Big Data, Big Problems & Big Potential

William C. L. Stewart
Assistant Professor of Statistics and Pediatrics

Nationwide Children's Hospital and
The Ohio State University

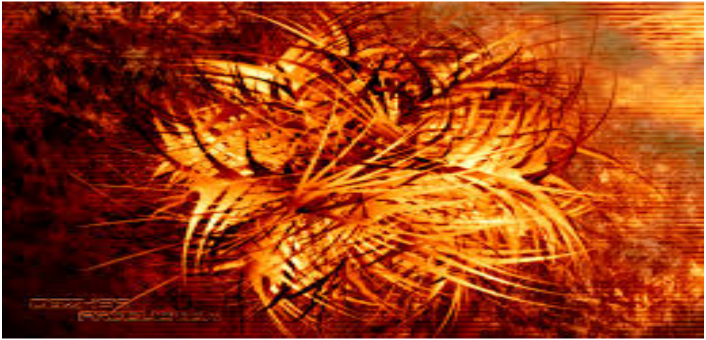
William.Stewart@nationwidechildrens.org

September 22, 2015

What Big Data looks like

Massive, Structured, Layered, Complex, and Dynamic

What Biomedical Big Data looks like to a Statistician



Highly unstructured, partly incomprehensible, partly unknowable, and in all honesty, a little dangerous.

Working Definitions...

Precision Medicine – is the customization of healthcare, with medical decisions, practices, and/or products and services that are tailored to individual patients. Molecular analyses (esp. the analysis of genomic and genetic data) are often used to select appropriate treatments and/or therapies.

Working Definitions...

Big Data – is a broad term for describing data sets that are so large and/or complex that traditional data processing applications are no longer adequate. Challenges inherent to Big Data are: analysis, capture, curation, queries, sharing, storage, transfer, visualization, and privacy. Presently, any data set with terabytes of data (i.e. trillions of bytes) would typically be considered Big Data.

Working Definitions...

EMR – an electronic medical record (aka EHR or EPR, although distinctions are now being made) is the collection of records for a single patient created by providers from specific encounters in hospitals and ambulatory environments. Typically, an EMR is generated and maintained within an institution.

A National Goal to Improve Public Health

Amass a large collection of EMRs so that, when used in conjunction with our current understanding of medical genetics/genomics, a new era of precision medicine can emerge.

A National Goal to Improve Public Health

Amass a large collection of EMRs so that, when used in conjunction with our current understanding of medical genetics/genomics, a new era of precision medicine can emerge.

Through the Medicare/Medicaid EHR Incentive Programs, the United States has sent a clear (i.e. $30B) message that paper records are out, and EMRs are in.

And given the continued support for genetic/genomic research at the NIH (National Institutes of Health), precision medicine is well on its way to becoming the new face of public health †

Google Translate (GT)

According to FJ Och, the creator of GT, one needs a bilingual text corpus of about 200 million words and two monolingual corpora each of about a billion words to accurately translate from one language to another. Presently, GT services 2,550 language pairs!

Interestingly, GT does not rely on grammars, semantics, and vocabulary. Instead, it assumes that a phrase in one language has probably been said before in another.

By exploiting the commonality of all human communications, GT yields better translations than rule-based methods, which suggests that languages are complex and at some level, unstructured.

Analogies to Biomedical Big Data

Because human health is complex and (at some level) unstructured, could one's overall health improve substantially from the implementation of something analogous to GT?

Many proponents of Big Data believe that the answer is "Yes", and that the emerging field of predictive analytics will produce a biomedical analog to GT.

Analogies to Biomedical Big Data

Because human health is complex and (at some level) unstructured, could one's overall health improve substantially from the implementation of something analogous to GT?

Many proponents of Big Data believe that the answer is "Yes", and that the emerging field of predictive analytics will produce a biomedical analog to GT.

Basically, a new patient should be able to benefit from the knowledge of health-related outcomes of similar patients that came chronologically before him/her. Put another way, with 7.3 billion people worldwide, the best treatment has probably already been administered to a similar patient somewhere.

An Early Attempt at Precision Medicine

Dr. J Halamka, a full professor at Harvard Medical School, used the relevant data from Harvard hospitals (totaling some 3,000 terabytes in 2011) to show us that precision medicine for certain health conditions is already a reality.

Specifically, 10,000 women with similar tumors, their treatment regimes, and their subsequent health outcomes were analyzed to construct a customized treatment regime for a 50-year old Asian woman with stage IIIA breast cancer. As with the other 10,000 women, this patient's tumor was also HER-2 negative, estrogen positive, and progesterone positive.

An Early Attempt at Precision Medicine

Dr. J Halamka, a full professor at Harvard Medical School, used the relevant data from Harvard hospitals (totaling some 3,000 terabytes in 2011) to show us that precision medicine for certain health conditions is already a reality.

Specifically, 10,000 women with similar tumors, their treatment regimes, and their subsequent health outcomes were analyzed to construct a customized treatment regime for a 50-year old Asian woman with stage IIIA breast cancer. As with the other 10,000 women, this patient's tumor was also HER-2 negative, estrogen positive, and progesterone positive.

Dr. Halamka reports that the patient is "totally cured, and that everything is fine" †

Mathematical & Computational

The usual suspects (part I):

- Matrix inversion, maximization, multiple test corrections, cluster analysis, summation, integration, sequential analysis, adaptive designs, etc.
- Outliers, resampling techniques (e.g. bootstraps, permutation tests, Monte Carlo procedures, etc.), sensitivity analysis, etc.

Data Processing

The usual suspects (part II):

- Volume, Velocity, and Variety (Laney 2001 of MEGA Group)
- Natural Language Processing (NLP), data transfers, data compressions, and data integration.

Issues that Affect Both

The usual suspects (part III):

- Missing Data
- Contradictory Data
- Data Imputation
- Data/Dimension Reduction

Access & Sharing

- HIPPA (Health Insurance Portability & Accountability Act)
- IRBs (Institutional Review Boards)
- Interoperability between health care providers

The Potential for Prime-Time Precision Medicine

We need to go beyond just the records of a single institution...

The Potential for Prime-Time Precision Medicine

EPIC has Care Elsewhere, and Care Everywhere modules for sharing patient records within and across their network.

EPIC manages nearly 54% of the medical records of US patients and almost 2.5% of the records worldwide.

Molecular and computer-aided diagnoses are also on the rise, and this too could greatly increase the impact of precision medicine.

Furthermore, Hadoop BashReduce, and Apache Spark are adequate for parallelizing (Kambhampati et al. 2013) and streaming certain analyses involving Big Data.

A Precision Medicine Revolution

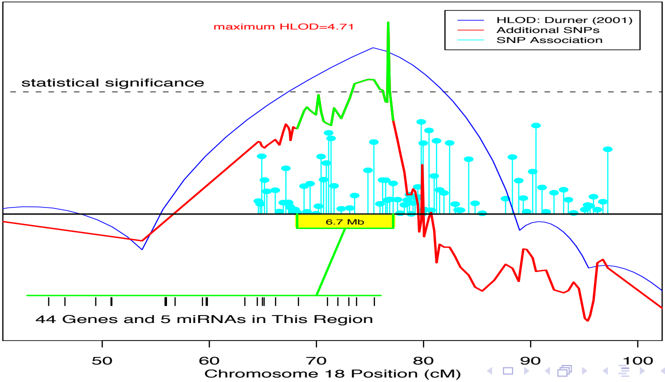
Presently, most of us think of precision medicine as an endpoint. However, if researchers were allowed *reasonable* access to biomedical Big Data, then precision medicine as we currently envision it, is ONLY the beginning!

Fueled by the generation of new knowledge at a relatively instantaneous rate, precision medicine will evolve rapidly over time.

As an example, let's consider the following real world scenario...

A Precision Medicine Revolution

- In a collection of idiopathic generalized epilepsy (IGE) families ($n \sim 140$), which took more than 10 years to recruit, diagnose, and genotype, Durner et al. (2001) identified chr18.



- Then, using sophisticated analyses that took another 5 years to develop (Stewart 2008, Stewart et al. 2010, Drill et al. 2011, and Cerise et al. 2013) we confirmed an association (Greenberg 2005) to malyic enzyme 2 (ME2).

POPFAM+

Gene	p-value
ME2	0.0054
RNF165	0.0093
RP11*	0.0116
EPG5	0.0145
SMAD4	0.0218
ZBTB7C	0.0245

Had this 15 year long project started today, with the aid of biomedical Big Data, it may have taken less than a month!

- Then, using sophisticated analyses that took another 5 years to develop (Stewart 2008, Stewart et al. 2010, Drill et al. 2011, and Cerise et al. 2013) we confirmed an association (Greenberg 2005) to malyic enzyme 2 (ME2).

POPFAM+

Gene	p-value
ME2	0.0054
RNF165	0.0093
RP11*	0.0116
EPG5	0.0145
SMAD4	0.0218
ZBTB7C	0.0245

Had this 15 year long project started today, with the aid of biomedical Big Data, it may have taken less than a month!

This will give rise to a precision medicine explosion.

Acknowledgements

Thank You!

Dr. Christopher Bartlett (PI), Dr. David Greenberg (PI, Director of Neurogenetics), Dr. Yungui Huang (Director, Research Data & Computing Services), Dr. Irina Buhimschi (Director, Center for Perinatal Research), Dr. Meng Wang (Post-Doctoral Fellow)

NATIONWIDE CHILDREN'S
*When your child needs a hospital, everything matters.*SM

THE OHIO STATE UNIVERSITY
WEXNER MEDICAL CENTER