Predictive Analytics
Perspective for Health TechNet
Joseph Bormel, MD, MPH
July 20, 2018
Predictive Analytics Underlies All Societal Trends

Societal

INTERNET TRENDS 2018

Mary Meeker
May 30 @ Code 2018

1) Users
2) Usage
3) Innovation + Competition + Scrutiny
4) E-Commerce
5) Advertising
6) Consumer Spending
7) Work
8) Data Gathering + Optimization
9) Economic Growth Drivers
10) China (Provided by Hillhouse Capital)
11) Enterprise Software
12) USA Inc. + Immigration

https://youtu.be/HdjcdZqODoE
Transportation as % of Household Spending = 14% vs. 14% (1972)... #3 Segment of $ Spending Behind Shelter + Taxes

Relative Household Spending

![Graph showing relative household spending categories and their expenditure in 1972, 1990, and 2017.]

Total Expenditure:
- 1972: $11K
- 1990: $31K
- 2017: $68K

Source: USA Bureau of Labor Statistics (BLS), Consumer Expenditure Survey. *Pensions + Insurance includes deductions for private retirement accounts, social security, and life insurance. **Other includes education and miscellaneous other expenses. Note: Results based on Surveys of American Urban & Rural Households (Families & Single Consumers). 1972 data reflects non-annual survey conducted by BLS + Census Bureau to adjust CPI. 1990 and 2017 Data Based on Annual Survey performed by BLS + Census Bureau. Healthcare costs include insurance, drugs, out-of-pocket medical expenses, etc. 2017 = mid-year figures.

The Reagan-Udall Foundation has formed a public-private partnership to provide access for private-sector entities, such as regulated industry, academic institutions, and non-profit organizations, to a system based on the Food and Drug Administration’s (FDA) Sentinel Initiative. This collaboration works with selected Sentinel data partners and the Harvard Pilgrim Healthcare Institute, functioning as the Analytic or Coordinating Center, to facilitate the analyses of medical product safety evaluations.

http://reaganudall.org/innovation-medical-evidence-development-and-surveillance
Analytics in Service of Better Health

Product/Service Market vs Right

Transparency (Cost, Quality, Access)

Science / Learning system

Goals and Incentives

Social Determinants (SDoH)

http://www.who.int/social_determinants/sdh_definition/en/
Successful Initiatives

Get The Data → Find Actionable Opportunities → Take Action
Successful Initiatives

May not exist
May not be accessible
Cost to capture (time, ease)
Competing tasks
to capturing the patient’s story
How do providers
choose what to record?
Satisfice (“good enough”)

See Data Limitations slide

The five rights include: the right information, to the right person, in the right intervention format, through the right channel, at the right time in workflow.

See “Taking Action” slide

THE EIGHT PILLARS

Pillar 1: Graphical models for prediction and diagnosis
Pillar 2: The control of confounding
Pillar 3: Do-calculus – An all-seeing oracle for predicting the effects of policies and interventions
Pillar 4: The algorithmization of counterfactuals
Pillar 5: Mediation analysis and the assessment of direct and indirect effects
Pillar 6: External validity and sample selection bias
Pillar 7: Missing data
Pillar 8: Causal discovery

See Judea Pearl paper:

Causal discovery

Satis-fice
/sadēs, fīs/ noun
1. accept an available option as satisfactory
2. "It talks about telling you not to just satisfy but to always look for the best"
Taking Action
requires workflow integration and designed orchestration dialogues

Figure 4. Example of a CIS-consuming SOA for CDS.

From AMIA2012 - Clinical Information System Services and Capabilities Desired for Scalable, Standards-Based, Service-oriented Decision Support: Consensus Assessment of the Health Level 7 Clinical Decision Support Work Group

Kensaku Kawamoto, MD, PhD,1 Jason Jacobs,1 Brandon M. Welch, MS,1
Vojtech Huser, MD, PhD,2 Marilyn D. Paterno, MBI,3 Guilherme Del Fiol, MD, PhD,1 David Shields,1 Howard R. Strasberg, MD, MS,4 Peter J. Haug, MD,5 Zhijing Liu, PhD,6 Robert A. Jenders, MD, MS,7 David W. Rowed, MBBS,8 Daryl Chertcoff,9
Karsten Fehre, DI,10,11 Klaus-Peter Adlassnig, PhD, MSc,10,11 A. Clayton Curtis, MD, PhD12
Common to make design mistakes

Cannot infer Prevalence from Visits

Need rates by SDoH

Reliability of fatal overdose data >> non-fatal

WNL = “We Never Looked”
Authorized Uses

Data use, privacy, confidentiality, trust vary widely based on model
## Authorities

<table>
<thead>
<tr>
<th></th>
<th>Public Health</th>
<th>Research</th>
<th>Survey</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Consent and</strong></td>
<td>No</td>
<td>Yes</td>
<td>Implied (voluntary)</td>
<td>No</td>
</tr>
<tr>
<td><strong>IRB</strong></td>
<td></td>
<td></td>
<td></td>
<td>Inaccessible Opt-Out Consent</td>
</tr>
<tr>
<td><strong>Collection informed</strong></td>
<td>No (exploratory)</td>
<td>Yes (confirmatory)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Hypothesis</strong></td>
<td></td>
<td></td>
<td></td>
<td>Late-binding with massive aggregation / inference*</td>
</tr>
<tr>
<td><strong>Duty</strong></td>
<td>dissemination including regulatory; PHA, CIPSEA</td>
<td>Publication</td>
<td>Publication</td>
<td>Proprietary only</td>
</tr>
<tr>
<td><strong>Cost Effectiveness</strong></td>
<td>inexpensive (massive resuse)</td>
<td>expensive (often one-off)</td>
<td>depends on design</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Scale</strong></td>
<td>exhaustive depending on data partners</td>
<td>inclusion criteria and availability/recruitment</td>
<td>respondents rate</td>
<td>all consumers on the grid</td>
</tr>
</tbody>
</table>
Dark Matter
Issues
DATA

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

APRIL 02, 2018

First, clarify your objectives and assess whether you have the right data to support these objectives.

Second, build plenty of time to execute data quality fundamentals into your overall project plan.

Third, maintain an audit trail as you prepare the training data.

Fourth, charge a specific individual (or team) with responsibility for data quality as you turn your model loose.
Dark Matter Issues

- Predictive analytics (including AI) often presupposes that in the available data is signal that can portent the future.

- Clinical and administrative (claims) data is always incomplete. Data is collection is not free. People are pragmatic. Recording observations is often unsafe.

- Adequate analytics would need to surface the darkness. For example, a gap analysis between ideal model data and what is present in our data would be needed to assess what is known and unknown with any confidence.

- For example, in patients suspected of possible opioid use disorder, what is their relevant history (e.g. use) and social determinate of health (e.g. homelessness, transportation, employment, diet, loneliness, substance use in the residence, incarceration in extended setting).

- Some essential data will almost always be dark, such as intention behind acts. (e.g. Why was that drug given or taken?)

“Roughly 80 percent of the mass of the universe is made up of material that scientists cannot directly observe. Known as dark matter, this bizarre ingredient does not emit light or energy.”

https://www.space.com/20930-dark-matter.html

“Dark” is synonymous with “invisible”.

Analyzing free text can reveal references to problems such as diseases, medications, allergies, anatomy, medical procedures, and history.

It is tempting to believe that this information is necessary and sufficient to characterize the patient for diagnostic and therapeutic purposes.

In practice, natural language processing is used to make sure that we are not losing available information. That’s very important.

It is rare that an adequacy analysis is performed. When they are performed, embarrassing issues of data integrity are surfaced which can threaten the initiative.
Data Capture Issues

- What gets **said** by patient?
- What gets **documented** by provider?
  - What was heard?
  - How hard is it to document. (Does that change with SMART on FHIR, or other)
  - How safe is it to document (e.g. liability, pre-existing conditions, employment, legal citizenship, insurance coverage associated with suicide, etc)
- **How** is it said?
  - e.g. “Suicide Attempt” versus “Intentional Overdose”
- Other important factors
  - **Reputation** Index of the information source
  - How **aligned** is information with models needed for purpose (e.g. Opioid Crisis)
  - **Jurisdictional** Issues (Federal, State, County requirements, definitions, even sales tax)
  - Meets needs of those documenting in terms of bidirectional, community (**internal use**)
Design Limitations
using the Drug Abuse Warning Network as a typical example

- There was little public documentation of DAWN’s data quality, and a few published studies raised concerns that measurement error in DAWN was substantial and systemic.
- Because DAWN monitors episodes, not individuals, data are misinterpreted if analysts view them as prevalence measures.
- There is an uncertain, and perhaps inconsistent, relationship between the number of DAWN episodes and the true level of drug abuse problems in a given area.
- Delays in reporting DAWN data undermined its potential utility as an early warning mechanism for emerging drug problems.
- DAWN data were often assumed to represent heavy or chronic drug users, although the validity of that assumption has never been assessed. (What was sampled?)

[In sum, there were problems with timeliness, quality, representation and sharing of data.]

From: 2002-08 Drug Abuse Warning Network-- Development of a New Design - orig 10.1.1.444.3232 .pdf, pages 11,12
Re-Designs are Common

Enhancing Surveillance Systems

The Evolution of BioSense: Lessons Learned and Future Directions

Deborah W. Gould, PhD¹, David Walker, MPH¹, and Paula W. Yoon, ScD, MPH¹

Abstract

The BioSense program was launched in 2003 with the aim of establishing a nationwide integrated public health surveillance system for early detection and assessment of potential bioterrorism-related illness. The program has matured over the years from an initial Centers for Disease Control and Prevention-centric program to one focused on building syndromic surveillance capacity at the state and local level. The uses of syndromic surveillance have also evolved from an early focus on alerts for bioterrorism-related illness to situational awareness and response, to various hazardous events and disease outbreaks. Future development of BioSense (now the National Syndromic Surveillance Program) includes, in the short term, a focus on data quality with an emphasis on stability, consistency, and reliability and, in the long term, increased capacity and innovation, new data sources and system functionality, and exploration of emerging technologies and analytics.

Keywords

BioSense, syndromic surveillance, emergency department data, public health surveillance

be reporting these data to BioSense. To speed up data acquisition, CDC established servers in large metropolitan hospitals, and data were sent directly to BioSense. By 2007, only about 10% of US EDs were reporting, and BioSense stakeholders became concerned with the slow rate of adding hospitals and increasing the coverage of EDs across the country. State and local public health officials expressed concern about CDC recruiting data sources directly rather than working through the public health departments. Additionally, because syndromic surveillance was a new and untested methodology, state and local health department epidemiologists raised questions about data quality, validity, and utility, and members of the US Congress expressed concerns about the number of false alerts and the cost and effectiveness of BioSense as an early warning system for bioterrorism.7,8
Re-Designs are Common

**Figure.** Evolution of the Centers for Disease Control and Prevention (CDC) BioSense public health surveillance system, 2002-2016. Abbreviations: ED, emergency department; ESSENCE, Electronic Surveillance System for the Early Notification of Community-Based Epidemics; NSSP, National Syndromic Surveillance Program; VA/DoD, Department of Veterans Affairs/Department of Defense.

ref: 2017-07-10-pub Deborah W. Gould, PhD et al - The Evolution of BioSense- Lessons Learned and Future Directions- 0033354917706954
OPIOID TREATMENT OPTIONS

1. Buprenorphine + Counseling
   (best option)

2. Methadone + Counseling
   (second best option)

3. Naltrexone + Counseling
   (option three)

4. Counseling Only
   (relatively poor outcomes)

5. Residential Treatment
   (high cost, high relapse rate)

6. Incarcerated in Treatment
   (not acceptable)
Opportunities for Surveillance

https://www.sentinelinitiative.org/sentinel-system-story
What is the Sentinel System?

One of the FDA's biggest jobs is to make sure drugs, vaccines, and medical devices are safe. FDA wants to know if patients get bad side effects from these products. To make it faster and easier to learn about problems, FDA created a special program called the Sentinel System.

How the Sentinel System Works

Sentinel System's 3 important parts

- Information: The system looks at billing claims and patient records.
- Expert Team: Sentinel works with scientists, doctors and computer experts.
- Computer Programs: They study large groups of patients who take the same medicine, or use the same device.

How does FDA use the information?

- FDA can choose to collect more information.
- FDA can provide updated safety information for patients and providers.
- If you have concerns about your own medical products, please contact your doctor.

Personal privacy

- No one at FDA or the Sentinel Operations Center has access to your name, address, or any other information that identifies you.
- For more information, visit sentinelinitiative.org.

Sentinel asks questions like:

- How many patients take the same drug?
- How many patients are getting bad side effects (swelling, bleeding, etc.)?
- Are side effects more common after taking one drug than after another drug that treats the same problem?

https://www.sentinelinitiative.org/sentinel-system-story
Sentinel uses a **distributed data approach** in which **Data Partners** maintain physical and operational control over electronic data in their existing environments.

The distributed approach is achieved by using a standardized data structure referred to as the **Sentinel Common Data Model**.

Data Partners transform their data locally according to the Common Data Model, which enables them to execute standardized computer programs that **run identically at each Data Partner site**.

Data Partners are able to **review the results of the queries before sending them back** to the Sentinel Operations Center. Queries are distributed and results are returned through a secure portal in order to preserve privacy. **The combined collection of datasets** across all Data Partners is known as the Sentinel Distributed Database (SDD). The figure below (next slide) illustrates Sentinel’s distributed data approach.
Distributed Database and Common Data Model

Data Partners (DPs) hold data in Common Data Model format:
- Enrollment
- Demographics
- Medical Utilization
- Pharmacy Prescriptions
- Diagnoses
- Procedures
- Laboratory Tests
- Vital Signs

Queries Distributed to Data Partners (DPs)

Sentinel Operations Center (SOC)

Query Results Reviewed and Returned to SOC (all direct identifiers removed)

Data transferred securely

https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model
Data Partners

1. Aetna Informatics
2. Blue Cross Blue Shield of Massachusetts
3. Department of Population Health Sciences, Duke University School of Medicine
4. Harvard Pilgrim Health Care Institute
6. HealthPartners Institute
7. Hospital Corporation of America
8. Humana, Inc., Comprehensive Health Insights
9. Kaiser Permanente Colorado Institute for Health Research
10. Kaiser Permanente Center for Health Research Hawaii
11. Kaiser Foundation Health Plan of the Mid-Atlantic States, Inc.
12. Kaiser Permanente Northern California, Division of Research
13. Kaiser Permanente Northwest Center for Health Research
15. Marshfield Clinic Research Institute
16. Meyers Primary Care Institute, a joint endeavor of Fallon Community Health Plan
17. Optum
18. Vanderbilt University School of Medicine, Department of Health Policy

https://www.sentinelinitiative.org/sentinel/data/data-partners
Timeline

2007
Congress passes Food and Drug Administration Amendments Act (FDAAA)

2008
FDA launches Sentinel Initiative

2009
FDA launches Mini-Sentinel Pilot

2011
Mini-Sentinel distributed dataset reaches 100 million lives mark mandated by FDAAA

2012
Mini-Sentinel has suite of reusable programming tools for routine queries

2016
FDA launches Sentinel System

https://www.sentinelinitiative.org/background
Access

“... in the unique position of being able to offer industry, academia, and researchers access to a system similar to Sentinel for evaluating safety signals, implementing post-market studies, and assessing the impact of risk management actions.”

https://www.sentinelinitiative.org/sentinel/reagan-udall-foundation-and-imeds

Data Partners find tremendous value in the internal use of their data and the Sentinel toolkit.
https://www.youtube.com/watch?v=r2axuu9swgA
Active Risk Identification and Analysis (ARIA): The U.S. Food and Drug Administration’s (FDA) active post-market risk identification and analysis system, which is comprised of pre-defined, parameterized, reusable routine querying tools, combined with the electronic data in the Sentinel Common Data Model. Because ARIA uses parameterized tools and a trusted multi-site distributed database that undergoes continuous quality checks and refreshes, safety analyses can be done more efficiently to conduct medical product safety surveillance to fulfill the mandate in the FDA Amendments Act of 2007.

Sentinel Collaborating Institutions: A network of Data and Academic Partners that work with the FDA and Sentinel Coordinating Center to provide access to both healthcare data and scientific, technical, and organizational expertise.

Sentinel Coordinating Center: The Sentinel Coordinating Center includes the Sentinel Operations Center (SOC), comprised of the Applied Surveillance, Scientific Systems, and Administration Divisions housed at the Harvard Pilgrim Health Care Institute (HPHCI), and advisory groups. Both the Sentinel Coordinating Center and the SOC are led by the Sentinel Principal Investigator at HPHCI.

Sentinel Data Partners: Data Partners in the Sentinel System include a diverse group of organizations including academic medical centers, healthcare systems, and health insurance companies. Sentinel Data Partners maintain physical and operational control over electronic data in their existing environments.

Sentinel Infrastructure: The underlying data infrastructure created to enable analysis within the Sentinel System. The Sentinel Infrastructure involves: 1) a distributed data approach in which Data Partners maintain physical and operational control over electronic data in their existing environments; and 2) a Common Data Model consisting of standardized administrative and clinical information across Data Partners. The Sentinel Infrastructure has the potential to allow analysis of the data for other purposes besides safety for the FDA or those outside the FDA.

Sentinel Initiative: A multi-year effort beginning in 2008 to create a national electronic system for monitoring the performance of FDA-regulated medical products to improve the FDA’s ability to identify and assess medical product safety issues.

Sentinel System: An active surveillance system that uses routine querying tools and pre-existing electronic healthcare data from multiple sources to monitor the safety of regulated medical products. Subcomponents of the Sentinel System include: ARIA, PRISM, BloodSCAN and STAT.

https://healthpolicy.duke.edu/sites/default/files/atoms/files/2018_sentinel_initiative_workshop_terms_final_1000_hours.pdf
Questions