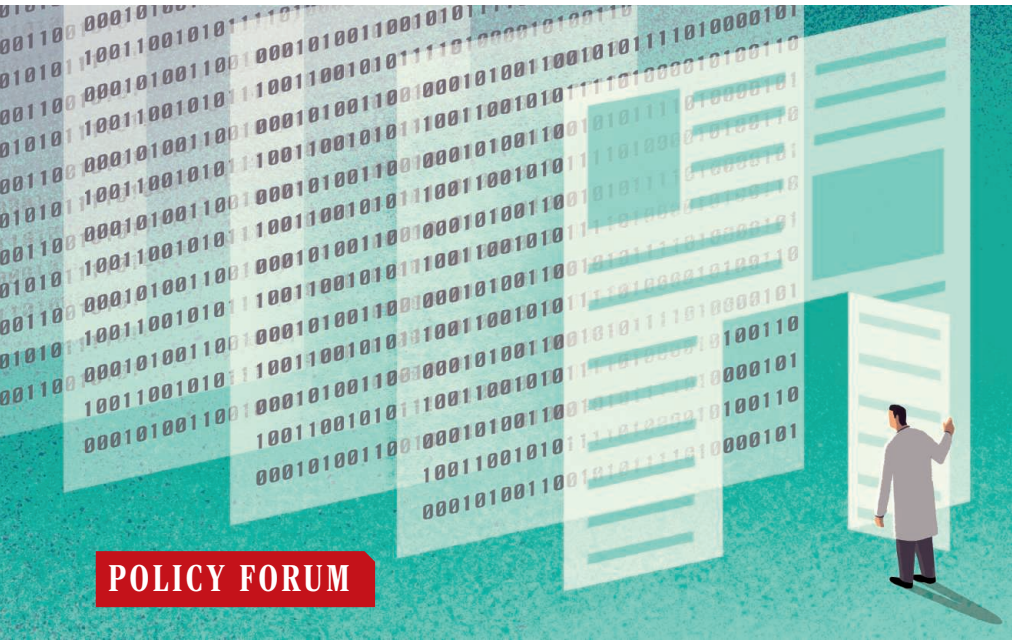

POLICY FORUM

REPRODUCIBILITY

Certify reproducibility with confidential data

A trusted third party certifies that results reproduce

By **Christophe Pérignon[1,2], Kamel Gadouche[3], Christophe Hurlin[2,4], Roxane Silberman[3,5], Eric Debonnel[3]**

Many government data, such as sensitive information on individuals' taxes, income, employment, or health, are available only to accredited users within a secure computing environment. Though they can be cumbersome to access, such microdata can allow researchers to pursue questions that could not be addressed with only public data (*1*). However, researchers using confidential data are inexorably challenged with regard to research reproducibility (*2*). Empirical results cannot be easily reproduced by peers and journal referees, as access to the underpinning data are restricted. We describe an approach that allows researchers who analyze confidential data to signal the reproducibility of their research. It relies on a certification process conducted by a specialized agency accredited by the confidential-data producers and which can guarantee that the code and the data used by a researcher indeed produce the results reported in a scientific paper.

[1]*HEC Paris, 78350 Jouy-en-Josas, France.* [2]*Certification Agency for Scientific Code and Data (cascad), 78350 Jouy-en-Josas, France.* [3]*Centre d'Accès Securisé aux Données (CASD), 91120 Palaiseau, France.* [4]*University of Orléans, LEO, 45067 Orléans, France.* [5]*French National Center for Scientific Research (CNRS), 75016 Paris, France. Email: perignon@hec.fr*

NATURAL TENSION

In general, research is said to be reproducible if the researchers provide all the resources such as computer code, data, and documentation required to obtain published results (*3*, *4*). Reproducibility has the potential to serve as a minimum standard for judging scientific claims (*5*, *6*). Recent promising initiatives to facilitate reproducible research include new research environments (e.g., Code Ocean, Popper convention, Whole Tale), workflow systems (e.g., Kepler, Kurator), and dissemination platforms or data repositories (e.g., DataHub, Dataverse, openICPSR, IEEE DataPort, Mendeley). The journal BioStatistics introduced a process in which an editorial board member aims to reproduce the results in a new submission using the code and data provided by the authors (*7*). Other examples of internal reproducibility assessments include the Applications and Case Studies of the Journal of the American Statistical Association (*8*) or the artifact evaluation process of the Principles and Practice of Parallel Programming (PPoPP) conferences. Alternatively, the reproducibility review can be outsourced to a third party, as in the partnership between the American Journal of Political Science (AJPS) and the Odum Institute (*9*).

Yet, despite the proliferation of such efforts, current computational research does not always comply with the most basic principles of reproducibility (*10*, *11*). Use of confidential data are often mentioned as a major impediment (*12*). Since the end of the 1970s, with the development of computers and a growing concern about privacy protection, legal frameworks regarding access to sensitive personal information have been reinforced in most countries. Though initially only direct identification (e.g., name, address, and social security number) was considered for defining confidentiality, the perimeter has been gradually enlarged to indirect identification that points to the use of multiple variables that potentially lead to a risk of identification (*6*). Given the extension of the legal spectrum of confidential data, a growing fraction of data used in science can fall into this category.

One possible approach to engaging in reproducible research with confidential data is to generate synthetic data by applying an information-preserving but anonymizing transformation to the initial data (*13*). By contrast, all Public Library of Science (PLOS) journals request that whenever data cannot be accessed by other researchers, a public dataset that can be used to validate the original conclusions must be provided. An alternative approach relies on improving the accessibility of restricted data for researchers—for example, the research passport proposed by the Inter-university Consortium for Political and Social Research, or the DataTags framework developed within Dataverse and the multidisciplinary Privacy Tools Project.

The natural tension between confidentiality and reproducibility could be alleviated by using a third-party certification agency to formally test whether the results in the tables and figures of a given scientific article can be reproduced from the computer code and the confidential data used by the researcher. A first attempt to implement such an external certification process is under way in France. Explicitly designed to deal with confidential data, the reproducibility assessment, like the original analysis, is conducted within a restricted-access data environment.

CERTIFICATION AND TRUST

In France, the Centre d'Accès Sécurisé aux Données (CASD) is a public research infrastructure that allows users to access and work with government confidential data

ILLUSTRATION: DAVIDE BONAZZI/SALZMANART

Downloaded from http://science.sciencemag.org/ on July 25, 2019

Published by AAAS

under secured conditions. This center currently provides access to data from the French Statistical Institute and the French Ministries for Finance, Justice, Education, Labor, and Agriculture, as well as Social Security contributions and health data.

The application process for researchers to get access to the CASD datasets takes around 6 months and requires a presentation of the research project before the French Statistical Secrecy Committee, which gathers data producers. Once the accreditation is granted, CASD creates a virtual machine allowing the researcher to access the specific source datasets required for the project, as well as the required statistical software. Remote access to the virtual machine is made possible thanks to a specific piece of hardware provided by CASD, which includes a fingerprint biometric reader. Since the inception of CASD in 2010, the question of allowing journal referees to get access to data used by researchers has been heavily debated. However, both the legal framework and the technical restrictions made intermittent and short-period access for referees difficult.

The Certification Agency for Scientific Code and Data (cascad, www.cascad.tech) is a not-for-profit certification agency created by academics (all coauthors of this paper) with the support of the French National Centre for Scientific Research, foundations, universities, and local governments. During initial meetings between cascad and CASD teams, they quickly understood the mutual benefit of joining forces to design a reproducibility certification process for confidential data. Thanks to this partnership, cascad was granted a permanent accreditation by the French Statistical Secrecy Committee to all 280 datasets available on CASD. This first-of-its-kind accreditation was motivated by the fact that cascad was providing a solution to the long-recognized problem of the lack of reproducibility of research based on confidential data. Also key for the approval was that the whole certification process remains within the CASD environment and that no data can ever be downloaded.

When an author requests a cascad certification for a paper, he or she needs to provide the paper, the computer code used in the analysis, and any additional information (software version, readme files, etc.) required to reproduce the results. Then, a reproducibility reviewer, who is a full-time cascad employee specialized in the software used by the author, accesses a CASD virtual machine that is a clone of the one used by the author. It includes a copy of the source datasets and of the author's computer code, as well as all software required to run the code. The reviewer executes the code, compares the output with the results displayed in the tables and figures of the paper, and lists any potential discrepancies in an execution report. In practice, such discrepancies can arise because of typos in the manuscript, numerical convergence issues, or differences in software package versions. This execution report is transferred to a cascad reproducibility editor, a senior researcher specialized in the author's research field, who ultimately decides on the reproducibility of the article. Lastly, a reproducibility certificate is sent to the author and is stored in the cascad database.

An example of a study recently certified by cascad is one that proposes a direct measure of tax-filing inefficiency in French cohabiting couples (*14*). The analysis relies on the *Echantillon Démographique Permanent*, an administrative dataset only available through CASD that combines information from birth, death, and marriage registers; electoral registers; censuses; tax returns; and pay slips. All the tables and figures of the paper have been reproduced by a cascad reproducibility reviewer from the source datasets, and Python scripts provided by the authors [see certificate (*15*)].

While complying with strict data confidentiality rules, the cascad-CASD partnership offers several advantages: (i) signal research reproducibility when data are confidential. The author can transfer the reproducibility certificate to an academic journal when submitting a new manuscript, similar to the reproducibility badges introduced by the Association for Computing Machinery; (ii) outsource reproducibility review. External certification enriches the peer review process, but this extra step is outsourced by academic journals, as in the AJPS-Odum partnership. The staff of the certification agency is specialized and has more time than editorial teams at academic journals; (iii) provide economies of scale to the research community. This model connects a data-provision organization (CASD, a single entry point to a large number of data producers) and a reproducibility certification organization (cascad, a single entry point to a large number of journals and researchers). The cascad-CASD model is a generalization of the standard reproducibility process in which a single researcher goes through the whole reproducibility process on his or her own, obtaining similar data and redoing the analysis; (iv) speed up reproducibility review. Any skeptical researcher can still seek to reproduce the work himself or herself, but unlike researchers who need to go through a 6-month application process, cascad reviewers benefit from a fast-track process (2 days) to access any data necessary to conduct the reproducibility assessment. The certification process is supposed to be completed within 2 weeks; (v) ease replication and robustness tests. Once certification is completed, the computer code and detailed information about the source datasets (metadata) can be publicly posted on the Zenodo archive and used to facilitate additional replication and robustness analyses.

Undoubtedly, the biggest challenge for any new certification service is to build trust. To build trust and increase credibility, cascad implements a transparent and detailed certification process. For each certification, all the actions, interactions, and problems that occurred during the process are recorded. Furthermore, all operations carried out within the virtual machine by the reviewer are recorded and traceable. Once the reviewing process is over, the environment is closed, sealed, and archived. The recorded operations and output can be referenced externally and shared with journal editors after proper accreditation.

Trust by data producers makes the process feasible, and trust by academic journals makes the process useful and worthwhile. Cascad has the trust of data producers at CASD, who want their data to be useful to society and accessible for reproducibility. Now cascad needs to convince researchers and journals to value its certificates. Overall, the experience with cascad thus far suggests that preserving confidentiality and privacy does not necessarily have to lead to opaque and nonreproducible research. ■

REFERENCES AND NOTES

1. L. Einav, J. Levin, *Science* **346**, 1243089 (2014).
2. C. Lagoze, L. Vilhuber, *Chance* **30**, 68 (2017).
3. J. B. Buckheit, D. L. Donoho, in *Wavelets and Statistics*, A. Antoniadis, G. Oppenheim, Eds., Lecture Notes in Statistics (Springer, New York, 1995), vol. 103, pp. 55–81.
4. V. Stodden *et al.*, *Science* **354**, 1240 (2016).
5. G. Christensen, E. Miguel, *J. Econ. Lit.* **56**, 920 (2018).
6. Y.-A. de Montjoye *et al.*, *Sci. Data* **5**, 180286 (2018).
7. R. D. Peng, *Science* **334**, 1226 (2011).
8. M. Fuentes, *Amstat News* (2016); http://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/.
9. T.-M. Christian, S. Lafferty-Hess, W. G. Jacoby, T. Carsey, *Int. J. Digit. Curation* **13**, 114 (2018).
10. A. C. Chang, P. Li, *Am. Econ. Rev.* **107**, 60 (2017).
11. V. Stodden, J. Seiler, Z. Ma, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2584 (2018).
12. J. K. Harris *et al.*, *PLOS ONE* **13**, e0202447 (2018).
13. B. E. Shepherd, M. Blevins Peratikos, P. F. Rebeiro, S. N. Duda, C. C. McGowan, *Am. J. Epidemiol.* **186**, 387 (2017).
14. O. Bargain, D. Echevin, N. Moreau, A. Pacifico, working paper (2019); https://doi.org/10.5281/zenodo.3241310.
15. https://doi.org/10.5281/zenodo.3256633.

ACKNOWLEDGMENTS

We are grateful to the anonymous referees and to conference participants at the 2018 Conference of European Statistics Stakeholders (Bamberg, Germany) and 2019 Advances in Social Sciences using Administrative and Survey Data (Paris, France) for their comments. We thank the French Statistical Secrecy Committee and its president Jean-Eric Schoettl, the French National Research Agency (ANR-10-EQPX-17), the French National Center for Scientific Research (CNRS), the region Centre-Val de Loire, and the HEC Paris Foundation for their support.

10.1126/science.aaw2825

Downloaded from http://science.sciencemag.org/ on July 25, 2019

Published by AAAS

Science

Certify reproducibility with confidential data

Christophe Pérignon, Kamel Gadouche, Christophe Hurlin, Roxane Silberman and Eric Debonnel

Science **365** (6449), 127-128.
DOI: 10.1126/science.aaw2825

ARTICLE TOOLS http://science.sciencemag.org/content/365/6449/127

REFERENCES This article cites 12 articles, 4 of which you can access for free
http://science.sciencemag.org/content/365/6449/127#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science* is a registered trademark of AAAS.

Downloaded from http://science.sciencemag.org/ on July 25, 2019