Real-life clinical data mining: generating hypotheses for evidence-based medicine

Jean-Emmanuel Bibault

Laboratory of Artificial Intelligence in Medicine and Biomedical Physics, Stanford University School of Medicine, Stanford, CA, USA

Correspondence to: Jean-Emmanuel Bibault, MD, PhD. Laboratory of Artificial Intelligence in Medicine and Biomedical Physics, Stanford University School of Medicine, Stanford, CA, USA. Email: jebibault@stanford.edu.

Provenance: This is an invited article commissioned by the Editorial Office, *Annals of Translational Medicine*.

Comment on: Zhou ZR, Wang WW, Li Y, *et al.* In-depth mining of clinical data: the construction of clinical prediction model with R. Ann Transl Med 2019. doi: 10.21037/atm.2019.08.63

Submitted Oct 11, 2019. Accepted for publication Oct 22, 2019.

doi: 10.21037/atm.2019.10.99

View this article at: http://dx.doi.org/10.21037/atm.2019.10.99

Zhou *et al.* report in this journal a series of methodological reviews of clinical prediction model construction in 16 sections. They thoroughly explain the concept, current application and construction methods required to create such models with real-life data. Evidence-based medicine is based on randomized clinical trials designed to answer specific questions, like the efficacy or toxicity of a treatment. The large number of parameters that need to be taken into account to treat patients makes it very difficult to design dedicated trials (1). Clinicians need to be able to use all the data generated by a patient, in order to better personalize the therapy. Electronic health records (EHR) could be used to create detailed phenotypes. At the same time, the similarity between patients included in research protocols and "everyday" patients is questioned (2). New approaches are needed, where we can use detailed phenotypic profiles of a large number of patients in order to define new treatment strategies. Here we will develop the concepts on which predictive models are built.

Data science and the concept of Big Data in medicine

The concept of Big Data is generally used to describe "datasets so large or complex that traditional data processing methods are inadequate" (3). The concept also relies on the definition provided by Gartner, which proposed to describe Big Data by its Volume, variety and Velocity (4). Additional "V's" have since been used, including Veracity, Value, and Variability (5).

Volume: how big is Big Data?

The size of big datasets, their complexity and diversity, make them difficult to store, retrieve, manipulate and analyze. New datasets are made available every day, yielding petabytes (10^{15} bytes) or exabytes (10^{18} bytes) of data.

Variety: where does Big Data come from?

Variables that should be considered and integrated into a Big Data analysis are heterogeneous by nature (6). They can be extracted from medical software or from electronic health records and include, but are not limited to:

- Clinical features (medical history, type, stage and grade of cancer, patient reported outcomes)
- Imaging features: number of lesions, size, volume, texture, radiomics features
- Molecular features: treatment sensitivity (7), proliferation and normal tissue reaction (8).
- Treatment features: drug dose and temporality, surgical techniques, procedure duration.

The available sources of data can be found across many levels and scales: from molecular, cellular data to whole individual or populations. This data can be either structured or unstructured. Structured data are well organized, and follow a consistent pattern. They are simple to enter, store, query and analyze. In healthcare, examples of structured

data are genomics data, laboratory tests results and data generated by treatment planning or record and verify systems. In medical specialties relying on informatics systems for treatment planning and delivery, data is already highly structured and can be easily extracted. However, this data can have very heterogeneous labels that will require time-consuming curation. On the other end of the spectrum, unstructured data has no predefined and does not conform to rows and columns. This kind of data can be extracted from various sources (medical notes, manuscripts, reports) and medical imaging (radiographs, computed tomography, and magnetic resonance imaging). Up to 80% of healthcare data is unstructured (9).

Velocity: how fast is Big Data generated and interpreted?

The Velocity concept of Big Data is related to how frequently the data is updated or the data's growth over time (10). This feature is of major importance in population or public-health datasets where receiving and analyzing data in real-time could improve the understanding of a phenomenon and its mitigation. Data updated in real-time has significant value, as it could leveraged for strategic decisions in a preventive of therapeutic manner. A system collecting a huge amount of data, but unable to analyze it rapidly would be useless in that setting.

Variability: how does Big Data change?

Variability is related to the heterogeneity of the data available, its completeness and how it may change over time. Analyzing data that changes dynamically represent a significant methodological challenge: data quality control procedures must be used, with decision on how to impute missing values or how to handle repeated measurements (3). For instance, gene transcription can be different in different organs and change over time. The context in which a genomics analysis is performed can be very important for the interpretation of the results.

Veracity: how accurate is Big Data?

The concept of Veracity relates to the fact that the quality of the data is very important: no matter how well the data will be used, if its quality is bad, the results will not be interpretable. Noise, redundancy, inconsistency can lead to significant bias and should be controlled before any analysis pipeline is used. Data needs to be cleaned in a faithful manner with rigorous integrity (11) before it can be used. "Dirty Data" will provide poor results: the use of flawed, or nonsense input data will produce nonsense output or "garbage". This is the GIGO principle: "Garbage In, Garbage Out", no matter how many patients are included, or how many variables are explored.

In order to minimize this effect, EHR standardization should be favored. Data heterogeneity will lead to methodological difficulties and unreliability. In that regard, the use of standard medical ontologies in reports is almost mandatory. There are currently approximately 440 medical ontologies. The most common are SNOMED (12), the NCI Thesaurus (13), CTC AE (14) and the UMLS meta-thesaurus (15).

Value: why is Big Data important?

Big Data has the potential to provide an understanding of complex conditions that rely on many different variables that have been resistant to analysis. The value also comes from the fact that it could generate new knowledge more quickly than the traditional scientific methods (16,17). Big Data is also unbiased by prior knowledge and holistic, since it is not limited to a single pathway or individual. In that sense, Big Data better capture the variability of biology or individuals (18). The size of big datasets, their complexity and diversity, make them difficult to store, retrieve, manipulate and analyze. New datasets are made available every day, yielding petabytes (10^{15} bytes) or exabytes (10^{18} bytes) of data. Today, the volume of data for an initial patient file will be around 10 GB, with genomics and imaging accounting for most of this amount.

Modelling can rely on traditional statistics or machine learning techniques. In any case, their main objective is to produce a model able to classify, predict, and estimate an outcome using known data. These models can encompass a higher number of parameters than the human can do (19). Machine Learning methods include artificial neural network, decision trees, random forest, support vector machine or bayesian networks, to name a few. One of the main limitations of the most advanced machine learning techniques lie in the fact that they are hardly interpretable.

In their article, Zhou *et al.* focus on logistic regression methods, that have the advantage to be easily interpreted and understandable by a human. They provide clear explanations on methods and hands-on examples with R code that will be very relevant for any clinician.

Acknowledgments

None.

Footnotes

Conflicts of Interest: The author has no conflicts of interest to declare.

Ethical Statement: The author is 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. Chen C, He M, Zhu Y, et al. Five critical elements to ensure the precision medicine. Cancer Metastasis Rev 2015;34:313-8.
2. Geifman N, Butte AJ. Do cancer clinical trial populations truly represent cancer patients? A comparison of open clinical trials to the cancer genome atlas. Pac Symp Biocomput 2016;21:309-20.
3. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics 2016;16:741-58.
4. Sagiroglu S, Sinanc D. Big data: A review. In: Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE; 2013:42-7.
5. Andreu-Perez J, Poon CC, Merrifield RD, et al. Big data for health. IEEE J Biomed Health Inform 2015;19:1193-208.
6. Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 2013;10:27-40.
7. Bibault JE1, Fumagalli I, Ferté C, et al. Personalized radiation therapy and biomarker-driven treatment strategies: a systematic review. Cancer Metastasis Rev 2013;32:479-92.
8. Okunieff P, Chen Y, Maguire DJ, Molecular markers of radiation-related normal tissue toxicity. Cancer Metastasis Rev 2008;27:363-74.
9. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data 2014;1:2.
10. Jee K, Kim GH. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013;19:79-85.
11. Rodríguez-Mazahua L, Rodríguez-Enríquez CA, Sánchez-Cervantes JL, et al. A general perspective of Big Data: applications, tools, challenges and trends. J Supercomput 2016;72:3073-113.
12. Systematized Nomenclature of Medicine - Clinical Terms - Summary | NCBO BioPortal [Internet]. [cited 2016 Mar 7]. Available online: https://bioportal.bioontology.org/ontologies/SNOMEDCT
13. National Cancer Institute Thesaurus - Summary | NCBO BioPortal [Internet]. [cited 2016 Mar 7]. Available from: https://bioportal.bioontology.org/ontologies/NCIT
14. Common Terminology Criteria for Adverse Events - Summary | NCBO BioPortal [Internet]. [cited 2016 Mar 7]. Available online: https://bioportal.bioontology.org/ontologies/CTCAE
15. Fact SheetUMLS® Metathesaurus® [Internet]. [cited 2016 Mar 7]. Available online: https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
16. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309:1351-2.
17. Viceconti M, Hunter P, Hose R. Big Data, Big Knowledge: Big Data for Personalized Healthcare. IEEE J Biomed Health Inform 2015;19:1209-15.
18. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform 2012;7:130-4.
19. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;28:4268-74.

Cite this article as: Bibault JE. Real-life clinical data mining: generating hypotheses for evidence-based medicine. Ann Transl Med 2019. doi:10.21037/atm.2019.10.99

