Data and Society
Lecture 5: Data and Health

2/16/18
Guest Lecturer: Dr. Leslie McIntosh
Announcements 2/16

• Op-Ed Drafts returned today after break.

• **Op-ed Finals due March 2.** Please turn in hardcopies of the draft and the final copy on March 2 at 9:00 a.m. If you will be doubling your draft grade, let Fran (bermaf@rpi.edu) know by March 1.

• Wednesday class February 21, starts at 8:30 a.m.

• Guest Speaker Dr. Leslie McIntosh today on data and health
<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday lecture</th>
<th>First Half of Class</th>
<th>Second Half of Class</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 17: NO class</td>
<td>January 19</td>
<td>L1: CLASS INTRO AND LOGISTICS</td>
<td>Presentation Model / Op-Ed Instructions</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 24: NO class</td>
<td>January 26</td>
<td>L2: BIG DATA 1</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>January 31: NO class</td>
<td>February 2</td>
<td>L3: BIG DATA 2 -- IoT</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 7: NO class</td>
<td>February 9</td>
<td>L4: DATA AND SCIENCE</td>
<td>4 Presentations</td>
<td>Op-Ed due Feb. 9</td>
</tr>
<tr>
<td>February 14: 3 Presentations</td>
<td>February 16</td>
<td>L5: DATA AND HEALTH / LESLIE McINTOSH GUEST SPEAKER</td>
<td>5 Presentations</td>
<td>Op-Ed drafts returned Feb. 21</td>
</tr>
<tr>
<td>February 21: 5 Presentations</td>
<td>February 23</td>
<td>L6: DATA STEWARDSHIP AND PRESERVATION</td>
<td>4 Presentations</td>
<td>Research Paper instructions</td>
</tr>
<tr>
<td>February 28: 5 Presentations</td>
<td>March 2</td>
<td>L7: DATA INFRASTRUCTURE</td>
<td>4 Presentations</td>
<td>Op-Ed Final due March 2</td>
</tr>
<tr>
<td>March 7: 5 Presentations</td>
<td>March 9</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 14: Spring Break</td>
<td>March 16</td>
<td>SPRING BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 21: NO class</td>
<td>March 23</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 28: 5 Presentations</td>
<td>March 30</td>
<td>L8: DATA RIGHTS, POLICY, REGULATION</td>
<td>4 Presentations</td>
<td>Research Paper due March 28</td>
</tr>
<tr>
<td>April 4: NO class</td>
<td>April 6</td>
<td>L9: DATA AND ETHICS</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 11: 5 Presentations</td>
<td>April 13</td>
<td>L10: DATA AND COMMUNICATION</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 18: 5 Presentations</td>
<td>April 20</td>
<td>L10: DATA FUTURES</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 25: 5 Presentations</td>
<td>April 27</td>
<td>L11: HOT TOPICS / TBD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dr. Leslie McIntosh

- PhD, MPH
- Research Data Alliance, Executive Director - US
- Ripeta, LLC, CEO
- Formerly Director for the Center for Biomedical Informatics and Professor for the Department of Pathology and Immunology at Washington University, St. Louis School of Medicine
- Biomedical Informatics, reproducibility researcher, global player in health data
A Journey through Biomedical Data using a Reproducibility Lens

Leslie D. McIntosh, PhD, MPH
Research Data Alliance, Executive Director - US
Ripeta, LLC, CEO

@mcintold
Disclosures

SafeT, LLC  Board Member
Asteris, LLC  Advisor
Ripeta, IIC  Founder
Team for Reproducible Research

NLM Supplement
Cynthia Hudson-Vitale
Anthony Juehne
Rosalia Alcoser
Xiaoyan Lui
Brad Evanoff

Research Data Alliance
Cynthia Hudson-Vitale
Anthony Juehne
Snehil Gupta
Connie Zabarovskaya
Brian Romine

RDA Collaborators
Andreas Rauber
Stefan Pröll
Funding Support

Washington University Institute of Clinical and Translational Sciences
NIH CTSA Grant Number UL1TR000448 and UL1TR000448-09S1

MacArthur Foundation 2016 Adoption Seeds program
Foundation through a sub-contract with Research Data Alliance
An Informatics Perspective
BDaaS
Biomedical Data as a Service

Biomedical Data Repository

Data Broker

i2b2 Application

Researchers
Move some of the responsibility of reproducibility
Challenges
Clinical recommendations based upon restricted use data pose challenges for research transparency and accessibility.
Given the protected nature of much biomedical research, what is needed to call our research reproducible?
How do we as biomedical researchers facilitate research reproducibility?
What *is* reproducibility research anyway?
Why do we care?
What is Reproducible Research?
Definitions

**Replicable** - independent people, collecting new data, and using same methods

**Reproducible** - independent people analyzing the same data

Why do we care?
Climategate

2009
Duke’s Precision Medicine Bust

2007 - 2011+
(Lack of) Reproducibility of Psychological Science
2015
Reproducibility of Cancer Biology 2017


https://elifesciences.org/collections/reproducibility-project-cancer-biology
<table>
<thead>
<tr>
<th>Paper</th>
<th>Conclusion</th>
<th>Focus of key experiment</th>
<th>Replication results</th>
<th>Citations</th>
</tr>
</thead>
</table>
Relative risk of Alzheimer's between men and women: Record corrected

https://www.sciencedaily.com/releases/2017/08/170828124531.htm

https://jamanetwork.com/journals/jamaneurology/article-abstract/2649260?resultClick=1&redirect=true
Privacy and Ethics
Personal Privacy
Population Vulnerability
“U.S. soldiers are revealing sensitive and dangerous information by jogging”
Back to Reproducibility
The purpose of this study is to:

- Assess the current research reproducibility practices of investigators
- Who use electronic health records (EHR) for secondary analysis, then
- Develop a framework
1. Do you have a well-formed, well-defined question?
2. Do you have data to answer your question?
3. If you get an answer, is the answer relevant or meaningful?
<table>
<thead>
<tr>
<th>Pertinence</th>
<th>Can the data answer the research hypothesis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>Is all the information available to reuse the data for a study?</td>
</tr>
<tr>
<td>Replicability/Reproducibility</td>
<td>Are the data of sound quality to <em>want</em> to reuse them?</td>
</tr>
</tbody>
</table>
How a business begins...
Scientific Method
Work-to-Date
Publication
Repeat: A Framework to Assess Empirical Reproducibility in Biomedical Research
BioMed Central (https://doi.org/10.1186/s12874-017-0377-6)

Repos:
https://github.com/ripeta
https://osf.io/ppnwa/

International Engagement
Research Data Alliance
100+ unique variables

5 categories
- bibliographic
- database & data collection
- data mining & cleaning
- data analysis
- data sharing & documentation
## Publication Overview and Bibliographic Information (21 items)

<table>
<thead>
<tr>
<th>Is the research hypothesis-driven or hypothesis-generating?</th>
<th>Hypothesis Driven</th>
<th>Hypothesis Generating</th>
<th>Unclear</th>
</tr>
</thead>
</table>

## Database and Data Collection (63 items)

<table>
<thead>
<tr>
<th>Publication states database(s) source(s) of data?</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Publication states database(s) source(s) of data in the following location:</td>
<td></td>
</tr>
<tr>
<td>Query methodology</td>
<td></td>
</tr>
<tr>
<td>*Does the shared query script for database contain comments and/or notations for ease of reproducibility?</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

## Methods: Data Mining and Cleaning (19 items)

<table>
<thead>
<tr>
<th>Does the research involve natural language processing or text mining?</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Is the text mining software application proprietary or open?</td>
<td></td>
</tr>
</tbody>
</table>

*If multiple applications were used, please select all options that apply.*

1. Proprietary
2. Mixed
3. Open
<table>
<thead>
<tr>
<th>Methods: Data Analysis (15 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the author state analysis methodology and process?</td>
</tr>
<tr>
<td>Does the author indicate the software used to develop the analysis code?</td>
</tr>
<tr>
<td>*Is the analysis software proprietary or open?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Sharing and Data Documentation (36 items)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the finalized dataset shared?</td>
</tr>
<tr>
<td>*Where is the finalized dataset shared?</td>
</tr>
<tr>
<td>Is there a clear process for requesting the data?</td>
</tr>
</tbody>
</table>

Ricochet Framework has been tested for inter-rater reliability and face validity.
Here we test: whether the published associations between steady-state warfarin dose and variants in warfarin pharmacogenes could be replicated in BioVU; how implementing published pharmacogenomic algorithms affects dosing error; and if an improved algorithm for African–Americans can be generated using variants associated with stable dose in this population.
Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

Cases were identified in BioVU, the Vanderbilt DNA Biobank, which accrues DNA samples extracted from blood remaining from routine clinical testing after it has been retained for 3 days and is scheduled to be discarded [27].

Data Source Stated = Yes

Data Source Cited = Yes
Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

The programming language was used for regression analyses, diagnostic-test calculations, and to implement and evaluate the algorithms (R Foundation for Statistical Computing, Vienna, Austria).

Software Stated = Yes
Software Cited = Yes
Software Version = No
Expanded Genetic algorithm.

Expanded Genetic:

\[
5.9487517 - 0.0073436353 \times \text{race (AA=0, EA=1)}
- 0.025161445 \times \text{age (in years)}
+ 0.058138499 \times \text{sex (F=0, M=1)}
+ 1.1848957 \times \text{bsa (kg/m}^2\text{)}
+ 0.068020571 \times \text{smoking status (nonsmoker=0, smoker=1)}
+ 0.058578086 \times \text{VTE indication (no=0, yes=1)}
- 0.10646416 \times \text{atrial fibrillation indication (no=0, yes=1)}
- 0.8142521 \times \text{amiodarone use (no=0, yes=1)}
- 0.64877338 \times \text{CYP2C9*2 (wt=0, heterozygote=1, homozygote=2)}
- 1.0601067 \times \text{CYP2C9*3 (wt=0, heterozygote=1, homozygote=2)}
- 1.9737831 \times \text{CYP2C9*6 (wt=0, heterozygote=1, homozygote=2)}
- 1.0622944 \times \text{CYP2C9*8 (wt=0, heterozygote=1, homozygote=2)}
+ 0.24749973 \times \text{CYP4F2 (wt=0, heterozygote=1, homozygote=2)}
- 0.31996754 \times \text{CALU (wt=0, heterozygote=1, homozygote=2)}
- 0.87262446 \times \text{VKORC1 (wt=0, heterozygote=1, homozygote=2)}
= \log[\text{weekly warfarin dose}]
\]
<table>
<thead>
<tr>
<th>id</th>
<th>article doi</th>
<th>journal publication</th>
<th>publication date</th>
<th>state data sources</th>
<th>state database location</th>
<th>database cited</th>
<th>analysis software</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>10.1186/s13054-015-1180-6</td>
<td>Critical care (London,</td>
<td>1/15/16</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>17</td>
<td>10.1016/j.urology.2012.11.002</td>
<td>Urology</td>
<td>1/30/13</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>24</td>
<td>10.1097/AJPS.0b013e3181f06b06</td>
<td>Clinical Journal of Pai</td>
<td>1/27/11</td>
<td>1</td>
<td>0</td>
<td>0 STATA</td>
<td>SAS</td>
</tr>
<tr>
<td>25</td>
<td>10.1016/j.jmig.2011.01.009</td>
<td>Journal of Minimally I</td>
<td>3/16/11</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>26</td>
<td>10.1097/OP.0b013e31828e92b0</td>
<td>Ophthalmic Plastic &amp;</td>
<td>1/6/13</td>
<td>0</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>27</td>
<td>10.1097/PSY.0b013e31821fb9f9a</td>
<td>Psychosomatic Medic</td>
<td>6/23/11</td>
<td>0</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>33</td>
<td>10.1016/j.clingeo.2014.02.016</td>
<td>Clinical Neurology an</td>
<td>2/25/14</td>
<td>0</td>
<td>0</td>
<td>0 SPSS</td>
<td>SPSS</td>
</tr>
<tr>
<td>35</td>
<td>10.1016/j.jpainsymman.2015.02.022</td>
<td>Journal of Pain and S</td>
<td>4/1/15</td>
<td>0</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>37</td>
<td>10.1148/radiol.14132418</td>
<td>Radiology</td>
<td>9/9/14</td>
<td>0</td>
<td>0</td>
<td>0 R</td>
<td>SPSS</td>
</tr>
<tr>
<td>39</td>
<td>10.1111/apt.13505</td>
<td>Alimentary pharmacc</td>
<td>2/4/16</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>44</td>
<td>10.1016/j.burns.2013.12.002</td>
<td>Burns: Journal of the</td>
<td>1/4/14</td>
<td>1</td>
<td>0</td>
<td>0 IBM SPSS</td>
<td>SAS</td>
</tr>
<tr>
<td>45</td>
<td>10.1016/j.clingeo.2015.10.004</td>
<td>Clinical Neurology an</td>
<td>10/8/15</td>
<td>1</td>
<td>0</td>
<td>0 SPSS</td>
<td>SPSS</td>
</tr>
<tr>
<td>46</td>
<td>10.1016/j.gyneco.2013.04.055</td>
<td>Gynecologic Oncolog</td>
<td>4/28/13</td>
<td>1</td>
<td>0</td>
<td>0 GraphPad</td>
<td>SPSS</td>
</tr>
<tr>
<td>47</td>
<td>10.1016/j.jamc.2014.06.006</td>
<td>Journal of American AI</td>
<td>9/27/14</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>49</td>
<td>10.1016/j.ijstrokecerebrovasdis.2015.06.043</td>
<td>Journal of Stroke and</td>
<td>6/27/15</td>
<td>1</td>
<td>0</td>
<td>0 SPSS</td>
<td>SPSS</td>
</tr>
<tr>
<td>50</td>
<td>10.1186/1472-6963-13-414</td>
<td>BMC health services r</td>
<td>10/14/13</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SPSS</td>
</tr>
<tr>
<td>54</td>
<td>10.1016/j.joms.2016.01.001</td>
<td>Journal of Oral and M</td>
<td>1/7/16</td>
<td>0</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>55</td>
<td>PMC4384267</td>
<td>Pharmacy practice</td>
<td>3/15/15</td>
<td>1</td>
<td>1</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>56</td>
<td>10.1186/1471-2318-14-75</td>
<td>BMC Geriatrics</td>
<td>6/17/14</td>
<td>1</td>
<td>0</td>
<td>0 STATA</td>
<td>SAS</td>
</tr>
<tr>
<td>58</td>
<td>10.1016/j.clinthera.2011.08.005</td>
<td>Clinical Therapeutics</td>
<td>9/3/11</td>
<td>1</td>
<td>0</td>
<td>0 SPSS</td>
<td>SAS</td>
</tr>
<tr>
<td>60</td>
<td>10.3399/jgp.15X866113</td>
<td>British Journal of Gen</td>
<td>8/1/15</td>
<td>1</td>
<td>0</td>
<td>0 STATA</td>
<td>SAS</td>
</tr>
<tr>
<td>63</td>
<td>10.1158/1055-9965.EPI-14-0487</td>
<td>Cancer Epidemiology,</td>
<td>7/30/14</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
<tr>
<td>66</td>
<td>10.1542/hpedc.2014-0085</td>
<td>Hospital Pediatrics</td>
<td>8/1/15</td>
<td>1</td>
<td>0</td>
<td>0 SAS</td>
<td>SAS</td>
</tr>
</tbody>
</table>
Manually curating manuscripts
Using NLP and neural network algorithms to mine data
Reproducibility in an Evolving EHR Data Repository
Next Steps

Refine RepeAT using mined literature

Incorporate more variables into tool

Incorporate data citability into WU EHR repository
Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Moving Biomedical Big Data Sharing Forward

1. Integrate the RDA recommendations for Data Citation of Evolving Data local EHR repository
2. Contribute all source code back to the i2b2 GitHub
3. Gather feedback about RDA WGDC-compliant i2b2 code from established i2b2 installations

MacArthur Foundation 2016 Adoption Seeds program Foundation through a sub-contract with Research Data Alliance
Implementation
R1 and R2 Implementation

PostgreSQL Extension “temporal_tables”

1

RDC.table

2

triggers

RDC.hist_table*

3

*stores history of data changes
ETL Incrementals

<table>
<thead>
<tr>
<th>RDC.hist_table</th>
<th>RDC.table</th>
<th>RDC.table</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sys_period</td>
<td>sys_period</td>
<td>sys_period</td>
</tr>
<tr>
<td>2016-9-8 00:00, 2016-9-9 00:00</td>
<td>2016-9-9 00:00, NULL</td>
<td>2016-9-9 00:00, NULL</td>
</tr>
<tr>
<td>Old Data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source Data

Update?

Insert
R3, R7, R8, R9, and R10 Implementation

1. PostgreSQL Extension “temporal_tables”

2. 
   - RDC.table
   - RDC.hist_table

3. 
   - functions
   - triggers
   - query audit tables
Data Reproducibility Workflow

TODAY

Data Broker

Query from View

Log Audit Data

Citation: Query Date
Query PID

SOME TIME LATER

Data Broker

Request Query Re-Run

Query PID

* tardis: http://p0c12.deviantart.com/art/TARDIS-Simple-Vector-612645056
Bonus Feature: Determine if Change Occurred
Other thoughts...
Other Information
Resources

Github - Research Reproducibility Assessment Tool
https://github.com/CBMIWU/Research_Reproducibility

Bibliography for NLM funded study
https://www.zotero.org/groups/biomedical_informatics_resrepro
Attribution

Bakery
https://www.flickr.com/photos/puthoor_photo/9615954804/

London Underground
https://www.flickr.com/photos/jvk/151830527/
www.flickr.com/photos/wttw/14763924/
http://bit.ly/1WL4k1U

UEA Climate Research Unit
By ChrisO - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=8730683

Reproducibility Duke University PMI Bust
Reproducibility of Psychological Findings
Doi: 10.1126/science.aac4716
Bibliography on Reproducible Research

14. Research Data Alliance, Data Citation of Evolving Data https://rd-alliance.org/sites/default/files/RDA-DC-Recommendations_150508_1.pdf
Questions
Break
Op-Ed Draft Grades

Final Op-Ed grade is Draft Grade + Final Grade or 2 X Draft Grade

DRAFT GRADES (out of 15)
Calculating your Grade So Far (GSF)

**GSF** (a %) =

Presentation grade / 15 [if you’ve given a presentation]  
+ Draft grade / 15  
+ Draft grade / 15 [if you’ll be doubling your Draft Grade]  
+ Participation Grade / 15 [Assume 15 if you will miss no more than 2 classes and fully participate in class]

*Example:* You got a 13/15 on your presentation and are doubling your draft grade of 12/15 and will participate fully in class. In that case, your **GSF** is

$$13/15 + 12/15 + 12/15 + 15/15 = 52/60 = 87\%$$
Discussion Article for February 23

Presentations
Presentation Articles for February 21


**Presentation Articles for February 23**


Presentation Articles for February 28


• “Librarians Saving the Internet”, Science Friday, https://apps.sciencefriday.com/data/librarians.html


Presentation Articles for March 2


Fran Berman, Data and Society, CSCI 4370/6370
Presentation Articles for Today


• “Healthcare is hemorrhaging data. AI is here to help.”, Wired, https://www.wired.com/story/health-care-is-hemorrhaging-data-ai-is-here-to-help/ [Matthew M]