JPMSA
JOURNAL OF THE PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

SPRING 2018 | IN THIS ISSUE:

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Table of Contents

Official Publication of the Pharmaceutical Management Science Association (PMSA)

The mission of the Pharmaceutical Management Science Association not-for-profit organization is to efficiently meet society's pharmaceutical needs through the use of management science.

The key points in achieving this mission are:

- Raise awareness and promote use of Management Science in the pharmaceutical industry
- Foster sharing of ideas, challenges, and learning to increase overall level of knowledge and skill in this area
- Provide training opportunities to ensure continual growth within Pharmaceutical Management Science
- Encourage interaction and networking among peers in this area

Please submit correspondence to:
Pharmaceutical Management Science Association
446 E High Street, Suite 10
Lexington, KY 40507
info@pmsa.net
(877) 279-3422

Executive Director:
Stephanie Czuhajewski, CAE
sczuhajewski@pmsa.net

PMSA Board of Directors

Karthikeyan Chidambaram, Genentech
President

Jin Tong, Novartis
Vice President

Jerome Su, Novartis
Treasurer/Secretary

Vishal Chaudhary, Sanofi
Professional Development Chair

Igor Rudychev, AstraZeneca
Research and Education Chair

Nuray Yurt, Novartis
Marketing Chair

Haibei Jiang, Bayer
Program Committee Lead

Arlene Little, Genentech
Program Committee

David Purdie, Genentech
Executive Advisory Council

Cynthia Dilley, UCB
Executive Advisory Council

Jason Carlin, Novartis
Executive Advisory Council

The *Journal of the Pharmaceutical Management Science Association (ISSN 2473-9685 (print) / ISSN 2473-9693(online))* is published annually by the Pharmaceutical Management Science Association. Copyright © 2018 by the Pharmaceutical Management Science Association. All rights reserved.

PMSA Journal: Spotlighting Analytics Research

Welcome to the sixth edition of the *Journal of the Pharmaceutical Management Science Association (PMSA)*, the official research publication of PMSA.

The purpose of the Journal is to promote and embody the mission of the association, by:

- Raising awareness and promoting the use of Management Science in the pharmaceutical industry
- Fostering the sharing of ideas, challenges, and learning to increase the overall level of knowledge and skill in this area.

The Journal publishes manuscripts that advance knowledge across a wide range of practical issues in the application of analytic techniques to solve Pharmaceutical Management Science problems, and that support the professional growth of PMSA members. Articles cover a wide range of peer-reviewed practice papers, research articles and professional briefings written by industry experts and academics. Articles focus on issues of key importance to pharmaceutical management science practitioners.

If you are interested in submitting content for future issues of the Journal, please send your submissions to info@pmsa.net.

GUIDELINES FOR AUTHORS

Summary of manuscript structure: An abstract should be included, comprising approximately 150 words. Six key words are also required. All articles and papers should be accompanied by a short description of the author(s) (approx. 100 words).

Industry submissions: For practitioners working in the pharmaceutical industry, and the consultants and other supporting professionals working with them, the Journal offers the opportunity to publish leading-edge thinking to a targeted and relevant audience.

Industry submissions should represent the work of the practical application of management science methods or techniques to solving a specific pharmaceutical marketing analytic problem. Preference will be given to papers presenting original data (qualitative or quantitative), case studies and examples. Submissions that are overtly promotional are discouraged and will not be accepted.

Industry submissions should aim for a length of 3000-5000 words and should be written in a 3rd person, objective style. They should be referenced to reflect the prior work on which the paper is based. References should be presented in Vancouver format.

Academic submissions: For academics studying the domains of management science in the pharmaceutical industry, the Journal offers an opportunity for early publication of research that is unlikely to conflict with later publication in higher-rated academic journals.

Academic submissions should represent original empirical research or critical reviews of prior work that are relevant to the pharmaceutical management science industry. Academic papers are expected to balance theoretical foundations and rigor with relevance to a non-academic readership. Submissions that are not original or that are not relevant to the industry are discouraged and will not be accepted.

Academic submissions should aim for a length of 3000-5000 words and should be written in a third person, objective style. They should be referenced to reflect the prior work on which the paper is based. References should be presented in Vancouver format.

Expert Opinion Submissions: For experts working in the Pharmaceutical Management Science area, the Journal offers the opportunity to publish expert opinions to a relevant audience.

Expert opinion submissions should represent original thinking in the areas of marketing and strategic management as it relates to the pharmaceutical industry. Expert opinions could constitute a review of different methods or data sources, or a discussion of relevant advances in the industry.

Expert opinion submissions should aim for a length of 2000-3000 words and should be written in a third person, objective style. While references are not essential for expert opinion submissions, they are encouraged and should be presented in Vancouver format.

Industry, academic and expert opinion authors are invited to contact the editor directly if they wish to clarify the relevance of their submission to the Journal or seek guidance regarding content before submission. In addition, academic or industry authors who wish to cooperate with other authors are welcome to contact the editor who may be able to facilitate useful introductions.

Thank you to the following reviewers for their assistance with this issue of the *PMSA Journal*:

Monal Chand
Associate Director, Business Operations, United Therapeutics Corporation

George A. Chressanthis, Ph.D.
Principal Scientist, Axtria

Mitch DeKoven, MHSA
Principal – Health Economics and Outcomes Research, Real World Insights, IQVIA

Ira Haimowitz
Vice President, Product Strategy, Crossix

Srihari Jaganathan
UCB

Jukti K. Kalita
Business Analytics, StratCo, Pfizer

Eze K. Abosi
Vice President, Analytics, Decision Resources Group

Ewa J. Kleczyk, Ph.D.
Vice President of Client Analytics, Symphony Health Solutions

Sudhakar Mandapati
Principal, Strategic Research Insights, Inc.

Ashish C. Patel
Commercial Lead and Medicare Data Scientist, CareSet Systems

JP Tsang, Ph.D. & MBA (INSEAD)
President, Bayser

Editor: Igor Rudychev
Sr. Business Insights Director, AstraZeneca

UPCOMING PMSA EVENTS

2018 PMSA European Summit
September 5 – 6, 2018
The Hilton Amsterdam
Amsterdam, Netherlands

2019 PMSA Annual Conference
April 14 – 17, 2019
Hyatt Manchester Grand
San Diego, California

Building an Oncology Data Visualization Platform to Leverage Integrated Patient Data and Analytics

Gita Mishkin, MPH, Principal, Brand Analytics and Center of Excellence, Symphony Health; Paula Fullman, MSOD, Vice President, Patient Analytics Center of Excellence, Symphony Health; Don Faust, Lead Consultant, Patient Analytics Center of Excellence, Symphony Health

Abstract: Understanding the patient journey and therapy utilization relies on familiarity with the granularity and complexity of an integrated dataset. It pulls together data from various resources, including, but not limited to, retail and mail order pharmacies, wholesalers, specialty pharmacies, hospitals, community-based offices, clinics and other healthcare delivery facilities, clinical registries, electronic medical records, and lab data. This allows pharmaceutical managers to develop strategy based on a holistic view of the brand (patient, payer, prescriber) experience without having to make leaps of faith across disparate data sources or to guess whether or not various data assets reveal different answers. To this end, the HealthCloud (an oncology data visualization application) allows for drag and drop insights and graphics for an otherwise analytically complex therapeutic area. Five milestones streamlined the process: (1) Identify key questions; (2) Develop business rules, including phases of implementation; (3) Design the core dataset needs; (4) Isolate data aggregations requirements to optimize performance; (5) Create dashboard for data visualization. The primary lesson was the importance of a development plan, including all the phases of release and roles and responsibilities.

Keywords: patient, integrated dataset, data visualization, oncology, dashboard, analytics

Background

Patient analytics

Patient analytics must be utilized as a foundation for driving healthcare monitoring and change, whether it be in market planning, marketing or operations. Four categorizations for patient analytics currently exist: (1) Prescriptive – identifies actions, (2) Predictive – examines likely scenarios, (3) Diagnostic – historical view of performance, and (4) Descriptive – what is happening now.[1] Typically, the pharmaceutical industry utilizes diagnostic or descriptive analytics to monitor brand activity or marketing trends. Data is pulled from a variety of different forums, including Centers for Medicare and Medicaid Services, health plans or claims data aggregators.

Integrated Data

Integrated data pulls together data from various resources, including, but not limited to, retail and mail order pharmacies, wholesalers, specialty pharmacies, hospitals, community-based offices, clinics and other healthcare delivery facilities, clinical registries, electronic medical records, and lab data. Integrated data utilizes a unique identifier to allow for linking of the information across all the different sources. By doing this, the addition of new data sources, such as biomarkers, becomes simpler. This type of data enables pharmaceutical managers to develop strategy based on a holistic view of the brand (patient, payer, prescriber) experience without having to make leaps of faith across disparate data sources or to guess whether or not various data assets reveal different answers. Understanding the healthcare dynamics within a particular disease state, such as a patient journey or therapy utilization, relies on familiarity with the granularity and complexity of an integrated dataset.

Data Visualization
Data visualization is a way to virtually present and manipulate data by using a business intelligence tool, such as Tableau® or Qlik®[2]. The visualizations vary in complexity from simple graphs and charts to heat maps and forecasting. The ability to drag and drop insights, filter by specific metrics or attributes and creating graphics specific to analytics of interest are all advantages to using a data visualization business intelligence tool. For the purposes of this specific initiative, Tableau® was utilized.

Oncology Therapeutic Area
Due to the complexity of the different tumor types and the medical regimens associated with each individual tumor type and the data complexity associated with various treatment delivery, the oncology therapeutic area was used to build the first HealthCloud. Oncology presented some unique opportunities to develop methodology for determining mechanisms for scaling across tumor types, such as starting with lower prevalence and characterizing complexities and nuances associated. For the purpose of planning the development of the dashboard, thorough research went into the definition of specific tumor types, including creating a comprehensive list of all the tumor types and estimating the potential magnitude of data, understanding what types of information was readily available via online resources, and engaging with a subject matter expert. In addition, the National Comprehensive Cancer Network® (NCCN) manages guidelines associated with many different tumor types, and their website was used as a primary source of information for each tumor type. These guidelines were accessed in order to identify the specific medications, develop treatment regimens, and understand physician and patient concerns.

Based on the Oncology Therapeutic Area review, multiple myeloma was the first tumor type to be included in the first phases of the HealthCloud development. This is based on an estimate of approximately 115,000 multiple myeloma patients found in the 2016 data utilized for this analysis. In addition, the NCCN has a detailed patient guidelines book for review of treatment regimens.

Methods
The Oncology HealthCloud team identified five key milestones in order to streamline the process of development: (1) Identify key questions; (2) Develop business rules, including phases of implementation; (3) Design the core dataset needs; (4) Isolate data aggregations requirements to optimize performance; and, (5) Create dashboards for data visualization. The implementation of each stage was accomplished by either a design team (responsible for defining the overarching results and analytics to be displayed) or a development team (responsible for creating the datasets or the data visualizations). Each team member has a unique skill set to allow for cross functional expertise and input.

First, the design team identified some key business questions for developing the Oncology HealthCloud. These questions defined the purpose of the HealthCloud, as well as the desired metrics and outcomes to be displayed in the dashboard. These questions examined patient demographics, such as age, gender, race/ethnicity and geography, baseline diagnosis, baseline procedure, prescription metrics, line of therapy, treatment regimen and source of business.

The next step was for the design team to develop the business rules. These were broken out by the specific phases of implementation. During this step, two items were completed: (1) a spreadsheet clearly outlining all the metrics,

attributes, and data filters requested for each of the analytics. 2) Visualization requirements for the analytics are best developed during this step as well. Identifying the types of graphics, whether maps or line graphs or tables, during this step will help streamline the following steps. The analytics document was reviewed in depth with the design and the development team.

The first phase contained several types of analytics, some basic and some more complex. Baseline counts, such as demographics of the patients and diagnosis, procedure, and prescription profiles (i.e. number of patients, number of claims, payment information) were created. Line of therapy, source of business, and plan control indices were also programmed into the first phase of the HealthCloud.

During step 3, the core dataset structure was designed by both the design and the development teams. Due to the magnitude of the integrated dataset used, it was important to identify specific fields to be included in a large data pull. This bolus of data was used as a large underlying dataset that could be housed on a data cloud. The large amount of data remains on the cloud with the ability to pull additional fields as needed for the HealthCloud. The development team created a technical document reflecting all the core dataset as requested by the design team.

Step 4 uses the data pulled from step 3. The magnitude and design of data pulled into a data visualization tool will directly impact its performance. Understanding the analytics designed for each phase, and how all of the different fields interact with each other, will help maximize the performance of the tool. One aggregated table gives the data visualization tool a smaller more limited view of the larger dataset. The data visualization tool does not need to search through a large dataset in order to find each requested field, but rather, a smaller, more specific dataset.

By step 5, the data visualization developer had the tools they need to create a workable dashboard. During this stage, ongoing meetings between the development and the design teams help streamline the process. The development team is responsible for creating the dashboard based on the analytics requirement document and the technical document. The design team was responsible for reviewing the dashboard as it is created and providing input for modifications, since they will present the dashboard to clients or internal stakeholders.

As each phase reached it final stages and the revisits to earlier steps slowed, the design team began to plan an internal training and roll out plan. To control the roll-out and change input, the design team chose a representative from each part of the commercial organization. This representative provided specific input on methodology and visualizations, which were later incorporated in the dashboards. Once modified, the design team conducted a larger training for the entire commercial organization. This included basic training on the data, Tableau® user training, and finally interpretation and management of the insights created in the HealthCloud.

Discussion

Many lessons were learned during designing and developing this data visualization platform. The primary lesson learned was how to optimize project management across a number of stakeholders with many competing priorities and a process that was prone to both forward and backward movement. Executing the methodology steps was an iterative process and revisiting a previous step for modification was a regular occurrence. To overcome this obstacle,

the design team created an itemized project timeline in Microsoft Project. This helped pinpoint delays, and allowed upper management to justify the need for additional resources to the project. In addition, the design team and the development team met once a week to review all the tasks required to create the dashboard. Future project phases and requirements could be started while waiting for resources to become available for some of the delayed steps.

A development plan clearly outlining all the phases of release and the roles and responsibilities of each team member optimized the design and development by capitalizing on individual knowledge and experience. The analytic plan clearly outlined the expectations for the output of the dashboard, and a technical document created a reference for database designers and the dashboard programmers.

Designed for internal use, development of this platform and tool often took a backseat to revenue generating work, and strong advocacy along with creative project management were required to keep this initiative moving forward. The project plan needs to provide for contingencies when priorities get shifted. The project leader frequently met with internal stakeholders' for their agreement of action in cases where developer priorities were shifted. The ability to remain flexible, to shift gears to take advantage of resources when they were available and continuous monitoring of the work flow were key to success.

The visualizations of the data also required a lot of review. A key learning early on was the difference in perspectives and knowledge about the data across the organization and the ability to align cross functionally to ensure that optimal utility within the visualizations required considerable time and effort along the way. There are several ways to represent data, and continuous review of the visualization helped provide ideas on how to alter the output to the most intuitive form. This also helped identify any issues with the data aggregations used to drive the visualizations. Often, restructuring the underlying tables was required to improve the display of information.

Conclusion

The oncology markets are complex and dynamic with new treatments and treatment paradigms evolving almost constantly. In order to effectively engage with and assist its clients who serve the oncology space, the need for an enhanced toolkit with the intention of providing proactive insights was identified and prioritized. The intention is to strengthen client knowledge and client relationships. As a market intelligence tool, it managed to make cumbersome, raw, integrated data more manageable with shorter analysis times and shorter time to insight. The analytical platform provides internal stakeholders and consultants with quicker access to patient, payer and prescriber related metrics for pharmaceutical and healthcare service related research. Client facing consultants are able to respond to client inquiries more quickly and to proactively offer insights to evolving market events. The faster insights allow for a better working relationship with clients or partners. In addition to helping provide faster information and insights to clients, this data, application and dashboards have also served as a training mechanism for employees.

As mentioned in the discussion section, the HealthCloud development and design is an iterative process. As each new tumor type is added to the dashboard, and as new data or analytics are added to the platform, the HealthCloud will continue to change and expand. After completing the first phase of the dashboard (one tumor type and a set of specific

analytics), this platform was rolled out to the larger internal organization, while still working on additional tumor types and analytics in later phases. To avoid misuse of the dashboard, or misrepresentation of the data contained within it, a large organization wide training was completed. While this maintained data security and integrity, it, equally important, allowed users to receive the proper education on the tumor types. The live user put the analytic platform through real use, and the real world feedback allows the development and design teams to modify the dashboard as necessary.

Each of the phases of analysis will grow increasingly more complex and include many more types of data and metrics. Lab data, biomarkers, and electronic medical records will all be added during the next phase of implementation. The analytics will be recalculated and recalibrated accordingly. Quality control reports will be created to help identify any changes in the underlying database. Persistency, compliance, length of therapy and medication possession ratio are all part of the phase 2 plans.

Although the steps involved in creating this dashboard were at times never-ending, internal stakeholders still maintained (and continue to maintain) the importance of this project.

About the Authors

Gita Mishkin, MPH, *is a data structure development and delivery professional in the pharmaceutical and life sciences industry. She has over 10 years of experience in healthcare research, consulting, development and analytics. She first joined Symphony Health in 2011 as a Practice Consultant with an emphasis on government relationships in 2011. Since then, she has touched all sides of the SH business, including strategic partnerships, data provider contracting and onboarding, and brand analytics. Currently, Gita is both a Principal on the Brand Analytics team working with clients to develop datasets and analytics with the primary purpose of driving the client's business forward successfully, and a Principal on the Patient Analytics Center of Excellence group tasked with project management and driving the development dashboards to showcase and monitor the Symphony Health data assets. Key areas of expertise include integrated data, designing datasets and analytics within the confines of client rules and regulations, and brand management.*

Paula Fullman, MSOD, *is a marketing analytics professional in the pharmaceutical and life science industries with more than 30 years' experience in research and consulting. Paula has been with Symphony Health since 2003 and more than 13 years with the Brand Analytics Practice where she partnered with numerous pharmaceutical and biotech companies throughout the country to define business issues and develop cohesive, targeted solutions. Since 2017 Paula has led the newly created Patient Analytics Center of Excellence at Symphony where she continues to work with clients and to expand the use of patient data and analytics across the Symphony organization. Key areas of expertise include patient and managed care analytics which assist in developing strategies and programs around brand management, marketing and reimbursement.*

Don Faust *is a pharmaceutical and life sciences professional with over 20 years of experience, covering research and consulting, project management, and client services/*

operations. Don joined Symphony Health in 2013 in the Brand Analytics Practice, most recently as Consulting Manager, working with various clients to help solve their business questions through problem identification, strategic solution recommendations and project design/implementation. In 2017, Don joined the newly formed Patient Analytics Center of Excellence group as Lead Consultant, helping both clients and internal teams advance the use of the many data assets that Symphony Health has to offer. Area of expertise includes the use of APLD data for brand management and marketing/market research uses.

References

1 *Four Types of Big Data Analytics and Examples of Their Use*. 13 09 2017. http://www.ingrammicroadvisor.com/data-center/four-types-of-big-data-analytics-and-examples-of-their-use.

2 *Data Visualization: What it is and why it matters*. 13 09 2017. https://www.sas.com/en_us/insights/big-data/data-visualization.html.

Improve HCP Email Message Response Through Personalized and Optimal Messaging Strategies Derived From Machine Learning Models

Anna Decker, Ph.D., Senior Data Scientist, Aktana Inc. and Marc-David Cohen, Ph.D., Chief Science Officer, Aktana Inc.

Abstract: Machine learning (ML) approaches are commonly used in retail settings and are starting to emerge in pharmaceutical brand management and sales operations. These approaches to predictive analytics can identify optimal messaging and interaction strategies that improve communication with healthcare providers (HCPs) by finding complex non-linearities in response functions that have many potential predictors. Non-parametric ML models are particularly helpful when insight into the functional form of the relationship between predictors and the response is *not* the main objective but quick, automated implementation in an environment where the number and type of predictors can change rapidly is important. Traditional methods based on parametric models are more valuable when a deep understanding of the shape of the relationship of predictors to predicted is desired and used to illuminate insights. For example, to capture complex non-linearities when using logistic regression, explicit transformations of the predictors are needed in contrast to random forest based models where hidden non-linearities are "found" by the modeling technique. Furthermore, fewer distributional assumptions are needed with ML models than with many more "traditional" approaches while predictive accuracy is often improved. We introduce these concepts and demonstrate some of these observations with an example. The main focus of this paper is to describe the problem of personalizing frequently changing messages to individual HCPs using an ML approach most suitable for its ability to be automated and to produce predictive accuracy that is comparable to other techniques. We show a progression of uses of random forest models that (1) provides an automated approach that predicts the next best message to send for each individual HCP and (2) identifies the segment-specific messages that HCPs are most likely to read. This model accounts for HCP covariates and historical HCP reactions to previously sent messages. Using this machine learned model of the joint distribution of the data, we simulate expected outcomes under hypothetical interventions to generate longitudinal predictions of message reaction. Our experience is that clients develop confidence when they understand the model from insight into these longitudinal predictions and then quickly proceed to full integration of the ML model scoring to gain its full potential. We demonstrate these methodologies with anonymized data from two clients and show estimates of expected improvements as well as actual doubling of email open rates by using these new techniques.

Keywords: machine learning, optimal HCP message strategies, simulation, big data analytics, improving email response rates, automating strategy execution

Introduction

Determining the optimal sequence of messages to send from pharmaceutical representatives to HCPs, whether they be via email, in person, or via a webinar, is challenging. Many factors influence the chance of a message being received and of interest, and even defining "message received" can be a challenge. Both the timing and content of messages are important factors for determining whether a message will be of interest and not dismissed out of hand. Here, the focus is on the order or sequence of messages that should be sent to maximize HCP interest and identifying the next "best" message to communicate.

Message sequencing is one aspect of a larger optimal messaging strategy. Understanding and optimizing the timing, content, and types of messages within and across communication channels is important for managing the relationship between HCPs and the brand and how valuable the brand is to the HCP. This article provides one optimization approach for message sequence with the specific example of email messages.

The utility of ML models in this approach is paramount. They are commonly used in estimating customer reactions in retail settings, such as with the use of Bayesian networks in direct marketing response models by Cui, Wong, and Lui or rule induction and neural networks in data mining by Bose and Mahaptra.[1,2] In the area of pharmaceutical brand management and sales operations there is an opportunity to use ML models for personalizing and automating marketing execution to ensure the most efficient and effective distribution of marketing resources.

Integrating Brand Strategy and Machine Learning

Typically, existing message strategies provide sets of approved messages that pharmaceutical sales representatives and others send or present to the HCPs. Collecting data on the messages delivered and the HCP reactions over time allows analysts to examine both the representative's adherence to this message strategy as well as assess the receptiveness of the HCP to each of the messages. For instance, for email messages, one can examine the chances of opening and click-through to other pages linked within the email text and use that reaction as a measure of HCP reaction to the message. Within a particular pre-specified set of messages, calculating the likelihood of a given message being opened or clicked through gives an objective ranking of the messages. This is a useful retrospective measurement of message effectiveness, and can be leveraged to provide prospective prediction of how a message will be received when delivered in the future. This is where predictive modeling becomes useful for directing future marketing and sales actions and execution. By building a model on top the collected data, including covariates that capture baseline HCP propensity to be more or less receptive to messages as well as the history of message sends and receptivity, a prediction of future receptivity can be generated for new messages.

Building a predictive model that performs well on both collected and new data is vital to this approach. Machine learning allows for more flexible modeling of the underlying joint distribution of covariates and features that capture historical HCP message behavior that more closely resembles the truth than a typical linear regression.[3] A variety of methods such as logistic regression, Bayesian generalized linear models, or recursive partitioning could be used for this modeling, however, many available ML methods are more effective at capturing complex nonlinearities in the data.[4,5] Additionally, it was important to take steps to protect against overfitting, i.e. the model working well on training data but not performing well on new test data.[3] Cross-validation is an effective method that we employ to avoid overfitting.[6] In particular, random forests are a ML technique that builds decision trees based on random resampling of the data and are inherently useful for this procedure since the decision to open or click a message is dichotomous and has an associated probability of occurring,[4] however, it can be used when there are other measures of HCP interest.

Given this type of predictive model, the probability of a message being opened or clicked given the HCP-level covariates and history of messages received can be estimated for a given message. Comparing the probabilities for different potential next

messages to send provides a method for making that choice that maximizes the HCP interest in this next message. This prospective prediction allows for more intelligent message suggestions and is able to adapt over time when new messages are added to the message set and in addition is applicable to new accounts as well. Thus, rather than an arbitrary sequence of messages that is the same for all HCPs, the algorithm provides a personalized sequence based on covariate data and historical message reaction data that optimizes the chance of an open or click.

Methods

Potential outcomes

Let W represent the matrix of baseline covariates in a data set of interest, A represent the send of a message of interest, and L represent the past sends and opens of other messages in the message set, and Y be an indicator of whether the message of interest has been opened or not. We are interested in modeling the conditional expectation of Y given the other variables, i.e. E[Y |A, W, L], which could be modeled using many approaches. Estimating the probability of a message open/click given the historical covariates and past open/click behavior of the HCP for each potential message allows for a ranking of the messages based on the predicted probability of open/click. The message with the highest probability can be promoted as the next best message to send.

In order to determine an optimal sequence of messages for a particular segment of HCPs or individual HCPs, an additional step was required. Using the modeled distribution of the probability of a message open, we deterministically intervene and set message send to be a "yes" for each model and then predict using this intervened data to obtain the probability of a message open under a potentially counterfactual message send.[7] This is a technique that is commonly used to estimate parameters motivated by the causal inference literature and provides an estimable statistical parameter that can make use of ML models. More formally, the so-called counterfactual outcome is calculated as

$$Y_1 = E[Y|A = 1, W, L]$$

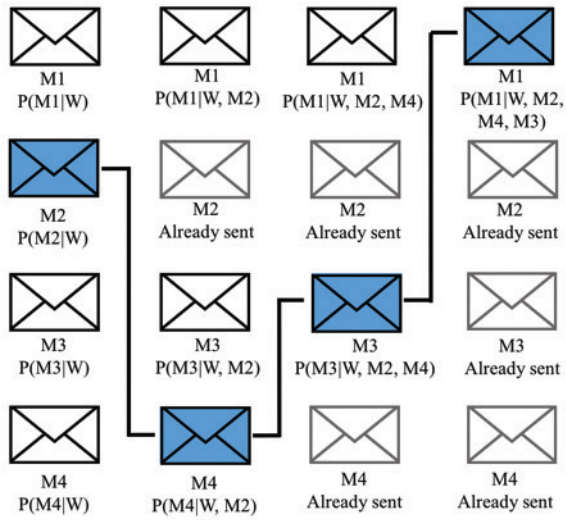
which is the average observed outcome for HCPs that did receive the message and the counterfactual outcome for HCPs that did not receive the message. This parameter corresponds to the treatment specific mean under a deterministic intervention and is well studied in the causal inference literature.[7-10] The predicted probability under this intervention measures how likely the message is to be opened if all HCPs received it.

Sequence Prediction

Computationally, the conditional probability of an open or click given the history of message sends and covariates can be fit using a variety of methods.[5,11] In our implementation, we utilized random forests to fit the conditional probability in order to avoid making unnecessary assumptions about the functional form of the relationship between the probability of a message being opened or clicked and the covariates and history of message opens. We fit a separate random forest model for each message given past observed data and use an ordered progression of models to generate the optimal sequence of messages to send for a finite message set. Note that these models may be built at the individual level (if there are enough data points), or at a pre-specified segment level, or for all HCPs.

From the initial message model and associated predictions, two copies of the intervened-upon data were created. In one copy, the counterfactual was that each HCP opened the first, highest

Figure 1: A Schematic of Message Sequence Optimization



likelihood message, and in the other, the assumption captured in the counterfactual is that each HCP does not open that message. These intervened-upon data sets were used in the random forest model to estimate the next most likely next message each HCP would open if sent. These predictions were then used to create the next pair of counterfactual data sets. This process creates a sequence for each HCP of messages that they are next most likely to open. Repeating this for each possible next message allowed us to rank the messages based on descending probability of an open given possible past send and open patterns. Applying a similar procedure across the entire message set, we obtained sequences of messages with their associated probabilities. At each step, the next best message to send is determined based on the history of sends and opens. This simulation allows for prospective sequence prediction and gives a potential optimal sequence. However, the send and open behavior may change for an individual HCP, so the observed sequence may deviate from what is predicted here. This process is summarized in Figure 1.

The determination of an optimal sequence requires a finite and small message set and assumes that messages are not sent repeatedly, although repeat sends could be accounted for. It also assumes a correctly-specified model.

The Utility of Random Forest

We elected to use random forest to model the conditional distribution because of its ability to detect complex interactions and non-linearities in the data. Random forests are ensemble learners that combine so-called "weak" learnings (decision trees). The algorithm resamples a percentage of the data and a subset of the predictor variables to build a decision tree, in which predictors are added based on how well they partition the data according to some objective function. Two parameters in random forests typically control the overall fit; one is the depth of the tree search and the other is the number of trees in the random ensemble. Care should be taken to match these to the dimensionality of the data to avoid overfitting.

To demonstrate random forest's capabilities, a simulation study was performed, where we simulated variables and had a known functional form that related the outcome to the predictors.

Table 1: A Comparison of Three Different Models of the Conditional Distribution of an Outcome Given a Set of Predictors Where the Relationship Is Complex

Model	Mean-squared error (MSE)
Simple regression	2398.2
Random forest	1235.2
Correctly specified model	1250.8

The simulation performed as follows:

$$X1 \sim \text{Normal}(\mu=5, \sigma=8)$$

$$X2 \sim \text{Normal}(\mu=0, \sigma=2)$$

$$X3 \sim \text{Binomial}(p = \text{mean}(X1 > 6))$$

$$X4 \sim \text{Binomial}(p = \text{mean}(X1 + X2 < 0.4))$$

$$X5 \sim \text{Normal}(\mu = \text{mean}(X3), \sigma=5)$$

$$Y = X1 + X1 * X2 - X3 + X5^2 + X3^3 + X5 * X3 * X1$$

The relationship between Y and the predictors was made deliberately complex to demonstrate the utility of random forest for detecting such a relationship. The performances of a simple linear regression, random forest, and a correctly specified model were compared using the mean-squared error (MSE). The expectation was that the correctly specified model would fit the best, i.e. have the lowest MSE, random forest would be able to fit some of the interactions and have the second-best MSE, and that simple linear regression would have the highest MSE. These expectations were met in our simulation. The results are summarized in Table 1.

For a relationship with complex interactions and a non-linear relationship, random forest performed almost as well as the correctly specified model while simple regression did not. Since we do not know the underlying functional form of the relationship between the outcome and predictors, it makes sense to choose a method that does not make any unnecessary assumptions about the functional form. In this example, while a careful study of the relationship between the predictors and targets might yield a correctly specified parametric model, the larger the number of predictors, and the more complex the relationship, the harder this structure is to uncover. Therefore, we advocate for the use of an algorithm that can learn the joint distribution empirically. Using this as the basis for the message sequence optimization procedure ensures that the conditional distribution is modeled as accurately as possible. When fitting these models, we output predictive performance metrics such as area under the receiver operating characteristic curve, accuracy, and sensitivity in order to assess model performance. In general, we have seen that using covariate data and send, open, and click history results in high predictive performance.

Limitations of the Current Approach

The method proposed above is based on historical data that have been collected in CRM (Customer Relationship Management) systems. Thus, brand new messages that have not been previously sent are required to undergo a period of being sent at random to gather send, open, and click data before a model can be built. We specified three possible strategies for these random sends: aggressive, moderate, and conservative. All three are based on the range of predicted probabilities of opens across all

messages for all HCPs (or segments of HCPs). The aggressive strategy sends the new message with high probability relative to this range, the conservative sends it with low probability relative to this range, and the moderate strategy sends the new message with a medium relative probability. These strategies allow the customer to prioritize new messages and control the amount of time required to collect the necessary data to build a new model. In our implementations of this approach, we found that a reliable model can be built once a message has been sent at least 20 times. Additionally, a text analysis of historical messages and their similarity to new message content could be used to guide initial send probabilities.

In addition to handling new messages, the method as proposed does not allow messages to be resent. However, a reasonable extension that we are implementing is to allow messages to be sent again under certain conditions, such as if a message was not previously opened or clicked or if a certain amount of time has passed. While the current method avoids message fatigue, where the same message is sent over and over, messages may need to be sent more than once before they are opened or clicked.

This approach requires a set of HCPs, a finite message set to be sent to that set of HCPs, and sufficient past data to support the construction of models for each message, i.e. a sufficient number of past sends, past opens, and past clicks. It is adaptive, in that models can be refit to account for new send, open, and click behavior, additional models can be built for new messages, and new HCPs can be added. It does not take into account other communication channels such as visit details, webinars, or phone calls. A multi-channel messaging strategy built on the same methodology described above could provide better overall insights into messaging strategies in the future.

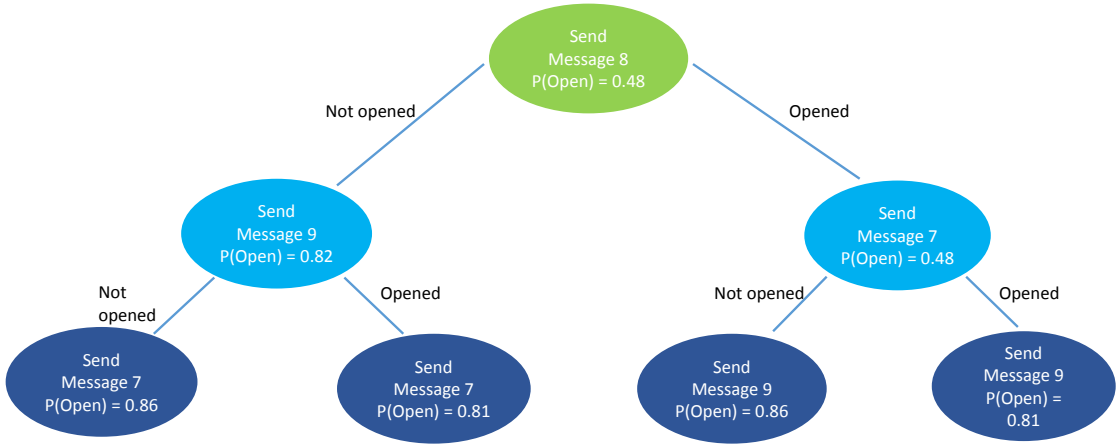
A Case Study

The message sequence optimization approach was applied to Customer A, who had rich historical data of email sends and opens. There were approximately 64,000 HCPs on which the model was built, which served as the basis for the message sequence optimization procedure for four specific messages. Demographic variables, segmentation variables, message characteristics, and observed history of message sends and opens were used to model the probability of a message open, generating a model for each message using a design matrix of several hundred predictors. These models of the conditional distribution for each message were then used in the message sequence optimization procedure. For this example, the four most commonly sent messages were modeled and possible sequences were determined using the sequential simulation described above. Starting each sequence with a different message generated a decision tree. An example of the results is shown in Figure 2. The tree depends on the first message sent in the simulation. For this tree, the optimal sequence is to first send Message 8, followed by Message 9, and finally Message 7, which has a joint probability of 0.34.

The optimal sequence can then be determined by choosing the sequence with the maximum joint probability of open. This sequence may be used in the planning stages before rolling out a new messaging strategy. The models may also be used to determine the next best message to send based on historical open and click behavior as well as demographic variables.

The message sequence optimization tool was recently deployed for almost 400 representatives who contacted approximately 1260 HCPs regarding a single product for Customer B. To preserve client confidentiality, we are unable to share detailed performance impact. However,

Figure 2: An Example Decision Tree That Comes Out of the Message Sequence Optimization Procedure for a Sequence of Three Messages



the tool has been in use for three months and so far, we have seen open rates double.

Conclusions

Using models fit with ML, it is possible to predict the next best email message to send to maximize message opening by HCPs. These models further serve as the basis of a sequence optimization procedure by estimating a parameter motivated by the causal inference literature and predicting open behavior under different proposed sequences. These models are able to be updated and refit daily to provide the most current next best message to send based on historical open and click behavior.

The utility of ML to this approach is vital. Random forest performed almost as well as a correctly-specified parametric model in a simulation study, and the message sequence optimization procedure relies on the conditional distribution of the probability of opening a message being modeled correctly. Other ML approaches may also be used to model the conditional distribution of the data such as support vector machines, neural networks, clustering methods such as k-nearest neighbors, or ensemble stacking methods.[3,12–18]

The example detailed in this article focused on the optimization of email message sequence. Since email messages have a clear response (open or click), this provided a clear outcome variable to model. A natural extension to this analysis would be to expand to other channels of communication such as visit details, webinars, or congress meetings. This expansion would complicate the optimization approach in two ways: (1): it would require a channel-agnostic measurement of HCP engagement and (2) it would require modeling the interaction between each channel. Additionally, this method determines the next best message to be delivered but does not determine the optimal timing of the next message. The timing of a message send can have a large impact on the probability of a message open since HCPs may have a higher or lower receptivity to a message depending on the time between events. For example, an email sent to follow up after an in-person visit might have a higher propensity to be opened than an unsolicited message. This further addresses variability in HCP open behavior and we currently use a separate model to determine when to send messages that is also based on a machine learned model of the joint distribution of the data.

Acknowledgments: We thank Gil Blustein, Wendong Zhu, and Holly Logue at Aktana for their helpful comments during our revisions of this paper. We are also grateful to our reviewers for their insights and edits to our initial draft.

About the Authors

Anna Decker, *Ph.D., Senior Data Scientist, Aktana Inc., is a statistician specializing in causal inference and semiparametric estimation using machine learning. At Aktana she builds machine learning models with an emphasis on thoughtful application of modeling and statistical parameter estimation to answer marketing questions. Previously, she was in a bioinformatics scientist at Genentech in research and development. She received her PhD in Biostatistics from University of California, Berkeley.*

Marc-David Cohen, *Ph.D., Chief Science Officer, Aktana Inc., is an experienced business leader with a background in operations research and statistics. At Aktana he leads the development of learning and insight generation capabilities. Previously he served as CSO at Archimedes Inc.—a Kaiser Permanente Company—and helped the company transform from HEOR pharmaceutical consulting to a products company focused on clinical studies and personalized medicine. Previous roles included VP of Research at FICO and multiple senior roles at SAS Institute where he initiated the SAS Marketing Optimization product.*

References

1 Bose I, Mahapatra RK. Business data mining — a machine learning perspective. *Inf Manage*. 2001 Dec 20;39(3):211–25.
2 Cui G, Wong ML, Lui H-K. Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Manag Sci*. 2006 Apr 1;52(4):597–612.
3 Friedman JH, Tibshirani R, Hastie T. The Elements of Statistical Learning. 2nd ed. Berlin: Springer Series in Statistics; 2001.
4 Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
5 Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008 Dec;2(4):1360–83.
6 van der Laan MJ, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. *Stat Appl Genet Mol Biol*. 3(1).
7 Robins JM. Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials [Internet]. Springer, New York, NY; 2000 [cited 2017 Nov 6]. p. 95–133. (The IMA Volumes in Mathematics and its Applications). Available from: https://link.springer.com/chapter/10.1007/978-1-4612-1284-3_2
8 Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986 Jan 1;7(9):1393–512.
9 Wang A, Arah OA. G-Computation Demonstration in Causal Mediation Analysis. *Eur J Epidemiol*. 2015 Oct;30(10):1119–27.
10 Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol*. 2017 Apr 1;46(2):756–62.
11 Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Taylor & Francis; 1984. 372 p.
12 Vapnik V. Estimation of Dependences Based on Empirical Data. Springer; 2006.
13 Specht DF. Probabilistic neural networks. *Neural Netw*. 1990 Jan 1;3(1):109–18.
14 Fix E, Hodges J. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. California Univ Berkeley; 1951 Feb.
15 van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1).
16 Burez J, Van den Poel D. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Syst Appl*. 2007 Feb 1;32(2):277–88.
17 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015 May 27;521(7553):nature14539.
18 Kim H-C, Pang S, Je H-M, Kim D, Yang Bang S. Constructing support vector machine ensemble. *Pattern Recognit*. 2003 Dec 1;36(12):2757–67.

A Way Forward: Leveraging Advanced Diagnostic Testing to Unlock the Value of Precision Medicine

Dave Smart, Director, Knowledge & Insights, Diaceutics; Steve Vitale, Managing Director, Marketing Team, Diaceutics; and Jeff Waldron, MBA, former Executive Director, PM Connective

Abstract: The need to contain rising healthcare costs is driving payers to adopt new models of reimbursement, such as value-based pricing. Advanced diagnostic testing, particularly with respect to therapies where the test result determines whether or not the therapy should be used (precision medicine) can potentially reduce costly inefficiencies in medicine by ensuring the right patient gets the right therapy, at the right time. However, there are a number of barriers to efficient testing which must be addressed in order to achieve this promise. This article identifies a number of potential barriers to effective commercialization of advanced diagnostic testing in the context of precision medicine and uses modelling to estimate the economic and personal costs such barriers impose on the healthcare system. Modelling suggests that nearly 80,000 patients per year in the US alone are lost to the benefits of precision therapies, at a cost not only to the pharma industry of in excess of $8Bn, but also to the patients in the form of inadequate medical care.

Keywords: precision medicine, diagnostic, commercialization, value

Payers are applying pressure for value-based pricing of pharmaceutical therapies and are already moving toward a pay-by-result model. They recognize that a significant percentage of drug treatments fail to achieve desired outcomes and are therefore trying to identify and reimburse for therapies that produce positive outcomes in patients while refraining from paying for those that fail. Advanced diagnostic testing offers both the insurance and pharma industries a long-term solution and a way forward.

This article identifies stakeholder concerns over value-based pricing of pharmaceutical therapies, the promise and potential of diagnostic testing in precision medicine, and the roadblocks and barriers that sometimes plague the diagnostic testing process. It also presents recommendations for how stakeholders can work together to maximize the value of diagnostic testing for providers, payers, consumers, suppliers, government and other vital stakeholders across the continuum of care, most importantly patients.

Bridging Stakeholder Agendas to Achieve Value-Based Pricing

Among the disparate agendas that enter the debate over value-based pricing of pharmaceutical therapies are the following:

Government price-crunching: Stakeholders may confront changes to the Medicaid best-price rule, where value varies depending on the indication. In the case of a low-value indication, Medicaid insists on a lower price; in the case of a high-value indication, the drug may be overpriced. Other payers tend to follow Medicaid in a drive to bring down costs.

Sky-high investments: As healthcare costs continue to escalate, pharma companies largely foot the bill for investment in new therapies. At

the same time, both payers and pharma companies are eager to extract higher value from healthcare—specifically from high-cost therapies that produce positive outcomes in patients. But pharma, which can invest $1 - 5 billion to develop and launch a new drug, must recoup its initial investment by pricing a drug to turn a profit, reinvest in innovation and cover the costs of those assets that fail prior to regulatory approval. Private payers aim to keep costs down to sustain profitability, while government payers must avoid added tax burdens.

International interests: The majority of pharma companies are multinationals concerned about marketing their products in a number of countries other than the US. Pricing deals struck between reimbursement agencies in non-US countries have prompted the U.S. administration to push for pricing that prevents what it views as U.S. consumer subsidization of drugs in other nations. Pharma must achieve a balance between deals made in other countries and prevailing trends within the U.S.

Association push-back*:* Associations like the American Society of Clinical Oncology (ASCO) and the American Medical Association (AMA) are likely to push back against allowing strict economic calculations to dictate the practice of medicine. Physicians and other providers will resist if payers insist on limiting reimbursement to less effective drugs that physicians would typically refrain from prescribing to their patients. This thinking is reflected in the Affordable Care Act (ACA), which prohibits the use of outcome calculations such as QALYs to set payment thresholds.

Engaged patients: Increasingly aware of breakthrough trends like liquid biopsies and genetic sequencing, more savvy patients and consumers need and want diagnostic testing and treatment. Now, however, they face escalating premiums, co-pays, and deductibles and worries over which tests and treatment their health plans will cover.

To move forward on value-based pricing, stakeholders, including pharma, consumers, payers, providers, researchers, labs and suppliers, must reach consensus on a definition of value and evaluate if therapies can and do achieve that value. They must decide which tests to make available to providers and consumers, and the level of performance required of these tests.

Looming large over stakeholder concerns is lackluster transparency. Transparency in pharma is lacking throughout the continuum that begins with manufacturers, employers, and payers and extends through PBMs, patients and consumers. Consumers, for example, neither know nor understand the rationale behind the prices of specific drugs. Instead, they reach out to a drug supply chain, fill their prescriptions and submit required co-pays.

Large health plans like United Healthcare, Anthem, Aetna or Cigna could negotiate prices with manufacturers, which would give members access to these therapies at lower prices. But most plans refrain from sharing such information with competing health plans and drug manufacturers. Therefore, the ultimate buyers—employers, patients and consumers—never learn the actual prices. Instead, they rely on plans and PBMs to manage pharmacy benefits. Transparency is absent because no entity wants competitors to identify revenue breakdown.

However, transparency may accelerate. Unfortunately, change rarely comes from within an industry. Instead new entrants with highly disruptive business models enter a market. They take business away from traditional enterprises and force them to adapt, much like online travel web sites replaced travel agents and online shopping attacks brick and mortar stores.

For example, disruption could come through Spark Therapeutics' newly approved gene therapy, which offers payers a three-pronged pricing program, including rebates on drug effectiveness at 30-to-90 days and again at 30 months. It leverages contracts with commercial payers and specialty pharmacies rather than with treatment centers. Additionally, payer reimbursement can be done via installment payments made over several years in collaboration with the Centers for Medicare & Medicaid Services.[1]

A second scenario driving enhanced transparency and value-based pricing is advanced diagnostics. Precision medicine provides an alternative to traditional, "one-size-fits-all" prescribing. Instead, advanced diagnostics aims to deliver the right treatment to the right patient at the right time, zeroing in on those patients who will likely benefit from a particular therapy.

Precision medicine represents an opportunity to reduce costly inefficiencies in medicine, including false positives and negatives, unnecessary treatment, and over and under medication.[2] This would slash the costs of so-called trial-and-error medicine while countering negative outcomes, including the fact that the top 10 highest-grossing drugs in the U.S. are effective between 1 in 25 and 1 in 4 of patients who take them.[3] Adverse drug events caused by imprecise medications are also a major cause of costly acute care admissions and readmissions.[4]

Understanding the Promise of Diagnostic Testing

Diagnostic testing comes with some good news and bad news. The downside of diagnostic testing is that drug prices may climb higher. The multi-billion dollar investment required to develop and market a blockbuster drug was once spread over millions of patients. Now, however, diagnostic testing will identify appropriate drugs for smaller, more highly targeted groups of patients. While diagnostic testing delivers access to more patients, it may pinpoint fewer patients who will benefit from a given therapy.

On the other hand, the use of diagnostic testing in clinical trials could foster smaller, faster and less costly trials—either by zeroing in on responder patients or broadening a trial's scope. And diagnostic testing comes with new hope for patients with rare diseases as researchers seek effective treatments where none exist.

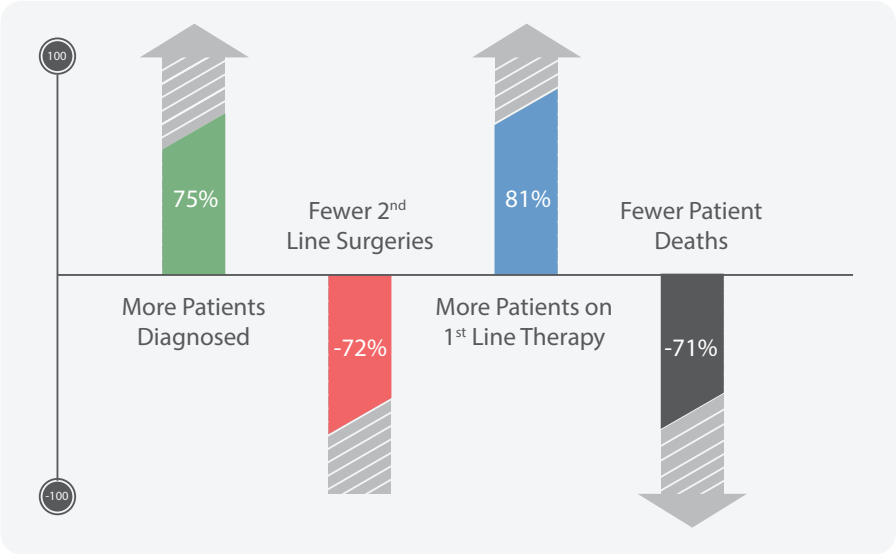
But cost could remain a barrier with new diagnostic tests emerging and becoming more mainstream. Next-generation sequencing is costly, and demand is high. Labs will likely respond with more innovation in the form of creative gene panel design, streamlined development, validation, workflows, and assessment of employee competence and workload.[5]

Other diagnostic testing advantages include the following:

Step Therapy: Diagnostic testing supports a way to avoid step therapy, often called "fail first therapy" by its detractors. Rather than waiting to see if a particular therapy is having the intended effect on a patient, quality diagnostic testing can identify the likely responders up front. This avoids unnecessary expenses related to administering a therapy that will likely fail to work. It also curtails the added strain put on a patient whose health is already compromised by the condition being treated.

Adverse Reaction Spotting: Testing can help determine a patient's risk for having an adverse event (AE). If the patient is a likely non-responder, testing can prevent the patient from being exposed to the risks of an AE

Figure 1: % Changes – Better Diagnostics & Earlier Rx Intervention



without receiving the benefits of treatment. If clinicians can determine if a patient could or will have an adverse reaction to a drug, they can save both on the costs of administering the drug and dealing with its side effects. If testing is unable to determine whether a drug will work or not, clinicians may be able to claim that the drug will precipitate an adverse event that could lead to a costly hospitalization.

Diagnostic testing can also support physicians in understanding how a patient metabolizes drugs, which may determine how a patient responds to a specific drug. Clinicians can modify the dose of a drug given to a patient according to how the patient will likely metabolize the drug (or metabolize a pro-drug to an active form). This ensures that the dose is safe and effective. For example, a clinician could recommend evaluating a patient's response to the antiplatelet drug clopidogrel (Plavix) by detecting variations in the CYP2C19 gene, which indicates how a patient will metabolize the drug.

Accelerated Treatment: Diagnostic testing also informs physicians in rapid-response patient treatment. The earlier clinicians can diagnose patients, the more effective a therapy may be. Armed with the results of diagnostic testing, physicians could treat breast cancer in its earliest stages rather than in stages three or four. Also, they could not prescribe drugs that are unlikely to work in the patient, but which could potentially have adverse side effects that threaten patient safety.

The promise of diagnostic testing is most evident in recent breakthrough work on melanoma where the PM Connective and its collaborators have unveiled the components of a model that demonstrates an opportunity to save nearly $1 billion in healthcare costs via earlier diagnosis and use of targeted therapies. By improving diagnosis, patients are identified earlier in their disease, presenting an opportunity to intervene with treatments that are likely to be more effective at these initial stages. While those diagnostic and prescription costs increase, they are offset by significant decreases in 2nd line surgeries often required in later stages of melanoma. Additionally, the rate of patient mortality is likely to decrease significantly.[9] (Figure 1)

The model emerged from a collaborative process facilitated by the Personalized Medicine (PM) Connective through a series of workshops and surveys. Among the model's focus areas are clinical practice and operational, diagnostics, pharmaceuticals, reimbursement, regulatory, and guidelines and education. Each focus area is supported with a series of near-term, medium-term and longer term goals, including improved clinical collaboration, new options of melanoma adjuvant therapy, consideration of clinical trials as a standard of care, and measurement of combination therapy versus monotherapy.

Also unique to the model is its analysis of the unique features of melanoma, which is typically discovered via inspection by the patient or a physician who classifies the melanoma and performs further diagnostics to identify effective therapies. Among the therapies is immunotherapy, which earned significant media coverage when former President Jimmy Carter was effectively treated with pembrolizumab (Merck's PD-L1 inhibitor Keytruda®). Unfortunately, immunotherapy may work in only 20 to 25 percent of melanoma patients whose cancer expresses the protein PD-L1. This offers diagnostic testing the opportunity to determine if immunotherapy or more traditional therapies are likely to generate more positive outcomes.

The nature of cancer, including melanoma, reveals the vast potential of diagnostic testing. Cancer is heterogeneous, not homogenous. A solid tumor biopsy may not represent an entire cancer. Some cancer may mutate, meaning that even if a therapy is initially effective, it may not be effective for an extended period of time. A clinician may have removed only some components of a cancer during a biopsy while missing others. And cancer can interfere with the effective functioning of the immune system. Immunotherapy can turn the immune system back on without being able to turn it off, which could cause the immune system to attack healthy cells in a phenomenon known as a cytokine-storm.

The future of diagnostic testing in cancer treatment is bright, as medicine transitions to liquid biopsies where clinicians can detect cancer biomarkers in the blood—or other bodily fluids—before making therapeutic decisions. As liquid biopsies come of age, benefits will become more obvious. Clinicians can bypass invasive surgery and conduct tests more frequently and at a lower cost. They can more easily track a patient's progression via the process of therapeutic drug monitoring and make an evidence-based decision to switch therapies. While liquid biopsy is still an unproven technology, experts believe that it will deliver more representative views of entire cancers. And while it can take years for solid tumors to form, cancer may show up in the blood more quickly, allowing clinicians to diagnose and treat cancers earlier and save patients' lives.[7]

The growing importance of diagnostic testing is also evident in the number of drugs released with companion diagnostics that reveal which patients will benefit from a specific therapy. Pharmaceutical companies are more commonly embracing companion and complementary diagnostics, with growing numbers incorporating diagnostics or biomarkers into their clinical trials. They recognize the potential for diagnostics to help prove the value of a drug, which also supports the interests of diagnostic suppliers that typically confront the barriers of reimbursement and lack of investment in test development.[8]

We are also seeing an evolution for diagnostic testing. In May 2017, the Food and Drug Administration (FDA) approved pembrolizumab for use against tumors that share a specific genetic profile. Rather than strictly basing effectiveness on the location in the body where the tumor may have originated, i.e. lung or

breast, the indication is based on whether the tumors express a biomarker referred to as microsatellite instability (MSI-H) or mismatch repair deficient (dMMR). This development allows physicians to treat a number of different cancers based on a single diagnostic test result.

Diagnostic Testing Barriers, Roadblocks and Bottlenecks

Diagnostic tests vary in the following ways:

- Analyte: e.g. protein, gene, organism, cell (or cell type)
- Sample: e.g. fixed, wax-embedded tissue section, blood sample, bacterial culture
- Analytic platform/technology: e.g. immunohistochemistry, next generation sequencing, immunoassay
- Location where the test is performed e.g. central lab, doctor's office, patient's bedside, patient's home.

Diagnostic testing needs to be used in an optimal way in order to fully unlock the value of targeted therapies. Optimization of a test may involve a number of factors such as quality, process and test performance which may impact different tests in different ways. These factors may include:

Physician knowledge and awareness: Physicians order diagnostic tests. However, if physicians never learn about or understand the rationale, value and availability of a diagnostic test, or if they assume that payers will refuse to cover the test, they may refrain from ordering it.

Questions about test quality: Physicians may be unconvinced of the quality of test results. Labs may lose samples or ask for larger samples. Two- to-five percent of lab results are returned with the notation QNS, meaning "quantity not sufficient." This translates into a demand for retesting and putting the patient through another biopsy.

Sample availability: Physicians may not be willing to bring a patient in for a second biopsy. Instead physicians will weigh the advantages of a first, second or even third biopsy against the disadvantages of putting a patient through added discomfort and inconvenience. A biopsy also heightens the risk of tumor cells spreading to other locations, so physicians may be hesitant in some instances to perform multiple biopsies.

Test availability: The majority of labs are concentrated in large urban centers. However, test availability also depends on the test's technical features. While a simple stain of tissue is widely available across the U.S., more esoteric next-generation sequencing-based tests may only be available in only one lab. For example, the CCR4 test for Maraviroc has a turnaround time of about 12 weeks. If physicians want a test called OncotypeDX they must send a sample to a single lab and wait for results by fax or e-mail.

Test turnaround time: Labs with exclusives on tests can apply economies of scale and ensure test quality. However, they sometimes struggle with bottlenecks that delay test results and treatment decision making. If a clinician needs to treat a patient within three-to-four days, but gets results back in 13 days, the patient will suffer the consequences. For example, patients with Acute Myeloid Leukemia (AML) are often gravely ill. If a physician has to wait more than 72 hours for a test result, the patient's life may be at risk, so treatment decisions may have to be made in the absence of an important diagnostic result that could indicate which treatment may be most effective. In some cases, clinicians never use a treatment—or don't use a treatment quickly enough—due to problems in lab turnaround times.

Test complexity: A growing number of companies offer test/algorithm combinations that involve examination of a 350-to-500 gene signature and application of a proprietary algorithm to interpret the patient's prognosis or outcomes of treatment. The algorithm becomes as important as the test because it mines evidence-based insights out of the data. However, the complexity of the test may lengthen turnaround times.

Reporting complexity: A related issue is reporting complexity. Complex results on multi-gene signatures may lead physicians to misunderstand what the test results mean, preventing them from properly interpreting, analyzing, and acting upon diagnostic test results. Some labs perform large-scale genomic analysis with complex algorithms. However, lab reports sometimes fail to provide adequate guidance and actionable information on which therapies physicians should prescribe. Interpretation of results is often left to the physician, which may be confusing. Some labs that do PD-L1 testing only perform the technical component while leaving the professional component of interpretation to the ordering physician. With a number of different assays available, and differences in indications and cutoff points for the various PD-L1 therapies, just getting a result may not be enough to make the correct decision for a particular patient.

Reimbursement: While reimbursement for diagnostic testing exists, it may be insufficient and will vary by test. In the case of more routine diagnostic tests like surgical pathology or immunohistochemistry, CPT codes are in place and reimbursement is a routine process. However, with some mid-scale genetic tests like BRAF and genetic mutation, the payer requests pre-authorization. The lab must ask the payer to determine the appropriateness of the test given the patient's diagnostic codes.

Reimbursement challenges every stakeholder. New diagnostic tests may not be reimbursed for 18 months-to-two years, while payers deem other tests "investigational" and "not medically necessary." If a payer denies reimbursement, and the physician fails to make a case for the test, the bill is left with the patient. And while pharmaceutical companies pay for a drug, depending on the market, they may not be able to subsidize a test.

Issues and financial impact: All of these problems come with economic consequences. A Diaceutics analysis reveals that inaccurate, delayed or incomplete test results for 13 cancer biomarkers prevent some 78,000 patients annually from receiving targeted treatments. Labs plagued with slow turnaround time and problems with sample adequacy and quality cause pharma companies to lose $8.3 billion in annual drug revenues, in the US alone.[9] (Figure 2)

Solutions to Accelerate and Ensure the Quality of Diagnostic Testing

Several actions promise to unblock the potential of diagnostic testing and address stakeholder interests, including increased clinician demand for testing, test expansion with improved quality, more predictable reimbursement, and greater transparency around test ordering and reporting of results.

Collaborate with organizations that can bring about innovation and transformation in precision medicine: The non-profit PM Connective was formed in 2014 to build a new model for personalized medicine at a disease specific level, linking advanced diagnostics with state-of- the art therapies. The PM Connective now has more than 300 collaborators in its network representing the diverse silos of healthcare—patients and their families, patient support

Figure 2

groups and foundations, individual physicians and nurses/caregivers, provider organizations and hospitals/treatment facilities, pharma companies, diagnostics firms, clinical investigators and research labs, payers, plan sponsors, and government regulators. The PM Connective accelerates collaboration of these companies and organizations around the financial and clinical outcomes of two diseases—malignant melanoma and early onset asthma – via the development, validation, and deployment of an integrative business model.[10]

Endorse the development of actionable insights on diagnostic testing: Stakeholders are eager to learn the costs associated with treatment strategies, including the role of a specific therapy within that treatment strategy, the conditions for which the therapy is used, and the likelihood of a positive or negative response or adverse event. The application of sophisticated analytics to testing and prescribing data can generate actionable insights on all aspects of testing and deliver the economic analysis required to document the cost and value of a therapy. Among the emerging data analytics trends are migration to cloud based analytics for fast results, increased scalability, cost control, enhanced security and strategy development; adoption of artificial intelligence, use of real-world evidence and anonymized patient-level data; and use of real time analytics.[11]

Ultimately, insights on the clinical and business performance of a therapy may surface as one component of a diagnostic test ordered by a clinician. While payers may prefer to postpone or restrict payment, diagnostic testing data from multiple sources will help ensure patients get the best treatment available. Publication of testing data from multiple sources, including clinical trials and real-world studies, will raise awareness of how and where clinicians should use a test, which will, in turn, drive demand for testing.

Promote the clinical and business performance of diagnostic testing via research and education. This step includes the design and implementation of prospective research and cost effectiveness studies through research partnerships and the dissemination of results both with payers, and with health care providers who shape demand for quality diagnostic testing. Education programs can help clarify the processes of test ordering while

serving as a platform to discuss the rationale, features, functions, benefits and results of existing and emerging diagnostic tests.

Support labs via technology, standards and guidelines: While companion and complementary diagnostics have moved forward considerably to support the significant wave of precision medicine therapies, enabling physicians, payers, labs and patients to reap considerable benefits, more work can be done. Labs and lab-related associations need expertise to design standardized procedures that will overcome the testing concerns outlined in this article—from tissue management and turnaround times, to test quality and results reporting. They also need support to select and implement technologies that will overcome time-consuming and costly tech hurdles.

Labs also need guidance in making data more useful to physicians, payers, consumers and other stakeholders. Interpretation of test results coming from the lab to the physician must be clear enough for clinicians to take effective action by treating or not treating a patient with a certain drug. Interpretation must also be clear enough for payers to understand what they are covering and reimbursing for the benefit of the patient.

Closing Thoughts

Going forward, diagnostic testing will continue to move therapy from one-size-fits-all to truly personalized medicine. Results will vary by patient, condition and therapy. Testing will determine both the course of treatment, likely outcomes and potential risk of adverse events. The goal is to create a data-driven balance that maximizes the benefits of diagnostic testing and treatment for all stakeholders—most importantly, the patients.

About the Authors

Dave Smart *is a biochemist with a PhD in neurochemistry and over 24 years of experience in the design, development, manufacture and commercialisation of diagnostics. He has worked extensively on modelling the impact of diagnostics on therapy in a variety of conditions and currently works as a Director for Diaceutics, specialising in the acquisition, analysis and understanding of diagnostic data as it relates to therapy utilisation and the patient journey.*

Steve Vitale *is an MBA with 20 years of experience in marketing throughout the pharmaceutical industry. He has lead commercialization efforts for a variety of new products and indications across multiple disease areas. As a Managing Director at Diaceutics, Steve has driven activity to help transform the industry model by enabling pharmaceutical companies to effectively commercialize and integrate diagnostic testing with targeted therapies, radically improving patient outcomes.*

Jeff Waldron, *MBA, has over 25 years of healthcare experience, working cross functionally in strategic, financial and operational capacities. He has developed expertise in reimbursement and access, while working to identify and resolve the potential barriers patients face. As Executive Director of the non-profit PM Connective, Jeff worked to better integrate the precision medicine business model across all stakeholders.*

References

1 Sagonowsky, E. Spark Sets Off Gene Therapy Debate with $850K Sticker on Luxturna. *FiercePharma*. 3, January, 2018. Washington, DC (Available from https://www.fiercepharma.com/pharma/spark-prices-gene-therapy-luxturna-at-850k-grabbing-top-spot-pharma-s-costliest-drugs)

2 Frost & Sullivan. How All Medicine Will Become Precision Medicine. 3, May 2017 Mountain View, CA (Available from https://ww2.frost.com/frost-perspectives/how-all-medicine-will-become-precision-medicine-5-evolving-trends-watch)

3 Schork. N. Personalized Medicine: Time for One-Person Trials. 25, April 2015, London, UK (Available from http://www.nature.com/news/personalized-medicine-time-for-one-person-trials-1.17411)

4 McLachlan, CYL, Yu, M, Ling, A, Jardine, JL. Adverse Drug Events Are a Major Cause of Acute Medical Admission. *Internal Medicine Journal*. 7, July 2014 (Available from http://onlinelibrary.wiley.com/doi/10.1111/imj.12455/abstract).

5 Baudhuin. L. The Real Cost of NGS. The Pathologist. January 2017. Knutsford, Cheshire, UK (Available from: https://thepathologist.com/issues/0117/the-real-cost-of-ngs/)

6 Keeling, P, Waldron J. A New Disease Level Precision Medicine Model, Starting with Melanoma. *Journal of Precision Medicine*. 2017 (cited 27 November, 2017) May, 2017. (Available from: http://www.thejournalofprecisionmedicine.com/archive-manager/a-new-disease-level-precision-medicine-model-starting-with-melanoma/)

7 Kaiser, J. Liquid Biopsy Promises Early Detection for Cancer. *Science*. 18, January 2018. Washington, DC (Available from: http://www.sciencemag.org/news/2018/01/liquid-biopsy-promises-early-detection-cancer)

8 Myshiko, D. Pharma R & D Is Now Biomarker Driven. *PharmaVOICE*. January 2016. Titusville, New Jersey (Available from: http://www.pharmavoice.com/article/2016-01-biomarkers/)

9 Clark, J., Smart, D., Keeling P. Patients Lost to Precision Treatment Due to Suboptimal Testing: US and EU Real World Evidence. First European Alliance for Personalized Medicine Conference. Personalising Your Health: A Global Imperative. 27-28 November, 2017. Belfast, UK (Available from http://eapmbelfast2017.com/wp-content/uploads/2017/11/EAPM_Programme-ProgrammeFinal.pdf)

10 Diaceutics. A New International Collaborative Has Been Formed to Help Remove Barriers Limiting the Promise of Precision Medicine. 7, August 2016. New York, New York (Available from: https://www.diaceutics.com/a-new-international-collaborative-has-been-formed-to-help-remove-barriers-limiting-the-promise-of-personalizedprecision-medicine).

11 Packer, S. Projected Pharmaceutical Data Trends 2018. *DataScience*. 25 January 2018. Research Triangle Park, North Carolina. (Available from: https://www.rxdatascience.com/blog/projected-pharmaceutical-analytics-trends-2018)

Estimating Sales Force Efficiency in the Pharma Industry

Ashish Ranjan Jha, Managing Principal, Analytical Wizards

Abstract: The field force continues to be the most important channel of promotion for pharmaceutical companies. Every year manufacturers earmark a substantial portion of marketing budgets for sales force visits to physicians. Tighter regulatory environment, restricted access to physicians and diminishing returns on detailing mean that the marketers are under pressure to improve the efficiency of these sales force visits. Knowing the efficiency levels at which medical representatives are operating is a good first step towards giving that extra boost to portfolio sales. Training and hiring efforts are more focused and effective when the strongest and the weakest links in the sales force are known to marketers. Given the constant pressure on cost, it might be prudent to direct efforts towards improving the efficiency of the existing workforce rather than simply increasing the size of it. Stochastic Frontier Analysis is a potent technique not only to measure efficiency but also to gain insights on ways to improve it.

There is a great deal of variation in the efficiency levels of sales representatives. Those in the higher quintiles of efficiency have 1.5 to 4 times higher efficiency than those in the lower quintiles on average.

Keywords: sales force efficiency, field force efficiency, Stochastic Frontier Analysis

Business Context

A leading pharmaceutical company was looking for ways to improve its sales. Amongst other options, it decided to review performance of its sales force team. Efficiency at the sales representative level was chosen as the metric to be evaluated, as was any difference in the efficiency level of representatives in different tenure bands. In other words, it wanted to understand how the efficiency of a sales representative changed as it tenured within the organization. For this study, business team clubbed sales force representatives in three different tenure bands viz. 0 to 3 years, 4 to 9 years and greater than 10 years. They wanted to depute their training resources where it mattered the most.

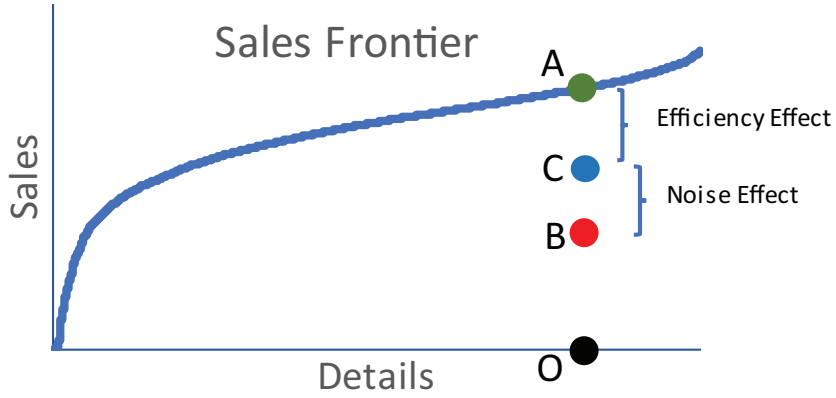
Data Description

The project required cross-functional collaboration between client teams. The Sales department and the Human Resources Development (HRD) department worked closely to ascertain the objective of the study and then to collate data accordingly. Data included sales operations provided with monthly brand sales, market sales, number of samples distributed, and calls data on each sales representative. HRD was provided with the tenure of the representatives within the organization. Data was limited to one country, one brand and was in a one-year time frame.

Challenge

Commonly used metrics for sales force efficiency include the total number of calls, number of calls per day, time spent per call, etc. But these are all surrogate measures that have nothing to do with the efficiency of sales representatives. Moreover, these measures do not take other factors into account e.g. market size, other forms of promotion, etc. Therefore, there is a need for a robust methodology which takes a comprehensive view of efficiency.

Figure 1



A benchmark is needed to measure efficiency. Standard ordinary least square regression techniques can be used to fit a curve between sales and input parameters. This regression curve can then be used as a benchmark. But a regression line passes through the mean and, therefore, it is a representation of average performance rather than the best performance. Moreover, regression measures each deviation from this fitted line as random noise. But, in the context of brand sales, we know that the deviation is caused by both random noise and the efficiency of the sales representative.

Another common methodology for measuring efficiency is Data Envelopment Analysis. Unlike least square regression, it measures efficiency with respect to the best performer. One problem with this approach is that it treats each deviation from the fitted line as a measure of inefficiency. But we know that deviations from the best frontier line have two components – one is obviously inefficiency, and the other is random noise in behavior which is inherent in any process.

Solution Design

Efficiency can be defined as the ratio of observed output and maximum possible output for a given level of input. But, considering the stochastic element of sales, this ratio should be calculated after accounting for the random noise component.

Refer to Figure 1. Point A lies on a curve, which is called the sales frontier. The sales frontier is the line that passes through the maximum achievable sales at all levels of possible inputs. Therefore, it can be said that point A is efficient if random noise has already been accounted for. Point B, on the other hand, is inefficient because it is lower than the sales frontier. Since sales has a stochastic component we should account for the random noise too. Point C is the level of sales post adjustment for random noise. Now, BC is the random noise effect, whereas CA is the inefficiency effect. Therefore, if noise effect can be segregated from efficiency effect, then efficiency can be derived as OC/OA. It turns out that the Stochastic Frontier Analysis incorporates both random noise and efficiency in the model separately, and therefore, it is quite suitable for estimating efficiency of the sales force team.

Solution Development

As mentioned above, Stochastic Frontier Analysis incorporates both efficiency and the random noise in the model. It is a parametric method which means one can model any functional form between the output, which is

sales in this case, and an array of inputs such as promotion activities and market size. Moreover, it treats the best performers as the benchmark and estimates efficiency of the rest with respect to this benchmark.

Stochastic Frontier Analysis subtracts a non-negative random variable, a measure of inefficiency in the process, from the standard production function and then employs a maximum likelihood estimation technique to estimate model parameters. The model can be represented by the equation below (see *Introduction to Econometric Production Analysis with R,* Arne Henningsen, 2015):

$$Actual\ Sales = f(x, \beta) * e^{v} * e^{-u}$$

Where,

$v \sim N(0, \sigma_v^2)$

$u \sim N^{+}(\mu, \sigma_u^2)$

While v can take any value on number line, u is a non-negative number in this equation.

This u is a measurement of inefficiency. Once this inefficiency term has been estimated for each sales representative, then one knows the overall efficiency at which the sales force is operating and the zones of inefficiency in the overall pool.

Sales can be assumed to be related with inputs in a Cobb-Douglas functional form. In other words, log of sales can be modeled as a function of log of inputs viz. promotional activities, market size, competition activity, etc.

Since it is a parametric approach, one can assume any functional relationship and model for any business assumption, so long as the functional relationship remains linear in parameters. For example, there may be territories with a low level of brand share historically. One would expect efficiency of sales representatives assigned to these territories to be low. In this study, after a sales force restructuring exercise, the sales team had assigned new sales representatives to such territories and wanted to make sure that new representatives were not penalized with low efficiency for traditionally low levels of brand sales in their respective territories.

Two approaches were proposed to accommodate this in the model. First, it was proposed to add prior year's sales to the model to act as a control for the historical level of sales. Alternatively, it was proposed to model panel data at sales representative and month level, instead of cross-sectional data, and evaluate efficiency for each month for each representative. The latter approach would allow the business to see if efficiency of a representative has been on an increasing or decreasing trend over time. With much deliberation, the company saw value in the second approach and therefore, panel data was modeled.

The final modeling equation was

$$Actual\ Sales = \beta_o * Market\ Sales^{\beta_1} * Details^{\beta_2} * Samples^{\beta_3} * e^{v} * e^{-u}$$

Model Result

A high level of variation was observed in the efficiency of sales representatives as derived from the model. Those in the highest quintile were approximately 2.5 times more efficient than those in the lowest quintile of efficiency. (Figure 2)

Almost a third of the territories with high market size were found to be served by those at below average efficiency level. This group was operating at an efficiency of only 0.43. There was a good opportunity to increase sales just by pushing efficiency of the sales force operating

Figure 2: Efficiency by Quintile

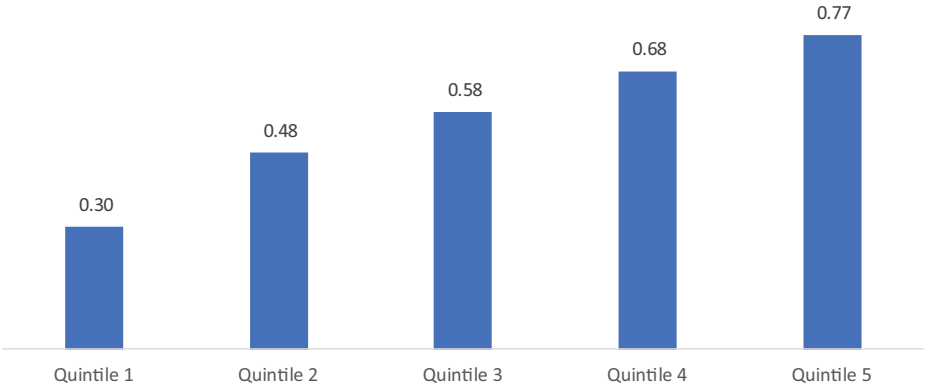


Table 1

		Efficiency	
		Low	**High**
Market Size	**Low**	0.40	0.68
	High	0.43	0.69

in this market segment. Analysis revealed that even a minor 10% increase in efficiency of this section of representatives could boost overall sales by more than 1%. This was seen as a big opportunity by the business team. (Table 1)

Though some decline was observed in the efficiency of representatives as they tenured within the organization, the difference was not found to be significant.

Conclusion

Given the prevailing market conditions, it is imperative for pharmaceutical companies to increase the return on investment they make on sales force visits. Stochastic Frontier Analysis is an effective technique to understand not only the overall level of efficiency at which the workforce is operating, but also to detect the zones of inefficiency and the underlying reasons driving that inefficiency. It models sales with all relevant input factors using the functional form of choice. Thus, it leaves a lot of control in the hands of users to model the type of relationship between sales and input factors that are in sync with business hypotheses. Unlike regression and data envelopment analysis, Stochastic Frontier Analysis assumes that any deviation from the most efficient sales frontier is due to both random noise and efficiency. And, therein lies its strength.

About the Author

Ashish Ranjan Jha, Managing Principal, Analytical Wizards, has years of experience in marketing analytics space serving global clients in Pharmaceutical and CPG industries in areas such as marketing mix optimization, sales force efficiency improvement, understanding and predicting customer behavior, personalized marketing, etc. He is skilled at utilizing statistical modeling and machine learning techniques to solve diverse business problems.

References

1 Henningsen, Arne. Introduction to Econometric Production Analysis with R. 2015.

Applying Real-World Evidence Data for Measuring Pharmaceutical Digital Media Programs

Ira Haimowitz, PhD, Vice President, Product Strategy, Crossix and Whitney Kemper, Sr. Director, Analytics Products, Crossix

Abstract: Pharmaceutical marketers can now leverage advanced clinical metrics found in real-world evidence data to:

1. Better understand the behaviors of audiences exposed to relevant media;
2. Improve digital DTC campaign optimization; and
3. Measure the direct-to-consumer incremental (exposed vs. control) impact on the patient journey.

This is demonstrated through specific case studies in the type 2 diabetes therapeutic category, showing enhanced insights and improved modeling accuracy.

Keywords: marketing optimization, real-world evidence, promotion evaluation, media measurement, health informatics, pharmaceutical marketing, digital media

Background

In recent years, improved technology and access to electronic health records (EHR) systems have enabled real-world evidence (RWE) data to be used for a range of healthcare applications. This data includes patient vitals captured during doctor visits, laboratory tests, and medical exam notes that describe symptoms and treatment responses in free text. The most common applications for RWE data are selecting institutions for clinical trial sites, modeling patients for clinical trial recruitment, and outcomes research measurement.[1,2]

More recently, RWE data has been applied to measuring direct-to-consumer (DTC) marketing programs for pharmaceutical brands. New audience quality measures for digital, television, point of care, and print campaigns based on RWE data have been developed.[2,3] For example, a type 2 diabetes brand team can now evaluate what percentage of those reached during their media campaign have elevated A1C levels indicating high blood sugar.

As an extension, this article describes two advanced applications of RWE data for media measurement. In the first application, in-depth insights are highlighted across key milestones along the patient journey and the effect of digital media exposure on patient decisions and timing. In the second application, RWE metrics are aggregated at the media referral source level to be used for a predictive model that forecasts future media performance in driving new patient conversions to brand, which is often brands' key media objective.

Digital Media Measurement and Optimization in Pharmaceutical Marketing

Digital marketing is a critical component of pharmaceutical marketers' DTC advertising campaigns. According to Kantar Media, pharmaceutical spending on digital ads was $515 million in 2016, roughly 10% of total media spend.[4] The analyses in this article focus on DTC digital display, mobile, and online video advertising that can be delivered through desktop, tablet, mobile web, or mobile

app platforms. With this focus, there is still significant spend; a single branded digital display and online video media campaign may entail an investment that ranges from $1 million to $5 million. Although this article focuses on digital display and video, there are similar measurement solutions that can be applied to evaluate paid search, website evaluation, and digital point of care, as well as healthcare professional digital media.

Over the past decade, the mix of digital publishers has become increasingly diverse and complex. Ten years ago, healthcare brand marketers usually funneled the bulk of their digital ad spend to endemic medical publishers that provide consumers with trusted sources with detailed health information by therapeutic category. Today, endemic publishers have also diversified themselves and now include niche patient disease-specific communities and websites that enable consumers to search for physicians, make appointments, and rate doctors' services. Digital ad networks are more diverse sets of websites for purchasing display advertising.

However, the true rise in display advertising has come from programmatic digital advertising, in which automated targeting algorithms and online auctions determine which consumer is served an ad. With programmatic capabilities, healthcare advertising can be served to the most relevant consumers anywhere they visit the internet, even on news, sports, or entertainment websites. Informative overviews of programmatic techniques can be found at the Interactive Advertising Bureau.[5] Specific providers have advocated that programmatic media is compatible with pharmaceutical privacy concerns and offers benefits of efficiency.[6] An analysis of a representative sample of digital campaigns showed that over two-thirds of digital impressions are generated from programmatic media.[7]

Given the breadth of DTC digital channels and tactics now available, healthcare advertisers want to know if they are spending their money wisely, and there is an increasing premium on media measurement and optimization. In the early 2000s, digital ad measurement was primarily assessed in two ways:

- Online primary research surveys of website visitors and media viewers: this was limited by long project durations where sufficient research sample accumulated, as well as its reliance on self-reported data rather than actual health transaction outcomes.
- Online "engagement" measured as click-through rates of digital ads and subsequent website visitor behavior after clicks: this was limited because only a small percentage of media viewers actually click on ads, and it uses rough estimates linking website activity to health outcomes.

More recently the pharmaceutical industry has changed its approach to addressing this need.[8] No longer is digital ad measurement based on analyzing click-throughs to websites. Instead, there are new methods that directly link media exposure to health metrics that are more closely aligned to brands' key marketing objectives, like reaching a qualified audience and generating new patient starts on their brand. In this way, digital measurement is now compatible with that of other prominent DTC media channels like television.

The health metrics that were evaluated for digital advertising, and all media channels, fall into three primary classes:

- **Audience quality:** the percentage of a media-exposed consumer audience that is relevant from a health history perspective. This includes prior diagnosis with the relevant medical condition, previous

treatment with medications in the relevant category, or prior results on relevant vitals or laboratory tests.

- **Intent to treat:** the percentage of consumers exposed to media that proceeds to visit a primary care physician and/or a relevant specialist within a one-month time frame after exposure.

- **Conversion:** the percentage of consumers that are exposed to media and later begin a new treatment for the advertised brand or within the relevant condition category (or sub-category) within three months after exposure.

Importantly, these metrics are evaluated net of a control group, matched by age, demographics, geography, and prior patient treatment history, in order to determine the incremental value driven by exposure to the campaign.

It is not sufficient to calculate these metrics in a one-time study at the end of a media campaign because a dynamic media channel, such as digital, requires ongoing monitoring and optimization. To optimize media spend, pharmaceutical marketers require two main components from their measurement system:

- **Granularity of Reporting:** measurement of results not only at the overall digital campaign level, but separately for each publisher. Instead, metrics *within each publisher* at the level of *placement groups* were measured. These are sets of display or video ad locations that share commonalities of targeting, format (mobile or desktop), and content theme. Placement groups are ideal for enabling marketers to maintain their spend with a critical publisher and maximize their impact on quality reach and post-media behavior.

- **Frequency of Measurement:** repeated evaluation and specific detection of trends, outliers, and market shifts. In digital measurement, all metrics were measured *weekly* as a standard.

Data Sources for Both Studies

For both studies, comprehensive distributed data network technology was leveraged, which includes clinical data (EHR, medical claims data, prescriptions, etc.), frequent shopper loyalty card data for OTC and packaged-goods purchases, consumer data (demographics, financials, interests, etc.), multi-channel media, and other data sets covering over 250 million U.S. consumers. A sample list of data attributes is shown in Table 1. Special attention was given to diagnoses from medical claims and lab results from EHR data.

Media data, in particular, was gathered using a tag-based implementation that tracks impressions from digital display, video, and mobile advertising. Media data is gathered at the overall campaign level, publisher level, and placement group level within publishers. Digital media impressions are matched to digital identities, which are anonymously matched to health data in a HIPAA-compliant way.

Each category of data attributes provides distinct benefits to digital measurement. Leveraging health data provides marketers with more insight into the consumers they reach through their campaigns and ensures that the right ads are reaching qualified audiences who are in line with the brand's objectives. Shopping data can be used to indicate conditions where OTC products are being used as alternatives or as supplements to prescription therapy—like in the case of allergies. Additionally, many health conditions are often treated with concomitant diet modification (e.g. low salt for hypertension, sugar-free for diabetes, and gluten-free for

Table 1. Data Attributes (U.S.)

Rx/Medical Claims/EHR Health Behaviors	Shopping	Consumer	Media
250MM+ patients Virtually all practitioners Updated daily	CPG: 90MM+ house-holds Food and Drug Products Updated daily	240MM+ U.S. adults 2,000+ different variables Updated quarterly	Multi-channel Consumer & HCP Updated daily, weekly, monthly
Patient • Age, gender, geo **Rx** • Date filled, product, quantity, refills **HCP** • HCP/prescriber/specialty, location, office visit dates **Diagnoses** • Diagnosis codes, lab orders & results **Payer & Cost** **Pharmacy Type**	**Item** • Date, product UPC, quantity, price **Store Type** **Shopping Basket** • Basket Size • Trips	**Demographics** **Financials** **Interests & Hobbies** **Media** • Propensity to buy over certain channels – internet, mail, phone, cell phone, magazine, TV, etc. **Buying & Shopping Activity** • Amount of spend in certain categories	• TV • Digital • Print • POC • CRM • Email • Direct Mail • Sales Calls

celiac disease). Consumer data allows marketers to evaluate media targeting based on demographic objectives. For example, a contraceptive brand can target younger females, whereas an erectile dysfunction brand can target middle-aged males. This reduces media spend waste by only focusing on audiences that are clearly part of the intended audience. Media data for these analyses is accumulated through a tagging-based system, which captures time-stamped exposure to advertising at the most granular "placement" level. This data is collected in a HIPAA-compliant, privacy-safe manner, and is ultimately matched to consumer health transactions behind firewalls of data providers. Combined, these data sources provide a powerful foundation to optimizing digital media campaigns.

Patient Journey

The goal of pharmaceutical marketing is to influence a patient to fill a prescription for an advertised drug. However, there are many steps a patient takes before he or she decides (or does not decide) to take the intended action. This series of steps is oftentimes referred to as the patient journey and is a key way to understanding patient behavior and how certain milestones influence the decision to fill a prescription. By tracking these stages, a marketer can better recognize where progress is being made and where there are roadblocks to the patient getting on therapy.

Using the example of a type 2 diabetes therapy, there is a clear progression that must be followed for the patient to fill a prescription. First, a patient must take the initiative to visit an HCP to discuss his or her health and broach

the topic of the advertised drug. For patients who are not already diagnosed with type 2 diabetes, an A1C test is usually administered. Depending on the results of that test, a patient may receive a new diagnosis of type 2 diabetes or prediabetes. Once a diagnosis has been made (or assuming a patient has already been diagnosed), the patient may be prescribed the relevant prescription or may get an earlier line therapy or a generic. In the latter case, there are follow-up visits, tests, and sometimes additional prescriptions.

Each touchpoint between the patient and HCP represents a step along the patient journey as well as an opportunity (or hindrance) to advance toward the brand marketer's ultimate goal. The effect of marketing on each step of the patient journey can be measured by comparing the actual health behavior of patients exposed to advertising to the behavior of those in a control group who are *not* exposed to the same advertising. In doing so, the marketer can understand the impact of marketing on patients' prescription filling behavior and where the marketing was effective (and ineffective) in moving the needle at each point in the progression of the patient journey. Additionally, marketing tactics can accelerate the timing for patients in taking their next steps along the treatment journey. This timing can be measured in days using transactional Rx, EHR, and medical claims data.

For example, a marketer may find that advertising is successful in getting patients to visit an HCP to discuss the possibility of them having diabetes, thereby increasing the rate of HCP visitation and the rate of patients receiving A1C tests and diagnoses. However, when it comes to the point at which an initial prescription is written (assuming the drug is a first-line therapy), patients may still receive prescriptions for competitor or generic drugs at the same rate as the control group. This would indicate that the advertising is effective in initiating a conversation with an HCP for new patients but not at influencing prescription-writing behavior. Had the marketer only considered the impact on conversion to their drug, they would have seen an increase in conversions but would have overlooked that this impact was happening due to more patients being diagnosed rather than an increasing share of prescriptions filled versus competitors.

The detailed descriptions of two research studies that use all of these data sets and employ the patient-level linkage between media exposure and post-media health behaviors including RWE data are below. These studies were also presented at the 2017 PMSA Annual Conference.[9]

APPLICATION 1: PATIENT JOURNEY POST-MEDIA EXPOSURE

Methodology

Over several type 2 diabetes-branded digital campaigns, key patient journey metrics were measured. Individuals exposed to digital advertisements were matched to EHR data with dates time-aligned relative to the dates of media exposure. Patient-level metrics related to doctor visits, lab tests, diagnoses, and ultimately treatment with a prescription were then calculated. These data preparation steps yielded a transactional data set of tens of thousands of patients, illustrated with a sample in Table 2.

Each record in this dataset represents a single patient in the study; columns are a combination of patient journey elements and timestamps, expressed as days relative to the first media exposure at day 0. For example, the first row demonstrates that the patient had a 7.2 A1C level 26 days before digital media exposure, without a formal type 2 diabetes diagnosis historically. Six days after media exposure, the patient visited a primary care physician, and 125 days (about 4 months) after media exposure, the patient visited an endocrinologist.

Table 2: Patient-Level Behavioral Data from the Patient Journey Analysis*

Pre-Exposure Last A1C Result	Pre-Exposure Last A1C Day	Pre-Exposure T2D Diagnosis Day	Post-Exposure PCP Visit Day	Post-Exposure Endocrinologist Visit Day	New Diabetes Diagnosis Day	Rx Conversion Day	Post Rx-Conversion A1C Day	Post-Rx Conversion A1C Result
7.2	-26	n/a	6	125	125	126	128	7.2
8.4	-7	n/a	56	76	74	67	74	8.2

*Data illustrative

At the endocrinologist, the patient received a diagnosis, and a prescription (brand omitted), which was filled the next day (day 126). The lab test result came back on day 128 and showed an A1C of 7.2 again.

Patient-level metrics were then aggregated and summarized at a population segment level to determine the overall patient journey behavior for exposed individuals.

In analyzing the patient journey, distinct paths were identified by tracking relevant healthcare behavior across all consumers exposed to media. Certain paths were chosen for further analysis based on how frequently they appeared in the data and if they were relevant to the brand. Given its nature, EHR data may not always capture 100% of a patient's longitudinal health behavior, which can lead to abbreviated or inconsistent paths for some patients. To prevent this from affecting the analysis, patient coverage and continuity eligibility rules were implemented for each patient to ensure that a complete, multi-stage path was being captured for each. Patients with insufficient data were removed from the analysis.

In order to measure the impact of the digital advertisements on this patient journey behavior, a control group was created to compare against the exposed treatment group. To do this, a group of individuals who were not exposed to the digital advertisements was identified. These individuals were similarly matched to EHR data. A matched pair control picking methodology was then used to identify patients who were similar to patients exposed to the digital advertising being analyzed. Patients were matched on both demographics (age, gender, geography) and healthcare characteristics (type 2 diabetes diagnosis, A1C level, Rx treatment, comorbidities).

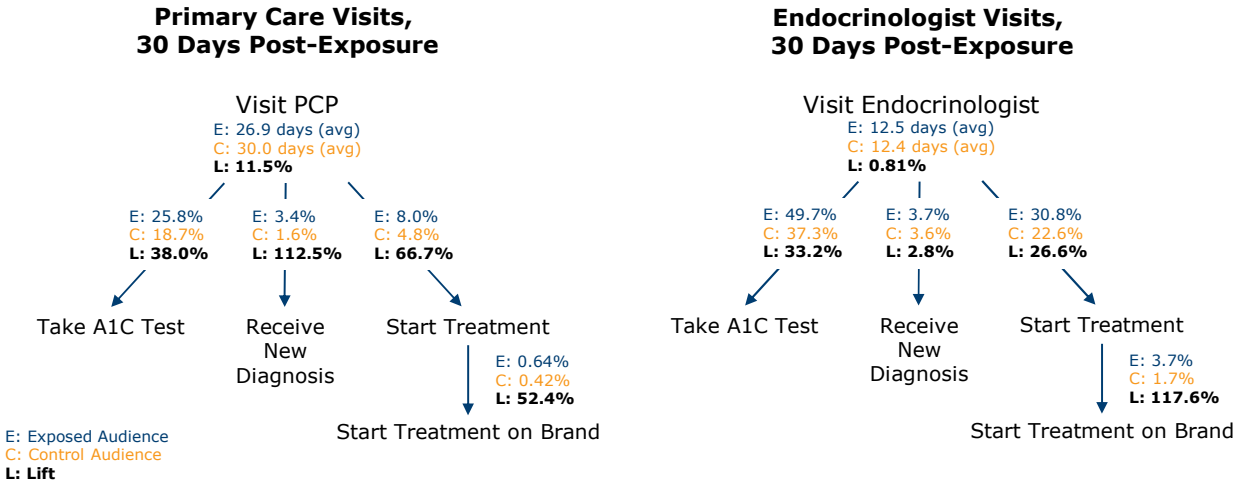
After selecting this control group, the same patient level metrics were calculated for this population. Population rates were then compared to those of the exposed group in order to measure impact. In addition to measuring the impact on patient journey behavior, this same data was used to measure the impact on conversion to relevant drugs being advertised.

Key Learnings

Through analyses like this, it was discovered that the addition of RWE data can dramatically improve the assessment of media campaigns. For example, there was a significant lift (in some cases over 100%) across the patient journey metrics for the audience that was exposed to the relevant digital media versus those who were not exposed. Figure 1 illustrates the detailed metrics regarding patient visits to primary care physicians and endocrinologists after being exposed to a diabetes medication campaign.

Consider for example, the leftmost branch of the Primary Care Visits tree of Figure 1. This

Figure 1: Media Exposure Impact on Actual Patient Pathways by Physician Specialty



indicates that those exposed to the digital media visited a PCP in 27 days, three days faster than those in the control group, which translates to a lift of 11.5% (note: the same timing difference was not detected for endocrinologists). Continuing down the tree, the exposed population received an A1C test at the PCP doctor visit 25.8% of the time compared to 18.7% of the time for the control group, a lift of 38%. Reviewing the middle two branches of the Primary Care Visits tree, there were also sizable lifts for the exposed population receiving a new diagnosis of type 2 diabetes, starting a new treatment, and starting a treatment on the advertised brand. Similar lifts were seen in test taking and Rx treatment for visits to endocrinologists by the media exposed population, as shown in the Endocrinologist Visits tree on the right.

APPLICATION 2: MEDIA-SOURCE PREDICTIVE MODEL OF FUTURE CONVERSION

Methodology

Similar RWE data has also been applied to improve forecasting and determine which digital media publisher would most effectively drive new patients to fill a brand prescription. A leading type 2 diabetes brand ran a digital advertising campaign across eight media publishers, which in turn were divided into 30 placement groups for the purpose of optimization. The marketers wanted to predict which of the 30 media sources would generate the highest end-of-campaign, new-to-brand Rx conversion using mid-campaign information about the publishers and the audience quality data of the 30 media sources.

In the middle of these consumer digital campaigns, publisher/placement group-level regression models were built using the variables illustrated in Table 3.

Each row in the table represents a placement group within a media publisher that was serving digital display or online video media for this advertising campaign. Two placement groups for an endemic healthcare publisher and a non-endemic programmatic publisher are illustrated. The columns represent specific metrics for that placement group at the mid-campaign milestone; the rightmost column provides the percentage of consumers exposed to that media source that converted to the advertised

Table 3. Regression Model Inputs

Placement Group	Endemic Flag?	Mid-Campaign Unique Reach	Mid-Campaign Log (Unique Reach)	Mid-Campaign Impression Frequency	Mid-Campaign Treating in T2D Category	Mid-Campaign Pre-Diabetes A1C Level Rate	Mid-Campaign Diagnosed T2D Rate	...	End Campaign Conversion-to-Brand Rate
Endemic Pub1 Desktop	1	123,027	5.09	5.2	12.40%	2.20%	18.40%		0.31%
Endemic Pub 1 Mobile	1	138,038	5.14	4.5	13.10%	3.00%	19.60%		0.24%
....									
Programmatic 3 Demo Target	0	1,148,154	6.06	12.9	10.20%	1.60%	16.40%		0.17%
Programmatic 3 Behavior Target	0	1,318,257	6.12	21.9	10.80%	1.40%	14.30%		0.13%

brand by the end of the campaign. For example, the first row shows that the desktop audience of Endemic Publisher 1 had a unique reach of over 123,000 consumers (log = 5.09) and an exposure frequency of 5.2. By mid-campaign, 12.4% of the consumers exposed via this media source were already treating on Rx in the type 2 diabetes category, 2.2% of this audience had a prediabetes A1C level, and 18.4% were already diagnosed with type 2 diabetes. By the end of the campaign, 0.31% of exposed consumers for this media source converted to the advertised brand within three months after their first digital media exposure.

The objective of the modeling at the placement group level across all media sources was to determine a predictive relationship between the mid-campaign media and audience quality attributes, and the end-campaign conversion-to-brand. This relationship would enable media planners to optimize among the various sources mid-campaign to select those placement groups most likely to have high conversion rates to brand.

Two models were evaluated:

- The first model used only publisher media delivery metrics and prescription history criteria of the media sources' respective audiences.
- The second model added additional RWE metrics, including the rate of taking the A1C lab test, uncontrolled threshold of the A1C level, and prior type 2 diabetes diagnosis.

Both models used a three-month lookahead rate of conversion to the advertised brand, modeled at the media placement group level. The two models were compared for goodness of fit as measured by an R-squared value. Performance was also evaluated for each model on a holdout sample of media sources.

Results

Comparison of the two predictive models (N=30) are summarized in Table 4.

This statistical model predicted future media-unit level conversion at the end of the campaign with a high model fit of R^2=0.69. This was a lift over the R^2=0.56 for a model that only used media-level treatment history. Likewise, the model that leveraged RWE data had a closer fit to the holdout sample. Additional details for the two models are illustrated in Figures 2 and 3.

Key Learnings

The addition of RWE variables to the regression forecasting model resulted in a significant

Table 4: Impact of Real-World Evidence Metrics on Predicting End-Campaign Media Conversion

	Baseline Predictive Model	Enhanced Predictive Model
Inputs (mid-campaign)	Treatment history on type 2 diabetes drugs Media impression and frequency levels	Treatment history on type 2 diabetes drugs Media impression and frequency levels A1C lab test rate Uncontrolled A1C rate Type 2 diabetes diagnosis rate
Outputs (end-campaign)	Conversion to new to brand Rx	Conversion to new to brand Rx
Goodness of Fit (R^2)	0.56	0.69

Figure 2: Original Model Predictiveness of End-Campaign Media Population Conversion

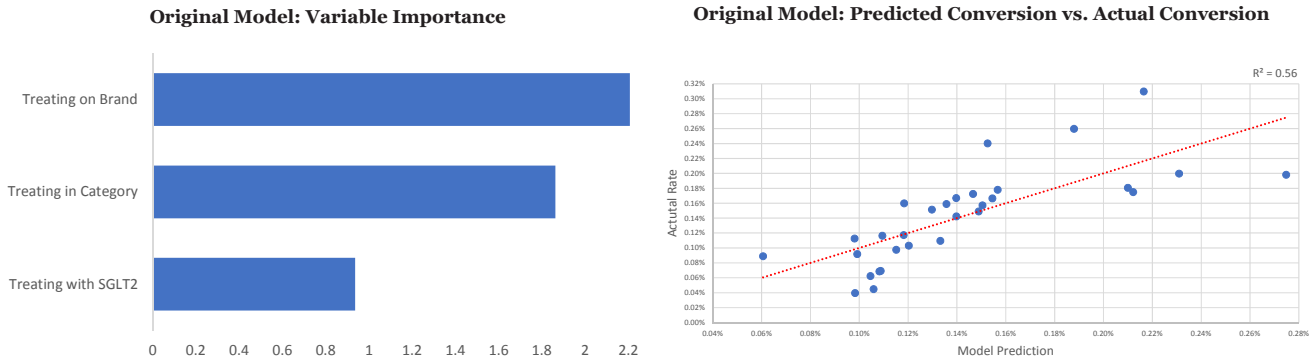
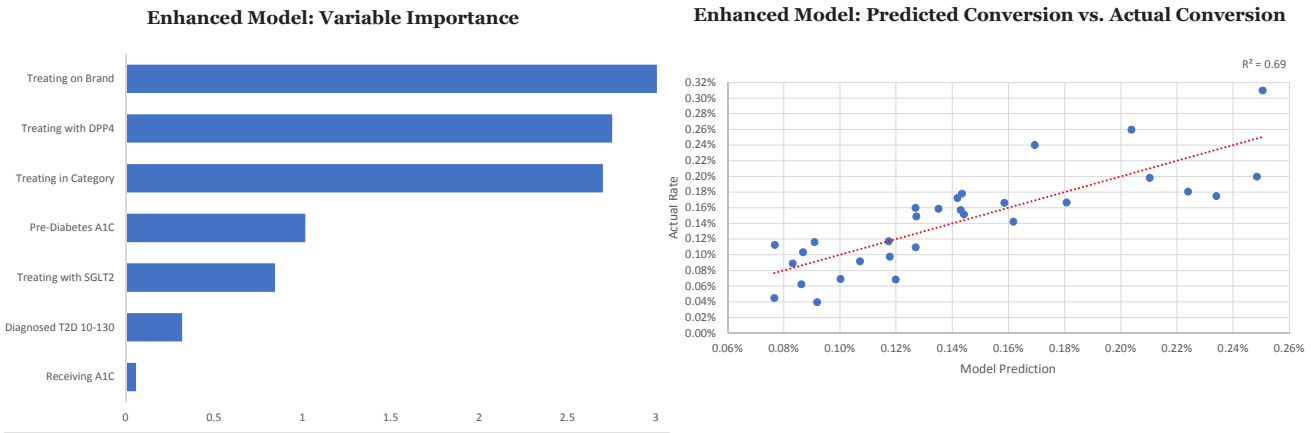


Figure 3: Enhanced Model Predictiveness of End-Campaign Media Population Conversion with RWE Data



increase in the goodness of fit in predicting end-of-campaign media-level conversion to the advertised brand. They were not quite as important in the model as prior treatment in category or brand but did show a significant contribution. The implication here is that healthcare digital advertisers assessing media performance in an advertising campaign should consider in-depth audience quality metrics, including whether their media sources are reaching patients with prior diagnoses and lab tests above key threshold values.

Conclusions and Future Work

These analytics approaches to evaluating consumer media campaigns are a significant advancement over previous approaches to campaign measurement in three primary regards:

- Direct linkage of media exposure to health behaviors, rather than surveys or website clicks.
- Real-world evidence data utilized for actual patient health history and post-media outcomes.
- Predictive power and frequent reporting for campaign optimization.

Using these insights, marketers have discovered the multi-faceted benefits of digital DTC campaigns, including a deeper understanding of how media exposure impacts patient behavior. They have also optimized these campaigns across different media publishers much faster and with more confidence. These studies are each based on the commercialized digital measurement platform called Crossix DIFA™ ("Digital Impact for Advertisers").

The authors are in the process of automating the predictive model analysis of media sources to a production capability. Then, any digital media campaign can be evaluated at a media source level (across publishers or placement groups) to get an early read on which media source will generate the most new patient prescriptions for the brand.

Similar approaches are available not only for consumer-focused media campaigns, but HCP-focused campaigns as well. Increasingly, pharmaceutical marketers are turning to non-personal promotion to complement sales force efforts. Much of this non-personal promotion is digital, with doctors, nurses, and other providers as the audience of the display, video, search, and email advertising. To meet this marketplace need, similar measurements have been developed for healthcare professional-focused digital campaigns that allow real-time optimization at the publisher level. Future research will investigate relationships underlying the interaction of media campaigns across both patients and healthcare professionals.

About the Authors

Ira Haimowitz, PhD, *serves as Vice President, Product Strategy, at Crossix. Ira helps accelerate new product innovation for Crossix's range of healthcare targeting and measurement solutions. Ira also enhances client delivery as a subject matter expert. He has more than 20 years of pharmaceutical (Pfizer and Organon), CPG, agency, and consulting expertise. Ira has published multiple articles, including authoring the book Healthcare Relationship Marketing. Ira received both his Ph.D. in Computer Science and B.S. degree in Mathematics from MIT. Ira is a Past-President (2006) and longtime board member of PMSA.*

Whitney Kemper, *serves as Senior Director, Analytics Products at Crossix Solutions, which he joined in 2009. He leads the development of Crossix's analytical frameworks as well as the implementation of new technology and data. Whitney's extensive experience with data and marketing analytics allows him to navigate the challenges of integrating and gleaning meaning from the wide array of healthcare and consumer data that Crossix employs. He holds a B.A. in Economics, English, and East Asian Studies from Vanderbilt University.*

References

1 Risser R, Saha S. Deep dive on real world evidence. Presentation at: PMSA Symposium on Real-World Data. San Diego. 2017 Jan.
2 Kumar K, Parikh J. Value proposition of diagnostic test data. *Journal of the Pharmaceutical Management Science Association*. Spring 2017, pp. 23-31.
3 Mittler J, et al. What does Crossix's latest expansion of connected health data actually mean for pharma marketers? *PM360*. 2016 Dec 20.
4 Lee J. Pharma's $6 billion annual ad spend targeted by media publishers. *Medical Marketing and Media*. 2017 Aug 21.
5 Interactive Advertising Bureau. Automation and programmatic [Internet]. 2017. Available from: https://www.iab.com/guidelines/programmatic-rtb/.
6 Neuner C. It's time for the healthcare industry to take advantage of programmatic. *Medical Marketing and Media*. 2017 Jan 22.
7 Haimowitz I. Five things you should know, but don't know about digital campaign measurement. Presentation at: Digital Pharma East Conference. Philadelphia. 2016 Oct.
8 Haimowitz I. Embracing the right metrics for digital success. *PM360*. 2016 June 20.
9 Haimowitz I, McKinley M. Enhancing pharmaceutical digital campaign measurement using real-world evidence data. Presentation at: PMSA Annual Conference. 2017 Apr.

Improving Accuracy in Rare Disease Patient Identification Using Pattern Recognition Ensembles

Tim Hare, Sr. Manager, CE-Analytics, Symphony Health Solutions; Pratibha Sharan, Senior Consultant, Commercial Effectiveness Team, Symphony Health Solutions; Ewa J. Kleczyk, PhD, Vice President of Client Analytics, Symphony Health Solutions; Derek Evans, Sr. Vice President & General Manager, Symphony Health Solutions

Abstract: Finding patients who are appropriate candidates for therapy has always been a primary goal for pharmaceutical marketing teams. However, reaching the right patient at the right time has never been more critical given the strategic shift toward highly specified, personalized therapies such as immunotherapy, gene therapy and rare disease markets. Reimbursement for these high-cost therapies typically requires extensive demonstration to payers that the patient meets clinical requirements. For pharmaceutical manufacturers, identifying these patients through traditional data sources alone is no easy task: many rare disease patients go undiagnosed or misdiagnosed for years, many rare diseases lack specific diagnosis codes, access to electronic medical records (EMR) is limited, and the results from lab or genetic testing cannot be seen in traditional reimbursement claims. However, by coupling the historic claims data of specific patient populations with machine learning techniques, an algorithm can be created to identify "high likelihood" future candidates. In this paper, we will review studies in which a model was used to predict which patients were potential candidates for orphan drugs to treat two very rare (< 5,000 cases of each known) diseases. In this disease state, patient identification is particularly challenging given the lack of definitive diagnosis codes and symptoms that mimic those of extremely common conditions. The full claims histories of patients receiving the therapy and randomly selected control patients were used to build, train, and test multiple predictive models (single tree, boosted tree, bagged random subspace trees, though any number of different algorithms might be suitable). Each of these predictive models reached high levels of out-of-sample positive predictive value (PPV) in distinguishing target patients from control. Two ensemble predictive models have been deployed which have identified patients at rates well above disease prevalence.

Keywords: rare disease, patient finding, advanced analytics, predictive modeling, healthcare claims data

Introduction

Appropriate rare disease patient identification represents a significant opportunity for both clinical and commercial stakeholders. While a single condition can affect fewer than 200,000 patients, it is estimated that 7% of the developed world's population suffers from one of 7,000 known rare diseases.[16] With 25-30 million affected individuals in the United States alone, a physician is likely to encounter at least one rare disease patient in their practice. These patients' path to appropriate diagnosis is long, with an average time of 7.2 years.[16] Minimizing the time to correctly diagnose these patients is critical, particularly for conditions in which there are therapies available to limit the progression or relieve key symptoms of the disease. Identifying highly likely patients prior to diagnosis can facilitate timely and targeted disease education efforts.

Examination of longitudinal patient level health claims data,[18,19,20] laboratory and EMR data, socio-demographic information, and linkage to physician attributes, enables a comprehensive understanding of a patient's

medical history – diagnosis, treatment and testing. In order to identify potentially undiagnosed rare disease patients, there are primarily two approaches: top-down searching [18, 21] and bottom-up predictive analytics. Clinical expertise can be leveraged to search for patients exhibiting known rare disease symptoms and presentations. Alternatively, the pre-diagnosis data of known patients can be utilized to build models that reflect underlying patterns in claims data.[8, 17] This approach removes human bias and accounts for the potential wide variation in patient presentation. High dimensional machine learning, in which computers learn without explicit programming from a very large number of variables, is well suited to this effort. However, further combining these machine learning algorithms into an ensemble, to offset weakness in any one approach, achieves a level of accuracy well beyond that of any of the individual ensemble member models. This article will review currently leveraged machine learning techniques for rare disease patient identification. It will also discuss case studies that demonstrate the value of such an ensemble approach.

Non-Parametric Machine Learning Complements Classical Parametric Statistical Approaches

The field of machine learning grew out of research in computer science,[4] and is defined as the development of "computer programs that automatically improve with experience."[2] These computer programs adapt during exposure to data, in effect learning from experience.[22, 23] Machine learning algorithms also learn which variables are important, and can be used with what would otherwise be considered to be a prohibitively large numbers of variables, known as high dimensional data,[13, 24] without any requirement that the investigator choose a subset. Importantly, non-parametric machine learning approaches such as those used here have no inherent form prior to exposure to the data. This reduces the bias in the final model.

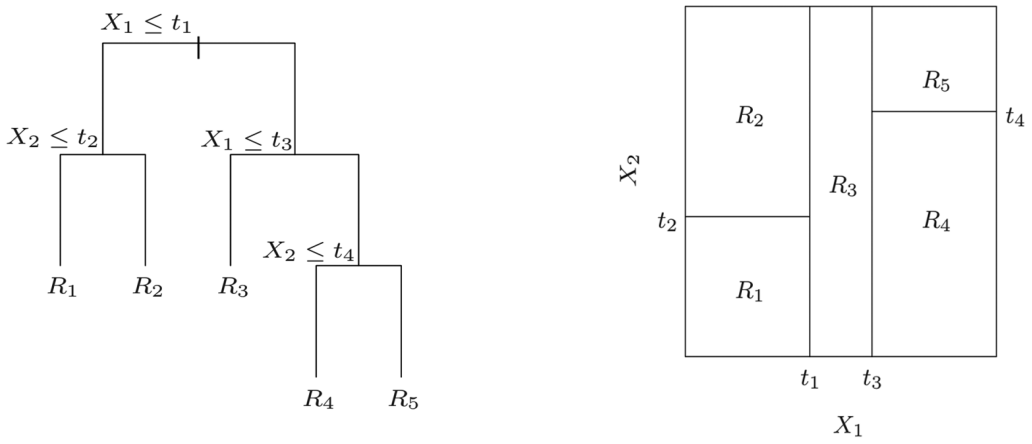
In contrast, classical statistical modeling relies on human agency in choosing a particular model form that fits the data best, as well as a reasonably small set of variables to be tested during modeling. The assumption is that the investigator has sufficient knowledge of the data to choose the correct model form and a small subset of variables that are likely to be useful. The model form is static, exists before exposure to the data, and does not change during exposure to the data. A lack of correspondence between inherent structure in the data and the chosen model form will result in poor modeling results, as will an incorrect or incomplete choice of variables to be tested.

For this reason, the non-parametric machine learning algorithms used here complement classical statistical approaches by offering additional flexibility as well as the ability to search directly in high dimensional spaces. However, a higher rate of over-fitting (the "bias-variance trade off")[3] can reduce model performance. Algorithm-specific error reduction methods ("regularization")[3] are often leveraged to reduce over-fitting. However, we proposed a different approach for rare disease patient identification, a simplified form of regularization using ensemble agreement.[3,5]

Evaluation of an Ensemble Approach to Machine Learning

Given patient rarity, large sales territories and promotional effort costs in the rare disease space, it is desirable to reduce false positive (FP) rates as low as possible. Precision or positive predictive value (PPV = TPR/(TPR+FPR))[1] can be increased when the true positive rate (TPR) is much smaller than the true negative rate (TNR). For example, if a population has a prevalence of TP of 0.1%,

Figure 1: Tree Algorithms



there are a very large number of TN: 99.9% TN, or 999 per 1000 patients evaluated. If a single model class error for the TN class is 25% (e.g. our FPR), and has a 0% FNR, precision equals 0.003984 or ~0.4% [1/(1+250)] despite this low FNR. That is, ~4 true patients for every 1000 patients screened can be expected on average.

Going beyond a single best model approach, we can leverage ensembles to compensate for the patient classification bias of any one algorithm. For example, under the simplifying assumption that each model misclassifies true positives independently, and given a false positive probability of P=0.25 for each model, the ensemble probability of 3 models making a misclassification error is $(0.25)^3$ or 0.0156, when all 3 models are required to agree. Our hypothetical ensemble-level false positive rate is now 1.56% and yields a positive predictive value of [1/(1+15.6)] = 0.0602 or 6.02%. Obviously the error rates may not be fully independent, and the precision gains will depend on the data; however for illustrative purposes, we see a ~15-fold higher ensemble PPV relative to single model PPV of 0.4%. This improved precision necessitates a higher false negative rate, a trade-off that must be considered when utilizing the ensemble approach.

Algorithms should be selected for participation in the ensemble so as to complement weaknesses. In the ensemble case studies described here, Tree, Random Forest, and AdaBoost algorithms were chosen for the reasons outlined below, as well as for computational efficiency on large, high dimensional data sets.

Tree Algorithms

Tree-based algorithms such as the RPART algorithm used here,[25] partition the feature space (independent variables) into cuboid regions by a series of binary splits, and assign constant values to all members of that region. A simple two-dimensional (two features: X1 and X2) tree is shown in Figure 1, along with a corresponding partition of the feature space (from Hastie, Tibshirani, Friedman).[11]

Tree models operate on a data set defined by N observations where each observation (xi, yi) for I=1,2,…,N is associated with P features, such that xi = (xi1, xi2, ….,Xip). Tree algorithms result in a piecewise-constant[10,11] functional representation. It can be helpful to consider the final functional form along-side the machine learning approach that produced it. The functional form for a tree model is shown below.

$$f(x) = \sum_{m=1}^{M} Cm\, I(x \in Rm)$$

A single tree can be viewed as composed of many models, or sub-models, called "nodes", with the particular model applied to any input, xi, being the rule set associated with the constant, Cm, for that region, m. For example, the region m=R1 in the figure above would be the model defined by X1 <= t1 and X2 <= t2, and the value Cm associated with that region would be the result f(x) associated with all points Xi that satisfy the following:

$$Cm\_hat = average(yi \mid xi \in Rm)$$

Tree-based algorithms can rapidly model large amounts of data (many observations) given its computational efficiency. They are also robust to high dimensional data (many features, where P >> N) because the algorithm only considers a single feature at each split within the overall feature space and as such is agnostic to the dimensionality of the data.

Error minimization for each region is implemented by defining two sub-regions parameterized by the variable under consideration (j) and the value of that variable to perform regional splitting (s), as shown in the definition below.

$$R1(j, s) = \{X \mid Xj \leq s\} \text{ and } R2(j, s) = \{X \mid Xj > s\}$$

Penalty function minimization takes place as a minimization of the sum of the two sub-regions, as shown below.

$$\min_{j,s}\left[\min_{c1} \sum_{xi \in R1(j,s)} (yi - c1)^2 + \min_{c2} \sum_{xi \in R2(j,s)} (yi - c2)^2\right]$$

The two interior minimizations of mean square error (MSE) are trivial in that it is the constant assigned to that region - the average for each region (c1 and c2).

$$c1\_hat = average(yi \mid xi \in R1(j, s)) \text{ and}$$

$$c2\_hat = average(yi \mid xi \in R2(j, s))$$

Strengths of the recursive binary tree algorithm include the conditionality of each node, the ability to analyze high dimensional data, and computational speed. Weaknesses include splits based only on the single best variable, sampling bias, and over-fitting.[3] Additionally, regions must be rectangular cuboid, eliminating non-rectangular, irregular convex, and concave regions from consideration. The trade-off between under-fitting versus over-fitting the data can be controlled within an individual tree by empirically selecting good parameters such as minimum group size to split, maximum number of levels, etc. The onus is on the data scientist to determine optimal settings through a series of experimental runs with checks against out-of-sample (OOS) data (data not used in developing the model).

Ensemble Regularization Using Tree, Random Forest, and Boosted Tree Algorithms

As noted above, individual trees can have high error rates, especially if optimal parameter settings are not experimented with manually. To reduce these error rates as part of a turn-key ensemble regularization approach, we look to find algorithms that are not likely to make errors in the same way as single tree algorithms do, while maintaining our focus on algorithms that can work directly with high dimensional data, and offer sufficient computation speed on large data sets. Random Forest reduces variance (random error) and bias (systematic error)

Figure 2: Confusion Matrix

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

by minimizing correlations between the tree variables and averaging over many trees. The process of producing Boosted trees, unlike the algorithms above, leverages individual record error tracking during the tree growing process. The motivation for combining these particular approaches as a form ensemble regularization is driven by the recognition that each algorithm has a distinct bias toward certain types of errors, which can be offset using ensemble regularization.

Results and Pertinent Case Studies

The results and case studies below will be described utilizing the follow terminology: Receiver Operating Characteristic (ROC) curve, out of sample (OOS) testing, confusion matrix and positive predictive value (PPV). A ROC Curve is plot of the true positive rate against the false positive rate in a classification model. A ROC plot shows the relationship between the two, as a function of the stringency of the classification threshold. Out of sample testing refers to data not used in training a model that is useful in gauging the performance of the model. A confusion matrix is a table that describes the performance of a classification model for a set of test data in which the true values are known (Figure 2).

PPV is equal to the ratio of true positives divided by the sum of true positives and false positives [TP/(TP+FP)]. Also known as precision, it is a measure of the purity of the assigned classes in a classification model.

Background

The goal is to provide the client with high probability leads for clinicians whose patients are likely to be undiagnosed rare disease patients. Expanding the population of patients on therapy drives revenue growth and also shortens the period of time between first contact with patients and correct diagnosis.

The ensemble methods outlined in the introduction were used to score patients based on secondary data, such as prescriber information, diagnostic information, procedures, and socio-demographic information. The particular mixture of base learning agents and ensemble algorithms was motivated by the rare disease patient identification context outlined in the introduction. That is, given budgetary parameters and sales force sizing, clients typically have us carry out ensemble modeling that focuses on FPR reduction, leading to higher

PPV at the expense of FNR increases. Put another way, this approach may not be feasible if one is concerned with capturing most of the patient population, rather than an accessible fraction of it. As well, a client may have limited sales force allocations, hence they prefer to focus on small volumes of very high probability patients, again, at the expense of FNR increases.

Results

Case Study 1: Leveraging Predictive Analytics to Expand Patients on Therapy

In this case, an ensemble approach involving three models, as detailed above was utilized to identify untreated patients pre-diagnosis within a rare disorder, which has very low prevalence world-wide. The models were trained on available patients, a total << 1000, identified within the US population. The ensemble approach identified a number of patients as high likelihood untreated patients with the rare disease, given the volume and probability cutoffs desired by the client.

However, assuming that all patients were evaluated and all true positives for the disease were given the opportunity to choose therapy, this represents a 172,000-fold improvement over random picking at the native prevalence. In fact, this rate is subject to many factors, such as whether a diagnosis is made, willingness to go on treatment, ability to pay for treatment, insurance status, limited physician detailing and physician response to the intervention. Hence it is possible that this rate may be higher.

The overall false positive rate for each of the individual ensemble member models was 14% (Tree), 4% (Random Forest) and 12% (ADA Boost). Given the volume of patients that must be screened, if one were to use a single model approach, seemingly acceptably small error rates are not insufficient where disease prevalence is so low. For example, a 4% false positive rate (if we choose the best performing model, Random Forest) applied to 320 million (~ United States population) screened patients results in 12.8 million false positives. However, the ensemble level error rate would be as low as (under the simplifying assumption of model error rate independence noted above) P(false positive Tree, Random Forest, AdaBoost) = (0.14)(0.04)(0.12) = 0.000672 or 0.0672%, an ~60-fold (4%/0.0672%=59.5) improvement relative to the best performing single model in the ensemble.

TREE model: ROC (AUC=0.84) plot. (Figure 3)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.78, Sensitivity or True Positive Rate: TPR=0.84, FPR = 0.19, FNR = 0.14

Random Forest Model: ROC (AUC=0.96) plot (Figure 4)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.97, Sensitivity or True Positive Rate: TPR=0.89, FPR=0.04, FNR=0.11

ADA Boost Model: ROC (AUC=0.97) plot (Figure 5)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.91, Sensitivity or True Positive Rate: TPR=0.86, FPR=0.12, FNR=0.14

Figure 3: Case Study 1, Tree Model

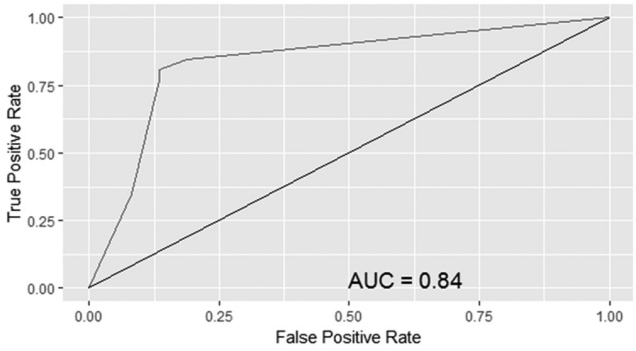


Figure 4: Case Study 1, Random Forest Model

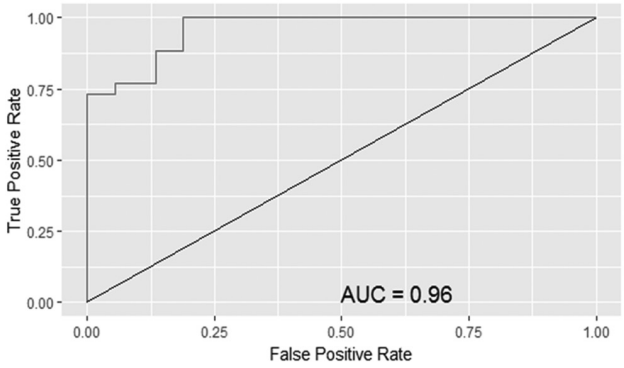
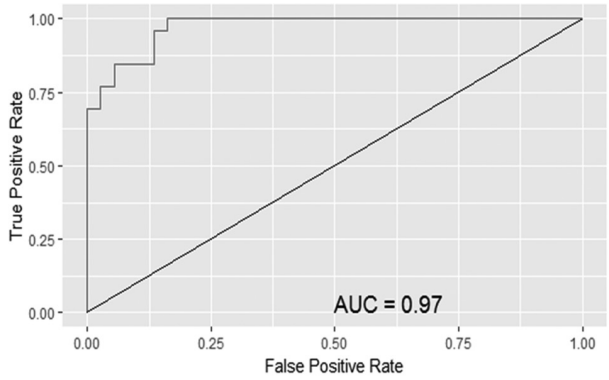


Figure 5: Case Study 1, ADA Boost Model



Of the ~9000 variables interrogated, 112 were implemented in the final ensemble. A portion of the variables found across the three modeling events (highly ranked by Mean Decrease in Accuracy from Random forest), were consistent with the therapeutic area and were significant when analyzed by Asymptotic Linear-by-Linear Association testing vs control (Table 1, P-value), and are shown in Table 1.

Case Study 2: Leveraging Predictive Analytics to Expand Patients on Therapy

In this case, an ensemble of three models as detailed above was utilized to identify untreated patients pre-diagnosis, within a rare disease population of approximately 1 in a million or rarer. The models were trained on available patients, where less than 5000 patients were identified within the US population.

We can calculate fold-increase based on all patients who were offered treatment, not just those who actually used the treatment. The rate of patient identification (model rate/ prevalence) represents a 257-fold improvement over random picking. In fact, this rate is subject to many factors, such as whether a diagnosis is made, willingness to go on treatment, ability to pay for treatment, insurance status, limited physician detailing and physician response to the intervention. Hence it is possible that this rate may be higher. Of the ~1775 variables interrogated, 10 were implemented in the final ensemble. Variables found across the three modeling events were consistent with the therapeutic area (data not shown).

Table 1

Variable Type	Variable Description	P-value
Demographic	AGE	7.28E-12
RX code	IMMUNITY RELATED	0.07817
RX code	CHOLESTEROL RELATED	2.48E-08
RX code	DIABETES RELATED	5.44E-07
RX code	FATIGUE RELATED	0.01128
RX code	LIPID RELATED	2.05E-06
RX code	DIABETES RELATED	0.0004229
RX code	INFECTION RELATED	0.02669
RX code	CARDIAC RELATED	0.003182
RX code	DIABETES RELATED	0.000158
Physician Level	PHYSICIAN SPECIALTY CODE 1	0.3762
Physician Level	PHYSICIAN SPECIALTY CODE 2	6.43E-10
DX code	DIABETES RELATED	2.01E-07
DX code	LIPID RELATED	0.0002046
Hospital/Clinic code	OUTPATIENT ACTIVITY RELATED	0.05494

Figure 6: Case Study 2, Tree Model

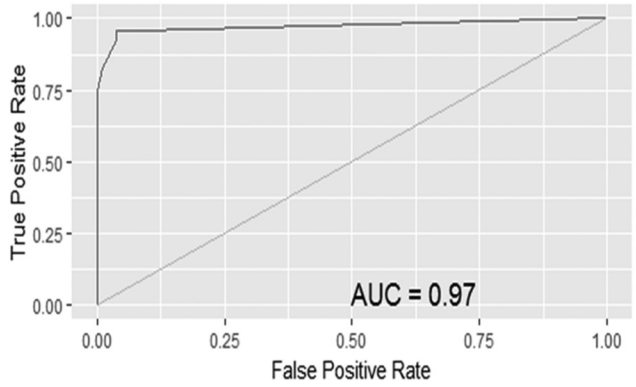


Figure 7: Case Study 2, Random Forest Model

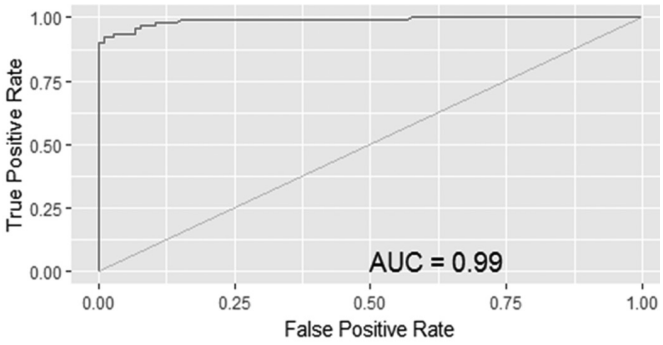
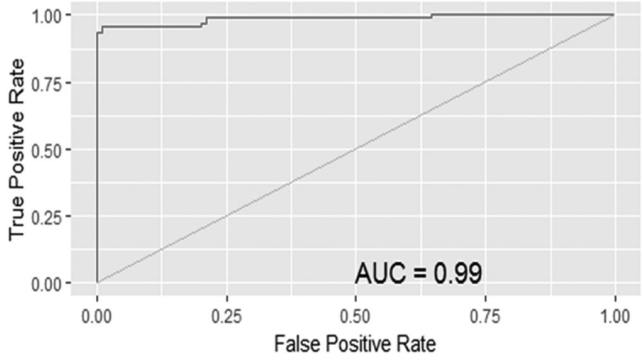


Figure 8: Case Study 2, ADA Boost Model



The overall false positive rate for each of the individual ensemble member models was 4% (Tree), 8% (Random Forest) and 4% (AdaBoost). However, the ensemble level error rate would be as low as (again, under the simplifying assumption of full model error independence) P(false positive for Tree, Random Forest and AdaBoost) = (0.04)(0.08)(0.04) = 0.000128 or 0.0128%, a ~313-fold reduction (4%/0.0128%=312.5) relative to either of the two best performing single models in the ensemble.

Tree Model: ROC (AUC=0.97) plot (Figure 6)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.96, Sensitivity or True Positive Rate: TPR=0.96, FPR=0.04, FNR=0.04

Random Forest Model: ROC (AUC=0.99) plot (Figure 7)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.92, Sensitivity or True Positive Rate: TPR=0.98, FPR=0.08, FNR=0.04

ADA Boost Model: ROC (AUC=0.99) plot (Figure 8)

True Positive Recovery vs False Alarms as function of model probability score

Positive Predictive Value: PPV=0.98, Sensitivity or True Positive Rate: TPR=0.96, FPR=0.04, FNR=0.10

Case Study 1 and 2: Results Comparison

There are two tunable parameters that were employed that affect the final patient discovery rates. The shape of the ROC plots above indicate that higher model probabilities are associated with lower false positive rates. In light of this finding, client preferences were to raise the probability threshold for classifying a patient within each model to require a higher score during classification of >=0.9 rather than the default of >0.5, hence the number of patients found reflects this. As well, "ensemble agreement" (within-patient) was also preferred, such that a patient would need to meet this higher threshold within all three models. These two parameter values impact volume and PPV.

In Table 2 and 3, we can compare the estimates above with out-of-sample PPV and FPR when class agreement is imposed. For example, no FP errors were made in the Case Study 1 ensemble until the model probability cutoff fell below 0.55, leading to a 100% PPV for this small (N less than 100) hold-out sample, for model cutoffs above 0.55. This therefore provides an opportunity to lower each individual model cutoff well below where one would have otherwise experienced large FP rates. In Table 3 we see that we would experience only 3,843 false positives at an ensemble cutoff of >= 0.9

An important aspect of the ensemble regularization process being proposed here is balancing the cost of the false negative rate increases coming from both a high probability cutoff and the ensemble member agreement requirement, against the value of being able to properly match some volume of high quality candidate patients to fit available resources. Since any threshold (even thresholds below 0.5) can be used in an ensemble regularization process, volume can be scaled up to fit larger force sales sizes, with added assurance from the ensemble regularization that FP rates will be

Table 2

CASE Study	Model cutoffs	Ensemble Agreement Imposed	Fold Enrichment from Randmom Picking (Ratio of Treatment % to Prevalence %)
1	0.9	YES	172248
2	0.9	YES	257

Table 3

CASE Study	Ensemble OOS FPR at >0.9 probability cutoff **	Ensemble FPR Reduction vs Best Model Single Model at >=0.9 probability cutoff	Estimated FP count base on United States Population (320M) and >=0.9 probability cutoff
1	0.000012	3333	3840
2	ND	ND	ND

** Due to small N, the FP was '0' for two of three models at >=0.9 probability cutoff. To avoid and unrealistic assessement a conservative non-zero value of 0.01 was substituted, actually inflating our FPR. ND: calculation not done

improved relative to any single model. Our Case Study 1 and 2 examples used very stringent cutoffs (>=0.9); however, the levels could be set at or below 0.5 as part of the ensemble regularization process that focuses on higher volume, higher FPR and lowers FN rates.

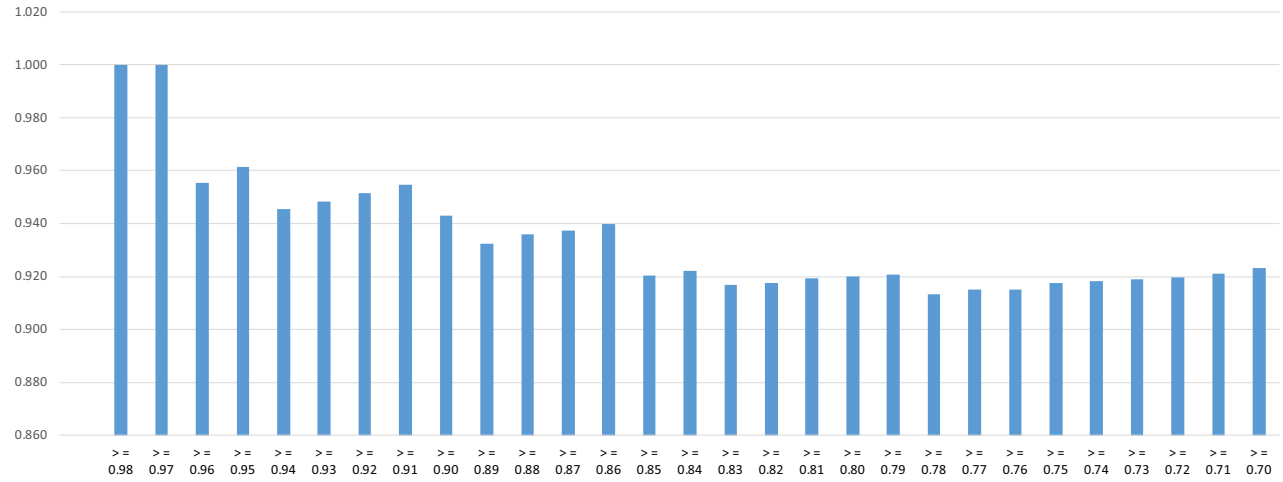
Case Study 3: Ensemble Regularization and Model Probability Cutoffs

Due to the very small hold-out sample set sizes for the rare disease case studies above, estimates of PPV as a function of cutoff are very sparse. For example, no FP errors were made in the Case Study 1 ensemble until the model probability cutoff fell below 0.55, leading to a 100% PPV for this balanced N of less than 100 hold-out sample size. In order to show a robust distribution of ensemble PPV as a function of model probability cutoff, we turn to a third rare disease case where a larger balanced N>1000 hold-out sample was available. As detailed in Figure 9, ensemble level precision remains high when models are applied to an out-of-sample data set even when the cutoff for each model is as low as 70% probability. The ensemble level error remains above 90% despite lowering each of the individual model stringencies to >=70%.

Recommendations

Identifying undiagnosed and untreated rare disease patients[8] represents a significant opportunity to shorten time to diagnosis. As noted in the introduction, minimizing the time to correct diagnosis for these patients is critical, particularly for conditions in which there are therapies available to limit the progression or relieve key symptoms of the disease. Identifying highly likely patients prior to diagnosis can facilitate timely and targeted disease education efforts. From a clinical perspective, predictive analytics represents a means to expedite

Figure 9: OOS Ensemble-Level PPV as Function of Probability Cutoff



A three (3) model ensemble showing ensemble-level precision (positive predictive value, PPV) on the Y-axis vs the model probability cutoff on the X-axis where all three models share the same cutoff.

early and accurate diagnosis and care. From a commercial perspective, identifying patients and their associated physicians supports ongoing promotional and educational efforts in a highly targeted fashion.

A predictive analytics approach to rare disease patient identification is accomplished most effectively through high dimensional[13, 14] machine learning.[4] In this process potential human bias is reduced as computers learn without being explicitly programmed and very large selections of variables are evaluated. Tree, Random Forest and Boost algorithms are well-suited to this task given their computational efficiency and utility in an ensemble error reduction process.

The ensemble approach to error reduction is particularly appropriate for rare disease, in which true patient rarity, large sales territories and promotional effort costs result in a low false positive (FP) tolerance. The cases in this article highlight the efficacy of this approach for two separate rare disease cases. Here we have employed this approach in building predictive model ensembles for two very rare diseases. In each case, high levels of ensemble predictive accuracy were achieved well beyond the level of any of the individual models. Precision, or positive predictive value, provides the best measure of value within the rare disease patient identification context, as it focuses on the purity of the pre-diagnosis rare disease patients we identify for clients. In this context, large volumes of false positives will accumulate when screening a large number of patients unless a model has a zero (or very near zero) false positive rate. Through the use of ensembles, false positive rates can be reduced by orders of magnitude below that of any single model, as long as we can tolerate a higher false negative rate. In the case of the very rare diseases modeled here, a prohibitively high false positive rate would have prevented implementation had we not used an ensemble approach.

Choice of ensemble members depends upon a variety of factors. The machine learning algorithms used here were chosen to maximize value to the client based on a consideration of budget, timeline, computational efficiency and

robustness against ultra-high dimensional data. The ability to interrogate algorithms post-modeling in order to set the stage for additional explanatory modeling was also a key factor, hence a single "greedy" (best variable at each binary split) tree algorithm that generates explicit rules allows one to look at all conditionality, and Random Forest generated "Mean Decrease in Accuracy" scores associated with key variables, both contribute information that can be used to choose subsets of variables for final modeling.

Additional exploratory modeling on subsets of variables using classical statistical approaches such as Logistic Regression would add value. That said, project overhead might not justify the additional investigation time and cost. Per this, the choice of models used here allows analysis to take place directly in ultra-high dimensional space (6,000-15,000 variables), helps streamline the work flow in a turn-key fashion, produces less biased models, and shrinks the variable space[6, 7] for explanatory modeling methods that require low dimensional data sets.

Recommended improvements to the above turn-key process that might add value would be combining patient finding models with additional differentiation models where possible. For example, the Case 1 disease is in fact a subset of many very similar conditions, and developing an additional model to differentiate between these different conditions might improve accuracy. As well, building out the ensemble by adding additional member models built within specific data silos, such as the diagnostic code space, the prescribing code space, etc., forces learning and variable selection within a restricted domain. This information might otherwise be lost or under-exploited when all data sources are combined during training, if stronger variables are associated with a single data silo.

A final point with regard to model selection has to do with diversity. While all three algorithms leverage a tree as the base learning agent, the strengths and weaknesses of each algorithm are very different. The complementary nature of these differences make the final ensemble approach particularly powerful, however any number of different algorithms types can and should be evaluated. For example using mixtures of tree-based algorithms, support vector machines, artificial neural networks might results in improved model precision.

About the Authors

***Tim Hare**, Sr. Manager, CE-Analytics, Symphony Health Solutions, is a statistician and computational biologist with 13+ years of robotic high throughput drug discovery experience, and 14+ years of pre-clinical biochemical research experience. Tim was Sr. Scientist in Informatics at a pharmaceutical firm prior to joining the commercial effectiveness team at Symphony. Tim has an extensive publication record and has received 10 industry awards. He has a bachelor's degree in Biology from Earlham College, and a master's degree in Applied Statistics from West Chester University of Pennsylvania, where his thesis focused on the application of ensembles of artificial neural networks in the prediction of drug-target interactions.*

Pratibha Sharan *is a Senior Consultant with the Commercial Effectiveness Team at Symphony Health Solutions in Conshohocken, PA. Pratibha has completed her MBA from the University of Pittsburgh and has over 5 years of experience working in functions such as Commercial Effectiveness, Marketing Ops, Market Research across big pharmaceutical companies. She is adept in analyzing different types of healthcare datasets, developing actionable insights and project management. She aspires to create continuous improvement, process efficiency and business value in using machine learning as a mainstream tool to drive client strategies.*

***Ewa J. Kleczyk**, PhD is a Vice President of Client Analytics with Symphony Health Solutions in Conshohocken, PA and an Affiliated Graduate Faculty for the School of Economics at the University of Maine in Orono, ME. She also holds a doctorate degree in Economics from Virginia Tech. Ewa leverages over 12 years of industry experience and a strong passion for healthcare to develop key insights from healthcare claims data. Going forward, she is excited about the new and expanded role of machine learning in providing more exact and refined optimization of sales and targeting strategies and resources, as well as their deployment timing.*

***Derek Evans**, Sr. Vice President & General Manager, Symphony Health Solutions, is an intuitive business leader with over twenty years of marketing, analytical, strategic and operational experience in the pharmaceutical and consulting industry. In his current position, Derek leads a multi-million dollar health informatics consulting practice, which is focused on helping improve the delivery of healthcare via real world evidence. Under his leadership, his practice has generated breakout growth for Symphony Health. Over the years, Derek has been recognized by Janssen, Pfizer, Daiichi Sankyo and Novartis for leadership on programs that have improved company profitability, service, delivery and product development.*

References

1 Petersen L, Wright S, Normand S, Daley J. (1999), Positive Predictive Value of the Diagnosis of Acute Myocardial Infarction in an Administrative Database. JGIM 14: 555-558.
2 T.M. Mitchel, Machine Learning, McGraw-Hill Publishing, ISBN-13: 978-0071154673
3 T.J. Hastie and R.J. Tibshirani, Introduction to Statistical Learning, Springer Publishing, ISBN-13 978-1-4614-7137-0
4 A.L. Samuel (1959), Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3:210–229
5 P. Sollich, A. Krogh (1996), Learning with ensembles: How over-fitting can be useful, In D.S.Touretzky, M.C.Mozer, and M.E.Hasselmo, editors, Advances in Neural Information Processing Systems, 8:190-196, MIT Press.
6 R. Genuera, J. Poggiab, C. Tuleau-Malotc, Variable Selection Using Random Forests, Pattern Recognition Letters 31(14): 2225-2236
7 A. Hapfelmeier, K. Ulm, A New Variable Selection Approach Using Random Forests, *Computational Statistics & Data Analysis*, 60:50-69
8 M. Khalilia (2011), Predicting Disease Risks from Highly Imbalanced Data Using Random Forest, *BMC Medical Informatics and Decision Making*, 11(1):51-
9 D. Wackerly, W. Mendenhall, R. L. Scheaffer (2008), Mathematical Statistics with Applications, Seventh Edition, ISBN-13: 978-0495110811
10 T.J. Hastie, R.J. Tibshirani (1990), Generalized Additive Models, First Edition, Chapman and Hall Publishing, ISBN-13: 978-0412343902
11 T.J. Hastie, R.J. Tibshirani, J. Friedman (2016) Elements of Statistical Learning, Second Edition, ISBN-13: 978-0387848570
12 T.K. Ho, (1998), The Random Subspace Method for Constructing Decision Forests, IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.
13 R.E. Bellman (1957), Dynamic programming. Princeton University Press. ISBN 978-0-691-07951-6., Republished: R.E. Bellman (2003), Dynamic Programming. Courier Dover Publications. ISBN 978-0-486-42809-3.
14 R.E. Bellman: https://en.wikipedia.org/wiki/Richard_E._Bellman
15 Kégl, Balázs, (2013), The return of AdaBoost.MH: multi-class Hamming trees, arXiv:1312.6086 Freely accessible [cs. LG]
16 L. Stokowski, W. Gahl (2015), Pursuing Elusive Diagnoses for Rare Diseases, https://www.medscape.com/viewarticle/840440
17 M.A. Vedomske, D.E. Brown, J.H. Harrison (2013), Random Forests on Ubiquitous Data for Heart Failure 30-Day Readmissions Prediction in Machine Learning and Applications (ICMLA), 12th International Conference on Machine Learning
18 J. Wennberg, N. Roos, L. Sola, A. Schori, R. Jaffe (1987), Use of Claims Data Systems to Evaluate Health Care Outcomes. Mortality and Reoperation Following Prostatectomy. *JAMA* 257(7): 933-936.
19 Y. Zhao, R. Ellis, A. Ash, D. Calabrese, J. Ayanian, J. Slaughter, L. Weyuker, B. Bowen (2001), Measuring Population Health Risks Using Inpatient Diagnoses and Outpatient Pharmacy Data. Health Services Research, 36(6): 180-193.
20 P. Hougland, J. Nebeker, S. Pickard, M. Tuinen (2008), Using ICD-9-CM Codes in Hospital Claims Data to Detect Adverse Events in Patient Safety Surveillance., in Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 1: Assessment), Rockville (MD): Agency for Healthcare Research and Quality; Publication No.: 08-0034-1
21 D. Bertsimas, D. Czerwinski, M. Kane (2013), Measuring quality in diabetes care: an expert-based statistical approach. SpringerPlus 2:226-
22 S. Marsland (2009), Machine Learning: An Algorithmic Perspective. Chapman & Hall Publishing, First Edition, ISBN-13: 978-1420067187
23 M. Negnevitsky (2005), Artificial Intelligence: A Guide to Intelligent Systems. Addison-Wesley Publishing, Second Edition. ISBN: 0321204662
24 N. Meinshausen and P. Buhlmann (2010). Stability Selection. *J Royal Statistical Society Series B*, 72(4):417–473.
25 R package RPART: https://cran.r-project.org/web/packages/rpart/rpart.pdf
26 R package RandomForest: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
27 R package ADABoost: https://cran.r-project.org/web/packages/ada/ada.pdf

Clinical In-Market Leading Indicators for Brand Performance

Nitin Choudhary, Principal, Symphony Health Solutions; Ewa Kleczyk, PhD, Vice President of Client Analytics, Symphony Health Solutions; Rajkumar Rajabathar, Engagement Manager, Symphony Health Solutions

Abstract: In today's world, there are multiple treatment options available for overall care of a patient in a disease state. For a pharmaceutical company, the need for effective product based strategy has gained importance. Traditionally, strategies have been derived from knowledge on disease states and focused on key stakeholders in a patient's journey. However, in many cases, there are key in-market factors that significantly influence a particular treatment choice. The patient's continuum of care, condition, nature of disease, therapy administration method, safety, efficacy, side effects, associated comorbidities, and concomitant medications play a role in selecting a particular treatment choice or switch. Additionally, evolution of therapies in the market, product adoption, and physician prescribing behavior, insurance coverage, and promotions are key considerations informing brand strategy.

This paper outlines a two-step approach to develop a cost-effective brand strategy, illustrated by a case study. First, using patient, physician, and payer information we explore significant drivers and barriers for brand initiation and continuation. Second, using a finite list of drivers and promotions data we identify and track leading indicators that can explain significant changes in brand performance. This can provide actionable levers for sales, marketing, market access, clinical and brand teams. Such levers can help interpret implications to forecast, enhance messaging, and optimize allocation of resources to derive more value. As a result, pharmaceutical companies can develop and ultimately execute a more product and patient centric, data-driven brand strategy.

Keywords: leading indicators, brand strategy, market diagnostics, statistical modelling, healthcare claims

Introduction

A leading indicator is a measurable economic factor that changes before the economy starts to follow a particular pattern or trend. Historically, leading indicators are used to gain some sense of which way the economy is headed. Investors use such indicators to adjust their strategy to benefit from future market conditions.[1] Unemployment rates and inflation[2] are key economic indicators that affect markets.[3] This technique has also been successfully applied in retail scenarios. Pinterest is an example of a leading indicator, the number of pins steadily increase as users are discussing within the social network and become engaged with an event or holiday.[4] Retailers who have periods of peak demand during holidays can predict performance using such an approach. Economic leading indicators such as state of the economy can help explain consumer buying power within a target segment.

The situation is no different in the pharmaceutical world, where early warning signs can be used to identify brand performance changes prior to actual change occurrence.[5] Pharmaceutical companies can use such early warning signs or leading indicators to develop and course correct brand strategy. However, predicting the future requires understanding of the past.

Figure 1: Conceptual Framework of Implications Across a Pharmaceutical Company

	SALES	MARKETING	MARKET ACCESS
STRATEGIC IMPLICATIONS	Design or inform high value targeting strategy Optimize allocation of resources	Design and deploy marketing campaigns specific to target audience	Inform market access strategy with MCOs
TACTICAL IMPLICATIONS	Identify patient mix for physicians to evaluate value of physician-rep interaction Communicate insights and train reps on messaging to specific type of physicians Cues for reps based on key moments of a patient journey	Design marketing tactics for targeting specific sub populations Identify messaging and channels for optimal reach Design scenarios for continuous improvement and return on investment	Identify plans with high rejections and reversals Deploy educational programs around cost and coverage Identify reasons for rejections and reversals
INSIGHTS	**New Initiations**	**Continuations**	
	Patients on commercial plans are more likely to initiate on product A Patients with Medicare plans are less likely to initiate on product A Patients between the age of 55 and 84 are likely to initiate on product A Patient taking medication X are more likely to initiate on product A Patients in middle income groups are more likely to initiate on product A Patients with Y medication are less likely to initiate on product A	Affordability is a key factor in patient continuations Patients on product A for more than 5 months are more likely to continue for 12 months Side-effects are not a key factor in patient discontinuations in this disease state Lower income groups are likely to switch away from product A Patients with product A reversed claims are more likely to switch to product C Patients less than the age of 65 are more likely to switch to product B Patients with higher CV risk are more likely to switch away from product A	

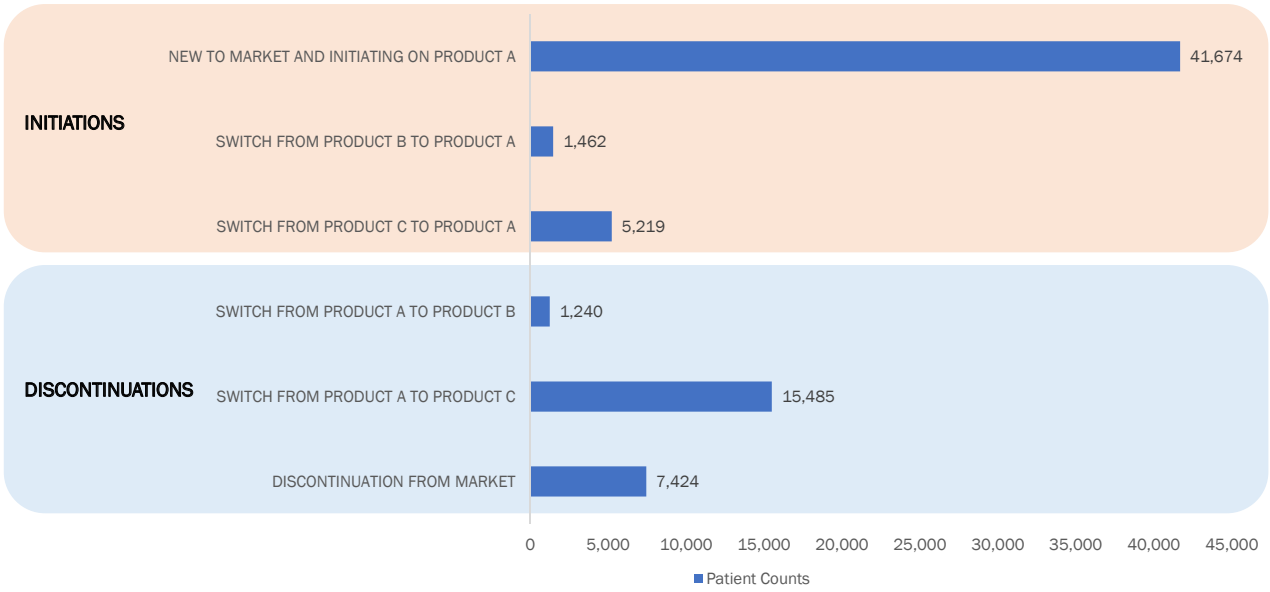
In an industry with rich physician and patient level longitudinal data, leading indicators can be identified using a data-driven approach. This paper discusses a quantitative approach leveraging statistical modelling techniques to identify and track leading indicators.

Key Considerations for Brand Strategy

In a competitive marketplace, a cost effective, product-centric brand strategy[6] requires tackling strategic questions, such as:

1. How to develop a more effective brand plan?
2. How to better utilize or understand disease state, patient behavior, and physician preferences?
3. How to improve on commercial assessment methodologies?
4. How to differentiate the brand in the market?
5. How to improve market access for brand?
6. How to optimally allocate promotional spend across channels, target audience and geographies?

Figure 2: Product Initiation and Discontinuation Cohorts for Observation



In order to address each strategic question, it boils down to gaining a better understanding of the brand in the market place. Interactions between physicians, patients, payers and treatment choices drive the evolution of the market for a particular disease state.[7] Understanding market dynamics can help identify leading indicators to quantify key dynamic questions such as:

1. How many patients are new to market, switch treatments and discontinue?
2. What is the treatment duration for each brand?
3. What are the key patient characteristics that are drivers and barriers for new adoptions and continuations of a brand in the market?
4. What are the significant, differentiating factors driving a particular brand's starts and stops as compared to competitors in the market?
5. What are the key differentiating leading indicators to predict brand volume?

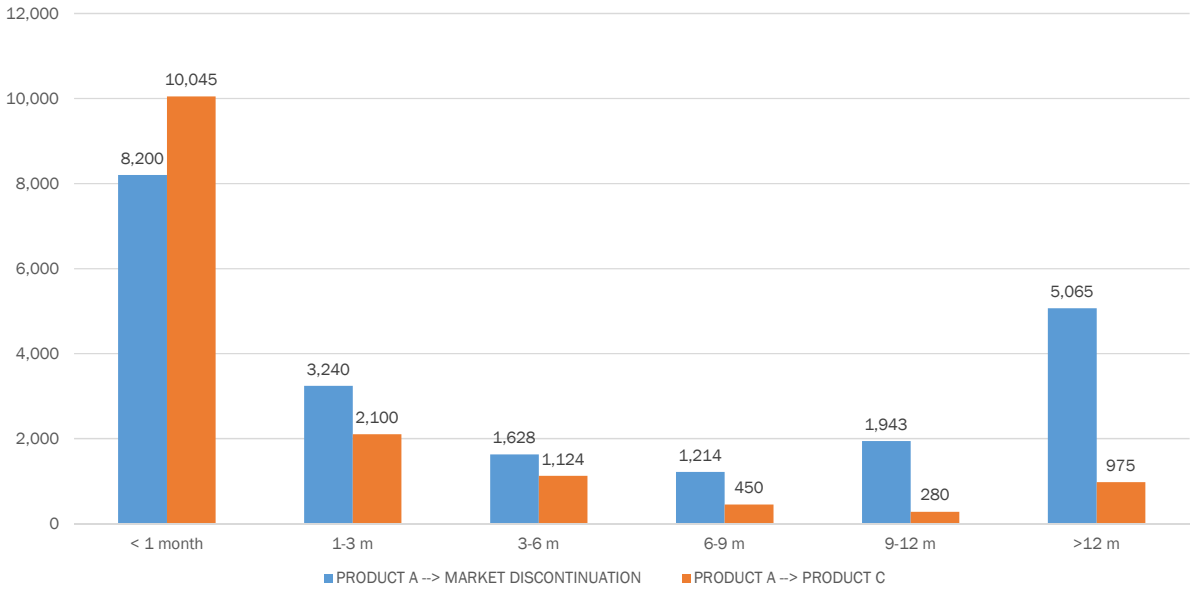
Such market insights can have implications across a pharmaceutical company, providing levers for sales, marketing and market access teams to act upon.[8,9,10] Figure 1 summarizes market insights from the case study and provides a conceptual framework of strategic and tactical implications.

Case Study

Introduction

A retrospective study consisting of approximately 200,000 patients initiating a treatment in a market were studied over a one year period ending October 2016. The market was defined based on three products – product A, product B and product C. Product A is of interest and product C owns the majority of the patient share in the market. As illustrated by Figure 2, the majority of patients are new to market. Switches away from product A to product C was a concern. In this study, physician preferences and promotions were considered to have minimal impact to the product losing share and thus, were not included in the analysis.

Figure 3: Product Continuation Cohorts for Observation



Discontinuation cohorts were further differentiated based on duration of therapy before discontinuation. Figure 3 illustrates durations of therapies on product A prior to discontinuation or switch. Most patients on product A switch or discontinue within one month of product initiation. Comparing cohorts based on duration of treatment can help understand key barriers and drivers for continuations. Some patients continue for more than 12 months whereas others discontinue after a few months on the product.

Analysis

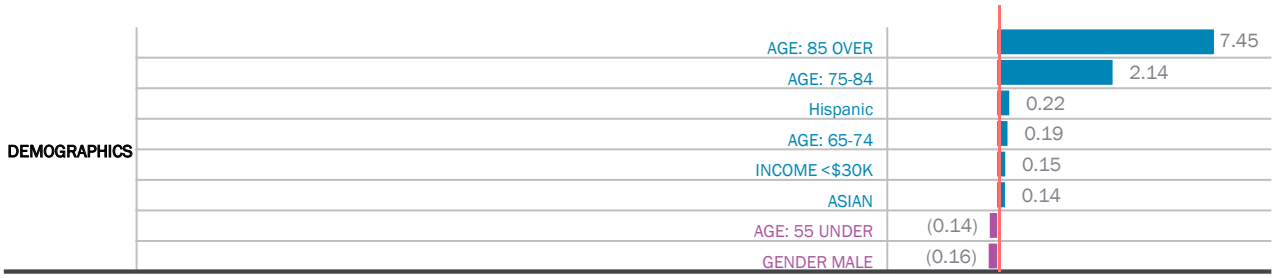
A retrospective analysis was conducted to compare new, continuing and discontinued patient cohorts. From longitudinal historical claims data, dimensions such as patient demographics, plan, diagnosis, other medications, and procedures[11,12,13] were studied 90 days prior to an initiation or discontinuation of product A. Analyzing patient cohorts helps uncover market dynamics. Six cohorts were analyzed:

1. New To Market (NTM) initiation of product A vs. product B
2. NTM initiation of product A vs. product C
3. Continuation of product A for less than a month vs. 12 months
4. Continuation of product A for 1-5 months vs. 12 months
5. Continuation of product A for 6-11 months vs. 12 months
6. Discontinuation of product A vs. switch to product C

Over 850 variables were developed to understand patient and payer dynamics between six cohorts of patients treated with product A. Univariate and bivariate analysis was conducted to assess validity of the data variables by ensuring at least 10% of the data is populated and to remove correlated variables. A shortlist of 50 features were fed into a logistic regression model. The end output was approximately 25 significant features (based on p-values) with higher odds ratios that are drivers or barriers for new initiations and continuations. Figure 5 illustrates odds ratios for significant patient demographics features (p value < 0.01) and high

Figure 4: Cohort Analysis Methodology

Figure 5: Significant Patient Demographics Features Differentiating Between New to Market Patients Initiating on Product A (Odds Ratios >0) and Product B (Odds Ratios <0)



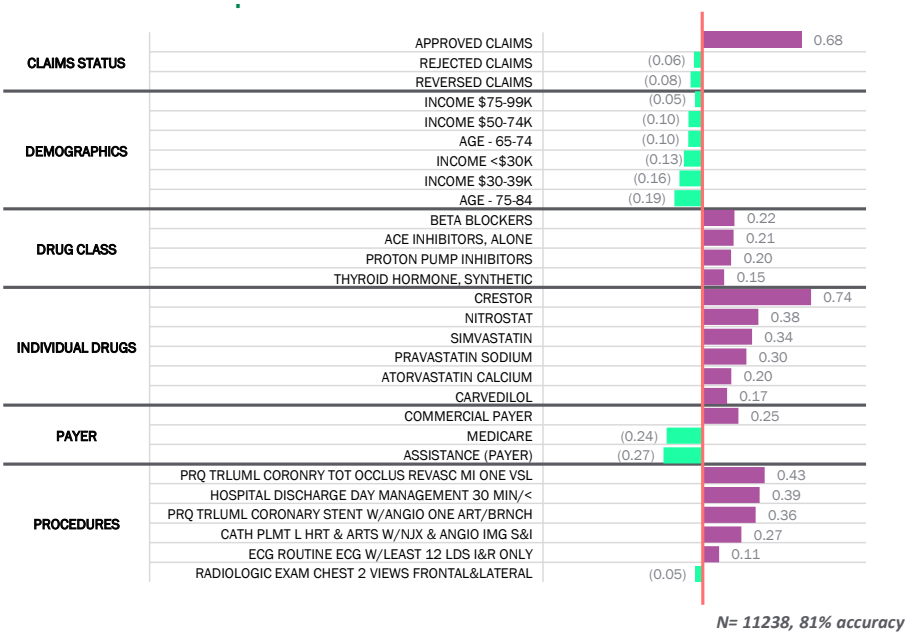
odds ratios. This helps identify that if a patient is new to market and over 85 years of age, that patient is more likely to initiate on product A as compared to product B.

Analyzing patients continuing on product A for one month of initiation are significantly different from patients continuing for 12 months. As illustrated in Figure 6, rejected and reversal claims, mid-income groups, and Medicare plans characterize patients discontinuing product within one month of initiation.

Analysis of six patient cohorts provided a comprehensive list of significant patient features affecting market dynamics.

1. Driving new initiations:
    a. Patients on commercial plans are more likely to initiate on product A
    b. Patients with Medicare plans are less likely to initiate on product A
    c. Patients between the age of 55 and 84 are likely to initiate on product A
    d. Patients on X medication are more likely to initiate on product A
    e. Patients in middle income groups are

Figure 6: Significant Patient Features Differentiating Between Patients Discontinuing Within 1 Month of Initiation (Odds Ratios >0) and 12 Months of Initiation (Odds Ratios <0)



more likely to initiate on product A

f. Patients with Y medication are less likely to initiate on product A

2. Promote continuance
   a. Affordability is a key factor in patient continuations
   b. Patients on product A for more than 5 months are more likely to continue for 12+ months
   c. Side-effects are not a key factor in patient continuations in this disease state
3. Arrest switches:
   a. Lower income groups are likely to switch away from product A
   b. Patients with product A reversed claims are more likely to switch to product C
   c. Patients less than the age of 65 are more likely to switch to product B
   d. Patients with higher CV risk are more likely to switch away from product A

Leading Indicators

Significant patient features were used to predict new and continuing patient volume depicted in Figure 7. About 25 significant features were derived from the driver analysis. Multivariate time series regression models can be used to predict time dependent variables such as prescriptions. A prediction model for new prescriptions and continuing prescriptions was built.

Figure 8 illustrates the overall analysis methodology for leading indicators.[14] A 5-step approach was used to select a final model for identification of leading indicators. Using patient features as inputs, a multivariate regression model was built on 70% of the data and validated on 30% of the data. Two models were built—one to predict projected new prescriptions and another to predict continuing prescriptions. The end output of the model

Figure 7: New and Continuing Patient Trend for Product A

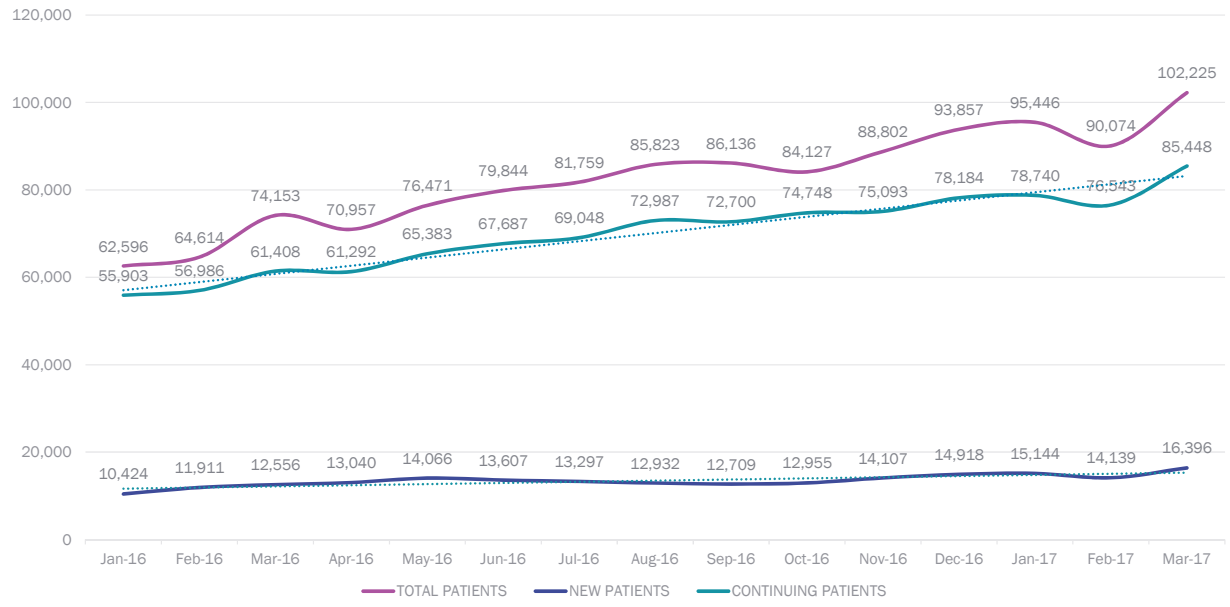
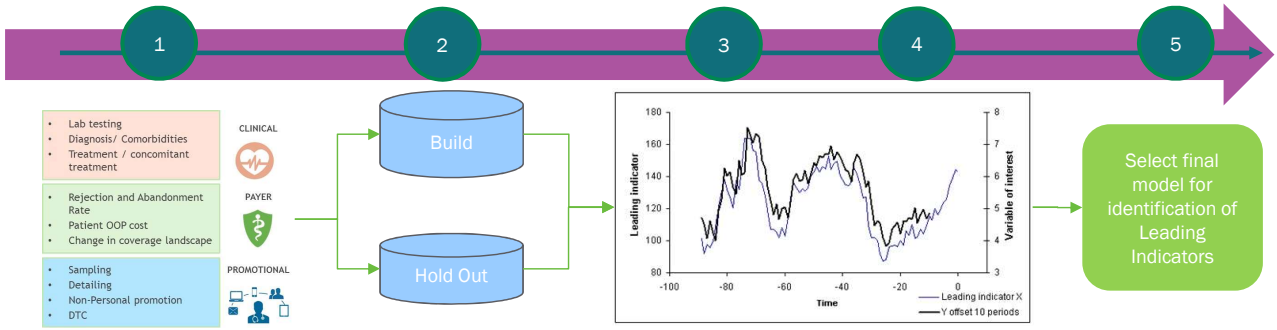


Figure 8: Leading Indicators Methodology



provides coefficients of each lead & lag variables based on the output equation. The coefficients of the equation quantify the impact of each variable to prescriptions. The models were more accurate in predicting continuing prescriptions; the out-of-sample error for new and continuing models were 4% and 1.5% respectively. Adjusted R-squared values were 0.63 for new initiations and 0.83 for continuing models. The equation of the new prescriptions prediction model:

Product A NRx = Intercept + 0.14*(DX_X_12) – 0.112*(DX_Y_12) – 0.04*(TX_Z_16) + 0.7 (seasonality component)

DX_X_12 = Number of patients diagnosed with X, 12 weeks prior
DX_Y_12 = Number of patients diagnosed with Y, 12 weeks prior
TX_Z_16 = Number of patients initiating treatment Z, 16 weeks prior

For new patients, some examples of leading indicators for product A are:

1. Population dynamics such as age, gender, geography and income

Figure 9: Clinical Leading Indicators for Initiation of a Brand Based on Patient Volume and Impact on Brand Performance

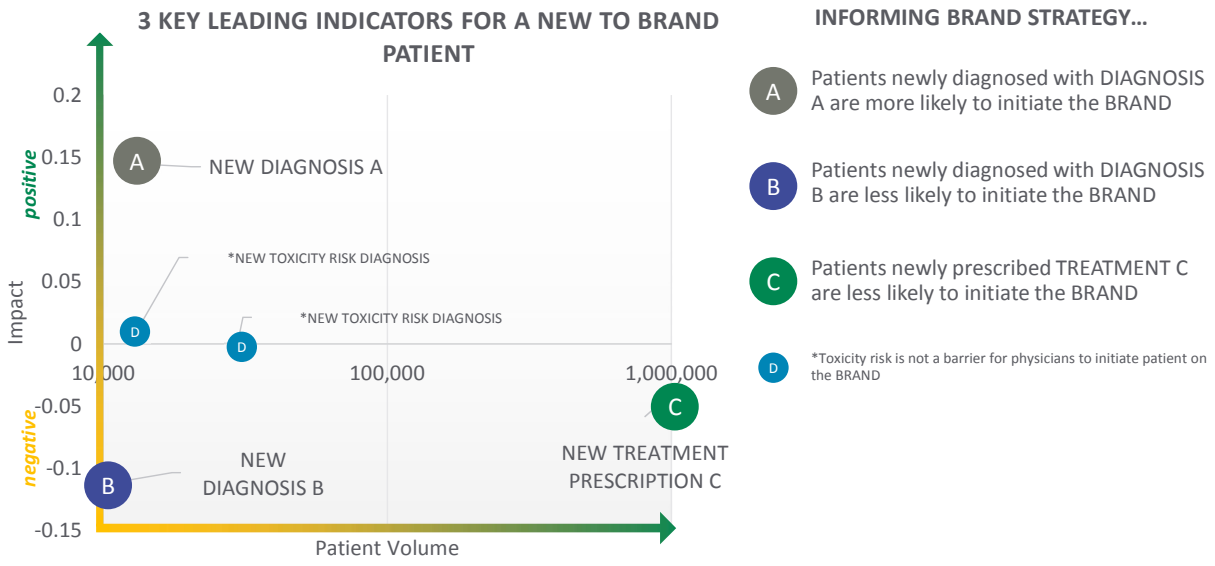
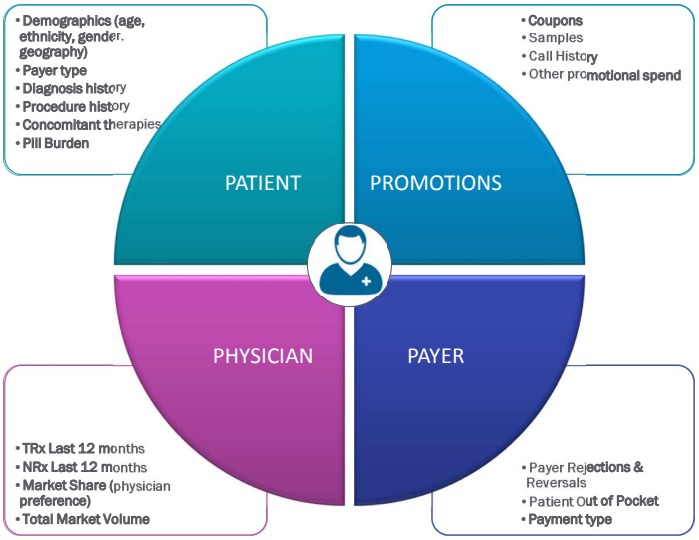


Figure 10: Comprehensive Set of Features to Consider for Leading Indicators



2. Diagnosis of certain comorbid conditions
3. Initiation of concomitant treatments

For continuing patients, some examples of leading indicators for product A are:

1. Medicare and Medicaid plans
2. Rejections and reversals
3. Patient severity, certain procedures administered

Such a data driven modelling approach validated key business hypotheses pursued as part of brand strategy. Leading indicators also provide a conceptual framework to develop, execute and incorporate feedback in brand

strategy. Brand strategy can be collaboratively executed through multi-channel tactics.

Results

Outputs from the statistical model were used as levers to inform brand strategy. Figure 9 illustrates examples of clinical leading indicators and their impact on brand performance. Diagnosis and treatments unrelated to the disease or product of interest were found to have a positive and negative impact on performance. Such indicators facilitated sales & marketing activities by enhancing product specific messaging.

Next Steps

Apart from patient features, there are several other factors that can be leading indicators for the brand. Physician preferences based on prescribing behavior and promotional spends affect brand performance. Figure 10 illustrates a more comprehensive set of variables that can be leveraged as leading indicators. Machine learning can also be used to train models for better accuracy.

About the Authors

Rajkumar Rajabathar *(RB), an engagement manager with Symphony Health Solutions, uses his 7 years of experience in providing high impact consulting services for pharmaceutical companies. He is passionate about using his knowledge of business, math and data & technology to solve complex business problems through a client-centric approach. With expertise from leading analytics for innovative growth strategy projects in the areas of: closed loop marketing strategy; improving R&D productivity through disparate patient data such as claims, EMR, prospective & retrospective clinical trials data; interactive analytics capability for commercial assessments; sales, marketing effectiveness & incentive compensation. With his experience, he is well positioned to guide his teams through strategy & analytics projects executed within the commercial effectiveness team. His deep knowledge of industry & analytics also enables him to develop new offerings for the team. In his prior role at Mu Sigma, he was recognized through impact & spot awards. RB holds a bachelor's degree in Aerospace Engineering from Indian Institute of Technology, Madras.*

Nitin Choudhary, *a Principal with Symphony Health Solutions, uses his 11 years of experience in providing analytical consulting services to create effective and innovative solutions to problems faced by the Commercial Effectiveness team. With expertise in leveraging patient level integrated data to solve problems in fields like Advanced analytics, Management Science, Field Analytics, Brand & Data strategy, Managed markets & Contracting, Reporting, and KPI & Dashboards, Nitin is well prepared to guide teams through the complex pharmaceutical market. His knowledge of the industry is demonstrated through accomplishments such as Symphony Consultant of the year, industry presenter at PMSA conference, and Innovator of the Year from Cognizant. Nitin holds a Masters degree in Mathematics and Computing from the Indian Institute of Technology Delhi.*

Ewa J. Kleczyk, PhD, *is a Vice President of Client Analytics with Symphony Health Solutions in Conshohocken, PA and an Affiliated Graduate Faculty for the School of Economics at the University of Maine in Orono, ME. She also holds a doctorate degree in Economics from Virginia Tech. Ewa leverages over 12 years of industry experience and a strong passion for healthcare to develop key insights from healthcare claims data. Going forward, she is excited about the new and expanded role of machine learning in providing more exact and refined optimization of sales and targeting strategies and resources, as well as their deployment timing. In the future, Ewa believes that past and future legislation will ultimately result in patients having greater involvement in their healthcare decisions, leading marketers to change the way they approach patients and the information they make available to them.*

References

1 Investopedia (IAC). Definition of a leading indicator. https://www.investopedia.com/terms/l/leadingindicator.asp
2 Philip. A. Klein. Leading indicators of inflation in market economies. *International Journal of Forecasting*. Vol. 2, Issue 4, 1986.
3 Investopedia (IAC). 4 key economic indicators that move the market by Brian Perry (Nov. 2017). https://www.investopedia.com/articles/fundamental-analysis/10/indicators-that-move-the-market.asp
4 American Marketing Association (AMA). How to use leading indicators to predict success and course correct for your brand by Jon Webb (May 2017). https://www.ama.org/publications/eNewsletters/Marketing-News-Weekly/Pages/predictive-insights-lead-indicators.aspx
5 Symphony Health Solutions. Leading indicators of brand performance (May 2013). https://symphonyhealth.com/wp content/uploads/2013/05/leading_indicators_sales_sheet.pdf
6 Boston Consulting Group (BCG), 5 ways to win commercially in pharmaceutical markets (Dec 2011). http://img-stg.bcg.com/BCG_Brand_Renaissance_Dec_2012_tcm9-108418.pdf
7 Mark Paich, Corey Peck, Jason Valant. Pharmaceutical market dynamics and strategic planning: a system dynamics perspective. *System Dynamics Review* Vol 27, Issue 1, Jan 2011.
8 Cognizant. Key Indicators: An Early Warning System for Multichannel Campaign Management (2013). https://www.cognizant.com/whitepapers/Key-Indicators-An-Early-Warning-System-for-Multichannel-Campaign-Management.pdf
9 PMLive. Pharma needs to evolve from customer centricity to stakeholder centricity (June 2017). http://www.pmlive.com/pmhub/healthcare_creative_and_design/owen_health/white_papers_and_resources/pharma_brand_strategy_2.0
10 PWC Strategy&. A critical makeover for pharmaceutical companies: Overcoming industry obstacles with a cross-functional strategy (Jan 2017). https://www.strategyand.pwc.com/reports/critical-makeover-pharmaceutical-companies
11 McKinsey and Company. How big data can revolutionize pharmaceutical R&D (Apr 2013). https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d
12 Forbes. How big data is changing healthcare (Apr 2015). https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/
13 Forbes. Data Silos: Healthcare' silent shame (Mar 2015). https://www.forbes.com/sites/davidshaywitz/2015/03/24/data-silos-healthcares-silent-tragedy/#244b9c4c2907
14 Fifth Joint European Comission (EC). Those unpredictable recessions by Sergey. V. Smirnov. https://www.oecd.org/std/leading-indicators/48985179.pdf

Deploying Machine Learning for Commercial Analytics

Jean-Patrick Tsang, PhD and MBA (INSEAD), President of Bayser Consulting

Keywords: commercial analytics, machine learning, panoramic contextual data, feature engineering

Introduction

Machine Learning is seeing unprecedented success across the board. Not a week goes by without our hearing of a new accomplishment or breakthrough. Let's mention only two. First, AlphaGo Zero. This is a Reinforcement Learning program developed by DeepMind, a UK-based company that Google acquired in Jan 2016. Zero refers to the fact that the program starts learning from scratch, without any human intervention. After 40 days of training on a regular laptop, AlphaGo Zero beats AlphaGo 100 to 0 and AlphaGo is the program that defeated the world champion at Go. Second, the self-driving car. As of Nov 2017, Waymo, Google's driverless company, started running autonomous minivans around Phoenix with no humans inside to grab the wheel should something go wrong. In just a few months, passengers will be invited to climb aboard the world's first driverless ride-hailing service. Waymo is arguably the most prominent contender but is far from being the only one. Eighteen companies are vying for a leadership position in the self-driving car market including GM, Ford, Daimler, Renault-Nissan, BMW, and Tesla.

If you are wondering about the relevance of all this to Commercial Analytics for Pharma, I have three key messages for you. First, Machine Learning is extremely relevant for Commercial Analytics and the stars are aligned. There are a few things that we need to get right though. Second, we identified eight problems that lend themselves to Machine Learning treatment that happen to be key problems for Commercial Analytics. They are high ROI problems in that with the proper deployment of Machine Learning, we'll be looking at very powerful solutions that may transform Commercial Analytics as we know it. They may even rewrite the agenda of the next generation of problems to tackle. Third, we'd like to share some of the lessons we learned from doing projects and some of the pitfalls to steer clear of. In a nutshell, data acquisition and feature engineering are key and they play a larger role than algorithm selection. Also, be wary of MINA (Missing Is Now Absent) as it can doom an otherwise perfect project.

Stars are Aligned

Now is an excellent time to get started. When one is having trouble getting up to speed, it's usually because of one of three reasons. First, poor mastery of the subject matter. Today, this can be fixed easily. There is an abundance of very good material on AI and ML ranging from the basic to the very advanced that is freely available on the web. If you are unsure of how back-propagation updates the weights of the synapses or why Stochastic Gradient Descent overshoots the local minimum or how Ridge regularization differs from Lasso, there are hundreds of sites that shed light on the matter.

Second, no good platforms to work with or they are exceedingly expensive. That's definitely not the case

here. There are several open-source platforms to choose from and you do not even have to shell out a penny. Here are the major ones.

1. Google's Tensor Flow - probably the most popular one; it's great for deep learning, mathematical computations, and reinforcement learning

2. Scikit-learn - a high-level framework built on top of Numpy and SciPy that supports both supervised and unsupervised learning

3. Spark MLib - a general-purpose library that provides algorithms for most use cases and can be used with Scala, Java, Python, and R

4. Facebook's Torch - a very friendly deep learning tool which owes its friendliness to Lua, a simple scripting language

5. Université of Montréal's Theano - an excellent low-level library for scientific computing based on Python that is often used with more user-friendly programs such as Keras, Lasagne, and Blocks

6. UC Berkeley's Caffe and Caffe2 - a special-purpose machine learning environment that comes with an abundance of pre-trained models for image analysis

7. Eclipse Deeplearning4j - a deep learning programming library written for Java that includes implementations of the Restricted Boltzmann Machine, Deep Belief Networks and Deep Auto-encoders, and more.

Third, no place to turn to when one is stuck. That's also not true. We have Github and it is definitely the go-to place. It is an open-source software clearinghouse and hosts all kinds types of projects. It is truly a great resource to turn to. Odds are an answer to your question may already be there.

Now that we are ready to get started, there are two basic questions we need to address. First, which algorithm to deploy for the problem at hand? Second, is the approach scalable?

The best resource to turn to when deciding which algorithm to deploy is undoubtedly Kaggle. Kaggle is a platform where companies offer prize money to solve predictive modeling problems. Its live leader board encourages participants to continue innovating beyond existing best practice. Competitions on the Kaggle site regularly attract over a thousand teams and individuals. It was founded in 2010 and was acquired by Google in March 2017. To date, there are more than a half million registered users from 194 countries. Heavyweight participants include IBM Watson's Jeopardy-winning team, Google's DeepMind, and the like. In a nutshell, Kaggle tells us what works and what does not. Techniques Kaggle winners employ time and again include Boosted Trees (XGBoost) for classification problems and CNN (Convolutional Neural Networks) for image analysis. Ensemble techniques are also deployed to provide a boost in performance, which oftentimes is just what's needed to snatch first place.

Actually, Kaggle was inspired by the Netflix prize. Back in 2006, Netflix was selling discs of movies and TV shows and needed to improve the accuracy of its movie recommendations. Netflix offered $1 million to anyone who could improve by 10% the predictive accuracy of Cinematch, its recommendation engine. That was an instant hit and people went crazy. Tens of thousands of participants started downloading the data, building models, and uploading their predictions. This crowd sourcing initiative was a godsend for Machine

Learning as it funneled the energy of tens of thousands of people into solving the predictive problem and improving tools and techniques in the process. The competition went on for three years and the $1 million bounty was finally awarded in June 2009 to BellKor's Pragmatic Chaos group.

The Netflix problem reminded the community of the relevance of matrix decomposition. The people-movie rating matrix, which by the way is a sparse matrix as a person only rates a fraction of available movies, can be expressed as the product of two matrices: a people preference profile matrix and a movie profile matrix. Expressed this way, one can more easily infer the rating of a person for a yet unrated movie by applying the person's preference profile to the movie profile of the yet unrated movie, using a simple dot product. This line of investigation led to several improvements in SVD (Singular Value Decomposition), the technique of choice for matrix decomposition, and resulted in more sophisticated versions of the SVD including the asymmetric SVD and SVD++.

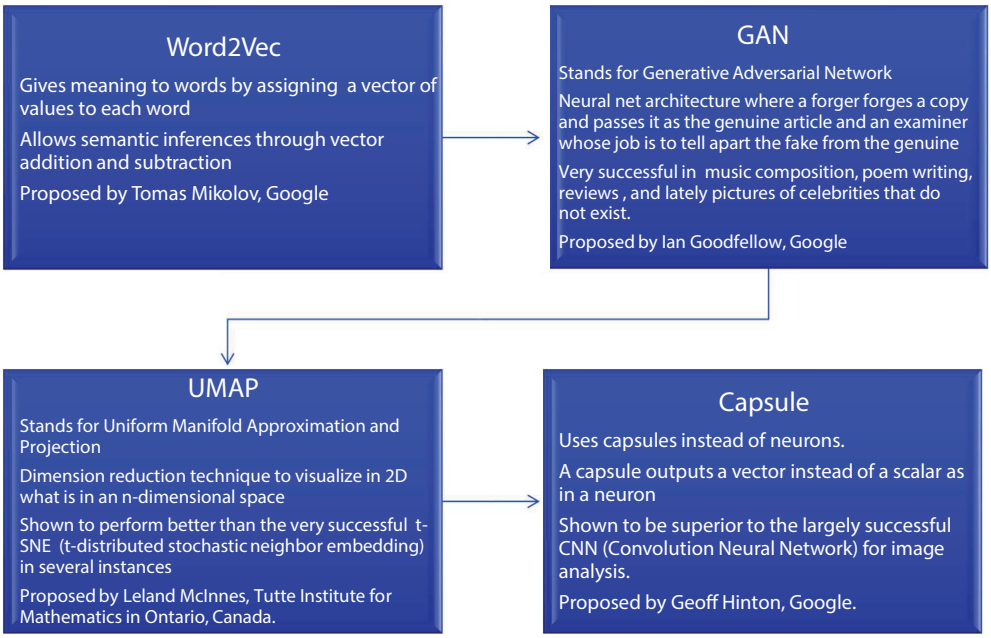
What about the scalability of the Machine Learning approach? For starters, many of the algorithms in the open-source platforms have already been deployed on very large data sets and scalability has not been an issue. In the unlikely event that scalability is an issue, there are two additional avenues one can pursue. One is to employ a cloud-based solution such as AWS (Amazon Web Services) which leverages parallel computing. The other is to deploy specialized hardware. NVIDIA just launched its Titan V graphics card which is explicitly designed for deep learning. The GPU card can be installed on a regular PC and costs about $3,000 as of this writing. In terms of performance, it is nine times faster than the previous generation at over 100 Teraflops with 12 GB of high-bandwidth memory.

A significant perk of getting involved in Machine Learning now is a large and active community. This means a constant supply of new tools and techniques. Let's mention the most notable ones.

Word2Vec is a technique that comes from Tomas Mikolov at Google. It consists of representing a word by a vector and that vector is inferred from the frequency of words that appear in the same sentence as the word of interest. As a result, similar words such as "engine" and "motor" will have similar vector representations. What's interesting is that we can use vector additions and subtractions to make semantic inferences. For instance, what is vector(Paris) - vector(France) + vector (Italy)? Well, it's vector (Rome). In the same vein, if you guess that vector(Woman) - vector(Man) + vector(King) = vector(Queen), you'd be correct. This technique is just what we need to convey meaning to a zip code: its population size, its socioeconomic profile, the presence of managed care, the influence of IDN's and the like.

More recently, Ian Goodfellow, currently with Google, came up with a novel architecture, the GAN (Generative Adversarial Network), which Yan LeCun, a prominent figure in deep learning, singles out as the most interesting idea in the last 10 years. The GAN consists of two neural nets that work against each other. One is the forger. It generates a forged copy of the real McCoy and tries to pass it for the real thing. The other is the examiner and its job is to tell apart the fake from the real one. The system has learned when the forger produces fakes that are so good that the examiner can no longer call them out. This technique has been successfully deployed to create music, poems, prose, paintings, and even pictures of celebrities that do not exist.

Figure 1: Examples of Recent Breakthroughs in Machine Learning



Then there is the recent proposal of the capsule network from the legendary Geoff Hinton, now a VP with Google. The capsule extends the neuron by having it output not a scalar but a vector to encode richer information. The capsule network also uses a dynamic routing mechanism to move information between capsules. The capsule network is meant to replace the hugely successful convolutional neural network. It captures the hierarchical relationships between object parts, which slashes the error rate of the convolutional neural network by a whopping 45%. The capsule network needs to see far fewer pictures to accomplish the same image recognition task. Also, flipping the picture upside down or presenting it from an angle does not bother the capsule network at all. The convolutional neural network, by contrast, is completely taken aback.

Another technique worth mentioning has to do with visualization. We need 2D visualization because we cannot see what is in n-dimensional space. The task at hand is to preserve neighborhood locality, which means that points that are close to each other in n-dimensional space need to be close to each other in the projected 2D space. As for points that are far apart in n-dimensional space, we do not really care how far apart they are in 2D provided that they are not too close to each other. A new technique has emerged to accomplish this task and it is the UMAP (Uniform Manifold Approximation and Projection). The claim is that it does better than the very successful t-NSE, a technique developed by Laurens Van der Maaten and Geoff Hinton. UMAP uses a fuzzy topological structure to model the information in n-dimensional space.

Another advantage of jumping on the Machine Learning bandwagon is that your programs may show improvement in performance with little intervention on your part. That's because the platform where your programs reside is constantly updating its functionalities to catch up with the latest advancements. In TensorFlow, for instance, if you replace the

keyword “GradientDescentOptimizer” with “AdamOptimizer” when making a call to the optimizer, the training time of your model will be much shorter. What’s remarkable is this benign change in keyword belies a tremendous amount of work punctuated with a string of breakthroughs. Let’s take a closer look.

Optimization is about taking a series of steps from an initial spot to land on the optimum. Needless to say, the smaller the number of steps, the faster the algorithm. In the plain-vanilla stochastic gradient, the size of the step is proportional to the slope, which means that on a plateau, the step size is very small and the algorithm very slow. If, however, we remembered the steepness of the slope we just rode down, we could use that momentum to move forward at a much faster clip. In other words, we could have the slope determine not the speed but the acceleration. Of course, we need to add some friction to ensure we do not exceed a terminal velocity, otherwise we’ll zoom past the optimum. That’s the idea behind momentum optimization. Interestingly, we can do better than that. Instead of taking the slope at the point where we are currently at, we can take the slope at a point a little further away in the direction of the momentum. This idea works well because in general the momentum points in the direction of the optimum. This improvement is known as the Nesterov Accelerated gradient.

Yet another strategy consists of fiddling with the learning rate. The learning rate controls the size of the step. Too small a learning rate and the algorithm takes forever. Too large a learning rate and the algorithm cannot find the optimum. AdaGrad uses the fact that the slope along one dimension may be steeper than the slope along another. It applies a decay factor to the learning rate and does so in such a way that the learning rate decays faster along dimensions of steep slope and slower along dimensions of gentle slope. Yet another improvement comes from RMSProp. While AdaGrad remembers the slope history of all the points that were visited, RMSProp only remembers the most recent ones. This results in a speedup while not overshooting the optimum.

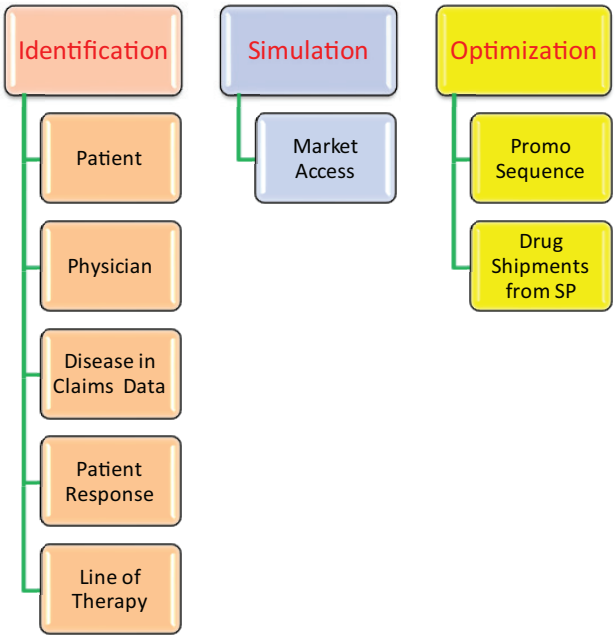
Finally, Adam which stands for Adaptive Moment Estimation leverages all these ideas and combines momentum optimization with RMSProp. It is simply the best optimization algorithm out there to date.

Eight Key Problems that Lend Themselves to Machine Learning

Below are eight problems that are central to Commercial Analytics and lend themselves to Machine Learning. Indeed, they all satisfy the three conditions for Machine Learning. First, there is a good amount of data. Second, a pattern exists and it’s not random. Third, the alternative rule-based approach to capture the pattern becomes quickly unwieldy, putting a lid on further improvement. What’s more, we can expect to see very powerful solutions come out of this new approach. They may even transform Commercial Analytics as we know it.

1. **Patient Identification** – Which patients are most likely to use our drug? This is a key question in Rare Diseases and Oncology where patients are few and the cost of therapy per patient per year is very high. We already have a fair amount of data today. They are Syndicated Claims data, EMR data, SP/SD data, and GPO data. Syndicated Claims data is a great resource as it describes the sequence of interactions of the patient with the healthcare system including doctor visits, lab tests and results, drugs prescribed, drugs administered in the office, surgical and

Figure 2: Key Problems that Lend Themselves to Machine Learning



other procedures, hospitalizations, and the like. EMR data is also a great resource even if does not identify the physician. It contains much richer data on the patient including line of therapy, lab results, vitals, family history, hospitalizations, physician notes, and the like. SP data is relevant because it provides a more complete view of our drug than syndicated data sources do. Also, it provides us with a yardstick to estimate the capture rate of competitive drugs in the syndicated data sources. Finally, GPO data informs us of drug usage in real time and can be used as site alerts for our reps to act upon.

2. **Physician Identification** – Which physicians will prescribe our drug? This question is always relevant since more prescribing physicians means more revenue, regardless of the stage of the drug in the product life cycle. Several factors determine if a physician will prescribe or not. First, the physician must have patients in need of the product. Since the number of eligible patients goes up and down over time, the physician may come across as fickle, making it difficult to predict who will prescribe at a given time. Next, the patient must be able to afford the out-of-pocket costs and that is in part a function of the insurance plan of the patient. Then, the physician must not dislike the drug, which is shaped by past experience, habits, and profile. This is captured by the residency program, hospital affiliations, involvement in clinical trials, speaking engagements, KOL status, consulting work done on behalf of pharma companies, where the physician is on the innovator-laggard spectrum, volume of patient referrals, and so on. Last, the physician may not respond to the promotional message unless it is delivered through the right channel for that physician. By the way, another complicating factor has to do with the tacit ROI assumption. We do not want to identify all physicians that will prescribe, only physicians that will prescribe within a promotional budget. At any rate,

relevant databases include patient level data, formulary access, physician profile, and channel preference.

3. **Promo Sequencing** – Once we have established which physicians to target, the next problem to address is execution. Take Dr. John Smith. Which of the following two sequences is more impactful: (1) C, C, NP, S, E, L or (2) L, C, NP, C, E, S where C stands for Call with sample, NP for No Promotion, S for sample, E for Email and L for Lunch? Actually, why limit ourselves to only those two sequences as there may be a third sequence of the same or lower cost that may be more effective? More generally, what is the optimal sequence for each physician given a promotional spend? There is one type of Machine Learning that works well for this type of problem and it is Reinforcement Learning. Reinforcement Learning sits in between Supervised and Unsupervised learning. In Supervised learning, there is a label or class for each example and our task is to find the label or class of a new example. In unsupervised learning, there is no such thing as a label or class. There are only examples and they need to be clustered along similarities that are to be uncovered from the data. Parenthetically, Reinforcement Learning is the workhorse algorithm behind AlphaGo, AlphaGo Zero, and the self-driving car. It's what Google was after when it shelled out $500 million in Jan 2014 for DeepMind.

4. **Patient Response to Drug** – Which drug will a patient respond to or show better response to? The converse is just as important: Which drug will a patient not respond to, not tolerate, or have an adverse event to? Either way, the underlying question is the same. Is there a patient profile for each type of response: Respond well, Respond, Do not respond, Do not tolerate, or Experience Adverse event. Claims data is helpful as it describes the drugs the patient has been on including combination and concomitant therapies, diagnoses, surgeries and procedures, lab tests, hospitalizations, and the like. The data is limited in that its description of the patient profile does not go beyond age, gender, ethnicity, and geography. EMR data offers a richer profile of the patient and is a great data asset to leverage. What's more, it has physician notes which may come in handy. The best data source though is arguably clinical trials data.

5. **Disease Identification from Claims Data** – Is this asthma or COPD? Type 1 or Type 2 diabetes? Bipolar or depression? Metastatic or not metastatic cancer? This question comes up whenever the drug has multiple indications or is used off-label. The diagnosis code can help resolve the matter but has its limitations. For starters, the diagnosis code may not be present in the claim. Also, the claim may indicate a made-up diagnosis to ensure that the patient gets the drug. This administrative workaround is employed when the Payer will only reimburse the drug for a specific indication and that's not the indication the physician had in mind. There are two business reasons that motivate the question. One has to do with Incentive Compensation. The typical scenario is the drug just got approved for a second or third indication. The sales force needs to direct its effort toward the newest indication, and, to that end, the pharma company rolls out an Incentive Compensation plan that only pays Reps for Rx's written for the new indication. The other reason is profit sharing, typically, between a startup that owns the molecule and a big pharma company that has an army of reps to promote the drug. Since only one company does the promotion, a natural arrangement is to split sales based on indication.

6. **Line of Therapy Determination** – This question comes up when we use claims data to figure out patient journey, and that's because claims data does not indicate when a line of therapy ends and another starts. It has to be inferred. This question also comes up with GPO data despite the fact that the GPO data indicates line of therapy. What the GPO calls first line of therapy may not be first line at all, simply the first time the GPO services the patient. The EMR is arguably the best source of line of therapy information although physicians do not always agree with each other, and this transpires in the EMR data. Overall, there is consensus for the most part. Another approach is to get medical experts in a room, show them several examples, and have them to articulate the rules that define lines of therapy. This may lead to heated debates but they'll get the job done. Now, you do not know how much of these rules is shared by the larger medical community and how much is specific to your handpicked experts. That's another reason Machine Learning is so appealing. As for the business questions that require a better understanding of lines of therapy, here are the common ones. What is the market share of our drug within a line of therapy? How fast do patients move through the different lines of therapy? If our drug is used in second line, who are the patients in first line that are most likely to move to second line and stand to benefit from our drug?

7. **Market Access** – It is well known that about half of formulary changes have no impact whatsoever on the prescribing behavior of physicians, which means that the other half does. This unleashes a series of questions. What type of formulary changes are material: changes in tier that lead to a significant difference in co-pay, Prior Authorization, Step Therapy, NDC Block, Quantity Limited, etc. Who are the Payers that are enforcing those changes and in which MSA's (Metropolitan Statistical Area)? In the traditional analytical approach, when we measure the impact of a change, we have to zero in on one change and assign the impact solely to that change. However, if we deploy a Machine Learning approach, the algorithm may factor in not only the change of interest but also changes that happened before, at the same time, and after the change of interest. By bringing in the context, the algorithm may more accurately predict the impact of the change we are contemplating through contracting with the Payer. Also, the Machine Learning approach will pick up spillover should there be spillover as it will be looking at the larger picture. The relevant data sources include patient-level data, physician-level prescriptions, and formulary changes.

8. **Shipment Optimization at the SP** – SP's face a major problem and that's costing pharma companies a lot. Indeed, SP's need to get approval from the Payer before they can ship the product to the patient. This approval process is very slow. On average, the time between writing of the prescription and shipment of the drug is in excess of 30 days. Patients cannot wait that long, so many abandon the prescription or end up using a different drug, resulting in significant loss in sales. Now, if the SP could predict which requests the Payer will grant, the SP could skip the wait and ship the drug right away. That would solve both the abandonment and switch-away problem. Why not use a Machine Learning algorithm to sort out which requests will be approved and which requests will be rejected? For starters, the SP has lots of data regarding which requests were approved, rejected, approved after the rejection is overturned, and rejected for good. For each of these cases, the SP has

information on the patient, the physician, the payer, the insurance plan, the prescription, and so on. Of course, there will be false positives. From time to time, the SP will be left holding the bag. It would have shipped the drug to the patient and the Payer would subsequently deny reimbursement. This prompts us to ask if this early shipment strategy will work. To be sure, it will if the accuracy of the algorithm is such that the new incremental sales dwarf the losses incurred by the false positives. To play it safe, the SP could choose to ship early only to patients where the probability of reimbursement is extremely high.

Lessons from Machine Learning Projects

What have we learned from the Machine Learning projects we've done? Four things. First, data is king. You will not get very far no matter how hard you try if you do not have the right data. Invest in getting the best data for the job. Second, do not underestimate feature engineering. Feature engineering unpacks information that is already available in the data. By making explicit what is implicit, it increases the predictive power of the classifier. This point is not fully appreciated though. That's because it is very tempting to embrace the romantic belief that if the information is in the data, somehow the algorithm will ferret it out. We wish that were true. Third, the algorithm. There are potentially several algorithms one may deploy for the task. There is no one algorithm that is good for all instances of a problem, otherwise there will be just one. Be open to the possibility that the best algorithm may not be your favorite and can even be one that you consider sub-par. In sum, explore and only then pick the algorithm. Fourth, beware of MINA! That's our acronym for Missing is Now Absent. We'll explain why it is so treacherous.

A. Data is King

The reason data is so crucial is because it is at the heart of how Machine Learning operates. In an expert system, for instance, we impart knowledge to the system by defining if-then rules that the system follows to draw inferences or take action when presented a new situation. In machine learning, by contrast, it is up to the system to figure out the rules it needs to deploy when presented a new situation. That's why it needs to see a lot of data. Obviously, the more data the better. Here is an example.

The task at hand here is to predict the prescribing behavior of physicians given their profile: age and gender, school attended, residency program, size of group practice, privileges in reputable hospitals, allegiance to pharma companies, indifference to drug pricing, role in patient referrals, star power as measured by paid-for trips, etc. We used boosted trees and got the AUC (Area Under Curve) to a very respectable 0.8.

Now, we all know that the Rx behavior of the physician is also contingent upon the behavior of the patients. For sure, the physician needs to put pen to paper but unless the patient hands over her money to the pharmacist, the prescription is not filled. It dawned on us that what was missing is a database that describes patient behavior at an aggregate level, which led us to develop a Panoramic Contextual database. It captures a whole array of dynamics that influence the prescription filling behavior of patients at the zip level and higher. They include:

1. Leading indicators of disease (cancer, cardiovascular, asthma, arthritis, mental health, COPD, CKD, etc.)
2. Incidence of Cancer (breast, cervix, leukemia, NHL, pancreas, prostate, bladder thyroid, etc.)
3. Exercise and fitness level (Fitbit data, fruits and veggie consumption, etc.)

4. Habits (hours of TV watching, soft drink consumption, smoking, e-cigarette, binge drinking, etc.)
5. Health Awareness (PAP smear, dentist visits, loss of teeth, etc.)
6. Education level (high school, associate degree, college, etc.)
7. Use of digital devices (computer, internet, etc.)
8. Taxes (gross income, taxable income, expected refund, etc.)
9. Crime (armed robbery, burglary, rape, arson, embezzlement, larceny, etc.)
10. Pollution (SO4, SO2, NO3, HNO3, NH4, Mg, Na, Ca, K, Cl, etc.)
11. Climate (UB Exposure, Precipitation)
12. Political Leaning (presidential voting results)
13. Business Presence – number and size of employers
14. Insurance coverage (e.g., Medicare Enrollment)
15. Road Traffic and Commuter Stress index

The enriched model now has access to both the profile of the physician and the aggregate dynamics of the patients of the physician to predict the prescribing behavior of the physician. We kept the Boosted Trees just as before and saw the AUC zoom past 0.9. Such a significant boost in performance is compelling evidence that it is worth investing in the data.

B. Feature Engineering

You have identified and leveraged all the relevant data assets you can lay your hands on. And still, the predictive power of your model lags behind. Somehow the model is not hitting on all cylinders. What's wrong? Well, there may be an issue with feature engineering.

Feature engineering is about making explicit what is implicit in the data. It unpacks information that is already available in the data through the creation of new variables from existing variables. Here is a colorful example from Kaggle. In one of the competitions, the task was to predict which car has the highest resale value. At first blush, anything could be a predictor: make and model, year, price of new, mileage, horsepower, weight, height, color, diameter of wheels, built-in GPS, AWD, and the list goes on. It turns out that the best predictor is the color of the car, but with a twist. Indeed, it has to be an unusual color for that type of car. If all medium-size sedans are, say, white, then yellow would do it. Rationale? People who purchase cars of unusual colors tend to be car buffs and they take very good care of their cars. As a result, the car is in such good condition that it fetches a handsome price at resale. For the record, this unusual color feature won first place. Note that this feature, namely, unusual color for the type of car, is akin to a standard deviation relative to a subset of the database.

Let's go back to the problem of predicting the prescribing behavior of physicians, and discuss a few feature engineering examples.

Say we are looking at an expensive drug. We'd like to have a predictor variable that captures the insensitivity of the physician to drug pricing. To that end, we can look at all the drugs the physician writes, rank order them by price, and look at, say, the 80th percentile. If that price is high, we can conclude that the physician is insensitive to drug prices. Another approach is to look at the share of branded drugs relative to generics.

The reluctance of a physician to prescribe a drug may have to do with the physician's financial involvement with other pharma companies, which as we know, is described in the Open Payments database (Sunshine Act). By looking at payments a physician perceives from pharma companies, we can develop an

allegiance index that indicates if the physician is strongly tied to one company or is open to developing new relationships with other companies.

It's always helpful to know who are the sought-after physicians. One way to do so is to look at the number of trips a physician takes on pharma's dime, and even at a breakdown of these trips by in-town, domestic, and international. Looking at year-on-year changes, we can also define features that describe how the star power of the physician is trending: rising, falling, steady, or wobbly.

Another great source of data for feature engineering is patient referrals. Looking at the data as a graph where nodes represent physicians and arcs referrals between physicians, we can establish how well a physician is connected to other physicians. Indeed, there is a whole host of centrality measures that we can deploy including degree, PageRank, eigenvector, closeness, in-betweenness, etc.

C. Algorithm

The choice of the algorithm should only be of concern once we are done with data acquisition and feature engineering. In other words, we are fully satisfied that we are deploying the best data assets for the job. Also, we have leveraged our domain expertise to the full and have made explicit all the key features that the model may need to do its job. Only then should we turn our attention to algorithm selection.

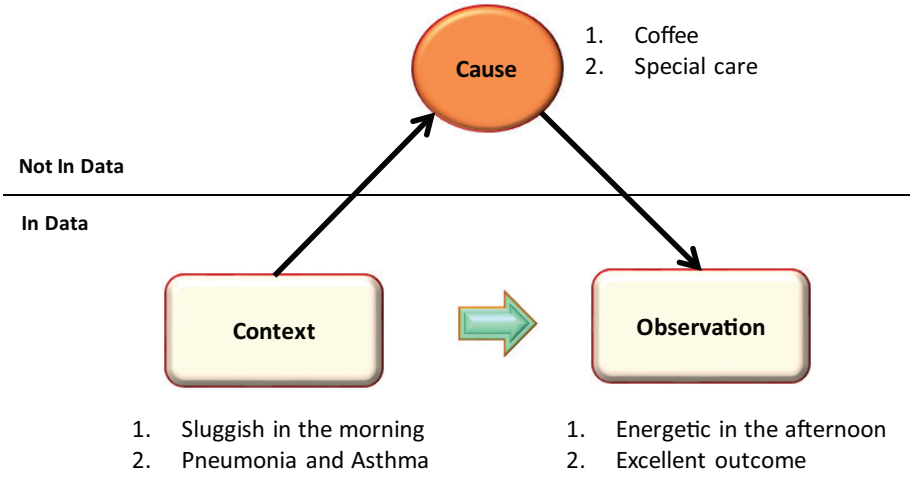
There are indeed quite a few algorithms to choose from. If we had to pick one right away, we'd probably start with Boosted Trees. In recent years, Boosted Trees won more Kaggle competitions than any other techniques. Before that, the reigning king was Random Forests and that's a good choice too. Before that, it was SVMs (Support Vector Machines) and that's not a bad choice either.

The fact of the matter is that each algorithm splits the n-dimensional feature space differently as it undergoes the process of separating the subjects to classify. Since the problem at hand distributes the subjects to classify in a very particular configuration in space, one algorithm is bound to work better than others. The issue is that algorithm may not be your favorite one, say, multilayered perceptron (MLP). It may even be an algorithm that you consider to be inferior (e.g., Naive Bayes) or not sophisticated enough (e.g., logistic regression) or of a different flavor than the one you are familiar with (e.g., kernelized SVM with radial basis functions). When that's the case, you would have missed the winning algorithm.

A study conducted recently at UPenn by Olson et al. compared the performance of 13 algorithms on 165 publicly available classification biomedical problems. Here is the finding. The top three algorithms are: Boosted Trees, Random Forests and SVM's. The bottom three algorithms are variations around Naïve Bayes: Bernouilli, Gaussian, and Multinomial. Also, for any of the 165 problems, one of the 13 algorithms came on top, which means that the "worst" algorithm (based on overall ranking) turned out to be the best for the problem at hand. What if that were the problem you were solving? Since you may not know ahead of time which algorithm is going to be the winner, a good policy may be to drop your prejudice and give all of them a chance.

In regard to the problems we worked on, and we did work on quite a few, there was always an algorithm that did better than others. However, not by much. When evaluating a model, we follow a procedure known as n-fold cross-validation. Here is how it works. Say we are looking at 10,000 subjects and n is 10. We first pull out the first 1000 subjects (1 to 1000) and train the model on the remaining 9,000 subjects. We test the model on the 1000

Figure 3: Problems Arise When What Was Missing in the Data Is Now Absent in the Real World



subjects that we pulled and that's one score. Next, we pull out the next 1000 subjects (1001 to 2000) and train the model on the remaining 9,000 subjects. We test the model on the 1000 subjects that we pulled and that's the second score. We repeat this process 10 times and take the worst of the 10 scores to be the score of the algorithm. Here's what we observed. For a couple of folds and sometimes several, the fold score of the runner-up algorithms is better than some of the fold scores of the winning algorithm.

Of course, the ideal is to identify the best algorithm for the job. The truth of the matter is that even if you miss and pick the second or third algorithm, things are not that bad. What this suggests is that you may be better off investing more time and energy in data deployment and feature engineering than sweating over algorithm selection.

D. Beware of MINA

MINA is an acronym we coined for "Missing is now Absent" to refer to a phenomenon that can wreak havoc in Machine Learning models. It is best explained with examples.

You observe over time that when a colleague is sluggish before lunch, the colleague is energetic in the afternoon. The next time you see a sluggish colleague in the morning, you predict that the colleague will be energetic in the afternoon. And you are right. One day, to your surprise, all your sluggish colleagues look drowsy in the afternoon. What happened? You discover that the coffee machine is broken and understand that the afternoon source of energy has been disrupted. Here's the point. The fact that the coffee machine is missing in your mental model is immaterial so long as it is there in real life. Problems start the day the coffee machine is absent in real life. Indeed, things take a different turn and you cannot explain why.

Here's another example from the triage of pneumonia patients in ER. The data suggests that patients that have pneumonia and asthma do extremely well and patients that have pneumonia but not asthma do just fine. As a result, the recommendation of the machine learning algorithm is to de-prioritize patients that have asthma. That's actually a very bad idea. The reason patients that have pneumonia

and asthma do extremely well is because they are high-risk patients and, as a result, are given special care. What's causing the great outcome is the care, not the asthma. The algorithm's recommendation misses the point and suggests getting rid of the special care. That's because care has been missing in the data all along.

In both cases, something momentous happened. The cause has disappeared in real life (coffee machine, special care) along with its implications. But to the data, nothing has changed. The model does not know about the change since the cause was never captured. As a result, the algorithm makes the same prediction as before, but this time it is off. (Figure 3)

The fix? Explain the prediction. Why are my colleagues so full of energy in the afternoon? Why do the pneumonia and asthma patients do so well? If the subject-matter expert cannot explain the recommendation based on a description of the situation as captured in the data, something important is missing (coffee machine, special care). In that case, we should refrain from following the recommendations of the Machine Learning model. Better be safe than sorry.

Conclusion

As discussed throughout this paper, now is a great time to get started with Machine Learning. The field is making progress by leaps and bounds. There is a vibrant community of practitioners across virtually all verticals. There are several open-source platforms to choose from and countless resources to turn to. There are also cloud-based solutions and specialized hardware should you require serious scalability. What's more, it's still early morning on pharma's clock and there are countless opportunities to seize.

Be ready for challenges. If what you read makes you feel you are lagging behind, take the write-up with a pinch of salt. Many who write about Machine Learning are not practitioners and have not wrestled with the myriad of problems that bedevil the task. So, they naturally paint a rosy picture and even though they do not mean to mislead, they do. Think about it. Who, apart from the practitioner, wants to hear an exposé of challenges and nuances that can only blur an otherwise perfect picture?

About the Author

Jean-Patrick Tsang *is the Founder and President of Bayser, a Chicago-based consulting firm dedicated to pharmaceuticals sales and marketing. JP has worked on 250+ projects to date including ROI optimization, data strategy, and segmentation & targeting. For the last few years, JP and his team have been focusing on Predictive Analytics using Machine Learning. JP publishes and gives talks on a regular basis and runs one-day classes on various subjects related to data and analysis. In a previous life, JP deployed Artificial Intelligence to automate the design of payloads for satellites and was the adviser of two PhD Students.*

JP holds a Ph.D. in Artificial Intelligence from Grenoble University and an MBA from INSEAD in France. He was also the Recipient of the PMSA Lifetime Achievement Award in 2015. He can be reached at (847) 920-1000 or bayser@bayser.com.

The Causes and Solutions of Chronic Drug Shortages in the United States: The Important Role of Better Analytics – A Commentary

George A. Chressanthis, Ph.D., Principal Scientist, Axtria Inc.

Abstract: A vexing public health policy issue plaguing the United States (US) pharmaceutical industry has been the existence of chronic drug shortages since 2000 with no end in sight. Chronic drug shortages persist despite attempts by the Food and Drug Administration (FDA) and federal legislation to remedy this problem. Drug shortages are severe enough to potentially cause adverse effects on patient health outcomes, changes from optimal treatment, and added costs to the healthcare system. Prior research shows drug shortages being caused by a lack of economic incentives, supply chain factors, manufacturing-quality problems, and managing regulatory expectations. However, the application of *better analytics* has recently been cited as needed to account for these effects and changes in market forces, business continuity planning, supply chain management, and improved insights into future demands through better forecasting to reduce the number and severity of drug shortages. This paper is a call to action to Pharmaceutical Management Science Association (PMSA) members to address this important health policy issue.

Keywords: chronic drug shortages, public health and business policy, pharmaceutical decision science analytics

1. Existence of Chronic Drug Shortages

Since 2000, one of the more vexing and troubling public health policy issues that has plagued the US pharmaceutical industry, yet receiving less deserving public news coverage than other industry stories, has been the existence of chronic drug shortages. A drug shortage is defined in which the "total supply of all clinically interchangeable versions of an FDA-regulated drug is inadequate to meet the current or projected demand at the patient level."[1] The peak year was 2011, with 251 drug shortages, 73% being generic sterile injectable drugs used to treat cancer, sepsis, and many other life-threatening conditions.[1] While the annual number has dropped, the issue of chronic drug shortages still persists, despite attempts by the FDA and federal legislation to remedy the problem.[2-3] The 2011-2014 period saw 456 situations of drug shortages severe enough to potentially cause adverse effects on patients and changes in treatment.[4] Health practitioners in office-based and hospital settings, as well as policymakers, have been alarmed at the adverse consequences to patients and higher costs to the healthcare system caused by persistent and prolonged drug shortages. The issue of chronic drug shortages in the US recently came to the forefront of health policymakers with discussions announced between the FDA and Pfizer regarding the shortage of numerous injectable medicines, including emergency syringes of epinephrine.[5] According to the FDA, manufacturing, distribution, and third-party delays were cited by Pfizer for the shortages.[5] Moreover, in October 2017, the FDA announced an initiative (with more long-term changes planned) to provide guidance to generic manufacturers on the most efficient way to develop complex difficult-to-manufacture medicines (e.g., injectable medications and

drug-device combination medicines) that are often singled-sourced even after patient expiration.[6]

This paper will briefly address the following two questions related to chronic drug shortages in the US:

- What are the key causes and solutions to the existence of chronic drug shortages in the US?
- Is there a role for the application of pharmaceutical decision science analytics to help mitigate the problem of chronic drug shortages?

2. Causes and Solutions to Chronic Drug Shortages

Five factors have been identified as driving the number of drug shortages in the US,[4] providing conclusions that closely align with insights also reported in the academic literature.[1-3,7]

1. *Market withdrawals*. A high percentage of drug shortages originate from single-sourced injectable generic manufacturers. Maintaining quality controls is difficult given the low margins received for producing more complex and costly injectable drugs. The marginal cost of production is far greater for manufacturing injectable drugs than traditional small molecule pills. Moreover, given the specialized nature of producing injectable drugs, manufacturing lines are not easily transferable to the production of other drugs. Thus, when a single-sourced manufacturer is shut down due to failure to meet FDA drug quality regulations, there is insufficient supply to meet demand, thus resulting in a shortage. Financial incentives, such as instituting an investment tax credit specifically targeted to generic manufacturers for maintaining the quality of production facilities, could be employed to encourage companies. Given the social costs of higher healthcare spending caused by drug shortages, such an investment tax credit could make economic sense when comparing net marginal social benefits to costs.

2. *Supply chain design*. Improvements in supply chain management by companies are necessary, especially by improved demand estimation for a product through better coordination of processes of sales, demand planning, inventory management, and production. Such process improvements would allow for more accurately estimating capacity requirements and establishing manufacturing redundancies to mitigate the effect of production breakdowns that occur in the supply chain system. The application of decision science analytics can provide greater clarity in these processes and will improve business planning.

3. *Purchaser-manufacturer incentives*. As alluded to above (factor 1), insufficient financial incentives are a major factor in contributing to drug shortages. The formation of guaranteed-volume contracts, or the ability to retain contracts, would allow for lessening the risks of investments in manufacturing equipment needed to produce these specialized medicines.

4. *Limited market insights into future demands*. The study referenced here and interviews conducted with pharmaceutical executives found that improvements are needed to obtain better information on expected demand.[4] Internal operations improvements are needed in the areas of sales and operations planning, demand forecasting, and market environmental information that affects external systems and programs. Again, this key factor points to a role in expanding the use of decision science analytics to help reduce drug shortages.

5. *Managing regulatory expectations*. Executive interview comments noted that regulations affected drug shortages given

production delays and higher costs to receive approvals for expanding manufacturing capacity or improving existing equipment. Further, many of the drug shortages involve older medicines developed 10-20 years ago, where government regulations prevent product and process improvements given the risks and costs. The key takeaway is that government policy must balance the marginal cost from added regulations versus the marginal benefit of imposing such rules.

3. Role of Pharmaceutical Decision Science Analytics

The preceding five factors listed as causes of drug shortages represent an important opportunity role of pharmaceutical decision science analytics to help find and implement solutions to resolve this problem.

1. Employ causal-based prediction models to determine the likelihood of single-sourced manufacturers from withdrawing from the market. These models can be developed and used by both public policymakers and individual companies to anticipate market disruptions due to market withdrawals.

2. Use data mining techniques to analyze and uncover previously unknown reasons for drug shortages. This is not to suggest relying on such techniques for prediction, but rather to provide insights into potential new reasons why drug shortages are occurring. New learning can then be adopted into taking steps using causal-based models to develop business and public policy to mitigate the likelihood of drug shortages.

3. Develop inferential-based cost and production function models to determine the effect of a lack of financial incentives in impacting manufacturing disruptions, and to estimate what changes in financial incentives are needed to minimize the likelihood of stoppages in production.

4. Improve demand estimation through the development of more granular geographic-based models that can increase the accuracy of supply chain design and associated business planning processes. Regional and metropolitan statistical (MSA) demand estimation models should be employed to improve demand accuracy. Significant intranational geographic variations in demand exist, such as variations caused by managed care plan dynamics, healthcare system design, macroeconomic factors, etc. The previous factors mentioned justify employing more granular demand estimation models.

5. Apply demand estimation algorithms found in (4) to machine learning to continuously update demand relative to production capacity to anticipate potential market shortages. Determinants of demand estimation can be done more quickly and regularly to assess potential conflicts with current production volume and capacity levels.

6. Improve demand forecasting and prediction models based on results from (4) and (5) as opposed to relying on naïve-based models. Using hold-out time periods can be used to test the accuracy of causal-based forecasting and prediction models. The main advantage of using causal-based models is that one can see what factors are associated with changing forecasts and predictions. Improvements in estimating and forecasting/predicting demand can be used to enhance the accuracy of inventory management models. These developments allow for better management control and understanding what policy variables can be employed (and by how much) to mitigate adverse effects from unanticipated changes in demand.

7. Include the effect from guaranteed-volume contracts and other means to retain contracts on purchaser-manufacturer incentives when conducting model specification in estimating cost and production functions. A similar specification addition should include effects from regulatory changes on drug cost and production. Analysis should also be conducted on specific drugs, and segmented by therapy class and drug age, to determine the existence of variations in systemic effects by key drug and market characteristics. When conjoined with demand models, these analytics can then be used to see more accurate changes in the likelihood of drug shortages and specifically what to do about it.

8. Develop and implement dynamic tool technologies to apply all previously-stated improved demand, cost, and production estimation and forecasting/prediction models for both public policymakers and pharmaceutical companies.

9. Enhance current data cloud information management that feed all the analytics previously stated. Adding data from non-traditional sources, such as claims and electronic health records, may be necessary to better understand the patient volume, factors associated with disease treatment, and the success of drug utilization on outcomes. This data can be analyzed to understand potential changes in future demand by measuring the success or failure of drug intervention in disease treatment.

4. Conclusions

The chronic problem of drug shortages in the US does not appear to be going away any time soon, despite numerous government policy efforts to address the problem. The social costs to our society caused by drug shortages, though not directly measured, are likely significant, in reductions in health outcomes and increased economic burden. Aside from the typical factors seen as causing drug shortages (as described above) is the growing realization how the application of improved decision science analytics could provide much needed information for manufacturers to reduce the problem through more effective planning. Typically, pharmaceutical decision science analytics address sales and marketing questions. However, there may be an expanded role for decision science analytics to help manufacturers institute more effective and efficient processes to mitigate the drug shortage problem that plagues the system.

"There's no such thing as a free lunch."
Milton Friedman (1912-2006); American economist and 1976 Nobel Memorial Prize in Economic Sciences

In conclusion, alleviating the problem of chronic drug shortages in the US requires resources, as implied by the above quote. Investments are needed to upgrade manufacturing/planning processes and supply chain management systems and the application of management science analytics to derive information needed to support these improvements to reduce the number and severity of drug shortages. Added incentives are also required for companies to continue bringing these life-saving medicines to the market. Lastly, companies should look to enhance capabilities in cloud information management, apply data from newer sources such as claims and electronic health records, employ machine learning and data mining, and deploy dynamic analytical tool technologies that can be leveraged with a range of pharmaceutical decision science analytics to help alleviate this national critical health problem.

About the Author

George Chressanthis *is Principal Scientist at Axtria, a big data and analytics company, since July 2016. This thought leadership role involves disseminating pharmaceutical research ideas on a wide variety of topics of importance to industry practitioners. His research focuses on key trends affecting pharmaceutical commercial strategic & operational issues and their intersection to HEOR/RWE modeling associated with measuring outcomes and their implications on changes in commercial analytics, new commercial model design, payer analytics, and public policy. He also conducts numerous workshops on the pharmaceutical industry within Axtria for training and development of employees. He spent almost 15 years working in the pharmaceutical industry from 1995-2009 after a long career in academia, with the majority of his time at AstraZeneca Pharmaceuticals LP US headquarters leading teams in support of sales force strategy, sales operations, and other commercial analytical and strategic functions. Other pharma industry experiences include roles at Wyeth-Ayerst Laboratories, IMS Health, and ZS Associates.*

He has had two academic careers. His second career involved holding full-time professorships in Healthcare Management and Marketing in the Fox School of Business and a secondary professor appointment in Clinical Sciences in the School of Medicine at Temple University from 2010-2016. His first career was as an academic economist from 1982-1995, eventually becoming a tenured full professor at Mississippi State University. He received his Ph.D. in Economics from Purdue University.

References

1 Donohue J, Angus D. National shortages of generic sterile injectable drugs: norepinephrine as a case study of potential harm. *JAMA*. 2017;1415-1417.
2 Mazer-Amirshah M, Goyal M, Umar S, et al. US drug shortages for medications used in adult critical care (2001-2016). *Journal of Critical Care*. 2017;41:283-288.
3 Chen S, Fox E, Hall K, et al. Despite federal legislation, shortages of drugs used in acute care settings remain persistent and prolonged. *Health Affairs*. 2016;35:798-804.
4 The Pew Charitable Trusts and International Society Pharmaceutical Engineering. Drug shortages: an explanation of the relationship between U.S. market forces and sterile injectable pharmaceutical products interviews with 10 pharmaceutical companies. Report from the The Pew Charitable Trusts and International Society Pharmaceutical Engineering, published online January 2017, available at http://www.pewtrusts.org/~/media/assets/2017/01/drug_shortages.pdf (accessed 9 February 2018).
5 HealthDay News. Lifesaving drugs from Pfizer in short supply: FDA. *HealthDay News*, published online 16 June 2017, available at http://www.health.com/healthday/lifesaving-drugs-pfizer-short-supply-fda (accessed 9 February 2018).
6 Fox Business. FDA acts to encourage generic competition for complex drugs. *Fox Business*, published online 2 October 2017, available at http://www.foxbusiness.com/features/2017/10/02/fda-acts-to-encourage-generic-competition-for-complex-drugs.html (accessed 9 February 2018).
7 Woodcock J, Wosinska M. Economic and technological drivers of generic sterile injectable drug shortages. *Clinical Pharmacology & Therapeutics*. 2013;93:170-176.