Published online: 2020-08-21

Untangling Data in Precision Oncology – A Model for Chronic Diseases?

Xosé M. Fernández
Institut Curie, Paris, France

Summary

Objectives: Any attempt to introduce new data types in the entangled hospital infrastructure should help to unravel old knots without tangling new ones. Health data from a wide range of sources has become increasingly available. We witness an insatiable thirst for data in oncology as treatment paradigms are shifting to targeted molecular therapies.

Methods: From nineteenth-century medical notes consisting entirely of narrative description to standardised forms recording physical examination and medical notes, we have nowadays moved to electronic health records (EHRs). All our analogue medical records are rendered as sequences of zeros and ones changing how we capture and share data. The challenge we face is to offload the analysis without entrusting a machine (or algorithms) to make major decisions about a diagnosis, a treatment, or a surgery, keeping the human oversight. Computers don't have judgment, they lack context.

Results: EHRs have become the latest addition to our toolset to look after patients. Moore's law and general advances in computation have contributed to make EHRs a cornerstone of Molecular Tumour Boards, presenting a detailed and unique description of a tumour and treatment options.

Conclusions: Precision oncology, as a systematic approach matching the most accurate and effective treatment to each individual cancer patient, based on a molecular profile, is already expanding to other disease areas.

Keywords

Artificial intelligence, big data, data science, electronic health records, precision medicine

Yearb Med Inform 2020:184-7
http://dx.doi.org/10.1055/s-0040-1701985

1 Introduction

Precision medicine requires streamlined software pipelines to handle vast amounts of information, yet it faces additional challenges as we embrace new technologies: high-throughput sequencing, improved medical imaging, wearable sensors, etc. Over the past two decades, the fastest technological advance in history has universalised access to genomics, prompting an increase in the number of national genome sequencing programmes [1]. Managing these data and their interpretation are the biggest challenges alongside safe storage and long-term preservation. Redundant data can clog our warehouses, but "data" is not synonymous of "information". Data scientists are required to curate noisy datasets. All this is ushering us to a new era of "precision healthcare", bringing human biology to centre stage as we integrate information about complex traits and susceptibility to disease in our healthcare systems. Information from a wide variety of sources is transforming personal health and challenging already overstretched health management systems. We must find the right way to access it without neglecting privacy and data provenance tracking [2].

Medical care is largely defined by clinical practice guidelines based on population-level data, however genomic medicine relies on an individual's genome information to guide personal strategies for disease diagnosis, treatment, and prevention. Cancer patients are now routinely stratified according to which treatment will be most effective for their tumour. The identification of clinically useful gene expression signatures can also be used to adjust a therapy regimen to reduce risk of toxicity, resulting in better patient survival. This transformation in patient management is not restricted to cancer, as we are starting to see similar approaches in other complex diseases [3].

Oncology has become increasingly data-driven. Genomic and molecular advances inform the development of targeted therapies replacing the traditional approach of describing tumours as a disease of the tissue of origin (*e.g.*, breast cancer, colorectal cancer, or lung cancer) and cell types (sarcoma, carcinoma). New technologies and computational resources, which were unthinkable a few years ago, have made this a reality: including gene editing [4, 5], immunotherapy [6-9], and artificial intelligence [10, 11]. As new information flows, more intriguing applications materialise beyond cancer [12].

2 Objectives

We are living a digital revolution driven not only by the abundance of data, but also by our capability to collect, store, and analyse this information. We often forget how much we rely on mathematical models to harness the data tsunami [13]. Each of our patients is systematically screened for a myriad of molecular information at the clinic. This generates terabytes of data per patient from which decisions must be taken regarding the best therapeutic options available. Integrating such data (most of it unstructured) requires computational methods that involve bespoke procedures. Contextual information is essential as little signs may hide the clue to the correct interpretation, in particular in cases when useful domain knowledge is already available.

Genes play a fundamental role in the functioning of life. Genetics turns into genomics as we start analysing the entire DNA in an organism instead of just a few genes. Between two and 10 novel mutations creep into our genome when cells duplicate their DNA [14]. The driving force behind inheritance and evolution will only be fully understood when due attention is given to DNA interactions [15].

3 Methods

Clinical records transcend their original purpose of keeping a record of disease progression and crucial information to support an optimal therapy. Hospitals have adopted EHR systems to hold mountains of paperwork [16-18]. Yet, clinicians favour flexible (unstructured) data entry methods. This requires therefore a careful strategy to capture that critical contextual information as we develop suitable tools.

ConSoRe is a tool to query medical notes, pathology reports, diagnoses, hard-to-find lifestyle data and structures, all the required information from an EHR system [19]. Processing unstructured medical notes with accuracy according to a predefined disease model, cancer, is automated [20]. It combines state-of-the-art text mining natural language processing (NLP) techniques with semantic knowledge graphs. It provides the necessary flexibility to enable physicians to quickly identify patients matching precise criteria (potentially reducing recruitment time for a clinical trial from years to just days or weeks). A disconnected patchwork of electronic information systems becomes queryable through a unique gateway, not far from the cancer Biomedical Informatics Grid (caBIG®) promoted by the NCI [21].

We are quantifying human health and disease with the help of artificial intelligence (AI) approaches. Large datasets are analysed helping us to discover new drugs and tailored treatments. However these applications in precision medicine can be severely hindered by the scarcity of data available in the training datasets. Indeed, we can find datasets containing nearly as many features as samples. When applied to population-based samples beyond the original clinical setting, these datasets will underperform due to distributional shift [22].

Another issue AI faces is that it cannot yet replicate the diagnostic process. A physician will order different tests sequentially throughout the period she's following a patient, any given test might be ordered due to the results of a previous result. So, when an algorithm is trained on retrospective corpora, the temporal dimension is removed and therefore the dependency within the dataset is often lost. Any such model subsequently produced will not include the related decisions which ultimately led to the original diagnosis [23].

The final aspect in this equation is accountability. Algorithms should be able to detect biases and therefore, they require robust and complete data [24, 25]. When we use mathematical models (*e.g.,* neural networks) to identify patterns, skewed collections will lead to biased models (data collections may contain inaccuracies and errors which should have been cleaned prior to be used for training models) [26, 27].

An essential aspect in oncology is to relate a detailed and unique description of a cancer to useful properties such as response to therapy or risk of relapse. However, amongst the largest public cancer cell line panels there is a poor representation of key mutations [28, 29], this means any model developed with these will be statistically underpowered.

We are building a digital ecosystem integrating new and existing technologies and data. We can investigate the potential of representation learning for cancer genomics to allow the Molecular Tumour Board to exploit the hierarchical and multi-scale nature of the data available.

4 Results

Ever since Hippocrates founded his school of medicine in ancient Greece some 2,500 years ago, observation, experimentation, and data analysis have been a core ethical principle of medicine. Precision medicine relies upon comprehensive data (and biobanks) on patient treatment and outcomes. Analysis of these data leads to improved models providing the basis for treatment, and for direct use in clinical decision-making. In fact, it is data from previous patients that will probably play the biggest role in making a current patient well again. It gives our treating teams the essential insights and knowledge on which to base their care. We aggregate data in warehouses, we have mentioned ConSoRe, but in France alone, we can find other models outside oncology applications such as Dr Warehouse and eHOP [30, 31]. Ethical and legal issues are paramount when developing these infrastructures, as it is unclear how samples (and data) might be re-used and whether any future uses were compatible with the original consent.

AI is unleashing an array of new approaches to healthcare, but we need to continuously benchmark any progress. Innovative technologies will only be widely adopted in medicine once they significantly improve outcomes for patients, and their implementation is ironed out. Solutions with potential for widespread adoption cannot be resource intensive to deploy and use, and should not be too complex. Manually annotated cohorts can be used to establish baselines to benchmark automated tools [32]. An example we have been using at Institut Curie, is ESMÉ cohort (grouping 30,000 breast cancer patients), a well annotated resource from the Unicancer excellence network of French Comprehensive Cancer Centres. Structuring medical records with ConSoRe can be compared to the work from an experienced team of curators.

Understanding how cancer arises requires more than converting biopsies into ones and zeroes, or lists identifying which genes are mutated in certain cancers. Molecular signatures bring us a step closer towards finding interventions to halt disease and enhance health. Patient stratification meets clinical practise, evidence reveals the language of the cell as each subtype may exhibit a predictable clinical phenotype.

Machine Learning (ML) is benefiting from robust platforms that enable scalable and reproducible computing on large datasets, however, quality is often the challenge [33]. Oncology datasets are often unsuited, as we are dealing with noisy and sparse data, various independent resistance mechanisms can operate [34]. Particular attention must be paid to avoid overfitting [35], when comparing results. The final model would only be as good as the data that was used to develop it and test it.

In a moment when the healthcare data economy is booming, places like Oxford, Paris, or Cambridge are teeming with start-ups promising to harness the power of data.

Precision oncology increases the range of treatment options, bringing quality improvements, but for only a relatively small number of people. Challenging the modern clinical trial paradigm with basket-trial approaches is blurring one of the hallmarks of medicine, the educated guess, without any pretentions of certainty.

5 Discussion

The potential of computers to transform the clinical decision process has long been recognised. We can trace medical informatics as an interdisciplinary research field back to the 1960s [36, 37]. Mycin, an *ad hoc* model with about 450 rules developed to diagnose blood infections, was one of the first expert systems [38]. ADM, a computer-assisted diagnostic system developed in 1972, covered 2,500 diseases with 22,000 signs and symptoms [39, 40].

Despite decades of research on the development of computer-based patient records the process has been hindered by the hope that difficult clinical problems might yield to mathematical formalisms [41]. Today, technology is placed at the fingertips of everyone; wearables offer an opportunity to capture first-hand data and address disease in the early stages. However, this comes with a risk of swamping already saturated health services with anxious individuals alarmed by false positives, following the adoption of devices promising real-time atrial fibrillation detection [42].

Genomics enables us to decipher and understand the blueprint of a living organism as we better understand biological systems at a molecular level. When the human genome was first assembled [43, 44], it would have been hard to predict that a few years afterwards the future in cancer would pass by single-cell sequencing. Today, we can sequence individual cells from biopsy samples or circulating tumour cells, enabling earlier diagnoses. Even when the cost of sequencing a cancerous tumour has dramatically dropped to affordable levels, the cost of understanding what then needs to be done remains considerably high [45]. Bringing our knowledge about the clinical implications of various genomic elements to the Molecular Tumour Board still requires substantial research investment [47].

In the precision medicine space we are often expected to assess not only how new tests can guide our decisions (companion diagnostics), but also which additional value they bring to the healthcare system (exploring drug repurposing). The ultimate goal in precision medicine is not only to treat patients based on their unique biology, but to get such better care without spending more money.

The publication of the first draft of our genome, and the beginning of precision medicine, concluded with the following words: "*it has not escaped our notice that the more we learn about the human genome, the more there is to explore*". Twenty years after this major milestone, we can only stress how true those words were.

6 Conclusions

Only those technologies showing genuine clinical utility will be widely adopted in medicine. ML enables us to extract knowledge from the outcomes of thousands of patients (billions in a global context) to inform care of each single patient. Structured information plays a critical role in medical decision-making. A central promise of ML in medicine is that each patient will benefit from the wisdom contained in the decisions made by nearly all clinicians as they will be based on the outcomes of billions of past patients. A corollary is that patients need to be informed that by sharing their data they are not only helping individuals today, but also future patients.

As Eric Topol says "Electronic health records have broken the backs of clinicians and made them into data clerks. So why would anyone in their right mind think that we could have a rescue through technology? (…) We've never had a technology that could actually give us the gift of time" [48]. We are not discussing digital alchemy, but augmented medicine through rigorous research that provides unequivocal benefit for patients.

We have seen that the potential value of computers in medicine is not something recent, but the development of digital ecosystems embedding information from EHRs allow us to streamline clinical queries across normalised medical records. Within this context, we expand our toolset beyond the hospital, extending the ever-increasing patient cohorts with new data types, opens the door to new exciting opportunities in Precision Oncology [49].

Any innovation must not only address clinical problems but should result in significantly improved outcomes for patients. It should not be too complex or resource intensive to implement and use, and should have the potential for widespread adoption and diffusion [50]. The emphasis is often put on data quantity when it should be on quality, which is inherently expensive as it requires human curation. Data can be augmented, but quality cannot be taken for granted.

There is a hard lesson to learn when wandering in the limits of science and medicine. Solutions must involve a team physician-scientist, otherwise we risk solutions will not be adopted. We may have elucidated the iconic double helix and have a better understanding of immunology, but we are still unable to save people from most forms of malignancy.

Acknowledgments

I wanted to thank the reviewers whose suggestions have improved this contribution. The greatest gratitude to the generosity of all our donors for the support provided to the Curie Foundation to beat cancer together. This work was supported in part by the European Union's Horizon 2020 research and innovation programme under grant agreement No 780495. We also benefited from SiRIC (grant INCa-DGOS-Inserm_12554).

References

1. Saunders G, Baudis M, Becker R, Beltrán S, Béroud C, Birney E, et al. Leveraging European infrastructures to access 1 million human genomes by 2022. Nat Rev Genet 2019:20(11):693-701.
2. Perakslis E, Coravos A. Is health-care data the new blood? Lancet Digit Health 2019;1(1):PE8.
3. Bodkin JA, Coleman MJ, Godfrey LJ, Carvalho CMB, Morgan CJ, Suckow RF, et al. Targeted Treatment of Individuals With Psychosis Carrying a Copy Number Variant Containing a Genomic Triplication of the Glycine Decarboxylase Gene. Biol Psychiatry 2019;86(7):23-535.
4. Cox DB, Platt RJ, Zhang F. Therapeutic genome

editing: prospects and challenges. Nat Med 2015;21:121-31.
5. Fuster-García C, García-García G, González-Romero E, Jaijo T, Sequedo MD, Ayuso C, et al. USH2A Gene Editing Using the CRISPR System. Mol Ther Nucleic Acids 2017;8:529-41.
6. Baruch K, Deczkowska A, Rosenzweig N, Tsitsou-Kampeli A, Sharif AM, Matcovitch-Natan O, et al. PD-1 Immune Checkpoint Blockade Reduces Pathology and Improves Memory in Mouse Models of Alzheimer's Disease. Nat Med 2016;22(2):135-7.
7. Gong J, Chehrazi-Raffle A, Reddi S, Salgia R. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. J Immunother Cancer 2018;6(1):8.
8. Schwartz M, Arad M, Ben-Yehuda H. Potential Immunotherapy for Alzheimer Disease and Age-Related Dementia. Dialogues Clin Neurosci 2019;21(1):21-5.
9. Tang J, Shalabi A, Hubbard-Lucey VM. Comprehensive analysis of the clinical immuno-oncology landscape. Ann Oncol 2018;29(1):84-91.
10. Topol EJ. High High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
11. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6(1):26094.
12. Agustí A, Hogg JC. Update on the Pathogenesis of Chronic Obstructive Pulmonary Disease. N Engl J Med 2019;381:1248-56.
13. Wigner EG. The unreasonable effectiveness of mathematics in the natural sciences. Communications on Pure and Applied Mathematics 1960;13:1.
14. Martincorena I, Campbell PJ. Somatic Mutation in Cancer Normal Cells. Science 2015:34:1483-9.
15. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E. The Human Cell Atlas. eLife 2017:6:e27041.
16. Rush R. Taking Note. N Engl J Med 2019;381:9.
17. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L. et al. The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians. Mayo Clin Proc 2019 doi: 10.1016/j.mayocp.2019.09.024.
18. Berwick DM, Hackbarth AD. Eliminating waste in US health care. JAMA 2012;307(14):1513-6.
19. Heudel P, Livartowski A, Arveux P, Willm E, Jamain C. ConSoRe: un outil permettant de rentrer dans le monde du big data en santé. Bulletin du Cancer 2016;103:949. French.
20. Heudel P, Durand T, Fervers B, Gomez F, Rivoire M, Bachelot T, et al. Data-mining of 110172 electronic patient records with the ConSoRe tool: An analysis of second primary cancer in a comprehensive cancer center. Ann Oncol 2018: Suppl 8:viii482.
21. Warden R. Impact of caBIG on the European cancer community. Ecancermedicalscience 2011;5:225.
22. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836-42.
23. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:1-17.
24. Babic B, Gerke S, Evgeniou T, Cohen GG. Algorithms on regulatory lockdown in medicine. Science 2019:366(6470):1202-4.
25. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019;25:433-8.
26. Lipton ZC, Steinhardt J. Troubling trends in Machine Learning scholarship. arXiv:1807.03341v2.
27. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28:231-7.
28. Smith SE, Mellor P, Ward AK, Kendall S, McDonald M, Vizeacoumar FS, et al. Molecular characterization of breast cancer cell lines through multiple omic approaches. Breast Cancer Res 2017;19(1):65.
29. Im JS, Herrmann AC, Bernatchez C, Haymaker C, Molldrem JJ, Hong WK, et al. Immune-Modulation by Epidermal Growth Factor Receptor Inhibitors: Implication on Anti-Tumor Immunity in Lung Cancer. PLoS One 2016;11(7):e010004.
30. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform 2018;80:52.
31. Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny M-L, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. In: Ohno-Machado L, Séroussi B, editors. MEDINFO 2019: Health and Wellbeing e-Networks for All. Amsterdam: International Medical Informatics Association (IMIA) and IOS Press; 2019. p1536-7.
32. Pérol D, Robain M, Arveux P, Mathoulin-Pélissier S, Chamorey E, Asselainh B, et al. The ongoing French metastatic breast cancer (MBC) cohort: the example-based methodology of the Epidemiological Strategy and Medical Economics (ESME). BMJ Open 2019;9(2):e023568.
33. Adamson AS, Wlech HG. Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard. New Engl J Med 2019;381:2285-7.
34. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nat Rev Cancer 2013;13:714-26.
35. Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: Users' Guides to the Medical Literature. JAMA 2019;322(18):1806-16.
36. Schwartz WB. Medicine and the computer. The promise and problems of change. N Engl J Med 1970;283(23):1257.
37. Shortlife EH, Axline SG, Buchanan BG, Merigan TC, Cohen SN. An Artificial Intelligence program to advise physicians regarding antimicrobial therapy. Comput Biomed Res 1973;6(6):544.
38. Wraith SM, Aikins JS, Buchanan BG, Clancey WJ, Davis R, Fagan LM, et al. Computerized consultation system for selection of antimicrobial therapy. Am J Hosp Pharm 1976;33(12):1304.
39. Lenoir P, Chalès G. Efficacité de l'ADM sur les succès et les couts du diagnostic. Medical Informatics 1980;5(4):309. French.
40. Lenoir P, Roger MJ, Frangeul C, Chalès G. Réalisation, développement et maintenance de la base de données ADM. Medical Informatics 1981;6(1):51. French.
41. Whiting-O'Keefe QE, Simborg DW, Epstein WV, Warger A. A computerized summary medical record system can provide more information than the standard medical record. JAMA 1985;254:1185.
42. Steinhubl SR, Waalen J, Edwards AM, Ariniello LM, Mehta RR, Ebner GS, et al. Effect of a Home-Based Wearable Continuous ECG Monitoring Patch on Detection of Undiagnosed Atrial Fibrillation: The mSToPS Randomized Clinical Trial. JAMA 2018;320(2):146-55.
43. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial Sequencing and Analysis of the Human Genome. Nature 2001;409(6822):860.
44. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science 2001;291(5507):1304-51.
45. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical Assessment Incorporating a Personal Genome. Lancet 2010;375(9725):1525-35.
46. Abernethy AP, Etheredge LM, Ganz PG, Wallace P, German RR, Neti C, et al. Rapid-learning System for Cancer Care. J Clin Oncol 2010;28(27):4268-74.
47. Basse C, Morel C, Alt M, Sablin MP, Franck C, Pierron G, et al. Relevance of a molecular tumour board (MTB) for patients' enrolment in clinical trials: experience of the Institut Curie. ESMO Open 2018;3:e00339.
48. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books; 2019.
49. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-40.
50. Bogers M, Chesbrough H, Moedas C. Open Innovation: Research, Practices and Policies. California Management Review 2018;60(2):5-16.

Correspondence to:
X. M. Fernández
ORCID: ID 0000-0001-7139-6215
Institut Curie
25 rue d'Ulm
75005 Paris, France
Tel : +33 (0) 156246277
E-mail: xose.fernandez@curie.fr