Responsible Data Sharing
breakout session summary

Julia Stoyanovich
Drexel University

data RESPONSIBLY

DaSH, October 5-6, 2016

Participants

- Julia Stoyanovich
- David Belanger
- Kathy Grise
- William Mcdermott
- Reynold Panettieri
- Cheickna Sylla
- Dimitri Theodoratos
- Fusheng Wang
- Judy Zhong

Data publishing

Privacy-Preserving Data Publishing: A Survey of Recent Developments, Fung *et al.*, ACM Computing Surveys, 42 (4), June 2010.

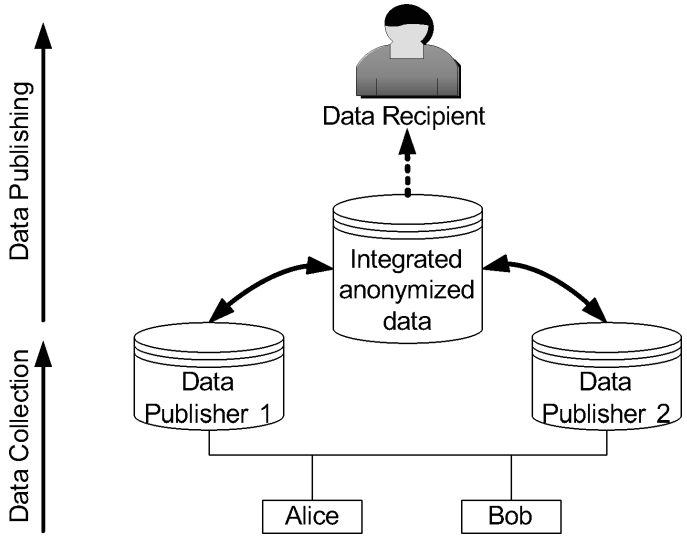


Fig. 4. Collaborative data publishing.

are we done?
do we stop at publishing?
do we worry about issues other than privacy?

Motivating example

- Bill Howe's slides: http://www.slideshare.net/billhoweuw/science-data-responsibly
- Alcohol study, Barrow Alaska 1979
- Methodological issues
- Ethical issues, with harms far beyond privacy
- Ethical rules generally followed, ethical principals grossly violated
- The specific ethical violations of this study would likely not happen today, but is the problem solved?

Stages of the data sharing pipeline

- collection
- integration
- cleaning & pre-processing
- analysis / meta-analysis
- publishing of datasets and of results of data analysis
- interpretation / interrogation

curation (manual / semi-automatic / automatic) permeates all stages

Who are the stakeholders?

- data provider (patient) - can ask to change data, but do not own it
- data collector / owner (healthcare provider, pharmacy, insurance company, government)
- data publisher (health informatics exchanges - HIE)
- data scientist
- the medical / scientific community
- the public /society

Building blocks

incentives

harms

responsibilities

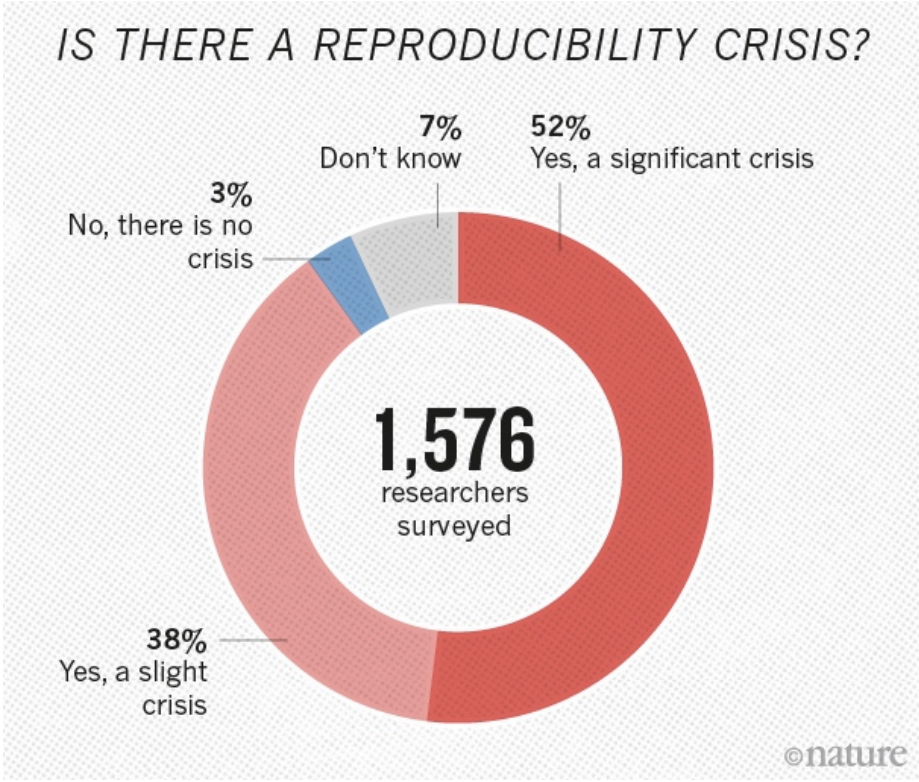
regulation

Harms: privacy, fairness, interpretation

- **privacy** violations - a classic risk of inclusion
- **fair** representation of patient cohorts - a risk of exclusion:
  - the goal is to ensure uniform quality of service / availability of treatment for different groups (ethnic, gender, disability etc)
- **interpretation** of results out of context - a risk to groups, members may not even participate in the study

Harms: junk science

1. low **data quality:** ambiguity, noise
2. **bias**: non-uniform coverage / lack of diversity / over-representation due to data collection, integration, cleaning, analysis
3. insufficient **sample size**
4. multiple hypothesis testing / **p-hacking**
5. blurring the line between **exploratory** vs. **confirmatory** research methodologies



http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Responsibilities

Responsibility for ethical conduct is shared by all stakeholders, specific harms should be mitigated by the **least cost avoider**

- data collector and publisher: due diligence in data cleaning and annotation - ensuring **veracity** and **interpretability** of the data
- data publisher: ensuring **privacy** of data providers (patients), providing information about **bias** (coverage / diversity / representativeness) of the data
- the act of sharing data compels the publisher to consider the **potential harms** that the data may bring
- typically there is **no legal responsibility** for subsequent use on the part of the publisher, but there are still **ethical concerns**

Responsibilities: data scientist

Responsibility for ethical conduct is shared by all stakeholders, specific harms should be mitigated by the **least cost avoider**

- making explicit that research hypothesis is appropriate for data, precisely **stating assumptions** and qualifying **applicability** of results (no “bait and switch”)
- **being skeptical**: it is unethical to trust data that “fell of the back of a truck”
- ensuring **transparency** of the data analysis process (enabling “analysis of the analysis”)
- ensuring **interpretability** of the results, in context!

Incentives

- **Publishing** & academic structures
  - academic structures introduce a lag in data sharing - the need to publish first is an incentive to withhold data
  - data generation / curation are not sufficiently valued, e.g., citing data is cumbersome (ongoing work on Data Citation @ Penn, see CACM 09/2016, https://www.youtube.com/watch?v=vTTgwvblA9s)
  - peer review should emphasize ethical data sharing: curated data + transparent / interpretable methods
- **Collaboration**: sharing data with someone who will recognize the contribution / usefulness / potential / beauty of the data
- **Funding**: research funders should fund ethical data sharing program
- **Training**: ethical data sharing training should be part of standard student research training

Regulation

- **Legal** and **policy** frameworks (IRB, FDA) play an important role, but are reactive by nature and have their limitations
- What is “data sharing malpractice”?
- Who is the police? - The courts, but are there additional steps prior to litigation, e.g., peer review, academic reputation, …
- Intentional vs. unintentional harm - legally there is no difference!

Action plan: towards a code of ethics

Are you, the stakeholder, acting professionally?

- Develop recommendations and guidelines that support effective and ethical sharing of both data and results
- Aspects of fairness, transparency, repeatability, interpretability are shared with (1) other areas of data-intensive science, and (2) the ongoing discourse about data-driven algorithmic decision making

Resources

- EFPIA and PhRMA: Joint Principles for Responsible Clinical Trial Data Sharing to Benefit Patients (http://transparency.efpia.eu/responsible-data-sharing)
- Data Science Association: Code of conduct (http://www.datascienceassn.org/code-of-conduct.html)
- American Statistical Association: Ethical guidelines for statistical practice(http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx)
- Certified Analytics Professional: Code of ethics / conduct (https://www.certifiedanalytics.org/ethics.php)
- ACM code of ethics, under revision (https://www.acm.org/about-acm/code-of-ethics)

Overflow

Discussion points: harms

- Give concrete examples of harms, covering each of the categories.

- How do we inform the stakeholders about the harms unethical data sharing?

Discussion points: methodologies

- What kinds of protocols, methodologies and tools are necessary to mitigate the harms, and support responsible data sharing?
  - annotation / curation
  - bias quantification at different stages
  - interpretation of data, processes and results - links to accountability / transparency / interpretability

Discussion points: incentives

- What are the incentives for ethical data sharing?
- To what extent do we rely on regulation?

Discussion points: technology

- What are some positive and negative examples of tools (w.r.t. usability), specifically in the healthcare domain, that address some aspects of responsible data sharing?

- Is data sharing a technical problem? Which parts of the problem can technology address? Is there a need for basic computer science research here?

Action plan: towards a code of ethics

- Develop a strategy for informing the stakeholders about the harms unethical data sharing
- Develop an education and outreach agenda
- Developing a set of recommendations and guidelines, aimed at the data publishers, data scientists, the medical / scientific community, that support effective and ethical sharing of both data and results