A Unified Modeling Approach to Data-Intensive Healthcare

IAIN BUCHAN
University of Manchester

JOHN WINN
CHRIS BISHOP
Microsoft Research

THE QUANTITY OF AVAILABLE HEALTHCARE DATA is rising rapidly, far exceeding the capacity to deliver personal or public health benefits from analyzing this data [1]. Three key elements of the rise are electronic health records (EHRs), biotechnologies, and scientific outputs. We discuss these in turn below, leading to our proposal for a unified modeling approach that can take full advantage of a data-intensive environment.

ELECTRONIC HEALTH RECORDS

Healthcare organizations around the world, in both low- and high-resource settings, are deploying EHRs. At the community level, EHRs can be used to manage healthcare services, monitor the public's health, and support research. Furthermore, the social benefits of EHRs may be greater from such population-level uses than from individual care uses.

The use of standard terms and ontologies in EHRs is increasing the structure of healthcare data, but clinical coding behavior introduces new potential biases. For example, the introduction of incentives for primary care professionals to tackle particular conditions may lead to fluctuations in the amount of coding of new cases of those conditions [2]. On the other hand, the falling cost of devices for remote monitoring and near-patient testing is leading to more capture of objective measures in EHRs, which can provide

less biased signals but may create the illusion of an increase in disease prevalence simply due to more data becoming available.

Some patients are beginning to access and supplement their own records or edit a parallel health record online [3]. The stewardship of future health records may indeed be more with individuals (patients/citizens/consumers) and communities (families/local populations etc.) than with healthcare organizations. In summary, the use of EHRs is producing more data-intensive healthcare environments in which substantially more data are captured and transferred digitally. Computational thinking and models of healthcare to apply to this wealth of data, however, have scarcely been developed.

BIOTECHNOLOGIES

Biotechnologies have fueled a boom in molecular medical research. Some techniques, such as genome-wide analysis, produce large volumes of data without the sampling bias that a purposive selection of study factors might produce. Such datasets are thus more wide ranging and unselected than conventional experimental measurements. Important biases can still arise from artifacts in the biotechnical processing of samples and data, but these are likely to decrease as the technologies improve. A greater concern is the systematic error that lies outside the data landscape—for example, in a metabolomic analysis that is confounded by not considering the time of day or the elapsed time from the most recent meal to when the sample was taken. The integration of different scales of data, from molecular-level to population-level variables, and different levels of directness of measurement of factors is a grand challenge for data-intensive health science. When realistically complex multi-scale models are available, the next challenge will be to make them accessible to clinicians and patients, who together can evaluate the competing risks of different options for personalizing treatment.

SCIENTIFIC OUTPUTS

The outputs of health science have been growing exponentially [4]. In 2009, a new paper is indexed in PubMed, the health science bibliographic system, on average every 2 minutes. The literature-review approach to managing health knowledge is therefore potentially overloaded. Furthermore, the translation of new knowledge into practice innovation is slow and inconsistent [5]. This adversely affects not only clinicians and patients who are making care decisions but also researchers who are reasoning about patterns and mechanisms. There is a need to combine the mining

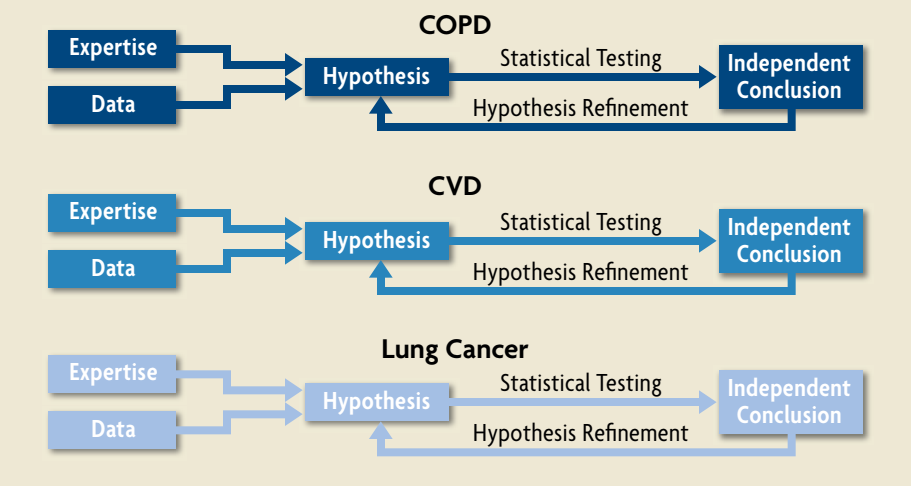

FIGURE 1.
Conventional approaches based on statistical hypothesis testing artificially decompose the healthcare domain into numerous sub-problems. They thereby miss a significant opportunity for statistical "borrowing of strength." Chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), and lung cancer can be considered together as a "big three" [6].

of evidence bases with computational models for exploring the burgeoning data from healthcare and research.

Hypothesis-driven research and reductionist approaches to causality have served health science well in identifying the major independent determinants of health and the outcomes of individual healthcare interventions. (See Figure 1.) But they do not reflect the complexity of health. For example, clinical trials exclude as many as 80 percent of the situations in which a drug might be prescribed—for example, when a patient has multiple diseases and takes multiple medications [7]. Consider a newly licensed drug released for general prescription. Clinician X might prescribe the drug while clinician Y does not, which could give rise to natural experiments. In a fully developed data-intensive healthcare system in which the data from those experiments are captured in EHRs, clinical researchers could explore the outcomes of patients on the new drug compared with natural controls, and they could potentially adjust for confounding and modifying factors. However, such adjustments might be extremely complex and beyond the capability of conventional models.

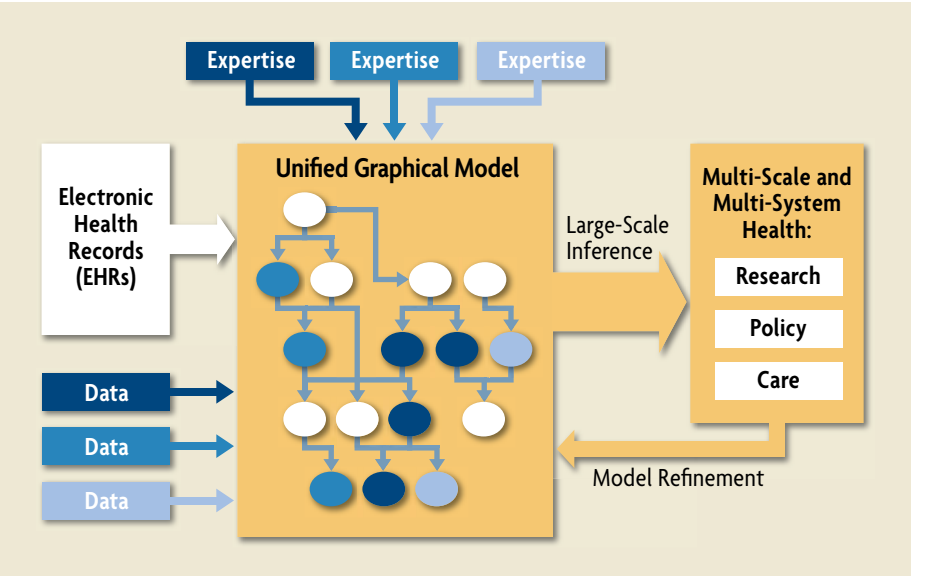

FIGURE 2.
We propose a unified approach to healthcare modeling that exploits the growing statistical resources of electronic health records in addition to the data collected for specific studies.

A UNIFIED APPROACH

We propose a unified modeling approach that can take full advantage of a data-intensive environment without losing the realistic complexity of health. (See Figure 2.) Our approach relies on developments within the machine learning field over the past 10 years, which provide powerful new tools that are well suited to this challenge. Knowledge of outcomes, interventions, and confounding or modifying factors can all be captured and represented through the framework of probabilistic graphical models in which the relevant variables, including observed data, are expressed as a graph [8]. Inferences on this graph can then be performed automatically using a variety of algorithms based on local message passing, such as [9]. Compared with classical approaches to machine learning, this new framework offers a deeper integration of domain knowledge, taken directly from experts or from the literature, with statistical learning. Furthermore, these automatic inference algorithms can scale to datasets of hundreds of millions of records, and new tools such

as Infer.NET allow rapid development of solutions within this framework [10]. We illustrate the application of this approach with two scenarios.

In scenario 1, an epidemiologist is investigating the genetic and environmental factors that predispose some children to develop asthma. He runs a cohort study of 1,000 children who have been followed for 10 years, with detailed environmental and physiological measures as well as data on over half a million of the 3 million genetic factors that might vary between individuals. The conventional epidemiology approach might test predefined hypotheses using selected groups of genetic and other factors. A genome-wide scanning approach might also be taken to look for associations between individual genetic factors and simple definitions of health status (e.g., current wheeze vs. no current wheeze at age 5 years). Both of these approaches use relatively simple statistical models. An alternative machine learning approach might start with the epidemiologist constructing a graphical model of the problem space, consulting literature and colleagues to build a graph around the organizing principle—say, "peripheral airways obstruction." This model better reflects the realistic complexity of asthma with a variety of classes of wheeze and other signs and symptoms, and it relates them to known mechanisms. Unsupervised clustering methods are then used to explore how genetic, environmental, and other study factors influence the clustering into different groups of allergic sensitization with respect to skin and blood test results and reports of wheezing. The epidemiologist can relate these patterns to biological pathways, thereby shaping hypotheses to be explored further.

In scenario 2, a clinical team is auditing the care outcomes for patients with chronic angina. Subtly different treatment plans of care are common, such as different levels of investigation and treatment in primary care before referral to specialist care. A typical clinical audit approach might debate the treatment plan, consult literature, examine simple summary statistics, generate some hypotheses, and perhaps test the hypotheses using simple regression models. An alternative machine learning approach might construct a graphical model of the assumed treatment plan, via debate and reference to the literature, and compare this with discovered network topologies in datasets reflecting patient outcomes. Plausible networks might then be used to simulate the potential effects of changes to clinical practice by running scenarios that change edge weights in the underlying graphs. Thus the families of associations in locally relevant data can be combined with evidence from the literature in a scenario-planning activity that involves clinical reasoning and machine learning.

THE FOURTH PARADIGM: HEALTH AVATARS

Unified models clearly have the potential to influence personal health choices, clinical practice, and public health. So is this a paradigm for the future?

The first paradigm of healthcare information might be considered to be the case history plus expert physician, formalized by Hippocrates more than 2,000 years ago and still an important part of clinical practice. In the second paradigm, a medical record is shared among a set of complementary clinicians, each focusing their specialized knowledge on the patient's condition in turn. The third paradigm is evidence-based healthcare that links a network of health professionals with knowledge and patient records in a timely manner. This third paradigm is still in the process of being realized, particularly in regard to capturing the complexities of clinical practice in a digital record and making some aspects of healthcare computable.

We anticipate a fourth paradigm of healthcare information, mirroring that of other disciplines, whereby an individual's health data are aggregated from multiple sources and attached to a unified model of that person's health. The sources can range from body area network sensors to clinical expert oversight and interpretation, with the individual playing a much greater part than at present in building and acting on his or her health information. Incorporating all of this data, the unified model will take on the role of a "health avatar"—the electronic representation of an individual's health as directly measured or inferred by statistical models or clinicians. Clinicians interacting with a patient's avatar can achieve a more integrated view of different specialist treatment plans than they do with care records alone.

The avatar is not only a statistical tool to support diagnosis and treatment, but it is also a communication tool that links the patient and the patient's elected network of clinicians and other trusted caregivers—for what-if treatment discussions, for example. While initially acting as a fairly simple multi-system model, the health avatar could grow in depth and complexity to narrow the gap between avatar and reality. Such an avatar would not involve a molecular-level simulation of a human being (which we view as implausible) but would instead involve a unified statistical model that captures current clinical understanding as it applies to an individual patient.

This paradigm can be extended to communities, where multiple individual avatars interact with a community avatar to provide a unified model of the community's health. Such a community avatar could provide relevant and timely information for use in protecting and improving the health of those in the community. Scarce community resources could be matched more accurately to lifetime healthcare needs,

particularly in prevention and early intervention, to reduce the severity and/or duration of illness and to better serve the community as a whole. Clinical, consumer, and public health services could interact more effectively, providing both social benefit and new opportunities for healthcare innovation and enterprise.

CONCLUSION

Data alone cannot lead to data-intensive healthcare. A substantial overhaul of methodology is required to address the real complexity of health, ultimately leading to dramatically improved global public healthcare standards. We believe that machine learning, coupled with a general increase in computational thinking about health, can be instrumental. There is arguably a societal duty to develop computational frameworks for seeking signals in collections of health data if the potential benefit to humanity greatly outweighs the risk. We believe it does.

REFERENCES

[1] J. Powell and I. Buchan, "Electronic health records should support clinical research," *J. Med. Internet Res.*, vol. 7, no. 1, p. e4, Mar. 14, 2005, doi: 10.2196/jmir.7.1.e4.

[2] S. de Lusignan, N. Hague, J. van Vlymen, and P. Kumarapeli, "Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research," *Prim. Care. Inform.*, vol. 14, no. 1, pp. 59–66, 2006.

[3] L. Bos and B. Blobel, Eds., *Medical and Care Compunetics 4*, vol. 127 in Studies in Health Technology and Informatics series. Amsterdam: IOS Press, pp. 311–315, 2007.

[4] B. G. Druss and S. C. Marcus, "Growth and decentralization of the medical literature: implications for evidence-based medicine," *J. Med. Libr. Assoc.*, vol. 93, no. 4, pp. 499–501, Oct. 2005, PMID: PMC1250328.

[5] A. Mina, R. Ramlogan, G. Tampubolon, and J. Metcalfe, "Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge," *Res. Policy*, vol. 36, no. 5, pp. 789–806, 2007, doi: 10.1016/j.respol.2006.12.007.

[6] M. Gerhardsson de Verdier, "The Big Three Concept - A Way to Tackle the Health Care Crisis?" *Proc. Am. Thorac. Soc.*, vol. 5, pp. 800–805, 2008.

[7] M. Fortin, J. Dionne, G. Pinho, J. Gignac, J. Almirall, and L. Lapointe, "Randomized controlled trials: do they have external validity for patients with multiple comorbidities?" *Ann. Fam. Med.*, vol. 4, no. 2, pp. 104–108, Mar.–Apr. 2006, doi: 10.1370/afm.516.

[8] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] J. Winn and C. Bishop, "Variational Message Passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, 2005.

[10] T. Minka, J. Winn, J. Guiver, and A. Kannan, Infer.NET, Microsoft Research Cambridge, http://research.microsoft.com/infernet.