Collaborative Learning on Private Data
Opportunities in Healthcare Research and Practice

Anand D. Sarwate
Rutgers ECE and WINLAB
Data-driven healthcare

Data-driven analyses will revolutionize the way we approach questions of human health

• smarter provisioning of services and resources
• predictive analytics for preventative care
• personalized medicine

And myriad other ways…
Challenges at many stages
Challenges at many stages

Healthcare data poses challenges at several different levels:
Challenges at many stages

Healthcare data poses challenges at several different levels:

• **Data collection**: small sample sizes, bias in data collection, lack of calibration/ground truth…
Challenges at many stages

Healthcare data poses challenges at several different levels:

• **Data collection**: small sample sizes, bias in data collection, lack of calibration/ground truth…

• **Data sharing**: legal and ethical issues around sharing personal health information (PHI), security challenges in cloud systems, standardization…
Challenges at many stages

Healthcare data poses challenges at several different levels:

• **Data collection**: small sample sizes, bias in data collection, lack of calibration/ground truth…

• **Data sharing**: legal and ethical issues around sharing personal health information (PHI), security challenges in cloud systems, standardization…

• **Data analysis**: data is often “big” in the wrong dimension, missing/noisy data, heterogeneity in quality
Data sharing can help

Data sharing in health research is different than open sharing or industry/academia sharing.

• Different regulations around human subjects for experimental data or for PHI in clinical data.

• Informed consent model allows subject-level and study-level privacy preferences.

• Data sharing is contingent and possibly transient.
Data to share depends on the specific application
Data to share depends on the specific application

Not all “healthcare applications” are the same:
Data to share depends on the specific application

Not all “healthcare applications” are the same:

- **Human health research**: research on how the body works, how disease happens, and how to mitigate it. 
  
  Data comes from *subjects* as part of a research study.
Data to share depends on the specific application

Not all “healthcare applications” are the same:

- **Human health research**: research on how the body works, how disease happens, and how to mitigate it.  
  Data comes from *subjects* as part of a research study.

- **Clinical care research**: research on how to improve how we deliver healthcare.  
  Data comes from healthcare centers and EMRs.
Data to share depends on the specific application

Not all “healthcare applications” are the same:

- **Human health research**: research on how the body works, how disease happens, and how to mitigate it.  
  Data comes from *subjects* as part of a research study.

- **Clinical care research**: research on how to improve how we deliver healthcare.  
  Data comes from healthcare centers and EMRs.

- **Population health research**: epidemiology and public health.  
  Data comes from surveys/crowd sourcing, or existing data sets.
Sharing structure varies according to the parties
Sharing structure varies according to the parties

- Institution-to-institution,
Sharing structure varies according the parties

- Institution-to-institution,
- Research group-to-research group
Sharing structure varies according to the parties,

- Institution-to-institution,
- Research group-to-research group,
- Institution to individual.
Sharing structure varies according to the parties

- Institution-to-institution,
- Research group-to-research group
- Institution to individual
- Institution to the public
Sharing structure varies according to the parties:

- Institution-to-institution,
- Research group-to-research group,
- Institution to individual,
- Institution to the public,
- ...
Privacy concerns often hinder data sharing

Example: Different research groups using the same type of measurements want to do a joint analysis.

• Sharing requires lawyers at each institution to generate and execute a Data Use Agreement.

• Resulting months of negotiation makes even small-scale collaboration too complicated.
Why not just anonymize?
Why not just anonymize?
Why not just anonymize?

[Sweeney 2001]
Why not just anonymize?

Public:
- Visit date
- Diagnosis
- Procedure
- Discharge record

$20:
- Name
- Address
- Party
- Last voted
- Date

Voter List:
- Sex
- Birthdate

[Netflixtm]

Movies

User 1

User 2

User 3

[Narayanan and Shmatikov 2008]
Why not just anonymize?

Sharing data can be problematic, even with good intentions.
Disease Association Studies

Correlation (R² values), Alice’s DNA reveals:
If Alice is in the **Cancer** set or **Healthy** set

[Wang et al 2009]
Disease Association Studies

Correlation (R² values), Alice’s DNA reveals:

If Alice is in the Cancer set or Healthy set

Sharing derivatives can be problematic, even with good intentions.

[Wang et al 2009]
Differential privacy (Dwork et al. 2006) is a framework for understanding the privacy risk associated with publishing/sharing functions of private data.

- Provides strong guarantees against the additional risk from sharing information.
- Many (interesting!) challenges in applying differential privacy in real systems.
- Not a silver bullet, but an enabling technology.
The Setting

Private data set $D$ is sanitized by a privacy-preserving sanitizer. The output is a non-public synthetic dataset and a public summary statistic. An ML model can be trained on the public data, and the privacy barrier ensures that the private data remains protected.
Differential Privacy

Participation of a person does not change outcome

Since a person has agency, they can decide to participate in a dataset or not.
Differential privacy means randomizing the computation

- Algorithms guarantee differential privacy by introducing noise or randomness in the computation.
- Privacy risk is quantified and a function of the noise/randomness.
- **Post-processing invariance:** privacy risk is incurred only at the computation — post processing doesn’t increase the risk.
- **Composition:** we can measure how multiple computations affect privacy.
Differential privacy in practice

Google: RAPPOR for tracking statistics in Chrome.

Apple: various iPhone usage statistics.

Census: 2020 US Census will use differential privacy.

mostly focused on count and average statistics
Differential privacy for data sharing
Differential privacy for data sharing

Differential privacy suggests that instead of sharing data directly, we should share access to data.
Differential privacy for data sharing

Differential privacy suggests that instead of sharing data directly, we should share access to data.

• “Private Data-as-a-Service”
Differential privacy for data sharing

Differential privacy suggests that instead of sharing data directly, we should share access to data.

- “Private Data-as-a-Service”

- **Engineering:** design computing systems, statistical analyses, and machine learning algorithms to incorporate privacy
Differential privacy for data sharing

Differential privacy suggests that instead of sharing data directly, we should share access to data.

• “Private Data-as-a-Service”

• Engineering: design computing systems, statistical analyses, and machine learning algorithms to incorporate privacy

• Application/practice: change the way in which researchers interact with data — data hygiene
Specific use cases

- iDASH Clinical Data Warehouse at UC San Diego
- COINSTAC system for collaborative neuroinformatics
Secondary use of clinical data

Hospital DB +

high risk

low risk
Secondary use of clinical data

- Original data was collected during clinical care.
Secondary use of clinical data

- Original data was collected during clinical care.
- Large number of records could contain important patterns of research interest.
Secondary use of clinical data

- Original data was collected during clinical care.
- Large number of records could contain important patterns of research interest.
- If a researcher has a hypothesis, they can request access to records.
Secondary use of clinical data

- Original data was collected during clinical care.
- Large number of records could contain important patterns of research interest.
- If a researcher has a hypothesis, they can request access to records.
- Current process of accessing the data is onerous.
Exploratory data analysis

Can we (safely) let researchers explore hypotheses before they “really” access the data?
Sharing access means new workflows

original data  "anonymized" data

Questions

Answers

original data

Questions

Results+noise

Answers

mechanism
Clinical Data Warehouse at UCSD with millions of patient records

- Allow querying by researchers using differentially private answers.

- Provides similar functionality to i2b2 and STRIDE systems but with quantified privacy guarantees.
Consortia in human health research

• Research consortia are common in many research areas involving human health:

• Foster collaborative research about a particular condition (Alzheimer’s, autism, breast cancer, etc.)

• Automated sharing is challenging, but this is changing. Goal: use privacy protections to encourage consortium growth.
Existing work: ENIGMA

ENIGMA has 30+ working groups on diseases, genomics, population variation, and methods. To do a study:

- Study proposal is approved by ENIGMA managers.
- Analyses performed on local sites and emailed to ENIGMA manager as Excel spreadsheets.

Manager has to perform “manual” meta-analysis.

“The ENIGMA Network brings together researchers in imaging genomics to understand brain structure, function, and disease, based on brain imaging and genetic data.”
The COINS System at MRN

- End-to-end system for managing data for studies on the brain
- 37,903 participants in 42,961 scan sessions from 612 studies for a total of 486,955 clinical assessments.
- Data from 34 states, 38 countries
- Partners with research consortia such as the Autism Brain Imaging Data Exchange (ABIDE)
Data is registered in the system and analyses are performed/aggregated automatically through message passing.

- Study is proposed specifying data needed.
- Local sites approve access to data.
- Analyses are run and aggregated automatically.
Current work on COINSTAC

• Currently implementing standard machine learning and signal processing algorithms for use in COINSTAC: ICA, IVA, SVM, k-means clustering, etc.

• Evaluating how differentially private message passing/computation affects performance.
Recap

These two systems are very different ways of sharing data for healthcare research.

- **Exploratory data analysis**: use privacy guarantees to prevent inadvertent disclosures while enabling useful interactive access.

- **Machine learning on distributed data**: use privacy guarantees to incentivize data holders to join research consortia.
Moving forward
Moving forward

In order for collaborative research/data sharing to “revolutionize healthcare” we need to understand the problem space a bit better:
Moving forward

In order for collaborative research/data sharing to “revolutionize healthcare” we need to understand the problem space a bit better:

• What are “canonical systems” that can be replicated?
Moving forward

In order for collaborative research/data sharing to “revolutionize healthcare” we need to understand the problem space a bit better:

• What are “canonical systems” that can be replicated?

• Where is privacy a bottleneck for data sharing and collaboration?
In order for collaborative research/data sharing to “revolutionize healthcare” we need to understand the problem space a bit better:

- What are “canonical systems” that can be replicated?
- Where is privacy a bottleneck for data sharing and collaboration?
- How can we explain the risks and benefits to patients/subjects/users/clients/customers?