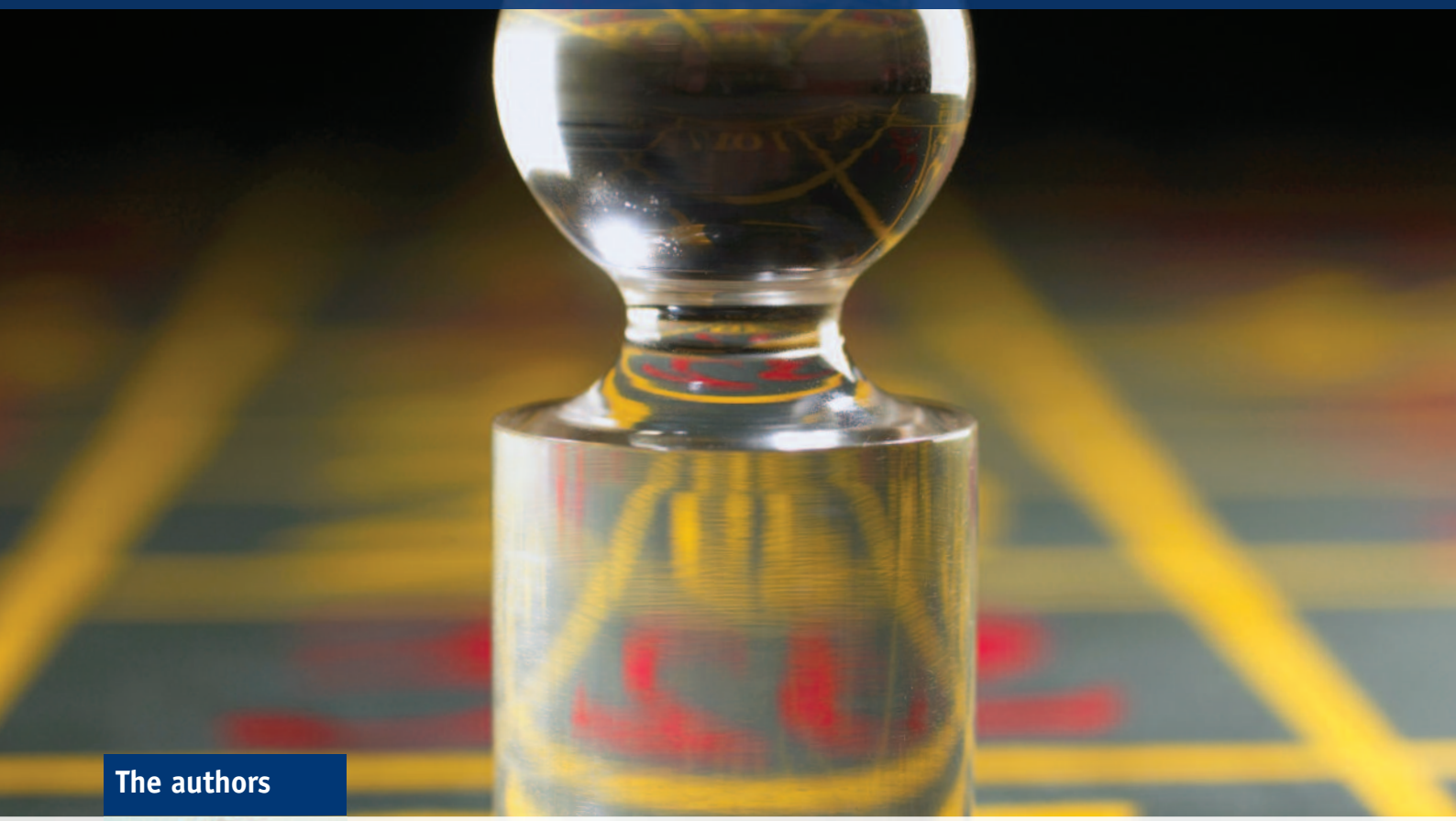
Predictive analytics promises to transform whole swathes of the healthcare sector, from early identification of at-risk individuals to reducing non-adherence through highly targeted, personalized interventions. But what exactly is predictive analytics? Where will it see the greatest application? And how does it differ from conventional statistical analysis?



The authors

John Rigg, PHD, MPhil, BA
is Senior Manager Advanced Analytics, RWE Solutions, IMS Health
John.rigg@uk.imshealth.com

Ben Hughes, PHD, MBA, MRES, MSC
is Senior Principal RWE Solutions, IMS Health
Bhughes@uk.imshealth.com

The promise of predictive analytics

A new crystal ball for healthcare?

The digitization of patient, hospital, prescription, biological and other vast data streams heralds a new era in the volume and complexity of healthcare information. As the data landscape continues to evolve so, too, must the tools employed to interrogate the data if its true potential is to be realized. Cue predictive analytics.

Predictive analytics is concerned with future or otherwise unknown events. Techniques in this field have progressed significantly in recent years, not least due to advances in computer processing, storage and retrieval. Increasingly, they are now being applied to address problems in healthcare.

In 2011, the Heritage Provider Network offered a grand prize of $3 million in an open competition for the best algorithm at estimating the number of days a patient would be admitted to a hospital within the next year, using historical claims data. The competition was intense, with advanced computing algorithms developed by data scientists, mathematicians, computer scientists, hedge fund managers and software engineers pitted against each other from 40 countries around the globe. Over 39,000 entries later, the prize was awarded in March 2013.

The aim of the sponsor, as articulated at the competition outset, captures the hopes of many for predictive analytics: "Once [the algorithm is] known, health care providers can develop new care plans and strategies to reach patients before emergencies occur, thereby reducing the number of unnecessary hospitalizations. This will result in increasing the health of patients while decreasing the cost of care. In short, a winning solution will change health care delivery as we know it – from an emphasis on caring for the individual after they get sick, to a true health care system."

The leading teams all used sophisticated analytical techniques known as machine learning algorithms. Machine learning, a branch of artificial intelligence, specializes in developing algorithms that are highly effective at identifying often subtle or hidden patterns in large volumes of disparate data. These algorithms can be remarkably accurate at predicting outcomes on new or unseen data. Generalizations or predictions are often made from traditional statistical analysis based on observational or retrospective data. However, predictive analytics, incorporating machine learning algorithms, can perform far better.

BROAD APPLICATION

It is not just in the analysis of hospital admissions where predictive analytics is helping to drive healthcare solutions. Although its uptake is still embryonic, predictive analytics is expanding across the spectrum of healthcare with applications such as: identification of patient risk factors; outcomes prediction; decision support; diagnosis; evaluation of treatment and pathways; clinical trial simulation; resource allocation; adherence; safety; product uptake; bill optimization; and fraud detection.

For example, there is a rapidly growing evidence base demonstrating the value of predictive analytics in diagnosis decision support for physicians. Waljee, et al, for instance, reported in 2010 how a machine learning algorithm substantially out-performed standard metabolite tests in predicting the clinical response of patients with inflammatory bowel disease on thiopurines.[1] This approach provides the potential for a low-cost, rapid alternative to metabolite measurements for monitoring thiopurine use.

continued on next page

Another example is the EuResist project, a 'real-life' predictive analytics solution. EuResist is an international collaboration designed to improve the treatment of HIV patients. A web interface allows physicians to specify patients' clinical and genomic data. The data is sent to the prediction engines and the combined response, displayed to the physician, includes various suggested treatments and a prediction of their effect on the amount of HIV in the blood. In 2009, the EuResist project was named as a Computerworld honors program laureate. This emphasizes the importance of developing predictive analytics solutions as part of a real-time system.

An alternative approach is highlighted in a recent paper by Agneeswaran, et al, which shows how real-time machine learning may be applied to aid physicians' interpretation of electrocardiogram (ECG) reports for arrhythmia detection.[2] The physician submits the ECG reading on-line to a cloud-based machine learning algorithm which then supports their decision making by indicating which arrhythmia classification is most likely to apply.

PREDICTING THE IMPACT OF PREDICTIVE ANALYTICS FOR LIFE SCIENCES

In a rapidly changing world, predicting the sectors where predictive analytics is likely to have greatest impact is itself a challenge perhaps best suited for a machine learning algorithm. Important ingredients in shaping its future will undoubtedly involve the regulatory environment, market conditions and the availability of high quality, integrated data. However, commercial incentives are likely to be the most telling drivers, particularly in areas where predictive analytics can make the greatest business impact.

Successful applications by life science could either play pivotal roles in supporting external key stakeholders or internal decision-making processes.

External application
External application will work both on an aggregate-level, such as forecasting and aiding healthcare systems and institutions strategy, and at a micro-level, including specific choice in care management. Prediction at the micro-level is solutions tailored to the requirements of individual patients. For example, various pilots in risk prediction and care management have demonstrated ~10-15% reduction in severe patient events. Other areas likely to see significant demand for highly accurate prediction algorithms include diagnosis, non-adherence and resource utilization. For life sciences in particular, a highly accurate prediction of adherence would allow stratified interventions in adherence or patient programs to significantly improve effectiveness and return on investment. However, capturing value will hinge on delivery infrastructure and engagement with health systems.

Internal application
Creating technology systems and necessary changes in business culture to action predictive analytics solutions in a timely, integrated and cost-effective manner may pose a greater challenge in many settings than creating the predictive algorithms themselves. The second area where much greater accuracy is required for internal decision making at a more aggregate level includes improving health economic and outcome models and their risk equations, or simulation for major decisions such as trial design. This is in addition to increasing the use of simulation in early discovery, where various mathematical and simulation techniques are used.

Healthcare also needs to expand its skills profile to enable effective predictive analytics solutions. Data scientists, computer scientists and mathematicians all have skill sets likely to face increasing demand from within this sector.

CHALLENGES

For all the promise predictive analytics offers, it, too, faces challenges. Perhaps foremost is the lack of transparency of many machine learning algorithms, referred to as 'black-box'. Creative visual representations and improved model diagnostics are helping end-users interpret output from machine learning algorithms. Nonetheless, opacity may to some extent always be the price that must be paid for a more powerful predictive solution.

PREDICTIVE ANALYTICS VERSUS TRADITIONAL STATISTICAL ANALYSIS

In 2001, the renowned Berkley statistician and predictive analytics advocate Leo Breiman argued: "The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems...If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools." [3]

Whilst not everyone would agree with this extreme stance, many would embrace a more empirically-driven focus to help address some challenges. So, how does predictive analytics differ from traditional statistical analysis using observational data? The following discussion may be of particular interest to those with direct experience of statistical research. Disparities between the two approaches may at times reflect not so much a paradigm shift but different points on the same continuum. Nonetheless, as shown in Figure 1, several differences characterize predictive analytics compared to traditional statistical analysis:

- A research objective designed to maximize prediction accuracy rather than draw inferences about associations between variables
- A scientific philosophy that embraces a data-driven inductive approach, rather than a hypothesis-driven deductive approach
- A choice of statistical model from a class of machine learning algorithms
- An analytical framework that uses separate samples for model development and evaluation/validation
- Evaluation metrics criteria focused on predictive accuracy of the model (eg, AUC - Area Under receiver Curve)

1. Research objective

In predictive analytics, the overarching objective is to maximize predictive accuracy. By contrast, variable 'importance' – the ability to draw inferences about the magnitude or significance of a particular variable (or variables) in relation to an outcome – is typically the primary objective in standard statistical analysis.

FIGURE 1: KEY DISTINGUISHING FEATURES OF PREDICTIVE ANALYTICS VS TRADITIONAL STATISTICAL ANALYSIS

	Predictive analytics	Traditional statistical analysis
Research objective	• Maximize predictive accuracy	• Draw inferences on variable associations
Scientific philosophy	• Data-driven, inductive approach	• Hypotheses-driven, deductive approach
Predominant modeling techniques	• Random forests, neural networks and model blending	• Regression, logistic regression, (sometimes Bayesian statistics)
Analytical framework	• Separate samples for model development and evaluation	• Single sample for model development and evaluation
Evaluation metrics	• AUC/model-level prediction accuracy	• R square, variable-level test statistics (p value, t-test)

continued on next page

In order for predictive analytics to live up to its promise, there must be willingness to embrace a methodological shift in the approach to statistical analysis.

The approaches converge somewhat in that predictive analytics often attempts to quantify variable importance (as measured by the contribution of variables to the predictive power of a model). Moreover, traditional statistical analyses frequently make generalizations – predictions. Nonetheless, differences in the primary research objective are a key distinguishing feature.

2. Scientific philosophy

Conventional statistical analysis generally starts with a clear idea about variable associations of interest. It is a hypothesis-driven approach; a deductive method of scientific enquiry. Predictive analytics, on the other hand, usually involves fewer prior expectations about variable inclusion and associations. Algorithms are allowed a 'free-reign' to 'mine' the data and find meaningful correlations. It is an inductive scientific philosophy. One notable commonality between the approaches is the data exploration phase which takes place in many traditional statistical analyses.

3. Modeling techniques

Unlike regular statistical analysis, predictive analytics routinely use so-called machine learning algorithms. There is a diverse array of algorithms, some of the most popular of which are based on decision trees – schematic tree-shaped diagrams showing how different combinations of variables (the branches of the trees) are associated with different values of the target variable. Hundreds or thousands of trees are usually created, each on a subset of variables and/or data, to form an ensemble decision tree model. Even if each tree is a poor predictor, combining the results from many trees can still produce a highly accurate model.

Random Forest is a widely adopted ensemble decision-tree approach. A random subsample of observations is used to grow each tree and a random selection of input variables are used to determine each node (the combination of input variables that denote a fork in a branch or a branch end-point).

AdaBoost (Adaptive Boosting), is another popular ensemble decision-tree algorithm. Each tree places a higher weight on observations misclassified by previous trees, thereby placing extra emphasis on predicting 'hard' observations. The final model is a weighted sum across all the trees.

Another strand of machine learning is inspired from advances in the mapping of biological neural networks. These artificial neural networks can provide powerful solutions for data characterized by highly nonlinear, interdependent associations.

Some of the most accurate predictive modeling solutions involve combining output from separate learning algorithms to produce a 'blended' model. These aim to capture the best features from different algorithms. This practice is in sharp contrast to the use of a single model in standard statistical analysis.

The label 'machine learning algorithm' is potentially confusing. Logistic Regression, possibly the most widely used modeling technique in regular statistical analyses, is itself a powerful learning algorithm in that parameter precision is increased with more data. Thus, machine learning tends to refer more loosely to 'non-standard' learning algorithms such as those mentioned above.

4. Analytical framework

A defining feature of predictive analytics is the use of a separate sample for model development and model validation/testing. This analytical framework helps overcome a phenomenon known as 'overfitting', where a model may describe accurately the data it is estimated (or trained) on, but has poor predictive accuracy on new or unseen data.

For instance, the analytical framework for predictive analytics may involve splitting the sample into three: a training, validation and test sample. A series of models is estimated on the training sample, including measures designed to reduce overfitting (a process known as regularization). Predictive accuracy of each model is assessed using the validation data and the preferred model identified. Finally, the test sample is used to assess the generalization/predictive properties of the chosen model. The test sample is necessary since a reliance on the validation data would violate the principle that data used in the final assessment of the model should not feature in model selection/training.

Not only does this analytical framework help minimize overfitting, it is a key component in producing good predictive solutions. Generalizations are often made from the results of traditional statistical analyses. Given the biases associated with overfitting, such conclusions stand to benefit from adopting a similar analytical framework.

5. Evaluation

Evaluation in predictive analytics is focused at the level of the model, rather than the typical variable-level focus in standard statistical analyses. Moreover, model performance is judged using metrics for prediction accuracy based on validation or test data (not the data the model was trained on).

A common metric in binary classification is the AUC. The curve, the Receiver Operating Characteristic, is a plot of the proportion of true positives classified by the model (True Positive Rate) on the vertical axis by the proportion of false positives classified by the model (False Positive Rate) on the horizontal axis, for different levels of the predicted probability. It provides a measure of the discriminatory power of the model. The AUC neatly ties in with diagnostic decision making since TPR is equivalent to benefits and FPR equivalent to costs. Thus, information about costs and benefits can be explicitly built into the evaluation criteria and used to select the optimal model.

FULFILLING THE PROMISE

Healthcare decision makers stand to benefit from application of predictive analytics solutions, from care management to trial optimizations. These are powerful solutions that can make best use of the growing volume of rich data across the healthcare sector. However, in order for predictive analytics to live up to its promise, there must be willingness to embrace a methodological shift in the approach to statistical analysis. Moreover, success will require applying the techniques to the right business problems, where specific choices can be made as a result of the prediction, and business processes are in place to execute those choices. While there is no playbook for this today, it will surely follow as evidence mounts highlighting the business impact of solutions based on predictive analytics •

Healthcare decision makers stand to benefit from application of predictive analytics solutions, from care management to trial optimizations - making best use of the growing volume of rich data across the healthcare sector.

[1] Waljee AK, Joyce JC, Wang S, Axena A, Hart M, Zhu J, Higgins PDR. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical Gastroenterology and Hepatology, 2010; 8:143-50

[2] Agneeswaran VS, Mukherjee J, Gupta A, Tonpay P, Tiwari J, Agarwal N. Real-time analytics for the healthcare industry: Arrhythmia detection. Big data, Sept 2013; 1(3): 176-182. doi:10.1089/big.2013.0018.

[3] Breiman L. Statistical modeling: The two cultures. Statistical Science, 2001; 16(3): 199-231