Data integration to counter epidemic threats

Faculty meeting Michelle Dunn, 12 September 2016

Wilbert van Panhuis, MD PhD
wav10@pitt.edu
Content

- The ongoing challenge of infectious disease epidemics

- Improving the use of information to counter epidemics: Project Tycho

- BD2K K01 research project: Data integration for global population health

- Work towards a Pitt Data Commons
How to handle the infectious diseases threat

Daniel M. Gerstein
© Updated 1245 GMT (2045 HKT) June 3, 2016

INFECTIONOUS DISEASES

As Ebola fades, a new threat
With health services devastated in the wake of Ebola, experts are bracing for a deadly measles outbreak in West Africa

The NEW ENGLAND JOURNAL of MEDICINE

Zika Virus in the Americas — Yet Another Arbovirus Threat
Anthony S. Fauci, M.D., and David M. Morens, M.D.

INFECTIONOUS DISEASES

An obscure mosquito-borne disease goes global
After racing through Oceania last year, the Zika virus is now spreading in the Americas
“An epidemic is one of the few catastrophes that could set the world back drastically in the next few decades. By building a global warning and response system, we can prepare for it and prevent millions of deaths.”

Bill Gates
Converting Big Data into public health

WE CONCUR WITH M. J. Khoury and J. P. A. Ioannidis (“Big data meets public health,”)

Big data meets public health
Human well-being could benefit from large-scale data if large-scale noise is minimized

Economics
Predicting poverty and wealth from mobile phone metadata

How big data could help stop the Ebola outbreak
Content

- The ongoing challenge of infectious disease epidemics

- Improving the use of information to counter epidemics: Project Tycho

- K01 research project: Data integration for global population health

- Work towards a Pitt Data Commons
Project Tycho: epidemic data more accessible

Unusable data

Easy-to-use data

Research
- Digitization
- Extraction
- Standardization
- Integration
- Annotation

US data: 1888-present
WHO data: 1952-2010

Wall street journal  Forbes magazine

www.tycho.pitt.edu
Data for 8 SE Asia countries: temperature -> dengue

Thailand  Laos  Cambodia  Vietnam

NIH DIRECTOR'S BLOG

Climate and Viral Illness: El Niño Event Linked to Dengue Epidemics
Posted on October 13, 2015 by Dr. Francis Collins

Van Panhuis et. al., PNAS 2015
Open access release of all data: www.tycho.pitt.edu

Active and growing user community

33,000 unique web visitors
200,000 page views
100,000 data queries
5000 datasets downloaded
“A University of Pittsburgh initiative called Project Tycho, for example, **unlocks CDC data** on contagious diseases which goes back all the way to 1888. Among other things, they've identified more than 100 million cases of contagious illness that were prevented by immunizations.” (Secretary Sebelius)

Reimagining Health Care Delivery: Remarks to the Aspen Institute Care Innovation Summit (Washington DC, Feb 27 2014)
Content

- The ongoing challenge of infectious disease epidemics

- Improving the use of information to counter epidemics: Project Tycho

- K01 research project: Data integration for global population health

- Work towards a Pitt Data Commons
K01 Specific aims

1. Standardize a wide range of datasets for the vector-borne diseases dengue and Chikungunya
   - + Cholera and HIV (supplement)

2. Develop logic-based algorithms that can automatically combine data required for a specific disease simulation

3. Develop a value-of-information-based method that can rank missing datasets by their usefulness for specific simulations

4. Training to become an independent investigator at the interface between population health and data science
Apollo project at University of Pittsburgh

- Methods and data design for standardization of epidemic data developed by the Apollo project

  - Directed by Dr. Mike Wagner in Pitt Department of Biomedical Informatics (K01 mentor)
  - Funded by NIGMS since 2012 (5R01GM101151)
  - Designed a standard data format for epidemic data, defined by an XML Schema Document (XSD)
  - Also standardized epidemic simulators and hosts data library and simulation web services: [https://github.com/ApolloDevel](https://github.com/ApolloDevel)
Apollo complex data type “Epidemic”

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>causalPathogens</td>
<td>[1..*] ApolloPathogenCode</td>
</tr>
<tr>
<td>epidemicPeriod</td>
<td>[1..1] EpidemicPeriod</td>
</tr>
<tr>
<td>administrativeLocations</td>
<td>[1..*] ApolloLocationCode</td>
</tr>
<tr>
<td>epidemicZones</td>
<td>[0..*] ApolloLocationCode</td>
</tr>
<tr>
<td>infections</td>
<td>[1..*] Infection</td>
</tr>
<tr>
<td>preEpidemicCensus</td>
<td>[0..1] PreEpidemicEcosystemCensus</td>
</tr>
<tr>
<td>populationSerologySurveys</td>
<td>[0..*] PopulationSerologySurvey</td>
</tr>
<tr>
<td>populationInfectionSurveys</td>
<td>[0..*] PopulationInfectionSurvey</td>
</tr>
<tr>
<td>infectiousDiseaseControlStrategies</td>
<td>[0..*] InfectiousDiseaseControlMeasure</td>
</tr>
<tr>
<td>caseDefinitions</td>
<td>[0..*] CaseDefinition</td>
</tr>
<tr>
<td>contactDefinitions</td>
<td>[0..*] ContactDefinition</td>
</tr>
<tr>
<td>epidemicCaseCounts</td>
<td>[0..1] EpidemicCaseCounts</td>
</tr>
<tr>
<td>caseLists</td>
<td>[0..*] CaseList</td>
</tr>
<tr>
<td>transmissionTrees</td>
<td>[0..*] TransmissionTree</td>
</tr>
<tr>
<td>relativeRiskDataSets</td>
<td>[0..*] RelativeRiskDataSet</td>
</tr>
<tr>
<td>causalPathogenIsolates</td>
<td>[0..*] token</td>
</tr>
<tr>
<td>references</td>
<td>[1..*] Reference</td>
</tr>
<tr>
<td>curator</td>
<td>[0..1] string</td>
</tr>
<tr>
<td>editHistory</td>
<td>[0..*] string</td>
</tr>
<tr>
<td>reviewedBy</td>
<td>[0..*] string</td>
</tr>
<tr>
<td>acknowledgements</td>
<td>[0..*] string</td>
</tr>
</tbody>
</table>
Re-represented information

Chikungunya Epidemics

Summary of Information in the Chikungunya Collection

1. 1952, Tanzania [JSON] [XML]
2. 2005, Reunion [JSON] [XML]
3. 2005, Mayotte [JSON] [XML]
4. 2007, Italy [JSON] [XML]
5. 2012, Cambodia [JSON] [XML]
Vision: Data-simulation ecosystem

“List all possible meals that I can cook based on the ingredients and recipes available”

“If I want to cook meal A, what ingredients do I need to buy?”

“List all possible epidemics that I can simulate based on the data and simulators available”

“If I want to simulate scenario A, what data do I need to collect?”
Population health data ecosystem @ Pitt

University of Pittsburgh

Graduate School of Public Health

School of Medicine Biomedical Informatics

Global scale population health information system

Epidemic simulations

Epidemic forecasting

Economic and decision modeling

Data mining and machine-learning algorithms

Integrate population, individual and molecular information

Commercial applications

Population health data commons

Data applications
Content

- The ongoing challenge of infectious disease epidemics

- Improving the use of information to counter epidemics: Project Tycho

- K01 research project: Data integration for global population health

- Work towards a Pitt Data Commons
Recommendations to the Provost on:

- Governance and Policy
- RDM Support Staff
- Data Management Plans (DMPs)
- Metadata Services
- Active Data Storage
- Data Analysis, Workflow & Visualization
- Data Sharing and Reuse
- Data Repository and Long-Term Preservation
- Computing and Information

Vision:
“The DMC envisions an environment in which research faculty, staff and students across all disciplines will optimally share their data and adopt open scholarship practices, behaviors and cultural norms (“data commons”), and which are facilitated by robust institutional data infrastructure and coordinated supporting services.”
Pittsburgh Genome Research Repository (PGRR)

DBMI Lead - Rebecca Jacobson, MD MSIS, CIO IPM
Uma Chandran PhD, Mike Barmada PhD & Adrian Lee PhD

- DBMI, Health Sciences Tissue Bank (HSTB), Institute for Personalized Medicine (IPM), UPMC Enterprise Analytics and UPCI driven effort
- http://www.pgrr.pitt.edu
- NCI TCGA data with DUA for over 50 Pitt investigators (and growing)
- Creating a computing environment for storage and analysis via Pitt Center for Simulation & Modeling (SaM), Pittsburgh Supercomputing Center (PSC) and UPMC
- Center piece for our BD2K Center for Causal Discovery ($11M) and PA State CURE grant (funded $5M) with CMU
- New 2.1 PByte SaaS linked to HPC @ PSC in partnership with CMU and IPM

Result = Computable Genotypes
Linking EHR Data with PCORNet PaTH, NCATS ACT & PMI HPO

Shyam Visweswaran, MD PhD and Mike Becich, MD PhD -

Result = a clinical research data warehouse @ Pitt/UPMC

PaTH – Geisinger, Hopkins, Pitt, Penn State, UPMC, Utah & Temple (11M)
ACT – 20+ CTSA sites–Data Harmonization led by Pitt (37M)
Fueled forward by Precision Medicine Initiative (150K patients to be enrolled)

Result = Computable Phenotypes & a Learning Health System

Pittsburgh Health Data Alliance (PHDA)
http://www.healthdataalliance.com

Three Pittsburgh institutions.
One goal.

The future of health care is in the data.

The Pittsburgh Health Data Alliance represents an unprecedented collaboration between Carnegie Mellon University, The University of Pittsburgh and UPMC.

By working together to turn data into improved human health, we will change the practice of medicine.

• Collaboration of CMU, Pitt, and UPMC Enterprises

• Two Centers created:
  • Center for Machine Learning in Healthcare
    • Led by Joe Marks in CMU School of Computer Science
  • Center for Commercial Applications (CCA)
    • Led by Mike Becich and Don Taylor
• $2M/yr in Early Stage
• $22M in follow on funding for successful projects
• Launch in July 2015
Vision of a Pitt Data Commons

Data Resources

- Project Tycho
- Apollo
- Epidemiology Data Center
- Investigator data
- PaTH Network
- ACT Project
- PMI
- Mesothelioma data bank
- PGRR

Data Commons Infrastructure

- Catalogue
- Search
- Etc.
- Virtual workspace for analytics and collaboration

Results

- Access to more data for research and training
- Easier collaboration
- Broader impact and recognition
- Reaching new disciplines
- Commercial applications
Acknowledgements

Faculty, Staff, and Students
Jeremy Avigad, Carnegie Mellon University • Bill Hogan, University of Florida • John Levander, University of Pittsburgh • Max Sibilla, University of Pittsburgh • Nick Millet, University of Pittsburgh • Anne Cross, University of Pittsburgh • Sharon Crow, University of Pittsburgh • Dan Bain, University of Pittsburgh • Marc Choisy, Institute de la Recherche pour le Développement • Derek Cummings, University of Florida • Ernesto Marques, University of Pittsburgh

Funders and Partners
National Institutes of Health BD2K (1K01ES026836), and Models of Infectious Disease Agent Programs (U54GM088491) • Benter Foundation • Bill & Melinda Gates Foundation (#49276 and #OPP1091931) • Brazil Ministry of Health • Cambodia Ministry of Health • Council of State and Territorial Epidemiologists • Digital Divide Data • Johns Hopkins University Bloomberg School of Public Health • Laos Ministry of Health • Pan American Health Organization • Pittsburgh Supercomputing Center • Taiwan Ministry of Health • Thailand Ministry of Health • University of Pittsburgh Department of History • University of Pittsburgh School of Information Sciences • U.S. Department of Health & Human Services • U.S. Open Government Initiative • Vietnam Ministry of Health • World Health Organization