Overcoming Barriers to Large-scale Biomedical Data Use

Eric Perakslis
Datavant

An AllerGen *Webinar for Research Success* (Bioinformatics series)

Eric Perakslis, PhD, delivered a webinar in AllerGen's Webinars for Research Success *series on April 9, 2018, discussing the challenges and opportunities of biomedical informatics for research. His main messages and a hyperlinked index to his presentation follow.*

THE INFORMATICS CHALLENGE

Something we do very well in medicine is study pieces of things: an x-ray, a genome, a microbiome—each a unique study in and of itself. In informatics, one challenge is to bring all this data together, to obtain a picture that incorporates all the different determinants of an individual patient's health. Sharing this kind of data across patients, in turn, which is a second challenge in informatics, builds data density and empowers greater insights. If each of us only has 20 patients, none of us is ever going to cure a disease; if together we have access to 4,000 patients, we have a better chance of doing so.

THE INFORMATICS OPPORTUNITY

The promise of biomedical informatics: Realizing the potential of biomedical informatics requires a multi-dimensional clinical data warehouse that can link clinical, biologic, and omics data; that allows us to ask questions of the data; and, ideally, that we can also mine hypothesis free. By the latter, I mean: let early versions of machine learning ask questions for us. If the data is smart, it should be able to suggest some questions.

When we accomplish this, one outcome is the redefining of diseases. Diseases once understood as singular are now considered multiple: breast cancer used to be considered one disease; now it is understood to have dozens of subtypes. When we redefine a disease, the drug targets, epidemiology, *etc*., also change very quickly. In this way, informatics enables patient stratification and drug target identification.

The TranSMART solution: One open-source solution to the challenges of biomedical informatics is a system I helped build called TranSMART. TranSMART is a cloud-based clinical data warehouse that links genomes, phenotypes, genetics, radiology reports—anything about a patient, in a very translational way; that is, it allows one to look across those different domains of data to see, for example, the clinical information and the genomics in one context. TranSMART works in conjunction with a system called i2b2; TranSMART specializes more in the clinical study data, and i2b2 in the health record data.

TranSMART was officially open sourced in 2012. Today, there are foundations that run TranSMART, and more than 60 million euros worth of clinical data partnerships are being hosted on this data-sharing platform that we basically built and gave away.

CyberSecurity: In the early days of cloud-based informatics, few anticipated how serious the healthcare cybersecurity threat would become. When I last researched the topic, healthcare data was still the most valuable data to steal, in terms of dollars per record.

However, the cybersecurity threat should not be considered an insurmountable barrier; it just demands the harnessing of expertise to meet it. There are always reasons not to do something, but medical science is too important, and data sharing is too important, to let cybersecurity make it un-doable.

De-identification: Another aspect of security is the securing of privacy by de-identifying data; that is, by taking steps to prevent a person's identity from being connected with their information. Generally, research networks strive to de-identify all their data.

However, in some situations, it may not be possible. With rare diseases, for example, where the patient phenotypes are so severe, where you genotype them and their living pedigree and aggregate so much data on them, it may not be possible to de-identify the data properly.

There may even be cause *not* to de-identify.

The Undiagnosed Diseases Network, funded by the National Institutes of Health Common Fund, runs its data fully identified.

As important as privacy and security are, if you have a gravely sick child, you may decide, for example, to tweet pictures of that child to find a second child with information that would help the first. In rare diseases, patients are often willing to share their data completely; they put their pictures up, they put their records up, because they are looking for help. There are amazing stories of parents getting together on Facebook and building a whole registry of patients.

Thus, decisions around data privacy and security, as with everything in medicine, is about weighing the benefit against the risk.

Ethics Approvals: When it comes to having an informatics project reviewed by an ethics or a privacy board, although there are always extreme cases, in general I have found that, if you make the right case and articulate the benefits well, people are reasonable. You need to engage well with patients, with clinicians and ethicists in your planning, however, to ensure your proposed strategy is well informed.

Local Data Sharing—Simple Solutions: The better organized everybody's data, the easier it is to share. People that want to share data will find a way, even without recourse to sophisticated technologies; but to succeed you need to make sharing as easy as possible.

For example, one local solution is to have a shared drive that is clearly organized, so people can readily find the files relevant to their needs. This can be as simple as establishing file naming conventions; for example, every file can be named by the first three letters of the researcher's name, a project number, and then the file type, so in an organization of 200 people, everybody knows what the file names mean.

There are amazingly simple, innovative things like this that you can do to make data easier to share. Steps like this, in turn, can help nudge resistant institutions toward implementing data sharing at an institutional level.

RESOURCES

Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease

Translational research platforms integrating clinical and omics data: a review of publicly available solutions

TranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform

Cybersecurity in Health Care

A Cybersecurity Primer for Translational Research

INDEX OF WEBINAR CONTENT

Pushing Data Uphill:
Overcoming Barriers to Large-scale Biomedical Data Use

Q & A

Available for this webinar: slides (in PDF) | video recording

***Eric Perakslis** (PhD) is Chief Science Officer at Datavant, a technology company dedicated to organizing and integrating the world's healthcare data, and Visiting Scientist at the Department of Biomedical Informatics at Harvard Medical School. He is a research, informatics, technology and R&D leader with over 20 years of direct experience in information technology, informatics, research, healthcare, government regulation, biotechnology and pharmaceuticals discovery and development. He was previously Senior Vice President Data Sciences at Takeda R&D and served as the CIO and Chief Scientist (Informatics) at the U.S. Food and Drug Administration.*