Using predictive analytics and big data to optimize pharmaceutical outcomes

Inmaculada Hernandez, Pharm.D., Ph.D., Department of Pharmacy and Therapeutics, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA.
Yuting Zhang, Ph.D., Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA.

Address correspondence to Dr. Hernandez (inh3@pitt.edu).

This article is part of a special AJHP theme issue on population health management. Contributions to this issue were coordinated by AJHP Editorial Advisory Board member Kenneth M. Shermock, Pharm.D., Ph.D.

Purpose. The steps involved, the resources needed, and the challenges associated with applying predictive analytics in healthcare are described, with a review of successful applications of predictive analytics in implementing population health management interventions that target medication-related patient outcomes.

Summary. In healthcare, the term big data typically refers to large quantities of electronic health record, administrative claims, and clinical trial data as well as data collected from smartphone applications, wearable devices, social media, and personal genomics services; predictive analytics refers to innovative methods of analysis developed to overcome challenges associated with big data, including a variety of statistical techniques ranging from predictive modeling to machine learning to data mining. Predictive analytics using big data have been applied successfully in several areas of medication management, such as in the identification of complex patients or those at highest risk for medication noncompliance or adverse effects. Because predictive analytics can be used in predicting different outcomes, they can provide pharmacists with a better understanding of the risks for specific medication-related problems that each patient faces. This information will enable pharmacists to deliver interventions tailored to patients’ needs. In order to take full advantage of these benefits, however, clinicians will have to understand the basics of big data and predictive analytics.

Conclusion. Predictive analytics that leverage big data will become an indispensable tool for clinicians in mapping interventions and improving patient outcomes.

Keywords: big data, medication management, pharmaceutical outcomes, population health management, predictive analytics

Am J Health-Syst Pharm. 2017; 74:1494-500

The recent shift from payment based on volume to value has brought the implementation of outcome-based reimbursement models to the U.S. healthcare system. Because chronic conditions account for over 85% of total U.S. healthcare costs and 99% of Medicare spending, most indicators used to adjust outcome-based reimbursement or payments measure the quality and efficiency of care as well as the clinical outcomes of patients with chronic conditions. Interventions that target pharmaceutical use and outcomes are of special interest, because improving medication use in the context of chronic disease has been shown to reduce downstream medical costs and, therefore, total healthcare spending. In fact, the Centers for Medicare and Medicaid Services and commercial insurers use measures of appropriateness of medication use in patients with chronic disease to adjust payments to providers. For instance, in the current Medicare accountable care organization model, several quality measures are related to medication management, including the use of angiotensin-converting enzyme inhibitor or angiotensin receptor blocker therapy in coronary artery disease.
use of β-blockers in heart failure, and use of antiplatelet therapy in ischemic vascular disease.

Population health management interventions that aim to improve medication use and pharmaceutical outcomes can take different forms. For example, 2 articles in this theme issue provide recent examples of using electronic health record (EHR) and claims data in identifying patients who could benefit most from pharmacist interventions. One of those articles describes how Geisinger Health System combines EHR and claims data to identify target patients for its medication therapy disease management program, including patients who have a diagnosis of atrial fibrillation and are using an oral anticoagulant agent, patients with multiple sclerosis, and patients with familial hypercholesterolemia; the other article describes how the Greater Rochester Independent Practice Association developed an algorithm that synthesizes information from real-time claims data, inpatient and emergency room admission data, and laboratory results to identify patients who could benefit most from a care management team. Another common application of predictive analytics in optimizing pharmaceutical outcomes is to identify patients who are most likely to be nonadherent to medications. Because of the popularity of this application, a wide range of commercial predictive analytics services that predict medication nonadherence are currently available in the market. These solutions can be contracted by health providers to predict which patients are at highest risk for medication nonadherence so that specific interventions can be targeted at those patients to improve treatment cost-effectiveness.

Regardless of the form population health management initiatives take, a key component is the use of electronic patient data in the identification of target patients. Because of the increased quality and granularity of electronic patient data and increased accessibility to that information, the data used in identifying target populations for these programs are drawn from extensive information on a large sample of patients; this information is usually referred to as big data.

**Concepts and terminology**

**Definition of big data.** The term big data usually refers to extremely large and complex databases that aggregate information of different types and scales, that collect data across time and distance from multiple sources, and that often require complex data processing applications. In healthcare, big data may include EHR, administrative claims, and clinical trial data as well as data collected from smartphone applications, wearable devices, social media, and personal genomics services. A recent literature review defined healthcare big data sets as those in which the product of the sample size and the number of variables is greater than log 7. Big data are therefore characterized by 3 main characteristics: an extremely large sample size, high heterogeneity, and high dimensionality (a large number of variables available for each unit of observation).

**Challenges associated with analysis of big data.** Because big data have high dimensionality and are derived from a large sample, the analysis of big data sets is associated with some major challenges, such as noise accumulation and spurious correlation. Noise accumulation is the accumulation of estimation errors when the prediction is based on a large number of parameters, which leads to poor classification or poor prediction. In other words, when the number of predictors is too large, the addition of more predictors does not improve the predictive power of the model but only leads to accumulated noise, which deteriorates the performance of the model.

**Predictive analytics can be used in identifying patients who can benefit the most from pharmacist interventions and in predicting pharmaceutical outcomes.** Predictive analytics can provide pharmacists with a better understanding of the risks of specific medication-related problems that each patient faces, enabling the delivery of interventions tailored to the patient’s needs.

As more patient data and better predictive tools become available, the role of predictive analytics in optimizing medication use will expand.
As high-quality patient-level data become available to providers and researchers, predictive analytics will be increasingly used in healthcare.

**Big data predictive analytics in healthcare.** To illustrate the application of predictive analytics to healthcare, here we will discuss a recent study conducted by the authors that identified risk factors for cardiovascular adverse effects of antidementia medications. In that study, the research group analyzed claims data on over 160,000 Medicare beneficiaries with Alzheimer’s disease receiving antidementia drugs. Specifically, the likelihood of having a cardiovascular adverse effect was predicted using data on all medications documented during the study period and all International Classification of Diseases, Ninth Revision (ICD-9) codes and Healthcare Common Procedure Coding System codes recorded in the claims data of each patient.

**Construction of a patient-level data set**

Healthcare providers have access to rich patient-level data, such as medical and medication records, laboratory values, claims data, and physician notes. However, these data files are often stored and coded separately. The first step in applying predictive analytics in healthcare is to construct a unified and often deidentified data set. When predictive analytics are used in population health management approaches, the data set is usually constructed at the patient level. However, different levels of data consolidation may be preferred for other applications of predictive analytics.

Specifically, the construction of the data set involves the following steps: (1) cleansing and normalization of the original data sets so that the information is consistently formatted across different data sources, (2) aggregation of multiple data sources into a unified data set at the patient level, (3) deidentification of the protected information according to Health Insurance Portability and Accountability Act privacy rules, and (4) validation of the process to ensure the accuracy of the data. While integrating the different sources of data, it is often useful to keep track of the origin of the data; this is important because there is likely to be a repetition of information across different data sources and because certain algorithms are better trained in specific types of data sets. For instance, information on diagnosis or healthcare utilization is likely to be captured in both EHR and claims data. The choice of the data source to use in the development of a model will depend on the objectives of the analysis. Whereas EHR data may be more suitable when performing studies that require granular clinical information, claims data may be preferred in some applications (e.g., those that require the use of algorithms originally trained with claims data).

In the study of antidementia drugs mentioned above, pharmacy and medical claims data from a 5% random sample of Medicare beneficiaries were analyzed. The claims data consisted of a variety of files, including beneficiary files with demographic characteristics and insurance enrollment information, institutional claims, provider claims, and pharmacy claims. In putting together the analytic data set for the study, we first had to identify relevant information contained in different data files and merge it into an aggregated data set at the patient level. We further had to select which data sources to use in the case of information repeated across data files. For instance, in defining an adverse effect, we had to decide whether to define it on the basis of institutional claims or provider claims, since both contain ICD-9 codes. In this case we used provider claims, because the setting where the adverse effect was diagnosed was not relevant to the study. Then we extracted all provider claims for each patient and examined whether any of the ICD-9 codes recorded in those claims were for adverse effects. Based on this information, we defined a variable (the occurrence of adverse effects). Through the creation of this variable, we aggregated claim-level information into beneficiary-level information, which was then added to a patient-level data set containing other relevant patient characteristics.

**Preparation for processing of the input data set**

**Reduction of dimensionality.** A crucial factor in the successful application of big data predictive analytics to a specific research question is to account for both the special features of big data and the main limitations of predictive analytics, such as spurious correlations and noise accumulation. Even if predictive analytic algorithms can handle a large number of input predictors, reducing the number of input dimensions can improve the performance of models with high dimensionality. The process of dimensionality reduction is based on the assumption that the available high-dimensional data set lies on or near a lower-dimensional manifold. In other words, it can be assumed that the available high-dimensional data represent multiple and indirect measurements of an underlying source. In these scenarios, dimensionality reduction techniques can be applied to transform the original high-dimensional data set into a compact lower-dimensionality expression of that underlying source.

In the aforementioned study, we explored potential interactions between antidementia drugs and concurrent medications for comorbid conditions whose use could increase the risk of adverse effects. Because we hypothesized that interactions would be similar within therapeutic classes, we aggregated medication-use information at the therapeutic class level. Doing this substantially reduced the number of variables needed in the model for each patient. In the original claims data set, each medication was recorded according to active ingredient instead of therapeutic class. If we had created a variable for each...
Figure 1. Steps in the development of big data–driven predictive analytics in the implementation of population health management interventions. First, multiple sources of data are aggregated into a unified data set at the patient level. Second, the observations and variables of interest for the specific study question are selected. Third, techniques to reduce data dimensionality are applied. Fourth, the data set is randomly split into training, validation, and test sets. Fifth, the training set is used for variable selection and model building. Sixth, the validation set is used to calculate performance measures for all models and finalize model selection. Finally, the test set is used to calculate the performance of the selected model. Blue boxes represent data sets, red boxes represent processes, yellow stars represent the models constructed, and the purple star represents the model selected. EHR = electronic health record.
active ingredient, we would have had thousands of variables. By aggregating active ingredients into therapeutic classes, the number of variables that represented medication use was decreased from thousands to 242. This is an example of the application of a dimensionality reduction technique called aggregation of variables to a data set in which the available information (active ingredients) represented multiple measures of an underlying lower-dimensional source of information (therapeutic classes).

After dimensionality reduction techniques are applied, there may still be a large number of predictors. In such cases, researchers may decide to eliminate covariates that are unlikely to carry useful information, such as those with little variation or low prevalence. These techniques are often considered as a special type of dimensionality reduction technique because they achieve dimensionality reduction by creating a subset of the original number of variables, as opposed to the previously described techniques, such as aggregation of variables, which do not establish subsets of variables but create new variables that compact the original information.

In the antidiabetes drug study, after using the aggregation of variables technique to reduce the number of covariates, we still had a large number of covariates (3,077 in this case). To decrease this number, we excluded covariates with a prevalence lower than 1%. In other words, we excluded variables for which 99% of the study participants had the same value. The rationale behind this process is that variables that show little variation are unlikely to carry important information.

**Splitting the data set.** One of the biggest limitations of predictive analytics models is their tendency toward overfitting. Overfitting occurs when a model is excessively complex and describes random error and noise rather than the underlying relationship. The overfitted model represents the data rather than predicting it. A linear model predicts better when applied to other, similar data sets. In order to mitigate the overfitting problem, the original data set can be randomly split into a training set, a validation set, and a testing set prior to model building. The training set is used for model building; the validation set is used for model selection, and the test set is used to assess the performance of the model selected.

**Model building, validation, and selection.** As with conventional statistical methods, the choice of predictive analytics model depends on the type of outcome variable (e.g., categorical versus continuous) and the research questions. Models commonly used in big data analytics include artificial neuronal networks, support vector machines, discriminant analysis, and classification trees. A discussion of the specifics of each of these algorithms is out of the scope of this article. Even though these sophisticated models have been shown to have higher predictive power than conventional regression models, they provide little insight into the relative influence of each parameter or the effect size. For this reason, machine learning models are often referred to as “black box” approaches, which means that they produce predictions but do not provide an understanding of how they do so. In other words, they give an estimate of the outcome for each observation but do not specify the function used in those predictions and do not provide information on the effect of each covariate. This limitation is particularly relevant in healthcare, because clinicians usually require an explanation of the underlying reasons for model predictions. Because of this limitation, a conventional regression model may still be the best model when estimates of the impact of each predictor are needed. For this reason, in the previously mentioned antidiabetes study, we conducted a logistic regression to evaluate how different predictors affected the likelihood of having an adverse effect. The output of logistic regression included odds ratios, which provided information on whether a variable increases or decreases the odds of experiencing an adverse effect and by how much. If we had used other, more advanced techniques such as neural networks or support vector machines, we would not have been able to quantify the impact on the risk of having an adverse effect of each predictor.

After the application of the dimensionality reduction techniques to a given data set, there may still be a large number of covariates. For instance, in the study of reference, there were still 650 variables after the application of the dimensionality reduction techniques mentioned. In such cases, it is necessary to further decrease the number of variables before the construction of multivariable regression models. One way to do this is through univariate selection. This mechanism involves the construction of univariate regressions that measure the strength of the association between each of the remaining covariates and the outcome. Then, only the variables that are strongly associated with the outcome (those associated with a p value below a prespecified threshold) are selected for inclusion in the multivariable regression. In the study of reference, we conducted 650 univariate regressions to assess the association between each of the covariates and the outcome. We ranked them by p value for univariate association and selected the top 2.5% of variables to be included in multivariable models. In the development of multivariable models in predictive analytics, it is common to build several models using different algorithms or different parameters and compare their performance. For instance, investigators can develop a logistic regression model, a neural network, and a support vector machine model using the training set and then select the best-performing model using the validation set. Commonly calculated performance measures include accuracy; area under the curve; specificity; sensitivity; positive and negative...
predictive power for binary outcomes; and mean absolute error, mean squared error, and median absolute error for continuous outcomes. Finally, the test set is used to calculate the performance of the selected model, which predicts the performance of the model if applied to other data sets. In the antidiabetes case study, we developed 4 logistic regression models with varying parameters using the test set and then selected the top-performing model using the validation set.

**Required resources**

**Human resources.** Different areas of expertise are needed to apply big data in predictive analytics. First, a data analyst with expertise in using different sources of data (e.g., EHR, claims data, genomics data, data from wearable devices or social media) is needed to manage and create analytic data. Second, a pharmacoepidemiologist or a pharmaceutical outcomes researcher is needed to develop study design, define predictors and outcome variables, and interpret results. Third, depending on the types of models to be developed, a computer scientist with some capability in programming languages and advanced predictive analytics may be needed to conduct advanced predictive modeling. Finally, a system administrator may be needed to create and maintain the infrastructure for data storage and analyses.

**Technology requirements.** The analysis and maintenance of a big data system require a complex infrastructure, including powerful computing machines, large storage space, and security protection. Even if storage is relatively inexpensive, the maintenance of this kind of infrastructure can cost thousands of dollars per year, especially with the required support of a system administration staff. Instead of building their own systems, many organizations, especially small ones or those with specific needs, rent pre-built storage and computing resources at cloud-based data warehouses from online providers.

**Discussion**

The use of predictive analytics in optimizing medication will likely transform current clinical practice. Specifically, these techniques will direct pharmacists’ interventions and other healthcare resources toward patients whose clinical outcomes can be improved the most as a consequence of such interventions. Furthermore, because predictive analytics can be used in predicting different outcomes, pharmacists will have a better understanding of the risks of specific medication-related problems that each patient faces. This information will enable pharmacists to deliver tailored interventions that target the medication-related problems that each patient is most likely to experience. For example, in a health system where predictions for medication nonadherence, drug interactions, or adverse effects are available for each patient, pharmacists would be able to identify which of these problems are most relevant to a certain patient and focus their time on providing recommendations to prevent them. In order to take full advantage of these benefits, clinicians will have to embrace these predictions in their current clinical practice, which will likely require some understanding of predictive analytics. Therefore, it is important to include the basics of big data and predictive analytics in the training of healthcare providers.

In addition to the need to train clinicians in this discipline, the application of predictive analytics in a health system faces 3 other major limitations. First, even if predictive techniques are developed to overcome the limitations associated with the use of conventional statistics in large databases, their results are still subject to a high chance of false scientific discoveries; this is because when a large number of variables are evaluated, some variables can be highly correlated with others by chance. Because of the high chance of false discovery associated with the use of large databases, and because of the methodological complexity of these methods, healthcare applications that employ predictive analytics often start as research projects. Once the methodological and interpretational challenges associated with the development of predictive analytics algorithms have been overcome, and once there is strong evidence that findings do not represent spurious associations, applications based on predictive analytics are deployed in clinical practice.

Second, the storage and analysis of large quantities of patient data require the development and maintenance of a complex and secure infrastructure with high computing power, as well as experts in multiple domains, including data management, computer science, epidemiology or outcomes research, and systems administration. Finally, the conventional procedures commonly used to assure privacy protection may not be sufficient to prevent patient identification, because the combination of high data granularity and extensive information on each observation can enable the reidentification of individuals.

As more patient data and better predictive tools become available, the role of predictive analytics in the optimization of healthcare resources will expand. The early adopters of these techniques will gain a competitive advantage among their peers.

**Conclusion**

Predictive analytics that leverage big data will become an indispensable tool for clinicians in mapping interventions and improving patient outcomes.

**Disclosures**

The authors have declared no potential conflicts of interest.

**References**

2. Hadi MA, Alldred DP, Briggs M et al. Effectiveness of pharmacist-led medication review in chronic pain


