White Paper

Harnessing the Power of Predictive Analytics for Population Health

Charles Engle, Ph.D., PMP

Predictive models as essential tools for effective case management, reduced costs and improved outcomes

MEDai
A LexisNexis Company

Health Care

Emerging Health Care Trends and the Effect on Payers

Population health is a term that has become in vogue since the passage of the Patient Protection and Affordable Care Act[1]. However, a precise, universally-accepted definition of the term has not yet been provided. For the purposes of this whitepaper, we will use the definition of Kindig and Stoddart given below.

Dr. David Kindig and Dr. Greg Stoddart, in their article in the **American Journal of Public Health** entitled "What is Population Health?" propose that **population health** as a concept of health be defined as "the health outcomes of a group of individuals, including the distribution of such outcomes within the group." These populations are often geographic regions, such as nations or communities, but they can also be other groups, such as employees, ethnic groups, disabled persons, or prisoners. Such populations are of relevance to policymakers. In addition, many determinants of health, such as medical care systems, the social environment, and the physical environment, have their biological impact on individuals in part at a population level[2].

"Predictive analytics encompasses a variety of statistical techniques from modeling, data mining and game theory that analyze current and historical facts to make predictions about future events."[3] Predictive analytics is an idea and a technology that has its roots in history, but has only been viable as a commercial concept since the advent of computers that allowed very large data repositories to be exploited. Thus, this powerful approach to forecasting risk is only about a human generation old, yet it has already revolutionized several business domains, including the nascent field of population health.

In this review, predictive analytics will be presented from a historical perspective to define the concept and place it in context for population health. Next, the various solutions that are offered to implement predictive analytics and extract the power of the technique will be discussed. Finally, projections of the future of predictive analytics as it pertains to population health will be discussed with a view toward stimulating further research directions and business opportunities.

"Predictive analytics encompasses a variety of statistical techniques from modeling, data mining and game theory that analyze current and historical facts to make predictions about future events." [3]

Overview of Methodologies

Historically, the first approach to predictive analytics focused on forecasting risk at an aggregate level. This is little more than the application of well-understood statistical techniques to health data. This technique is noted for its measure of central tendency; the aggregate analysis results in a "norming" effect and reduces the influence of outliers. The larger the population of the aggregate, the more accurate the prediction is likely to be. The disadvantage is that this "average" result is not specific for any single individual and the predictions apply to aggregate groupings.

The next approach focused on forecasting risk at an individual level. This approach is significantly more complex, but yields improved results. One statistical approach that assists in the prediction of results at the individual level is a technique called clustering, which affords sufficient discrimination to determine individual results due to the admission of more data values or variables called predictors, grouped according to some chosen clustering scheme. Given the use of computers, the additional complexity is manageable and the resultant predictions are more specific for any given individual.

Can these approaches alone make an impact on the top 2% (regression to the mean problem) and improve population health?

There was "growing evidence that prior utilization models or threshold-based models were not adequate for case finding. They often led to misallocation of resources, inefficiencies, and missed opportunities."[4] A new tool was needed to focus population health resources on those cases that would be best served by intervention. Predictive Models growing out of Predictive Analytics were considered the best approach for targeting resource allocation.

Predictive Analytics usually result in a Predictive Model which is the heart of the application of Artificial Intelligence to health care. Predictive Modeling is the process by which a clinical database is used to describe mathematically the likelihood of outcome events, given a set of values for variables representing a new patient. In short, a model is a somewhat simpler abstract representation of a real world process, often using mathematical formulae. The fact that it is somewhat simpler than the real process allows the human

One statistical approach that assists in the prediction of results at the individual level is a technique called clustering, which affords sufficient discrimination to determine individual results due to the admission of more data values or variables called predictors, grouped according to some chosen clustering scheme.

mind to comprehend its complexity in a way that is not possible when the whole process must be considered. This allows the human observer to reason about complex processes and to adjust them as necessary to maintain control. Without the ability to reduce the complex process to a model that can be observed and controlled, many processes would be difficult or impossible to maintain, especially health care outcomes which are inherently complex.[5]

Modeling is a way to represent real-world processes with abstract entities such as mathematical formulas to reduce the overall complexity of the problem and represent the problem in a form comprehensible to humans. Thus, it seeks to present the essential elements of the problem without the mundane unnecessary details. As such, manipulating the model provides an acceptable alternative for manipulating reality. While the essence of the problem is captured however, the details that add to complexity are left out. This means that the model is not necessarily 100% representative of what will happen in the real-world. Consequently, the complex discipline of modeling, both the creation and manipulation of the model, is equally an art and a science. Practitioners require expertise with the underlying science and experience with the domain as an art form to produce truly representative models that can be controlled and analyzed.

There are several basic approaches to modeling that have been studied and used successfully. Early in the development of predictive models rules-based approaches were tried and enjoyed some success. Indeed, their success encouraged the development of alternative modeling capabilities using a variety of techniques from the field of artificial intelligence. Pure artificial intelligence techniques such as neural networks are very complex, but in the right domain, they are extremely powerful. Recently, one of the most successful approaches has been to blend these technologies to derive the best of both worlds. When should which modeling technique be used?

Approaches that use **rules-based** techniques have been in place for many years. They have shown themselves to be very natural to use and very effective within their known limitations. The technique is so well established that within a stable domain for which it is suited, it produces highly regarded models that are very efficient.

Modeling is a way to represent real-world processes with abstract entities such as mathematical formulas to reduce the overall complexity of the problem and represent the problem in a form comprehensible to humans.

However, this technique does have two drawbacks that limit its effectiveness in general. First, it cannot perform better that the rules and assumptions underlying it, i.e., the rules are generated by an expert, and so far as the expert encodes knowledge, the model is effective; but when the model must deal with new data outside the scope of the rules, it is mostly ineffective. Second, this technique cannot adapt to new circumstances or the analysis of data outside the scope of its rule-base by adding data. i.e., this technique cannot learn as it is used.

The second general technique in use today is more recent and more complex. **Artificial Intelligence** systems, such as neural networks, have the ability to perform well as a modeling technique with the added advantage that they can "learn" from more data. These systems are non-linear and very flexible. They can be used without a complete knowledge of the underlying data and will associate relationships between and among the data while they are executing. This provides these types of models with the capability of adapting to the data they are processing much as a living organism learns to adapt to its environment. The result is that even with noisy data these models are very effective, much more so than rules-based techniques. While all models are sensitive to missing data or to extreme outliers, these models have sufficient flexibility to adapt to the data to somewhat compensate for the noise. The additional complexity in their creation and the sometimes inability to determine why the model is producing certain results has tended to limit the application of this technique universally. Sometimes these models are over-trained and that limits the flexibility and their capability to generalize about the data. Further, these techniques are more difficult to explain to an untrained consumer without a strong background in mathematics.

Finally, there are the models created with a **blended technique.** This is a relatively recent innovation where the two previous approaches are blended to attempt to maximize the good points and minimize the weak points in each. Blended technology is the considered combination of traditional linear statistical techniques with non-linear artificial intelligence techniques.

Using the approaches previously described, automated tools can be developed that create models with more accuracy than either technique alone and with sufficient transparency to be able to

In the Elderly and Disabled Medicaid populations, the spending per capita can reach close to an average of $19,000 per capita, annually.

explain why a prediction has been made. This allows the underlying root causes, called drivers, to be exposed. Additionally, this technique lends itself to safeguards that limit overfitting.[7]

In 2002 and again in 2007, the Society of Actuaries produced a report comparing the predictive accuracy of different modeling techniques as implemented by various vendors. An analysis of the results of the accuracy of predictions for certain high risk prediction models was provided. The 2002 study in particular showed some statistically significant differences in the results of the models. Recall that benchmark accuracies and theoretical maximums (called $R^{(2)}$) are typically in the 0.15–0.20 range[8] as published in earlier documents from the Society of Actuaries.

The 2002 study sponsored by the Society of Actuaries[9] compared results across several predictive modeling vendors. All vendors used traditional rules-based modeling. Using a validation set consisting of 61,580 members from the health plan for Sentara Health care (based in Norfolk, VA), $R^{(2)}$ values were calculated and compared. The highest $R^{(2)}$ value amongst the systems published in this article is 0.15. Using the same Sentara data, one vendor obtained an $R^{(2)}$ value of 0.31 using Artificial Intelligence (AI) modeling. This result was further improved to 0.34 by predicting per member per month (PMPM) charges. This demonstrates another strength of AI – the ability to effectively use and predict outliers in the model. The true R^2 values reflect this ability, and place emphasis on the outliers. While many scientists argue this is a weakness of the $R^{(2)}$ statistic, it should be considered a strength of the $R^{(2)}$ statistic in health care, since the high cost patients are those that require identification and sophisticated population health approaches.[10]

Overview of Data

In addition to the methodology chosen, the freshness of the data must also be considered. The first generation of predictive models focused on processing medical claims and pharmacy data on a quarterly basis. This cycle was chosen because of the need to receive the data, clean the data (health data is notoriously noisy and "dirty" – inconsistent, contradictory or missing), then process the data. Given the normal cycles in hospital quality, it was considered sufficient to receive updated data quarterly so that trending could be established. The result was that the data was sometimes

In 2002 and again in 2007, the Society of Actuaries produced a report comparing the predictive accuracy of different modeling techniques as implemented by various vendors.

considered "stale" when it was under review. Changes in policy or order sets took at least a quarterly cycle to show the resulting effects, so it was not timely to act upon the data.

For this reason, the second generation focused on processing medical claims and pharmacy data on a monthly basis. With improved automation of data extraction, cleaning and processing, the data could be refreshed on a monthly basis to achieve more timely results.

This shift in cycles allowed for more effective cause-effect relationships to be established and provided trending data that was more actionable. The net effect was improved quality in hospitals and improved population health capabilities.

While the periodicity of the data is increasing, improving the quality and actionability of the information obtained from the predictive models, there is still room for improvement in the data. Future generations need to focus on processing additional data sources like Health Reimbursement Arrangements (HRAs), lab data, biometric, and so on. Unfortunately, this introduces another potential problem – matching the data in disparate files to the correct patient, given the use of initials and nicknames. Fortunately, there is a relative simple solution requiring the creation of a master patient index so that a given patient can be identified with incomplete or partially incorrect data and matched to the rest of the data from whatever input source. The addition of more data from more sources enhances the resultant prediction by providing more data for modeling purposes allowing for the potential to increase the accuracy of the result and the specificity of the prediction.

Finally, with the adoption of Electronic Medical Records/Electronic Health Records (EMR/EHR) more real-time data is available and this needs to be incorporated into the predictive models and the population health solutions. Historically, predictive models are built using data that is relatively stale, perhaps three months old. Even when the frequency of data refresh is thirty days, the models are still using data that may not be relevant – especially for patients currently in the hospital. Since the data in EMRs is current, near real-time, what if this data could be used to refresh the predictive models? This would allow for near real-time predictions to feed into population health systems while the patient is still admitted to the hospital.

Finally, with the adoption of Electronic Medical Records/ Electronic Health Records (EMR/ EHR) more real-time data is available and this needs to be incorporated into the predictive models and the population health solutions.

The periodicity of data has been shortening over time to the point where it is now almost real-time. This timeliness can be used to feed population health systems to effectuate more rapid care for patients to provide for better outcomes.

Proposed Solutions

In the beginning of the application of predictive modeling to health care, the solution was often to run the available data through groupers to consolidate perhaps multiple patient encounters into a single episode of care. Using this data, the predictive models generated flat files of predictions that could be searched to determine where to allocate scarce population health dollars. This was relatively simple to do and yielded sufficient results to make the approach viable. Each client was provided with a CD each quarter that contained a "cube" that contained their patient data and predictions that could help target the patients most likely to benefit from population health intervention. There was sufficient robustness in the ancillary data and predictions to filter out non-impactable patients and to prioritize resources so as to provide service where it was most likely to do the most good.

However, this approach was not as timely as some clients desired and it was still too manual in the sense that the delivered "cube" had all of the data, but it took an experienced and trained case manager to ferret it out. The next breakthrough came from exploiting the internet and providing a web-based service as opposed to a CD solution. This allowed for the addition of significantly more "processed" data and more and better reporting of the data that was held. Predictions are provided on a single-member basis instead of the previous grouping and ranges. Evidence-based guidelines could be created and member compliance against those guidelines could be ascertained and reported. Physicians' compliance to evidence-based best practices could also be reported. Significantly more data about an individual member could be stored because the data was no longer limited to a CD distribution method. Indeed, patient-centric databases with terabytes of data were now available to store, relate, and exploit for more efficient manipulation and data mining. More patient history could be retained and risk information related to patient compliance could be updated and assessed.

Each client was provided with a CD each quarter that contained a "cube" that contained their patient data and predictions that could help target the patients most likely to benefit from population health intervention.

Thus, the transition from a solution as a product on a CD to a service on the web was a natural and beneficial step for case managers. As a service, updates to data or predictions can be made more frequently. In addition, errors and omissions can be updated when discovered. Whole new features were added to the service on a regular basis such as the identification of impactable members based on predictions arising from more data and more data associations. The identification and targeting of specific types of members with specific characteristics is possible with customizable filters on the data and the predictions made in real-time.

The application of predictive modeling to population health can impact profitability in four ways:

- **Identifying the best opportunities for early medical management** – Accuracy in forecasting high-risk members is essential since health plans generally have limited resources for case management. It follows that knowing which members are the highest risk for specific diseases, allowing for early intervention to prevent or mitigate those diseases, benefits the patient as well as the population health program.

- **Evaluation of pricing models and premium setting** – Predictive models forecast next year's costs for each individual member. Given accurate predictions at the individual member level allows premiums to be evaluated and set based on these predictions. This is also beneficial to both the patient and the population health program.

- **Identifying the drivers of high risk** – In addition to predicting high risk, most predictive models identify the root causes of the risk. These root causes are the drivers that underlie the predictions. The knowledge of these key drivers enables the population health program to focus medical management resources on the specific diseases/processes that most impact future costs.

- **Profiling and benchmarking** – Population health data is profiled in order to understand the relationship between patient outcomes, treatment methodologies, and population health resources. Profiling can be done by disease, physician, line of business, employee, etc. using a series of standard reports combined with three-dimensional data mining tools that allow for ad hoc analysis and drill down to the individual member level.

The identification and targeting of specific types of members with specific characteristics is possible with customizable filters on the data and the predictions made in real-time.

Next Steps

Since population health is the goal and predictive modeling is the latest, most powerful tool to support this goal, what does the future hold? There are many potential avenues for improvement in the methodologies used, the data acquired, the sources of data exploited, and the interpretation of the results by case managers. Some of the more near-term innovations are:

- **Understanding if members are ready to make changes (Motivation Index)**
- **Predictions in near real-time for readmissions, etc.**

In addition, some longer term problem areas for which solutions must, and eventually will, be found include:

- **Profiling future events that should be prevented**
- **More specificity and granularity in the identification of impactable members**
- **Identification of the optimal population health strategy for a specific member with a specific disease at a specific stage of the disease**

One significant weak-link in the exploitation of predictive models for identifying impactable members is the propensity for the member to be willing to heed the advice and guidance of the care manager. It does no good to be able to identify a member who is non-compliant and who would benefit from case management and to continue to use resources trying to motivate or check compliance for an identified impactable member, if that member is not willing to comply with treatment protocols.

A new and exciting adjunct to the predictive modeling capability is the ability to determine the likelihood that a member will comply with suggested treatments and medications. This capability is called a **motivation index**, a measure of the propensity for a member to be susceptible to treatment and to comply with case management interventions. It can be divined by adding additional complexity to the predictive models so that the models not only identify non-

There are many potential avenues for improvement in the methodologies used, the data acquired, the sources of data exploited, and the interpretation of the results by case managers.

compliant, impactable members, but also determine the likelihood of their compliance (i.e., their motivation index) to suggested remedial treatment or medication in a population health setting. This idea is only now starting to be proven effective in the predictive analytics community and the rollout of this concept embodied in commercial predictive models is already underway.

Another innovative opportunity on the near horizon is the capability to determine potential unfavorable outcomes for patients while they are still in the hospital. This is possible by taking the existing data on the patient from as many sources as possible, including lab results, pharmacy history, patient specific data, physician notes, nursing notes, etc. and assembling them into the Electronic Medical Record, then using this data combined with the historical data for this patient and other patients to forecast the likelihood of re-admission within 31 days, or of post-discharge infection, or any of several other possible outcomes. Armed with this foreknowledge, the case manager can intervene prior to the discharge of the patient to effectively eliminate or reduce the likelihood of the potentially unfavorable outcome by proper pre-release education or other interventions. Further, this allows the case manager to be more focused on follow-up visits, etc. to react to indications of non-compliance or other mitigating circumstance. The innovation here is the application of near real-time forecasting to patient's data during their admission.

A longer-term view of the application of predictive modeling to population health recognizes the need for three new innovations that would significantly impact population health for members. The first of these is the profiling of future events that should be prevented. These are the drivers for high costs members and include such things as complications for some diseases, the future occurrence of a disease, etc. The current models for this are not adequate and are very primitive in their ability to do this. For example, if a member has asthma, what is the likelihood that this will result in the complication of pneumonia? What is the probability that a Congestive Heart Failure will become a Myocardial Infarction, or pancreatitis will become acute pancreatitis? A better understanding of the data will result in more definitive models to profile these future events.

A longer-term view of the application of predictive modeling to population health recognizes the need for three new innovations that would significantly impact population health for members.

The second innovation needed is based on the fact that the current state of the art is far away from the identification of impactable members with specificity. There is a need for more accurate and specific potential savings estimates.

The current state of the art predicts acute costs and then uses an experiential factor to interpret the amount of that cost that is preventable. For example, experience may show that 70% of the acute costs for a group of members are preventable with the proper intervention. Thus, if the acute cost is first predicted from the data available, this 70% experiential factor is used to derive a secondary prediction of the preventable cost. This does not take into account that there are some diseases or some stages of some diseases that do not lend themselves to prevention. Some things cannot be prevented. Therefore, a means to predict the specific members for intervention who will have the most impact based on the prevention of their disease is a future innovation under active research and discussion today.

The third innovation needed is the ability to predict which of several possible population health strategies would be best for a given patient with a given disease at a given stage of the disease. Currently, the care managers select the best intervention for an impactable member based on their experience and the data they are provided. Since there are many nuances in the application of a specific intervention for population health, sometimes a slightly different approach may have a significant impact on the quality of life and the cost savings for the management of a member's care. Unfortunately, today we do not have the data to know what range of population health alternatives are available nor do we have the results of a given population health solution. When this data is provided, predictive analytics will be able to determine what intervention is best at a specific time for a specific patient and help to optimize their care. This step will help to potentially automate the population health process and make it available at less cost to more members.

Another possible innovation is the re-targeting of the approach of population health. Currently, population health is centered on high risk members, especially those with high impactability. What if population health could, through automation, be expanded to include those less-than-high risk members who are readily

Currently, the care managers select the best intervention for an impactable member based on their experience and the data they are provided.

impactable? It may be possible to have even earlier interventions while the impact of the disease is minimal to allow for identifying and treating diseases early in their life cycle. Perhaps something as simple as a targeted mailing to those with the potential to have some disease, but for whom a high risk determination is not justified could provide significant return on investment. Perhaps we can even prevent some diseases from occurring at all or at least mitigate the disease effects by early intervention in non-high risk members, all because we can automate the selection of a particular population health solution based on the member's profile. By detecting early certain diseases, preventing or delaying certain diseases from happening, and maintaining a healthier population, we may potentially reduce health care cost significantly, and generate a greater return on investment for population health.

These are some of the remaining challenges for predictive modelers in their quest to provide meaningful support for population health. Time will tell when and how these needs will be met.

Conclusion

Population health is an essential tool for controlling health care costs and positively impacting the outcomes of members. In order to be effective, case managers need to be able to identify and target those members with the most potential to improve their outcomes so as to focus scarce resources in the most effective manner. Predictive Analytics as embodied in predictive models can greatly assist in this process. As effective as predictive models are today, the near future is set to see an explosion in their use and enhanced capabilities. Predictive models have become an essential beneficial support for population health.

These are some of the remaining challenges for predictive modelers in their quest to provide meaningful support for population health. Time will tell when and how these needs will be met.

Bibliography/Citations

1 PUBLIC LAW 111–148–MAR.23, 2010, 42 usc 18001.
2 Kindig, David and Stoddart, Greg, What is Population Health?, American Journal of Public Health, 2003 March; 93(3): 380–383.
3 http://en.wikipedia.org/wiki/Predictive_analytics
4 Predictive Modeling in Disease Management, 2nd Edition, ©2007 HCPro, Inc. p. 4
5 Extracted from Chapter 9, written by Dr. Charles Engle, of the book Innovation-Driven Health Care: 34 Key Concepts for Transformation, Dr. Richard L. Reece, ©2007, p. 108
6 "Noisy data is meaningless data. The term has often been used as a synonym for corrupt data. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text." For more information and the author of this quote, see http://searchbusinessanalytics.techtarget.com/definition/noisy-data.
7 In statistics, overfitting is fitting a statistical model that has too many parameters. An absurd and false model may fit perfectly if the model has enough complexity by comparison to the amount of data available. Overfitting can lead to spurious or incorrect associations. Overfitting is generally recognized to be a violation of Occam's razor.
8 Gumming, R. B., D. Knutson, B. A. Cameron, and B. Derrick. 2002. A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations. A research study sponsored by the Society of Actuaries. May 24, 2002.
9 Ibid.
10 This paragraph adapted with permission from Predictive Modeling in Health Plans by Randy Axelrod and David Vogel, 2003.

For more information:

Call 800.446.3324 or visit
www.medai.com

About MEDai

For over 20 years, MEDai, has been developing and delivering medical and health information leveraging the industry's most data and technology solutions. MEDai started with one simple idea: the incredible volume of data that flows through health care organizations holds the secret to improving care and reducing costs. The key to success is being able to turn that data into action. MEDai combines sophisticated data mining technology with advanced analytics to enable health care organizations to make better decisions and improve the quality of health care outcomes and operational efficiencies.

MEDai
A LexisNexis® Company