EDITORIAL

BIG DATA AND PREDICTIVE ANALYTICS IN HEALTH CARE

Vasant Dhar
Editor-in-Chief

Abstract

Predictive analytics show great promise in health care but face some serious hurdles for widespread adoption. I discuss the state of the art of predictive health-care analytics using the clinical arena as an example and discuss how the outputs of predictive systems could be made actionable through differentiated processes that encourage prevention. Such systems have the potential to minimize health risk at the population and individual levels through more personalized health-care delivery.

The terms "predictive analytics" and "health-care analytics" in Google Trends show an impressive growth in interest since 2011. But what is the status of predictive analytics in health care at the current time? To what extent are predictive analytics actually being used in health-care practice? For several years now, there have been dire warnings of health-care spending exceeding 20% of gross domestic product (GDP) by the year 2025.[1] While much of the increase in cost is attributed to medical technology which saves, extends, or improves lives, the rise in costs of health care is nevertheless alarming. Big data and analytics can provide an important part of the solution for curbing rising costs and improving health care. There's hope for change. But what will it take to make it happen?

To put things in perspective, healthcare technology saw an intensive period of research during the 70s and 80s in the area of expert systems such as Internist/Caduceus[2] and Mycin.[3] Internist demonstrated impressive performance in terms of differential diagnosis accuracy across the field of internal medicine, but didn't make it into widespread use. The underlying technological infrastructure at the time didn't allow such sophisticated systems to merge with the practices and workflow of physicians. The systems were also handcrafted, making them brittle despite attempts to separate the representation of domain knowledge and data from the generic reasoning machinery. Incentives for use were few. If anything, there were concerns about malpractice.

We are now in a data-rich environment that enables new possibilities for health care in the clinical arena as well as cost cutting and operational efficiency. The costs of genomic sequencing have dropped dramatically, and continue to fall, aided by advances in data engineering driven by cloud computing. It is becoming increasingly common to do sequencing on cases involving various cancers including melanoma,[4] and new discoveries are being made on the basis of such data, for example, genomic tumor assessment revealing DNA alterations that are driving cancer growth and in identifying appropriate treatment options.[5] Genomic studies have also found new associations, such as that between the gene *LRRK2*, mutations of which are believed to be associated with Parkinson's disease and Crohn's disease, raising the possibility that the two diseases share common pathways.[6]

Big data is now the grist for knowledge creation. Instead of handcrafted knowledge bases for diagnosis and prediction, we have lots of data at the individual level from health-care system use, clinical trials, real-time monitoring, and various other sources.[7] Machine learning algorithms can discover useful patterns for prediction and explanation as well as cost reduction. Indeed, this area has been a hotbed of activity over the last few years.[7–10] It is entirely possible that the new knowledge bases discovered through data will lead to a renaissance in knowledge-based systems through the identification of causal relationships that can be used to "reason" about problems in the spirit of expert systems such as Internist.

In the remainder of this commentary, I use an example from the clinical arena to illuminate some of the key challenges that must be addressed in order for predictive analytics to be used in health care on a widespread basis. The basic reasoning and calculus applies to a range of problems across health care, where predictive systems are used to take or design actions.

A core challenge in making predictive systems operational arises from the fact that predictions are often wrong, and the costs of being wrong and the benefits of being right are difficult to quantify. A standard measure of performance of binary classification problems is the area under the receiver operating characteristic (ROC) curve (AUC) as shown in Figure 1, which graphs the true positive rate (TPR) versus the false positive rate (FPR). If a system predicts no better than random, its performance is illustrated by the diagonal line and the AUC is 0.5, exactly half of the total area, whereas a perfect predictor would have an AUC of 1. In this example, the AUC is roughly 0.8, which is illustrative of the performance of current day machine learning–based systems for predicting diabetes at least 6 months in advance.*

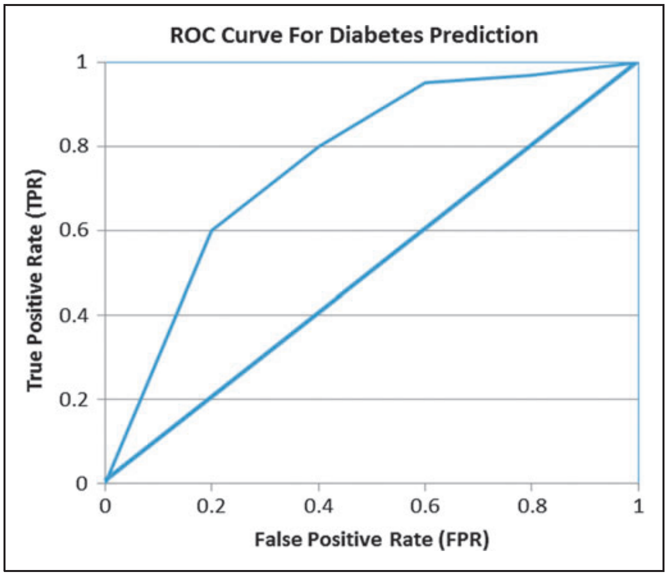

FIG. 1. Receiver operating characteristic (ROC) curve for diabetes prediction.

One can calibrate a predictive model to operate at any point of the curve: getting higher true positives entails tolerating higher false positives. In the two "confusion matrices" below, A+ and A− correspond to the number of actual positive and negative cases and +P and −P denote positive and negative predictions from a model. The matrices show true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) corresponding to two such points on the curve, namely 0.2 and 0.4 on the FPR axis.

TABLE 1.

	+A	*−A*	*Totals*
+P	**80 (TP)**	**200 (FP)**	280
−P	**20 (FN)**	**300 (TN)**	320
Totals	100 (p+)	500 (p−)	600

TABLE 2.

	+A	*−A*	*Totals*
+P	**60 (TP)**	**100 (FP)**	160
−P	**40 (FN)**	**400 (TN)**	440
Totals	100 (p+)	500 (p−)	600

In the first matrix, $TPR = TP/(TP + FN) = 80/(80 + 20) = 0.8$. $FPR = FP/(FP + TN) = 200/(200 + 300) = 0.4$. The false negative rate $FNR = 20/(20 + 80) = 0.2$ and $TNR = 300/(300 + 200) = 0.6$.

In the second matrix, $TPR = TP/(TP + FN) = 60/(60 + 40) = 0.6$. $FPR = FP/(FP + TN) = 100/(100 + 400) = 0.2$. The false negative rate $FNR = 40/(40 + 60) = 0.4$ and $TNR = 400/(100 + 400) = 0.8$. In general, $FNR = (1 - TPR)$ and $TNR = (1 - FPR)$. Another useful metric, $precision = TP/(TP + FP)$, measures what proportion of the cases predicted as positive are actually positive.

The first predictor is more aggressive in getting more of the positives correct, but at the expense of more false positives. Which is better? It depends on the costs of being wrong, **C**(−P,+A) and **C**(+P,−A) and the benefits of being right, **B**(+P,+A) and **B**(−P,−A). For diabetes, **C**(−P,+A) would be the cost associated with a missed diagnosis, where serious cases could be amputation and loss of vision. Costs associated with predicting diabetes incorrectly would be **C**(+P,−A), which might include follow-up tests such as A1C, clinician time, or unnecessary medication such as Metformin. The equation below shows the relationship:[11]

$$\textbf{Expected Value} = p(+)^{*}(TPR^{*}\mathbf{B}(+P, +A) + FNR^{*}\mathbf{C}(-P, +A)) + p(-)^{*}(TNR^{*}\mathbf{B}(-P, -A) + FPR^{*}\mathbf{C}(+P, -A)), \text{ or}$$

$$\textbf{Expected Value} = p(+)^{*}(TPR^{*}\mathbf{B}(+P, +A) + (1 - TPR)^{*}\mathbf{C}(-P, +A)) + p(-)^{*}((1 - FPR)^{*}\mathbf{B}(-P, -A) + FPR^{*}\mathbf{C}(+P, -A))$$

Determining **C**(−P,+A) and **C**(+P,−A) is particularly difficult. How much is that lost eyesight worth? And what is the impact of the wrong or unnecessary treatment? Furthermore, if we regard these costs as infinite, we will clearly go bankrupt.

*Personal communication with David Sontag at New York University (NYU). Results discussed here are illustrative only (and slightly conservative), based on research collaboration between NYU and Independence Blue Cross.

The reality, despite the difficulty of assigning costs, is that the false positives of current-day systems are still somewhat high relative to the true positives for such systems to be used unconditionally as a basis for action at the population level. In the more aggressive classifier in Table 1 above, for example, we have a "precision" of 80/280 or 28.57%, whereas the more conservative one in Table 2 has a precision of 60/160 or 37.5%, with the majority of predictions still incorrect. Applied unconditionally, many people would be mistreated[12] while the proportion of false negatives—people suffering adverse consequences that should have been avoided—is still quite high.

In general, when the "base rates" of a disease are low, that is, its prevalence across the population is low, and the costs of being wrong are high, a predictive system faces a high bar before it becomes actionable, even though the system performs significantly better than random. One way toward actionability from predictive analytics is to differentiate the overall population into finer segments according to the risk levels predicted by the model and tune the outreach at the individual level according to the risk level. For lower risk segments, for example, we could conduct lighter outreach actions such as promoting awareness and caution in the population predicted to be at risk, realizing that a majority of these cases will be false positives. This makes sense: do no harm, but inform and sensitize segments of the population that are at risk. For individuals predicted to be at higher risk, we might consider a more aggressive outreach and testing, including tracking compliance of people on medication, especially those exhibiting comorbidity. Information technologies are making it easier to implement these various levels of touch and can funnel more fine-grained information at the individual level into the predictive system.

"WE SHOULD EXPECT MORE AND FINER GRAINED DATA ON INDIVIDUALS TO LEAD TO MORE ACTIONABLE MODELS AT THE INDIVIDUAL LEVEL."

We should expect more and finer grained data on individuals to lead to more actionable models at the individual level. At the current time, we have relatively sparse data on individuals in the health-care systems, for example, an average of a dozen or so "transactions" per year for pre-diabetics and even fewer for normal people. For slow progressing diseases such as diabetes, this makes discrimination harder. If models become better through more granular data gathered via devices such as mobile devices, recorders, and telemedicine, actionability can become progressively more aggressive, targeted, and personalized.*

Improved monitoring might also go a long way toward understanding and addressing other behavioral issues related to habits. People don't change habits easily. We are unable to alter personal behaviors easily due to a variety of factors ranging from time availability, money, and established routines. There is a considerable amount of attention on using technologies and incentives to alter behavior, but this remains a serious impediment to improving health-care outcomes.

The holy grail, of course, is prevention rather than cure. Better predictive systems and monitoring will get us part of the way, but the health-care system must also have incentives that lead providers to put a high value on prevention. In the old days of "high touch" care a physician knew you and your family personally, thereby injecting a social element and "caring" into the process. In contrast, industrialized health-care of today is largely "process driven" and less personal, intended to be delivered in a scalable manner. The inevitable consequence has been little incentive for the provider to be proactive and really "care" with prevention in mind instead of treatment. A major challenge facing health care is one of aligning economic incentives of health-care providers toward prevention. Steps are being taken in this direction in the Affordable Care Act, where certain payments are tied to concrete markers such as lowering readmission rates and other proxies for preventive care. Coupling such policy actions with better technologies that include personalized predictions and recommendations could go a long way toward achieving better outcomes in health care. The challenge will be to accomplish this without sacrificing privacy to the detriment of the individual.

*Independence Blue Cross has communicated having success in terms of engagement and intervention with individuals at risk for congestive heart failure (private communication with Somesh Nigam, Chief Informatics Officer at IBC).

References

1. Congressional Budget Office. The long-term outlook for health care spending, introduction and summary, November 2007. Available online at www.cbo.gov/ftpdocs/87xx/doc8758/MainText.3.1.shtml
2. Pople HE, Jr. Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. In Szolovits, P., ed. Artificial Intelligence in Medicine. Westview Press, Boulder, Colorado. 1982.
3. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Math Biosci. 1975; 23:351–379.
4. Pleasance et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010; 463:191–196.
5. Berger M, et al. The genomic complexity of primary human prostate cancer. Nature 2011; 470:214–220.
6. http://en.wikipedia.org/wiki/LRRK2
7. http://mucmd.org/conference-information.php

8. Wang X, et al. Unsupervised learning of disease progression models. 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 2014.
9. Maguire J, and Dhar V. Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: Data-driven predictive analytics in healthcare. Health Systems 2013; 2:73–92.
10. http://healthaffairs.org/blog/2014/07/08/new-health-affairs-july-issue-the-impact-of-big-data-on-health-care/
11. Provost F, Fawcett T. Data science for business. New York: O'Reilly Media. 2013.
12. Ioannidis J, Goodman S, Greenland S. Why most published research findings are false: Problems in the analysis. PLoS Medicine 2007; 4:e168.

Vasant Dhar
Editor-in-Chief
Big Data