Perspective

Big data in health and disease: re-processing information for discovery and validation

Roseanne Yeung[1], Enrico Capobianco[2]

[1]Division of Endocrinology, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Canada; [2]Center for Computational Science, University of Miami, FL, USA

Correspondence to: Enrico Capobianco. Center of Computational Science, University of Miami, Gables One Tower 600, 1320 S. Dixie Highway Suite 600K, Loc: 2965, Coral Gables, FL 33146, USA. Email: ecapobianco@med.miami.edu.

Abstract: A lot has been already said about the emerging role of big data in health and disease. Large scale data efforts are increasingly being undertaken in response to the advent of Personalized and Precision Medicine and in association with both the "omics revolution" and the Electronic Health Records centrality. big data have demonstrated that their complex characteristics bring both strength factors and bottlenecks to research problems widely identified, analyzed and reviewed across many sectors of medicine and public health. As the most significant feature of big data is "variety", and this implies heterogeneity, our knowledge in complex disease contexts may substantially benefit from the fusion of different data types when a major role is assigned to harmonization and interoperability strategies. We discuss of an example, diabetes.

Keywords: Dark matter; electronic health records (EHR); diabetes

Received: 24 January 2019; Accepted: 28 February 2019; Published: 14 March 2019.
doi: 10.21037/jmai.2019.03.01
View this article at: http://dx.doi.org/10.21037/jmai.2019.03.01

Introduction

"Invisible dark matter makes up most of the universe—but we can only detect it from its gravitational effects" (https://home.cern/about/physics/dark-matter).

Cross-border thinking between medicine and physics is inspired, among other factors, by the parallel centered on the "Dark Matter" term. Of popular use in Physics, this term has been applied to genomics [see (1), and the references therein], within the non-coding RNA context (https://www.encodeproject.org/, https://www.gencodegenes.org/) (2).

Both Dark Matter and ncRNAs recall a type of complexity which is hard to decipher. A possible association with big data is justified by data characteristics like volumes, dimensionality and heterogeneity that increase the uncertainty of descriptive and analytical inference procedures. In genomics, Dark Matter refers to patterns that while appearing pervasively detected remain largely ignored because of the speed of technological changes. These are sustainable but require timely and efficient methods, say experimental, quantitative and analytical, but also qualitative (from annotations and validations).

In principle, acknowledging the existence of complex synergistic effects at cellular level has offered motivation to scrutinize the genomes further in depth, as "epigenetics" suggests (3). While genomics clearly represents a golden mine of information to understand complex mechanisms at a biological level, health and disease issues present many implications and challenges. In particular, the ultimate goal in many complex diseases is advancing knowledge with regards to effective therapies rather than radically curing.

Methodological aspects

Big data implications

Consider the example of diabetes mellitus, a well-studied complex pathology defined across a broad continuum of factors including genomics, biochemical measurements, clinical history, patient reported outcomes, and resource utilization, just to name a few (4). Better technological and methodological tools are needed to capture all such factors, and to connect them. In turn, this may lead to meaningful patterns improving the assessment of health outcomes.

Electronic health records (EHR) may offer this type of contribution (5), and several studies have started to appear [see for instance (6), in which detection of T2D phenotypes received large benefit from EHR *vs.* more traditional data].

Big data imply that emerging information technologies enable health providers and patients to collect and analyze data from multiple entry points. But a few questions remain: (I) How do we best capture knowledge from a myriad of data sources and sheer data volumes? (II) How do we resolve redundancy, gaps, systematic errors? (III) How do we establish the meaningfulness of data to analyze and cross-reference further downstream?

In the medical and overall healthcare context data circulation involves many subjects, and mainly clinicians at one extreme and patients at the other extreme. What is currently lacking is the identification of EHR with a novel type of asset. Exactly as a customer's profile may orient product sales towards that customer and others with similar attributes (community), a patient's profile can offer information on the needs of an individual or a group with specific characteristics that may guide preventive or curative solutions. And behind such needs there are economic aspects too, involving goods like drugs, food, lifestyle aspects etc.

Big data bottlenecks

Will healthcare systems be able to adapt rapidly to new standards and the medical practices scale to data-empowered clinical outcomes? Furthermore, will such aims possibly lead to enhancement of people engagement?

Back for a moment to complexity. A typical challenge in complex systems is the identification of "states". Of interest is the identification of person-driven trajectories across a spectrum of pathophysiological conditions (7). States can be defined as components or building-blocks of such trajectories: through them, one can try to measure change-of-state dynamics, and establish whether the trajectories have a few characteristics, namely: (I) temporal consistency, i.e., the same sequential order of states is observed in groups or communities; (II) persistence or transitory nature, i.e., an extended timeframe is covered thus allowing stationary behavior; (III) anomaly of patterns, identified as those deviating from regular ones.

Translated into clinical terms, epidemiologists and clinicians will face novel disease classification and population risk stratification methods that now must assimilate vast amounts and varieties of data, calling for a refinement of the analytics designed for the treatment of life-course evolving medical complexity. Akin to (I) defining data standards from varieties or sufficient statistics from big data heterogeneous volumes, and (II) making interoperability concrete and actionable, a key aspect is to establish "minimal information units" that may be considered those containing structured and unstructured information. This step is needed to assess significance and allow benchmarking and cross-referencing of multilevel evidence and results.

Similarly to the 'qubit' in quantum systems, decision makers—from clinicians to informatics experts—must decide especially upon consistency, robustness and scalability of methods that capture routine clinical data and enhance the clinical workflows. This component will hugely impact the efficiency of big data processing. In summary, by assuming that a better understanding of some of the big data organizing principles is a priority, the next step is to move on with facing contextual complexities (say disease-related, or specific to a reference environment or health facility etc.).

This below is a list of open questions.

(I) How well data linkages can be revealed? We need to establish the role of granularity of inherent structures, hierarchies, clusters, within data types, and also linkages between data varieties (networks and …. networks).

(II) What metrics serve the best possible scopes of analyses, how measuring similarity/dissimilarity in object (non-Euclidean) spaces? It is key to understand more of information embedding techniques designed to reach novel mappings (i.e., from images to networks, from text to trees, etc.).

(III) What types of synthesis measures could be established? It is destined to grow the centrality of association rule learning to assign statistical significance to data processed by machine learning techniques (say, blockchains in health, considering that more than 40% of healthcare organizations believe that interoperability of EHR will accelerate blockchain implementation, according to an Hyperledger's survey).

Preliminary results

Use case: big data standardization

We recall (8) the recommendations that AMIA (American Medical Informatics Association) recently proposed to NITRD (Networking and Information Technology Research and Development) program about increasing and

improving of quality and efficiency of health data exchange:

- Making sense of all new data types (EHR, IoT, devices, etc.);
- Focus on health IT interoperability testing;
- Training in foundational informatics;
- Policy development as driver of technology optimization.

More specifically, critical data areas are: (I) granular specifications and metadata solutions, considered key to facilitate data re-use; (II) interoperability testing to ensure that standards may allow variations; (III) portability to help patients engagement by better usability and built-in safety. In parallel, other guidelines come from ONC (the Office of the National Coordinator for Health Information Technology) (9). By promoting EHR interoperability, ONC identified 4 areas causing variation in measurement standards: (I) architecture, say cloud *vs.* client-based for instance (II) development level, say granularity depth regarding measurement (III) data access (end user friendly) and (IV) variability in the way a standard is used.

Use case: diabetes

Below, a short list of prioritized aspects relevant to developments and advances in this field.

Diabetes registries

Encourage establishment of minimum standards for each center to create a diabetes patient registry such that de-identified data can be collected for existing or new registries.

EHR

Facilitate effective and efficient use in diabetes and ordering systems by sharing creative solutions, protocols, and order sets among members to make best use of currently-existing EHR systems. Provide representation as a single voice to request specific EHR improvements that impact diabetes care.

Standards

Agreement on how to capture important clinical characteristics affecting diabetes management, with examples including: diet, exercise, progression of symptoms, hypoglycemia and hypoglycemia awareness—thus making some good use of the data, by exploiting common features and improving the level of governance.

Data governance

Effective diabetes data governance requires more accurate geo-localization and intercommunication between data sources. Profile descriptors and assessment standards need to be refined in view of emerging risk factors and clinically relevant variables.

Patient engagement

A new model of data participation and share is required. Clinician-patient engagement is only one example within a web populated by other subjects. It must become clear what the influential connecting nodes are and what roles they play. The data flow from multiple sources need to be reconciled and harmonized. Many EHR do not have practical functionality to enable panel management. Relevance to the design of patient histories is needed on the basis of their disease and comorbidity trajectories. Patient education/empowerment is rudimentary in foundation builds—EHRs should better support patient communication and use of multimedia.

Development of clinical decision support systems

User interfaces in EHR tools are often overwhelmingly busy, and relevant data lie hidden within such systems. The salient features should be easily identified to drive the user across diagnosis, treatment, and prognosis. A convergence to complex diabetes population maps is expected at the end of an integration process inspired by personalized view (bottom up). The traditional way of dictation has now been replaced by typing, but many face problems with data input, diminishing the quality of data entry. Also, current EHR make it difficult to access data, and even "holding the data hostage" practices are present, counter to medical practice standards where clinicians and administrators are increasingly tasked with population management.

Discussion and concluding remarks

The concept of big data reveals a complexity that goes far beyond data types or technicalities involved (mining, analysis, etc.). It is the reference context that makes a difference and allows to generate hypotheses. These need to be tested and validated, to be found clinically relevant. We considered diabetes as an example. In diabetes, causality is to a large extent well understood, and a multitude of factors have a defined role. Other complex diseases offer more uncertainty, especially when causal relationships are often not identifiable. Therefore, big data could start to have an impact in diabetes first, and then reach other contexts (10).

Big data may help in establishing stability of some

associations between variables and define feature heterogeneity across patient sub-populations beyond epidemiological stratifications. Technologies are main influencer of the central role of big data. The gains here are to be seen in the superior predictive power coming from modeling big data according to precision tools such as decision support systems, in which automated rules are given and algorithmic learning goes through sophisticated machine learning and Artificial Intelligence methods. These developments will enforce the practical benefits of big data by making them of large use to the non-expert field profiles, especially clinicians not willing to engage in methodological issues but eager to make these tools part of the learning healthcare machine.

Approximately 98% of the genome has been recognized to be something else than genes, i.e., genomic dark matter. Pseudogenes, long non-coding RNA, transposons, and many other biotypes exist and play a role in complex cellular regulation mechanisms. Knowing what alterations affect their functions and the kind of synergistic effects that we can measure on gene networks are most likely the useful information we need to advance in biomedicine. Beyond genomics, connecting all dots at 360 degrees requires cross-referencing all the types of information now available to us. Big data in the clinics are now becoming available, and more are on the way. The complexity associated to them is also destined to grow, and it is easy to predict a proliferation of methods: then, it seems wiser and better to focus on maximizing their interpretability.

Acknowledgments

The Authors would like to thank their colleagues for stimulating discussions.
Funding: None.

Footnote

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai.2019.03.01). EC serves as an unpaid editorial board member of *Journal of Medical Artificial Intelligence*. The other author has no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Capobianco E, Tsinoremas N. Health or Disease - Why does “Dark Matter” matters more? J Investig Genomics 2014;1:3-4.
2. Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Front Genet 2015;6:2.
3. Greally JM. A user’s guide to the ambiguous word “epigenetics”. Nat Rev Mol Cell Biol 2018;19:207-8.
4. Shah VN, Garg SK. Managing diabetes in the digital age. Clin Diabetes Endocrinol 2015;1:16.
5. Capobianco E. Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective. Clin Transl Med 2017;6:23.
6. Anderson AE, Kerr WT, Thames A, et al. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J Biomed Inform 2016;60:162-8.
7. Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Comm 2014;5:4022.
8. ONC Seeks Improved Interoperability Standards Measurement. Available online: https://ehrintelligence.com/news/onc-seeks-improved-interoperability-standards-measurement
9. About-ONC. Available online: https://www.healthit.gov/topic/about-onc
10. Delgado AP, Brandao P, Narayanan R. Diabetes associated genes from the Dark Matter of the human proteome. MOJ Proteom Bioinform 2014;1:86-92.

doi: 10.21037/jmai.2019.03.01
Cite this article as: Yeung R, Capobianco E. Big data in health and disease: re-processing information for discovery and validation. J Med Artif Intell 2019;2:5.