University of Pennsylvania
ScholarlyCommons

2018 ADRF Network Research Conference Presentations

ADRF Network Research Conference Presentations

11-2018

Reproducibility Confidentiality Data Access

Lars Vilhuber
Cornell University

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2018

Vilhuber, Lars, "Reproducibility Confidentiality Data Access" (2018). *2018 ADRF Network Research Conference Presentations*. 8.
https://repository.upenn.edu/admindata_conferences_presentations_2018/8

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2018/8
For more information, please contact repository@pobox.upenn.edu.

Reproducibility Confidentiality Data Access

Abstract
The recent concern about the reproducibility of research results has not yet been robustly incorporated into methods of providing and accessing administrative data, casting doubts on the validity of research based on such data. Reproducibility depends on disaggregating and exposing the multiple components of the research - data, software, workflows, and provenance - to other researchers and providing adequate metadata to make these components usable. The key worry is access: the authors of a study that uses administrative data often cannot themselves deposit the data with the journal, thereby impairing easy access to those data and consequently impeding reproducibility. This suggests a critical role for administrative data centers. We argue, that data held by ADRF do have attributes that lend themselves to reproducibility exercises, though this may, at present, not always be communicated correctly. We describe how ADRF can and should promote reproducibility through a number of components.

This presentation is available at ScholarlyCommons: https://repository.upenn.edu/admindata_conferences_presentations_2018/8

Reproducibility Confidentiality Data Access

Lars Vilhuber

Cornell University

Funding acknowledged under NSF-#1131848 (NCRN) and a grant from the Alfred P. Sloan Foundation.
The opinions expressed in this talk are solely the authors, and do not represent the views of the U.S. Census Bureau, the American Economic Association, or any of the funding agencies.

Replicability and Reproducibility

Bollen et al (2015), NSF Report

- Reproducibility
  ("narrow" in J Appl Econometrics, "pure replication" Hamermesh, sometimes called "replicability")
  - Same data, same code
  - [Programming] aspects of a "unit test"
- Replicability
  ("wide" in J Appl Econometrics, "scientific replication" Hamermesh, sometimes called "reproducibility")
  - Different data and/or
  - Different implementation and/or
  - Different assumptions?

Journals requiring replication archives

- Concerns about reproducibility
  - in economics, going back to the early 1980s (Dewald Thursby Anderson **1986**)
- Data (and code) requested to prior to publication
  - In the American Economic Association, since 2005
  - J Applied Econometrics since 1988 (continuously!)
  - Others have joined the fray

AMERICAN ECONOMIC ASSOCIATION

Membership About AEA Log In

Journals Annual Meeting More +

Home › Journals › AEA Journals: Policies › Data Availability Policy

Data Availability Policy

It is the policy of the American Economic Association 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.

Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AEA website. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.

Key attributes of a replication archive

- **Accessible** data
- Available **without any additional information** or help from the authors
- Clear **data provenance**
- Well-structured, comprehensible, **functioning** programs

Progress

Journals

- have implemented **verification** of submitted code and data during the editorial process (AJPS with Odun Institute; JASA)
- **highlight** data and code access criteria (badges: OSF)

- maintain **lists of acceptable third-party repositories** (and embed some of those within the submission workflow) (Nature, CoreTrustSeal)
- **interlink** with collaborating repositories to highlight authors' (and repositories') contributions to the data component of a scholarly work (Elsevier + ICPSR)

Ongoing elsewhere

Services

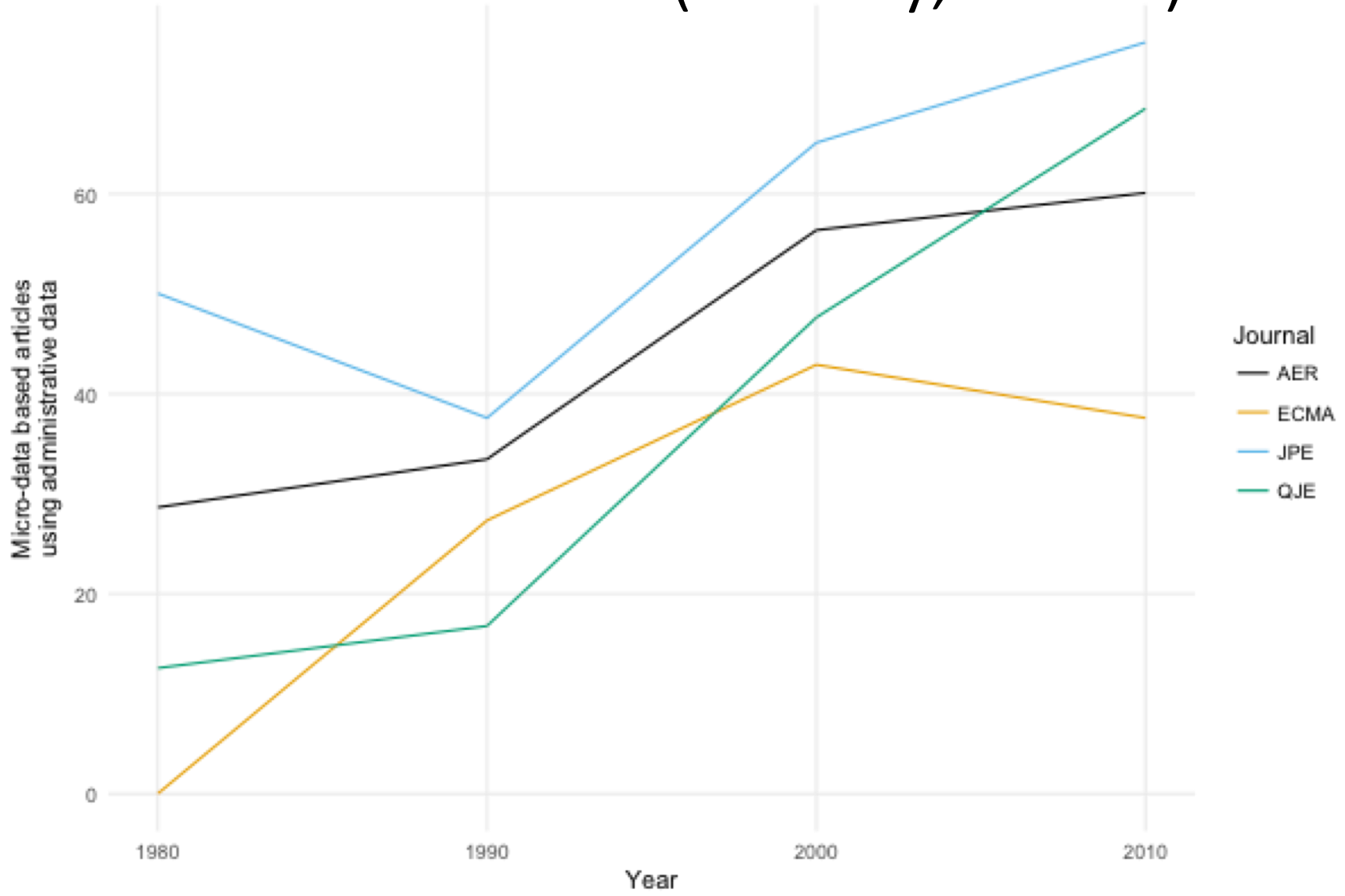
- For free or low-cost archiving
  (**openICSPR, Zenodo, figshare, dryad, Dataverse**)
- Online computational capsules
  CodeOcean (object includes OS, verification)

(note: I did not include Github, Gitlab, etc. – on purpose)

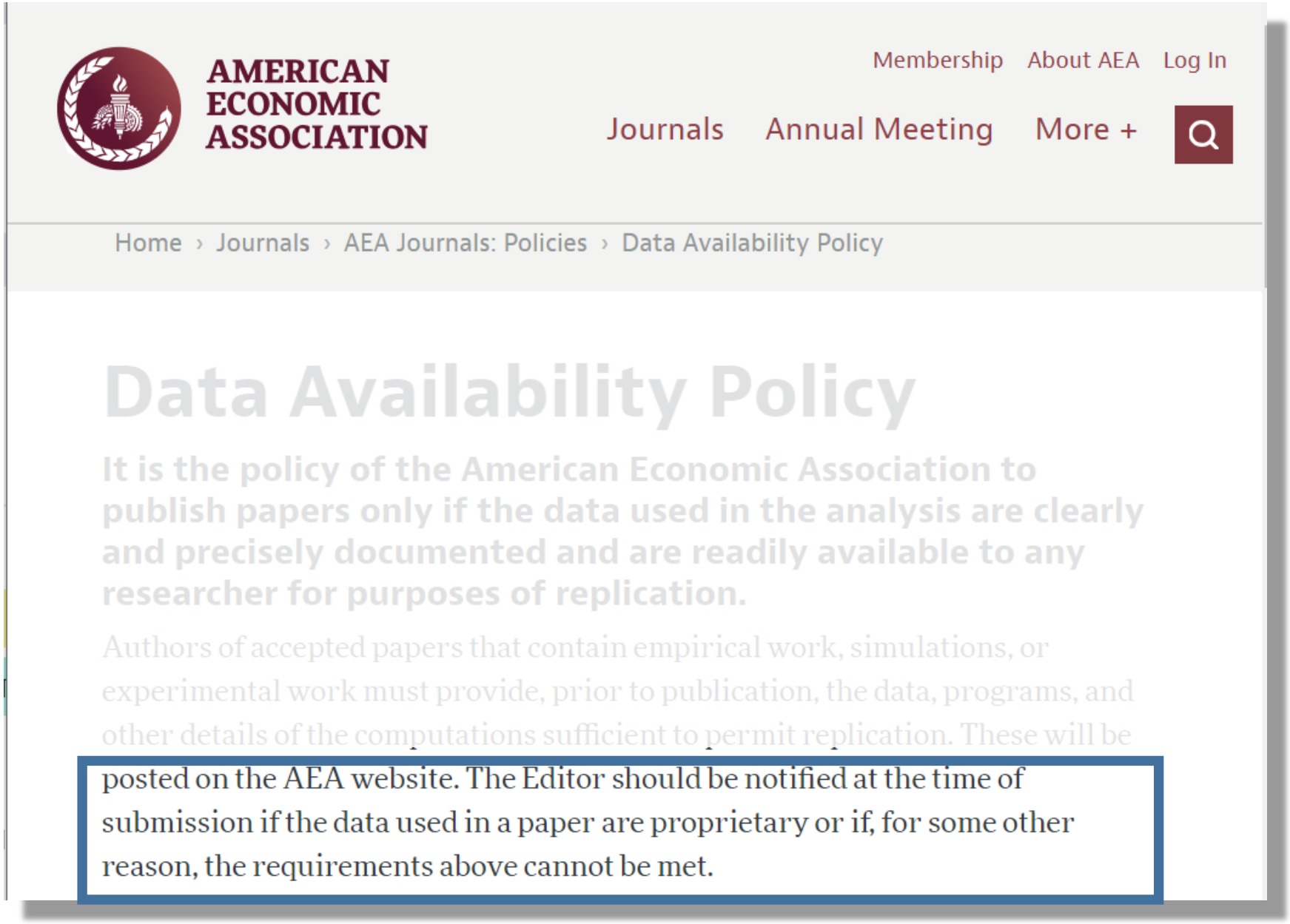
So what do you do with

Confidential data?

Use of admin data (Chetty, 2012)




CORNELL UNIVERSITY
FOUNDED A.D. 1865


AMERICAN ECONOMIC ASSOCIATION
Membership
About AEA
Log In
Journals
Annual Meeting
More +
Home › Journals › AEA Journals: Policies › Data Availability Policy
Data Availability Policy
It is the policy of the American Economic Association 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.
Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AEA website. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.

Broad adherence to AEA policy

Journals (Publisher)	Type of policy	Archive	Confidential data
AER and Journals (self)	AEA	Journal website	Exemption
QJE (OUP)	AEA	Dataverse	Exemption
ReStud (OUP)	Generic + assistance	Journal website	Exemption
ReStat (MIT)	Own	Dataverse	"… way to apply for data…"
J Applied Econometrics	Own	Own (Queens, 1988-)	Exemption
Econometrica	Own	Journal website	Exemption with "…reasonable effort…"
JOLE (Chicago)	AEA	Journal website	Exemption
JPE (Chicago)	AEA	Journal website	Exemption
JMCB	Own (barebones)	Journal website	--

Number of papers with “proprietary” data

Table 3: Type of Confidential Data

	Admin local	Admin National	Admin Regional	Private Commercial	Private Other	Total	
2009	0	2	1	0	1	4	/8
2010	2	8	0	4	3	17	/35
2011	2	9	4	1	0	16	/36
2012	1	10	2	0	2	15	/40
2013	2	2	1	4	2	11	/38
Total	7	31	8	9	8	63	/157

Table 4: Type of Access to Confidential Data

	Formal	Informal Commitment	Informal No Commitment	No Info	Total
2009	4	0	0	0	4
2010	2	3	9	3	17
2011	3	0	10	3	16
2012	12	1	0	2	15
2013	1	2	8	0	11
Total	22	6	27	8	63

Alternate view

Key insight (Lagoze and Vilhuber, 2017)

Research in restricted-access environments cycle is already a part of replicable workflow:

- Programs are verified by third-party prior to release
- Accompanying documentation is needs to detail creation of results
- Results are typically logged
- (person conducting replication or reproducibility exercise)
- This serves the same function as the typical “README” from replication archives!
- Allows for assignment of DOIs!

Moving forward in economics

- Move away from “open access only” and “deposit at journal”
- Systematically require that data be in a *trusted* repository
  - Trusted can be controlled but not (too) discriminatory access
  - Clear documentation of access policy, retention policy
  - **Uniform treatment of open access, licensed, and confidential data**
- Systematically (but with sampling) verify reproducibility of code
  - Straightforward for OA and “easy-access” data
  - Less obvious, but not impossible for restricted-access (may be part of the process of releasing disclosable results)
- Provide incentives and guidance to researchers to incorporate replicability into scientific workflow right from the start (not an afterthought)

Our recommendations to ADRFs:

- Access is already non-exclusive, and satisfied the reproducibility criteria
  - **Document the access protocol clearly and in a citable, persistent fashion. Ensure access protocol is transparent and predictable**
- Provide to the researcher the information already present in administrative database of RADEs (access rights)
  - **Provenance of input data to which the researcher has access (data citation, metadata)**
- Systematically release the programs used to generate the output
  - **Code archives for replication inside confidential areas**
- Catalog the result files, possibly certify that they were generated from the programs
  - **Provide a provable chain for the results in journals**

Our recommendations to researchers:

- With the help of ADRFs
  - **Document the access protocol clearly and in a citable, persistent fashion.**
- Data that you are use for your research
  - **Keep track of (input) data you use and (output) data you generate in a clear way (ideally with reference to public persistent web pages!)**
- Systematically request the programs used to generate the output
  - **Create code archives in public areas**
  - **Write code in a way that facilitates release!**
- Clearly identify the ADRF who provides the access
  - **Citable documents and websites**
  - **Cite their grants!**

Challenges

- Creating a public catalog of data assets with
  - persistent identifiers,
  - best-practice data citations,
  - public metadata
- Possible challenges in certifying that released results were created with the identified code
  - Often is a tabletop exercise (“plausibly generated by the code”)
  - Facilitated if researchers follow reproducible workflows!

Benefit

- Data in RADEs can become FAIR (Force11) at short notice:

  - **Findable**
  - **Accessible**
  - **Interoperable**
  - **Reusable**

thank you

lars.vilhuber@cornell.edu