JPMSA JOURNAL OF THE PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

SPRING 2020 | 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.

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

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

PMSA Board of Directors

Jin Tong, Novartis
President

Vishal Chaudhary, Sanofi
Vice President

Igor Rudychev, AstraZeneca
Treasurer/Secretary

Nuray Yurt, Novartis
Research and Education Chair

Haibei Jiang, Bayer
Marketing Chair

Nathan Wang, Janssen
Program Committee Lead

Mehul Shah, Sanofi
Program Committee

Scott Hull, Abbive
Program Committee

Jennifer Soller, Genentech
Professional Development Chair

Karthikeyan Chidambaram, Genentech
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 © 2020 by the Pharmaceutical Management Science Association. All rights reserved.

PMSA Journal: Spotlighting Analytics Research

Welcome to the eighth 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.org.

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:*

George A. Chressanthis, Ph.D., Axtria

Mitch DeKoven, MHSA, IQVIA

Adam Dubrow, Crossix

Ewa J. Kleczyk, Ph.D., Symphony Health Solutions

Dan Lee, DSI

Sudhakar Mandapati, Strategic Research Insights, Inc.

John Qu, Ph.D., Novartis

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

Editor: Nuray Yurt, Novartis

Key Drivers for Successful Patient Event Prediction: Empirical Findings on What Matters and to What Extent

Srinivas Chilukuri, Principal Data Scientist, ZS Associates and Sagar Madgi, Senior Data Scientist, ZS Associates

Abstract: Observational healthcare databases, such as administrative claims and electronic health records, present rich data sources for knowledge discovery from patient longitudinal histories. One such use case is the prediction of various events across the patient treatment journey, such as diagnosis and therapy initiation, progression or discontinuation.

If implemented well, patient event prediction models enable several applications in the commercial (predictive customer targeting, patient services design) and research (target patient universe determination, trial site selection) domains. However, owing to the richness, complexity and nuances in the data, there are several things to get right when it comes to model design. For instance, selection of right data set and sample size, length of medical history, prediction time window, modeling parameters, type of features (recency, frequency, sequence); and mechanism of feature generation (knowledge-driven vs. automatically generated).

In this paper, we present empirical findings on how these considerations weigh on model performance and downstream utility, drawing upon results from a diverse set of use cases spanning multiple therapy areas.

Keywords: Patient Event Prediction, Machine Learning, Knowledge Features, Data-Driven Features, Prediction Window

1. Background

1.1 Introduction and Motivation

The focus of the pharma industry is increasingly turning to specialty products for treating niche conditions. For the brands to succeed in this environment, it is imperative to identify the right patient at the right time. This necessitates predicting patient events ahead of time, which can then inform several downstream applications such as clinical trial planning, predictive customer targeting, personalized patient assistance, etc. Traditionally, these have been approached from a clinical perspective (e.g. $CHADS_2$, MELD etc.) but such scores are not available across the spectrum of events and so there is a need to build prediction models.

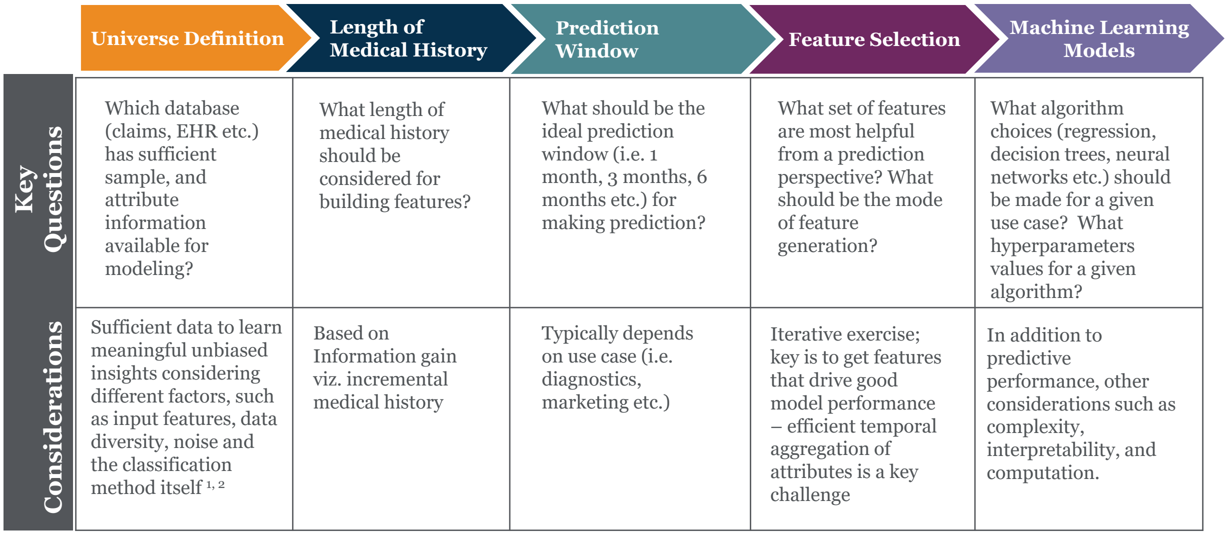
However, it is non-trivial to set up, operationalize and derive business value from patient event prediction models. This is because several distinct aspects must come together into a coherent machine learning pipeline for such efforts to be successful. This paper aims to discuss the critical success factors and provide empirical guidance for brand/analytics leaders and data scientists who would be undertaking such endeavors.

1.2 Prediction Modeling Components and Scope of the Paper

A typical patient event prediction model involves the following key ingredients:

- Right patient data set
- Representative patient sample
- Optimal length of medical history
- Optimal prediction time window
- Exhaustive model features (hypotheses underlying events)
- Suitable machine learning model and corresponding hyperparameters

Figure 1: Key Steps in Setting Up a Patient Event Prediction Model



	Universe Definition	Length of Medical History	Prediction Window	Feature Selection	Machine Learning Models
Key Questions	Which database (claims, EHR etc.) has sufficient sample, and attribute information available for modeling?	What length of medical history should be considered for building features?	What should be the ideal prediction window (i.e. 1 month, 3 months, 6 months etc.) for making prediction?	What set of features are most helpful from a prediction perspective? What should be the mode of feature generation?	What algorithm choices (regression, decision trees, neural networks etc.) should be made for a given use case? What hyperparameters values for a given algorithm?
Considerations	Sufficient data to learn meaningful unbiased insights considering different factors, such as input features, data diversity, noise and the classification method itself [1, 2]	Based on Information gain viz. incremental medical history	Typically depends on use case (i.e. diagnostics, marketing etc.)	Iterative exercise; key is to get features that drive good model performance – efficient temporal aggregation of attributes is a key challenge	In addition to predictive performance, other considerations such as complexity, interpretability, and computation.

Each of these steps involve making choices across several potential options (see Figure 1 for examples). This makes the whole process complex and time consuming.

Domain experts can hypothesize parameters for each step in the process that will lead to the best model performance; however, in most cases, determining these parameters is a matter of conjecture and presents combinatorially complex possibilities. Typically, multiple iterations are run to understand the effect of varied permutations and combinations of choices made in each step on model accuracy before finalization of the model. This is a time-consuming and laborious process, even for experts.[3]

Essentially, there's a trade-off then between model performance and resources (time and cost) that can be expended on improving the model accuracy. The cost associated with the exploration is steep and may increase rapidly with number of combinations, without a corresponding increase in performance. It is critical to find the right balance between exploration and performance.

To this end, this paper seeks to present empirical findings that can be used as reference for finding the right balance between the two. These will be based on observations derived from a set of event prediction experiments implemented for different use cases and therapy areas. Our focus is not on understanding the causal relationships between input variables and outcomes but more around understanding how choices for a specific set of parameters (observation windows, variables, etc.) affect model performance. Specifically, we seek to understand how the following considerations weigh on performance in patient event prediction:

1. **Length of medical history** – What is the right length of time window over which a patient's prior medical events should be considered?
2. **Prediction window** – What is the optimal time window to making predictions so they are actionable for the end users and address the business need at hand?
3. **Feature classes and types** – What kind of features matter across use cases and therapy areas? What is the relative importance of features across concept domains (diagnosis, medication, procedure, etc.) and type (recency, frequency, sequence, etc.) of features?
4. **Mechanism of feature generation** – how does mechanism of feature

Figure 2: Use Cases Across Patient Journey

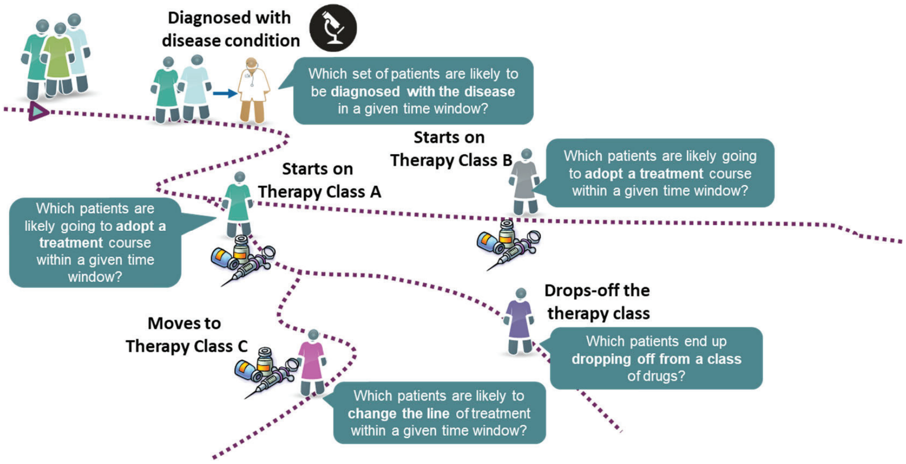
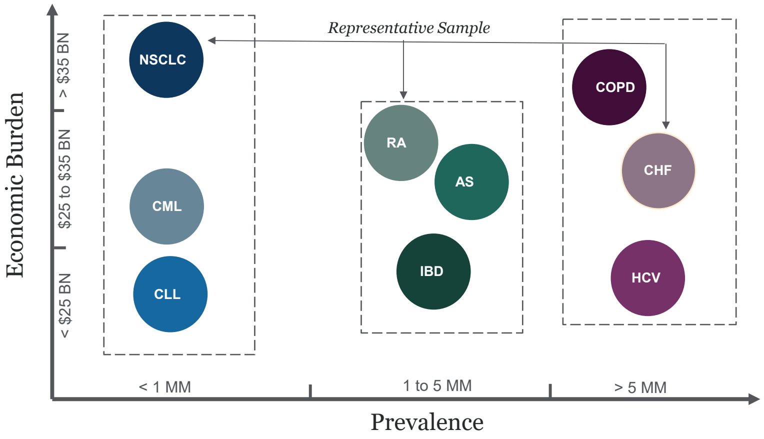

Diagnosed with disease condition
Which set of patients are likely to be diagnosed with the disease in a given time window?
Starts on Therapy Class B
Which patients are likely going to adopt a treatment course within a given time window?
Starts on Therapy Class A
Which patients are likely going to adopt a treatment course within a given time window?
Drops-off the therapy class
Which patients end up dropping off from a class of drugs?
Moves to Therapy Class C
Which patients are likely to change the line of treatment within a given time window?

Figure 3: Disease Areas by Prevalence[7-13] and Economic Burden [14,15-21]



generation (clinical knowledge vs. data-driven) weigh in on prediction accuracy?

These experiments will be conducted using a claims database, a de facto standard choice for event prediction algorithms, given their ubiquity across applications and higher accuracy.[4,5] While recent literature suggests that EHR does add more predictive power[6], exploration of impact of adding EHR data will be out of scope for this study, and will be assessed in a subsequent paper.

In the rest of this paper, we describe the experimental setup and results, concluding with a discussion and assessment of findings that could be used to guide the setup of event prediction models in general.

2. Experimental Setup

2.1 Prediction Use Cases and Disease Areas

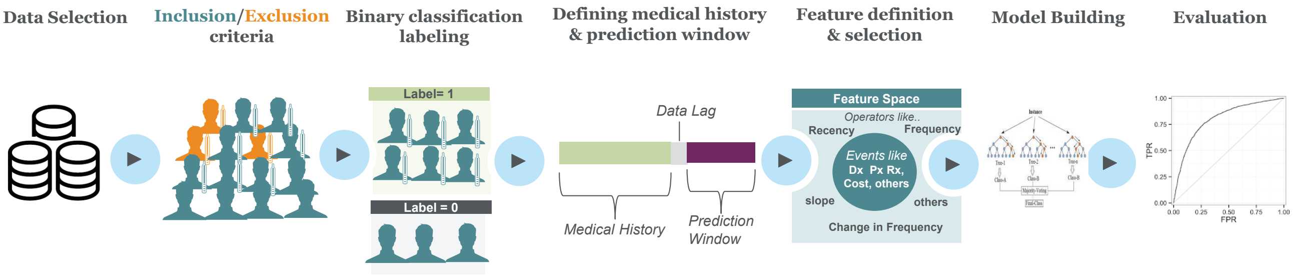
We focus on four major events occurring through a patient's journey (see Figure 2) viz. diagnosis of a condition, treatment adoption, treatment progression (change in line) and treatment drop-off.

From a disease area perspective, the experiments will focus on disease areas which a) are representative of low prevalence to high prevalence scenarios, and b) where event prediction models are relatively important for healthcare stakeholders such as payers, providers and pharma, given the burden imposed by these diseases. To this end, we believe the following diseases can serve as a good representative set (see Figure 3).

Table 1: List of Disease Areas and Use Cases for Experimentation

Prediction Use Case	Oncology (NSCLC)	Immunology (RA)	Primary Care (CHF)
	Identify patients who are likely to		
Disease Diagnosis	be diagnosed with metastatic NSCLC	be diagnosed with RA	be diagnosed with CHF
Treatment Adoption	adopt an EGFR drug	adopt an anti-TNF drug	adopt an ACE-inhibitor drug
Treatment Change (Line Switch)	switch to 2nd line	switch to 2nd line biologic	switch to 2nd line
Treatment Drop-off	drop off from an EGFR drug therapy	drop off from an anti-TNF therapy	drop off from an ACE-inhibitor therapy

Figure 4: Typical Patient Event Predict Model Development Pipeline



- Oncology – Non-small cell lung cancer (NSCLC)
- Immunology – Rheumatoid arthritis (RA)
- Primary care – Chronic heart failure (CHF)

Building on the above choice of use cases and disease areas, experiments will be conducted across the set of event prediction models listed in Table 1.

The effect of varying length of medical history will be tested for all the event prediction models listed above, as it is varied from 6-36 months. The impact of varying prediction window (from 1 month to 6 months) will be tested for models built with 12-month medical history.

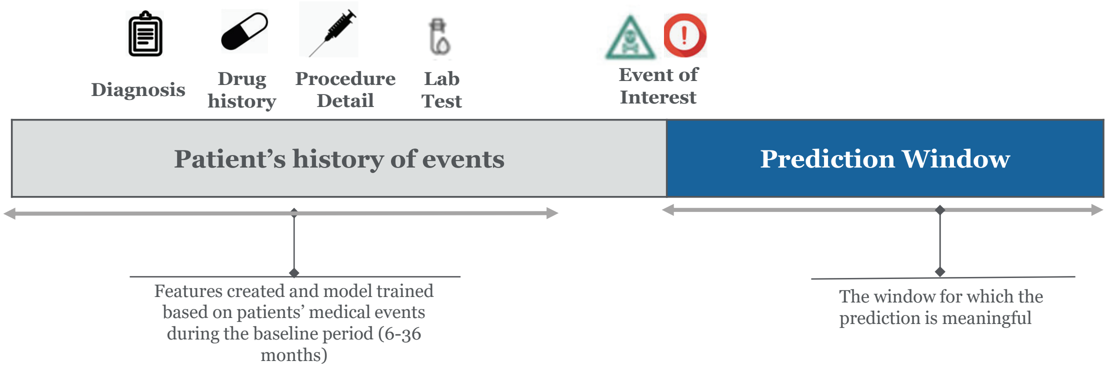
The impact of knowledge-driven vs. data-driven features on model performance will be assessed for diagnosis use cases across all the above disease areas; these are the cases where existing research gives us a good baseline for building knowledge-driven features. This will be done for models built using 12-month medical history.

Figure 4 shows the typical machine learning pipeline for building patient event prediction models. The same steps have been followed in our experimental set up.

2.2 Data

The experimental results in this paper are based on Optum's de-identified Clinformatics™ DataMart US healthcare claims database for NSCLC, RA and CHF during the time period 2012-2018.

Figure 5: Specifying the Prediction Problem


Diagnosis
Drug history
Procedure Detail
Lab Test
Event of Interest
Patient's history of events
Prediction Window
Features created and model trained based on patients' medical events during the baseline period (6-36 months)
The window for which the prediction is meaningful

2.3 Specifying the Prediction Problem

The prediction problem will be generalized as following, i.e., given a target cohort of patients with certain medical history, what is the probability of a patient experiencing an event of interest in the given prediction window? The target population and outcomes will be defined based on a set of inclusion/exclusion rules (such as occurrence of certain diagnosis, medication, labs etc.).

Additional details around inclusion/exclusion criteria and sample size for each model are available in the appendix.

2.4 Feature Definition and Selection

Variables derived from demographics, symptoms, comorbidities, drugs, procedures, visits and other observations recorded prior to the anchor date (see Figure 5) will be used for feature creation across experiments. To enable rapid experimentation, an automated, intelligent feature generation and selection framework will be utilized for agile discovery of relevant features. Additional details around this framework are available in the appendix.

For experiments involving knowledge-driven features, features will be defined based on existing research and expert opinion.[22,23,24,25] Details around these knowledge-driven features are available in the supplementary data file (at http://www.pmsa.org/_resources/journal/2020/ZS-KeyDrivers/KeyDrivers.xlsx).

2.5 Model Building

Labeled data, along with features, will be split into train and test sets in a 70:30 ratio. XGBoost, a state-of-the-art machine learning model, will be utilized for model training. We prefer XGBoost over more complex models such as Artificial Neural Networks (ANN) given our focus on explainability in addition to prediction performance. Appropriate hyperparameter tuning will be performed to ensure optimal learning. Additional details around the models are available in the appendix.

2.6 Model Evaluation

The models will be validated on test dataset, defined in the previous phase. The area under the receiver operating curve (AUC) will be used for assessing the performance of models across different scenarios of medical history, prediction window and mechanism of feature generation. See Figure 6.

In general, the higher the AUC the better a model performs. A random model without any predictive power generally results in a 50% AUC. On the other hand, a perfect model would result in a 100% AUC. While it varies by use case, typically AUC of 70% is considered good and 85% or above is considered excellent.

Feature importance will be assessed using the adjusted F-score. Additional details around

Figure 6: Area Under the Receiver Operating Curve (AUC)

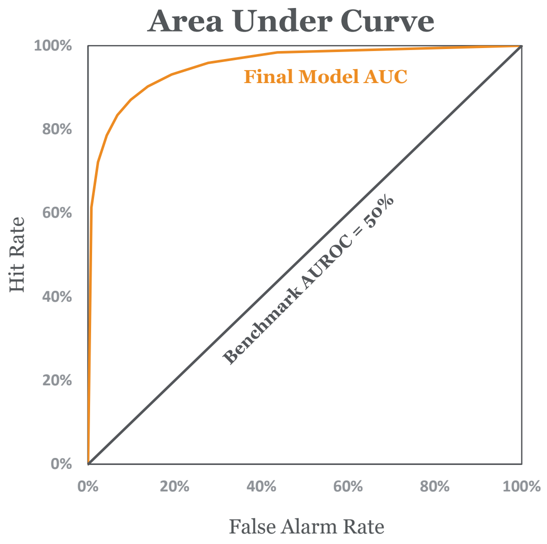
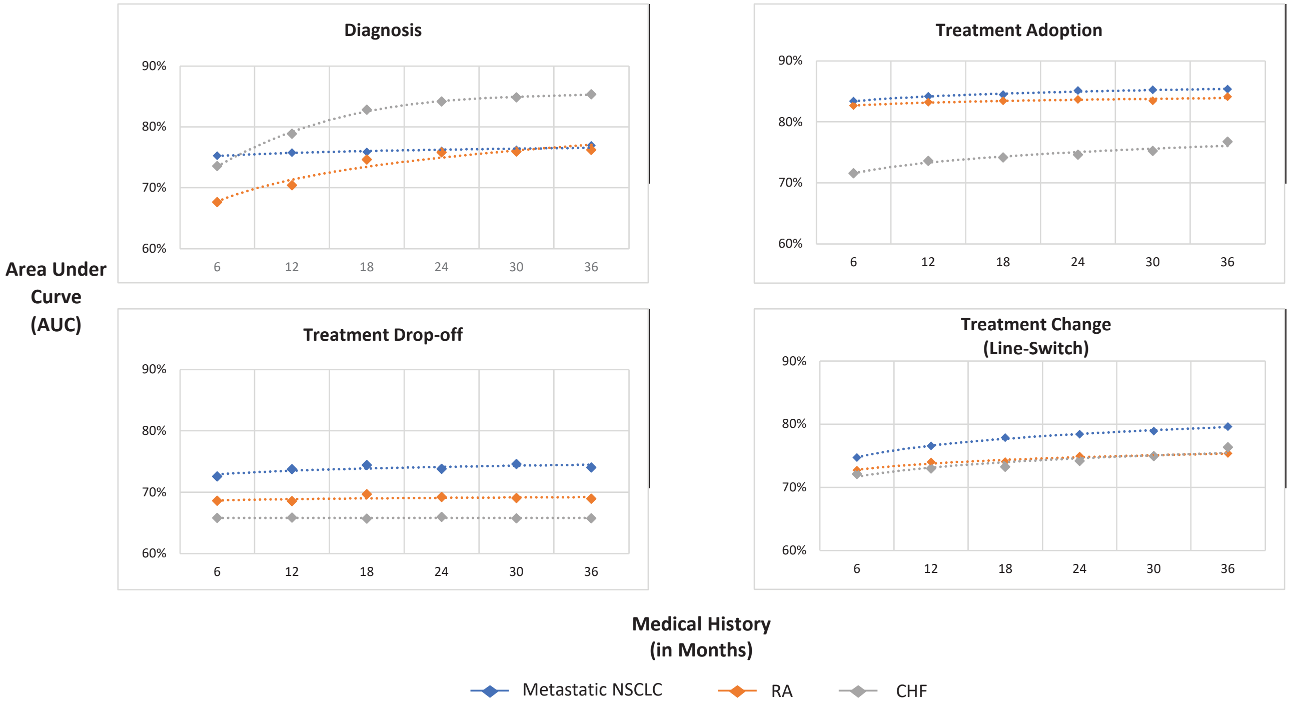


Chart 1: Prediction Performance (AUC) vs. Medical History (in Months)



AUC, and evaluation for various scenarios, are available in the appendix.

3. Results

3.1 Length of Medical History

Except for treatment drop-off and mNSCLC diagnosis, we note a monotonic lift in AUC across all use cases as medical history is varied from 6 months to 36 months. The gain, however, seems to be marginal after a certain length of history – 6 months in case of mNSCLC diagnosis and treatment drop-off vs. 18 months in other use cases. See Chart 1.

Chart 2: AUC for Different Prediction Windows

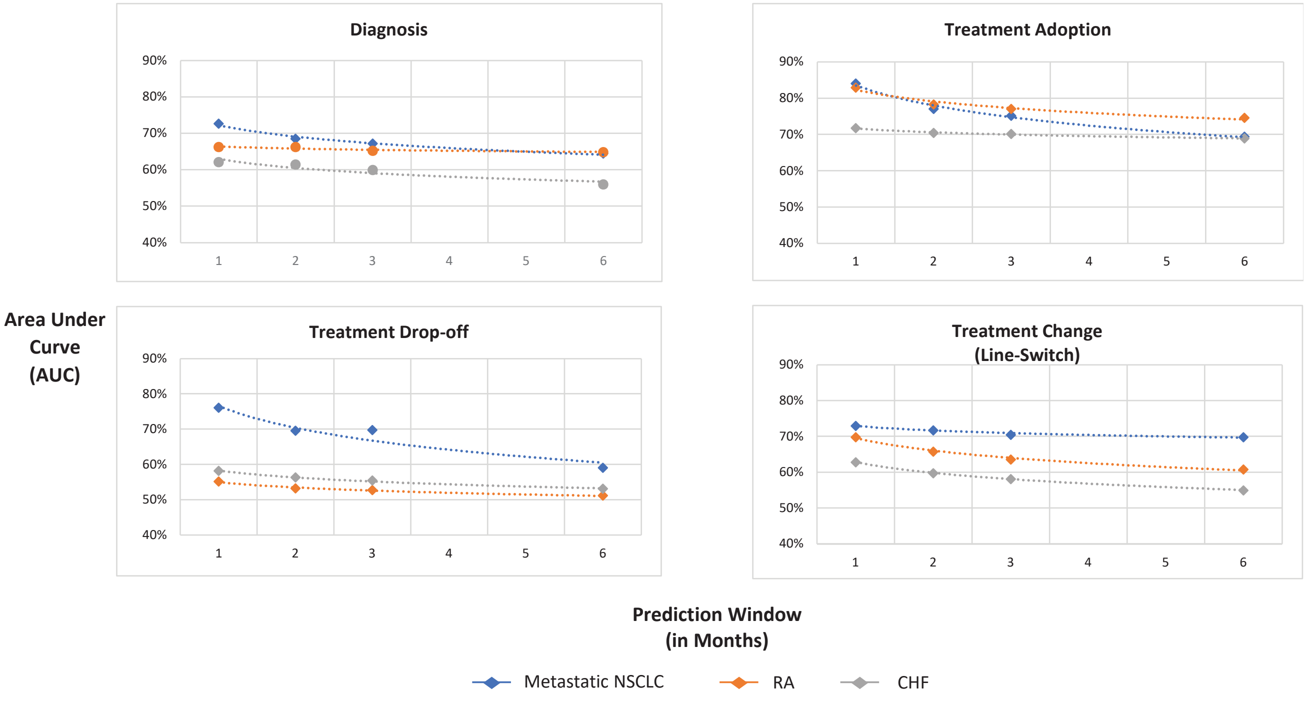


Chart 3: Relative Importance of Concept Domain Features Across Different Use Cases

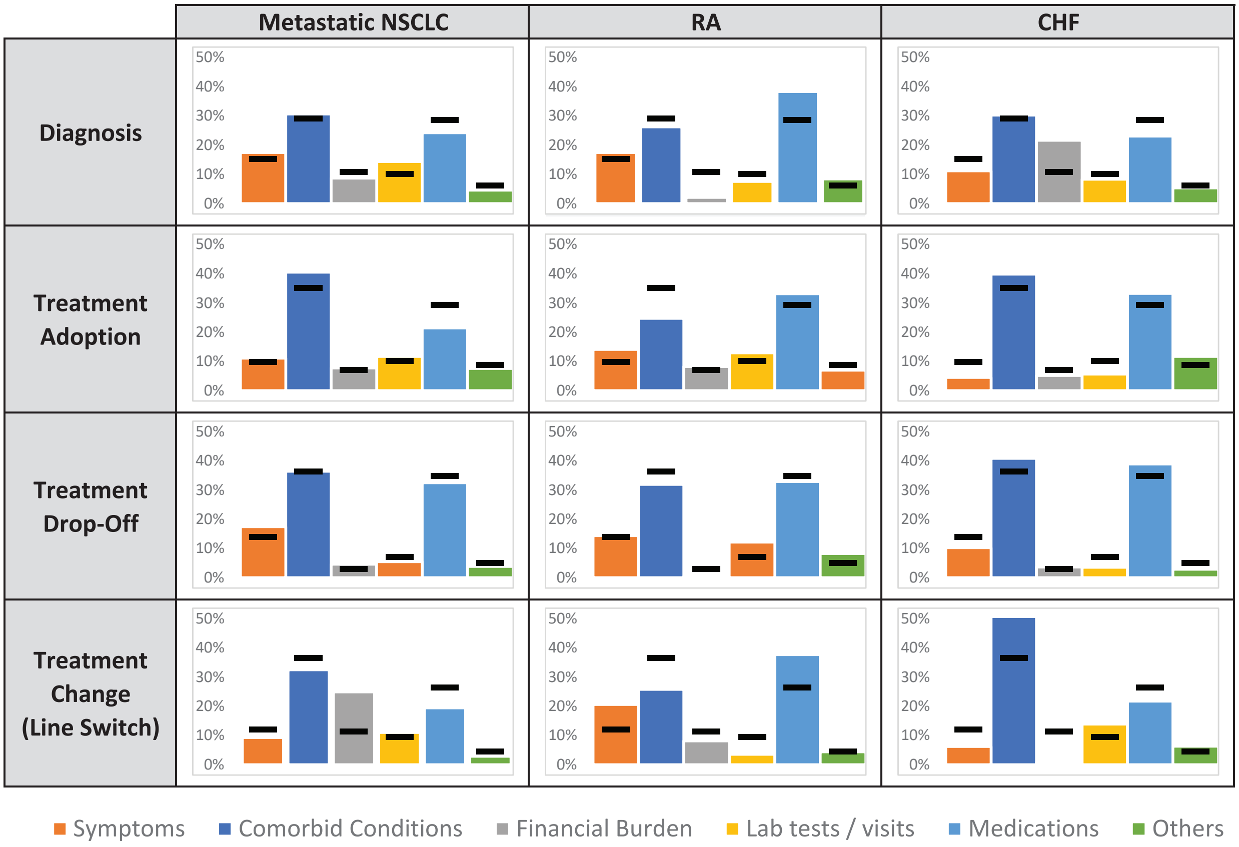


Chart 4: Time-from-Anchor Distribution of Recency Concept Domain Features

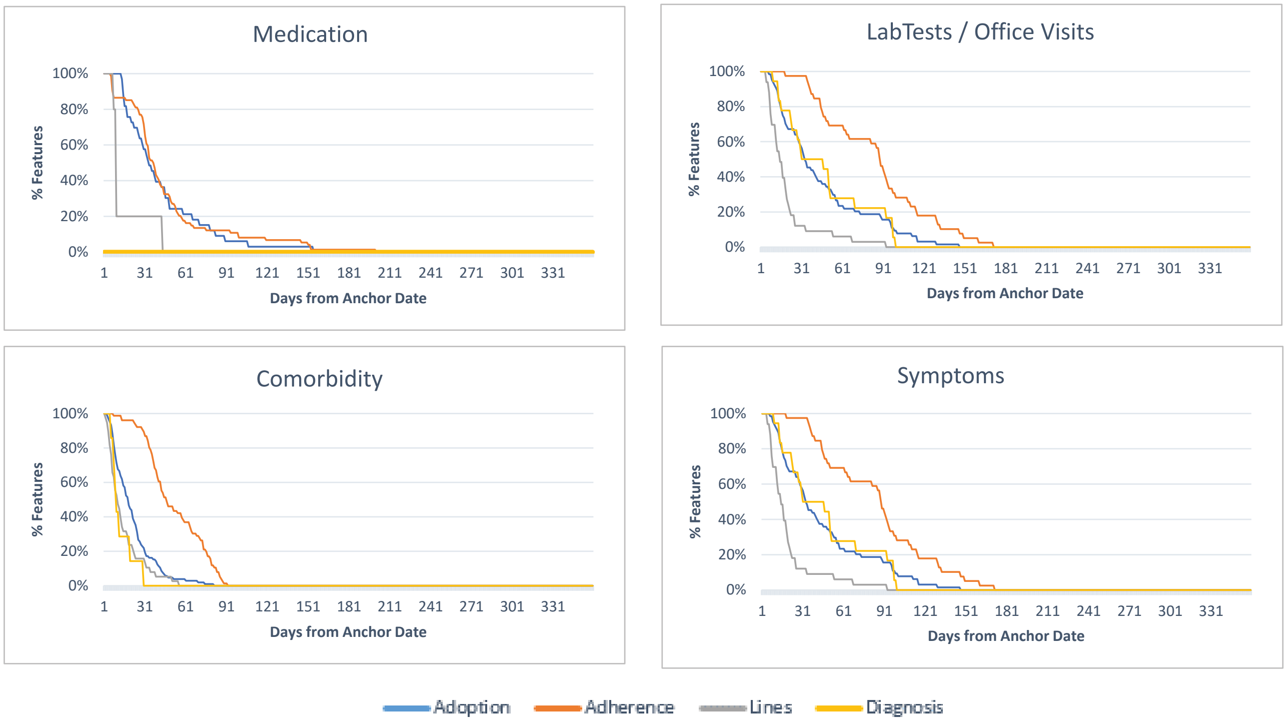


Chart 5: Time-from-Anchor Distribution of Frequency Features

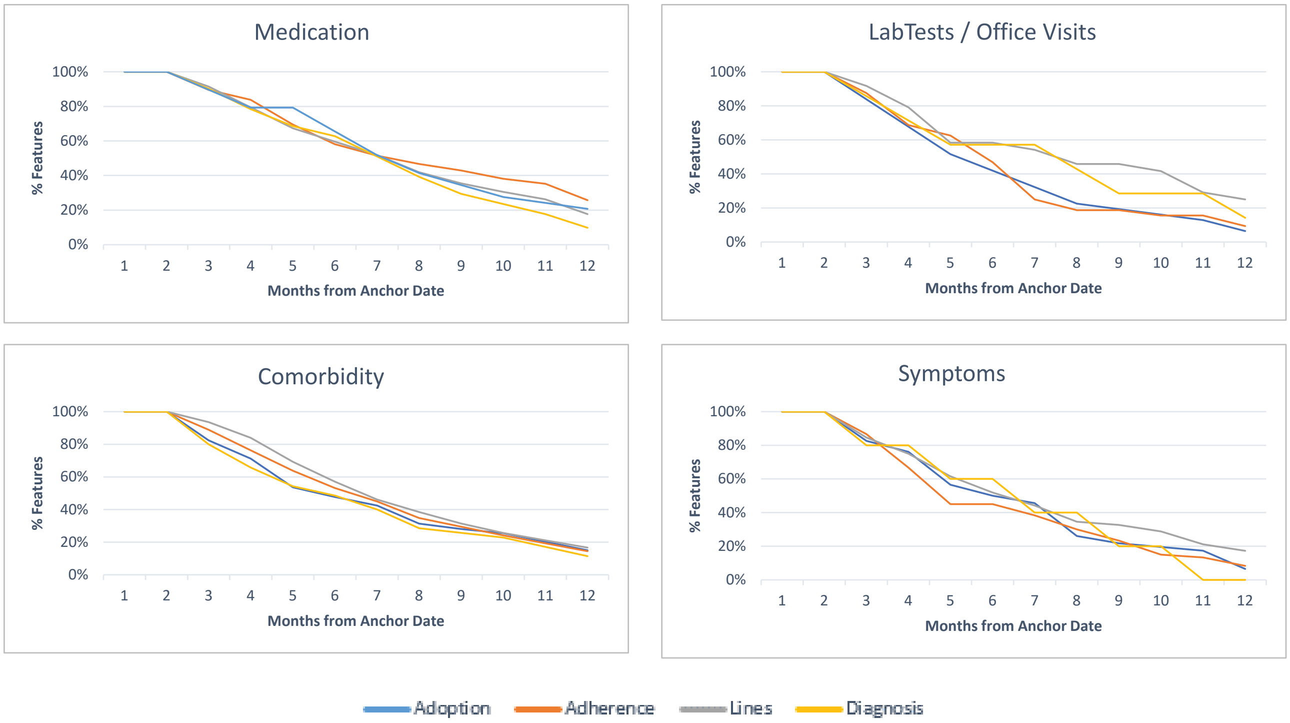
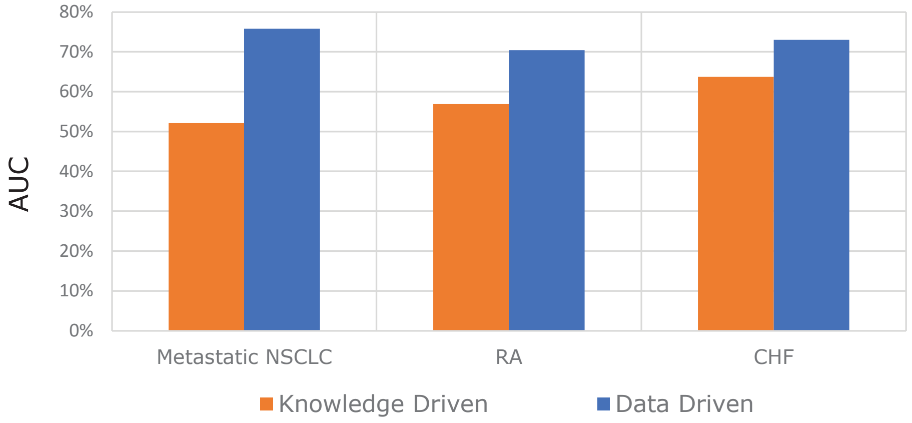


Chart 6: Mechanism of Feature Generation



3.2 Prediction Window

These results indicate that AUC peaks at the first month across therapy areas and use cases, gradually declining as the window extends from the first month to the sixth month. The decline is pronounced across all cases, except for CHF diagnosis and treatment adoption. See Chart 2.

3.3 Relative Importance of Features

We note that comorbid conditions invariably contribute to highest feature importance followed by medications, across all use cases. The trend is slightly distinct in case of treatment change; comorbid conditions still contribute to highest feature importance across all diseases but are followed by financial burden features (for NSCLC, CHF) and symptoms (for RA). See Charts 3, 4, and 5.

Frequency (or occurrence of events) of comorbidities, medications, symptoms and labs contribute to highest feature importance, followed by recency. Financial burden metrics are invariably associated with metrics showing change across time, such as trend and averages.

3.4 Mechanism of Feature Generation

In terms of model performance, we note that data-driven features significantly out-perform knowledge-driven features across all disease areas. It is interesting to note however that the features generated from data-driven approaches do capture knowledge-driven features in a different form and they tend to be among the top predictors. See Chart 6.

4. Discussion

From our investigations on event prediction models based on patient claims data, we have the following observations.

4.1 Length of Medical History

For most use cases, we see that there is no significant gain in model performance beyond a certain medical history (even for the few cases of exceptions, the incremental gains are diminishing and arguably don't justify the increased cost of implementation). In fact, models built using data from 6 to 18-months of medical history yield close to best performance in almost all use cases. The likely explanation for this is that the recent history of events has the highest impact on patient outcome event, and the impact decays as the window expands, leading to negligible impact of events past a certain date.

Prior research done in provider settings corroborates our findings that recent data contributes more to predictive accuracy than more data. Chen et al[26] indicated a half-life of four months for clinical data relevance and Min et al[27] observe no difference in prediction performance on records with a one-year observation window or a full history.

Based on this, we recommend a 12-month medical history for patient event prediction modeling as that seems to be the sweet spot that enables achieving good prediction performance as well as eases down the data preparation and computational complexity.

4.2 Prediction Window

We note that the prediction performance decreases as we increase the time window for prediction, i.e., we can predict well for the next 1-mo, but not as well for the next 3-mos (70-80% of 1-mo) and for 6-mo the prediction is as good as a coin flip in most cases. This is likely because as the prediction window expands, we will have less availability of predictor events which most often occur closer to predicted event.

Prior research around diagnostic prediction models corroborate this. Kleiman et al[28] noted an inverse relationship between the length of the prediction window and the quality of the model. They attribute this observation to a decrease in the number of patients available, smaller amounts of data and the importance of patients' recent health state on their immediate future.

The above findings indicate that it is extremely important to set up models for shorter prediction windows rather than longer (6-mos or more). This would require the modelers to educate and set the right expectations with the business stakeholders, as we have often seen a desire from them to predict as much ahead as possible. However, once the trade-offs are clear, a mutually workable solution could be developed.

4.3 Relative Importance of Features

Features derived from comorbidities and medication variables seem to play a prominent role in driving model performance across all use cases, except for treatment line change for NSCLC and CHF. These are closely followed by symptoms, labs/visits and financial burden metrics. The pre-ponderance of comorbidities and medications in driving model performance may have to do with the fact that these are more indicative of underlying disease conditions and thereby capture latent information more readily than other sets of features.

In terms of temporal distribution, we note that a high proportion of features associated with a time component, such as frequency and recency, seem to be concentrated within a window of 1-mo to 3-mo prior to the anchor date, with only a minority of features spanning to 6-mo and in some cases 12-mo window. This is in line with observations from the medical history, wherein it was observed that history beyond a 6 to 12-mo window doesn't add substantial incremental lift.

4.4 Mechanism of Feature Generation

We note that auto-extracted, data-driven features drive higher prediction performance (AUC – 0.75, 0.7,0.73) vs. knowledge-driven features (0.521, 0.57, 0.64). The conclusions driven by prior research, however, are mixed on this topic. Min et al[27] report improvement of model accuracy when data-driven features are added alongside handcrafted features, suggesting that while knowledge-driven features are powerful, data-driven features do help in improving model accuracy. Tran et al[29], however, suggest that the auto-extracted disease-agnostic features from medical data can achieve better discriminative power than carefully crafted comorbidity lists.

Typically, in the feature generation phase, analysts tend to rely on prior clinical and disease knowledge to craft features and test them iteratively, retaining features with the highest predictive power. These handcrafted features aid in getting to a certain baseline model accuracy;

however, incremental lift in model accuracy may necessitate additional features, the discovery of which is non-trivial given the span of feature space. Advances in machine learning are making this computationally feasible, allowing for search across the entire feature space (which might span hundreds of dimensions) and identifying the most relevant features, which can then be used for driving incremental model accuracy. Therefore, we recommend using a combination of knowledge-driven, along with auto-extracted, data driven features for good predictive performance.

Another key advantage afforded by data-driven features is their ability to handle concept drift.[30, 31] Models built using knowledge-driven features are more susceptible to performance degradation, given that these features are usually static in nature, and don't capture changes to underlying patterns in healthcare databases. In contrast, auto-extracted data-driven features can capture these changes with every refresh of database, given their very definition. Periodic iterations of automated feature generation algorithms can generate a newer set of data-driven features that can capture their patterns more readily.

5. Conclusion and Future Work

In this paper, we have presented results from experiments testing the impact of medical history, prediction window, different features and mechanism of feature generation on prediction performance. We believe these provide valuable benchmarks that can be utilized by data scientists and analysts while building patient event prediction models.

In terms of future work, we would like to make these conclusions more generalizable by expanding the experiments to cover a much wider range of use cases and therapy areas. Secondly, we would like to incorporate additional structured and unstructured data available in sources such as EHR, to assess the lift in predictive performance.

APPENDIX: Glossary

$CHADS_2$ – Score for Atrial Fibrillation Stroke Risk
MELD – Model for End Stage Liver Disease
NSCLC – Non-small cell lung cancer
CHF – Chronic Heart Failure
RA – Rheumatoid Arthritis
EGFR – Epidermal growth receptor factor
TNF – Tumor Necrosis Factor
ROC – Receiver Operating Curve
AUC – Area under the ROC curve

Specifying the Prediction Problem

Inclusion/Exclusion Criteria for Cohort (see Table 2)

Details for other use cases are available in the supplementary data file, at www.pmsa.org/ _resources/journal/2020/ ZS-KeyDrivers/ KeyDrivers.xlsx.

Patient Pool

(see Table 3)

Feature Discovery and Extraction

As outlined earlier, varied feature permutations can be created out of the combination of different concept domains (diagnosis, procedure, medication, demographics, etc.), time windows and aggregators (recency, frequency, change, sequence, etc.). To allow for agile feature discovery across concept domains and time, an intelligent feature generation and selection framework will be utilized for experiments involving prediction window, medical history and mechanism of feature generation. The framework utilizes a unique

Table 2: Inclusion/Exclusion Criteria for Cohort

Use Cases	Metastatic NSCLC	RA	CHF
Diagnosis	Patient Universe: Patients with at least one NSCLC diagnosis	Patient Universe: Patients with at least one RA diagnosis	Patient Universe: Patients with at least one CHF diagnosis
	Outcome Label 1: Patient with at least one metastasis diagnosis	Outcome Label 1: Patient with the first RA diagnosis in 2017	Outcome Label 1: Patient with the first CHF diagnosis in 2017
	Outcome Label 0: All other patients with only NSCLC diagnosis	Outcome Label 0: Patient with no RA diagnosis till 2015	Outcome Label 0: Patient with no CHF diagnosis till 2015
	Anchor Date: The first secondary diagnosis for Outcome Label 1. The last event (Rx / Px / Dx) for Outcome Label 0	Anchor Date: The first RA diagnosis for Outcome Label 1. The last event (Rx / Px / Dx) for Outcome Label 0 (till the end of 2015)	Anchor Date: The first CHF diagnosis for Outcome Label 1. The last event (Rx / Px / Dx) for Outcome Label 0 (till the end of 2015)
	Medical history: 1, 2, 3 years (1080 days)	Medical history: 1, 2, 3 years (1080 days)	Medical history: 1, 2, 3 years (1080 days)
	Additional inclusion/exclusion criterias: 1.) Excluded patients with NSCLC diagnosis before 2014 2.) Excluded patients with secondary diagnosis before NSCLC diagnosis	Additional inclusion/exclusion criterias: 1.) Excluded patients with RA diagnosis before 2014	Additional inclusion/exclusion criterias: 1.) Excluded patients with CHF diagnosis before 2014

Table 3: Patient Pool

Use Cases	Disease Areas					
	Metastatic NSCLC		RA		CHF	
	Train	Test	Train	Test	Train	Test
Diagnosis	42,747	10,687	56,775	24,332	154,952	66,408
Treatment Adoption	5,956	2,553	6,481	2,778	210,234	90,100
Treatment Drop-off	1,493	641	11,673	5,004	90,830	38,927
Treatment Change (Line Progression)	3,078	770	16,681	7,150	156,006	66,860

feature construction and selection architecture, enabled by evolutionary algorithms, that allows creation and testing of an exhaustive feature set across combinations of different domains, time windows and operators. Overall, this framework achieves feature selection in two broad steps:

1. Aggregator functions (recency, frequency, change in frequency, slope, change in slope) are applied on raw features available in the database, allowing the creation of multiple permutations and combinations across time.
2. Iterative selection and testing of these features via genetic algorithms across multitudes of generations, ensuring only the “fittest” features survive at the end of the iterations.

Model Building

XGBoost, an optimized distributed gradient boosting decision tree (GBDT) Python package,

Table 4: Model Hyperparameters

Model Parameters	Metastatic NSCLC				RA				CHF			
	Diagnosis	Treatment Adoption	Treatment Drop-off	Treatment Change (Line Switch)	Diagnosis	Treatment Adoption	Treatment Drop-off	Treatment Change (Line Switch)	Diagnosis	Treatment Adoption	Treatment Drop-off	Treatment Change (Line Switch)
max_depth	3	3	3	3	3	3	3	3	3	3	3	3
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n_estimators	100	100	100	100	500	250	100	100	800	100	100	100
verbosity	1	1	1	1	1	1	1	1	1	1	1	1
objective	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic	binary: logistic
booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
tree_method	auto	auto	auto	auto	auto	auto	auto	auto	auto	auto	auto	auto
n_jobs	1	1	1	1	1	1	1	1	1	1	1	1
gamma	0	0	0	0	0	0	0	0	0	0	0	0
min_child_weight	1	1	1	1	1	1	1	1	1	1	1	1
max_delta_step	0	0	0	0	0	0	0	0	0	0	0	0
subsample	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
colsample_bytree	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
colsample_bylevel	0.7	0.7	0.7	0.7	0.7	1	0.7	0.7	0.7	0.7	0.7	0.7
colsample_bynode	1	1	1	1	1	1	1	1	1	1	1	1
reg_alpha	0	5	10	0	0	2	10	0	0	0	0	0
reg_lambda	1	20	30	1	1	1	10	1	1	1	1	1
scale_pos_weight	1	9	2	1	7.5	1	13	6	1	5	6.5	1
base_score	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

is used for modeling. This model is tuned by optimizing parameters such as number of trees (n_estimators), maximum depth of the tree (max_depth), regularization parameters (lambda & alpha) and others. Training and test AUCs are assessed for any evidence of overfitting to ensure build of a stable model.

Our choice of XGBoost as model of choice is driven by experience. In our experience, for event prediction using real world data, while Recurrent Neural networks (RNNs) have the potential, XGBoost provides almost equivalent performance with a higher degree of interpretability.

This is corroborated by prior research like Wang et al[3] where post investigation of different machine learning models for readmission prediction concluded that deep convolutional networks (CNN) and recurrent neural networks (RNN) barely help; the additional performance uplift provided by such models is not commensurate with the complexity added to the model.

Model Hyperparameters

(see Table 4)

Model Evaluation

Area Under the Curve (AUC)

The area under the curve (AUC) is a widely used metric for assessing performance of a machine learning model in a classification problem. Essentially, it calculates the probability of assigning higher outcome risk to a randomly chosen patient with the outcome vs. without the outcome. It is typically generated by plotting the model's True Positive Rate (TPR) against 1-specificity. AUC is a well reported benchmark and doesn't depend on probability thresholds making the comparison unbiased. Other unbiased benchmarks such as AUPRC exist but are not widely reported across publications.

Prediction Window

For assessment of the impact of prediction window on model accuracy, area under the

Table 5: Model Performance

Use Case	Medical History (in Months)	Metastatic NSCLC	RA	CHF
Diagnosis	6	75%	73%	78%
	12	76%	78%	83%
	18	76%	81%	86%
	24	76%	82%	87%
	30	76%	82%	87%
	36	77%	84%	88%
Treatment Adoption	6	83%	83%	72%
	12	84%	83%	74%
	18	84%	83%	75%
	24	85%	84%	75%
	30	85%	83%	79%
	36	85%	84%	79%
Treatment Change (Line Progression)	6	75%	73%	72%
	12	77%	74%	73%
	18	78%	74%	73%
	24	78%	75%	74%
	30	79%	75%	75%
	36	80%	75%	77%
Treatment Drop off	6	76%	69%	73%
	12	77%	69%	73%
	18	75%	70%	73%
	24	76%	69%	73%
	30	76%	69%	73%
	36	77%	69%	73%

curve (AUC) metrics will be calculated while the window is varied from a period of one month to six months in one-month length.

Length of Medical History

To understand the effect of length of considered medical history on the model performance, AUC metrics will be calculated for each model iteration while the medical history is varied in semester-length windows from a semester to a period of three years.

Concept Domain and Types of Feature Classes

The contribution of the class of features by concept domain (diagnosis, medication, co-morbidities etc.) and aggregator type (recency, frequency, sequence, etc.) will be evaluated by disease area and use case to assess if a certain class of features are dominant. Feature importance will be assessed via adjusted F-score metric available from model outputs using data from a specific window.

Mechanism of Feature Generation

AUC metrics will be compared for models utilizing features from the knowledge-driven and data approach calculated for a 12-month medical history window.

Model Performance

(see Table 5)

About the Authors

Srinivas Chilukuri *is a Principal data scientist with ZS Associates where he leads the Artificial Intelligence Center of Excellence in New York. He has 15+ years of experience in applying machine learning solutions across various industries, primarily healthcare.*

Sagar Madgi *is a Senior data scientist with ZS Associates where he leads the Real World Data Artificial Intelligence Lab in Bengaluru. He has nearly 10 years of experience in applying machine learning solutions in healthcare.*

References

1 Van Niel TG, McVicar TR, Datt B (2005) On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sens Environ* 98: 468–480

2 Dobbin KK, Zhao Y, Simon RM. How large a training set is needed to develop a classifier for microarray data? *Clin Cancer Res*. 2008 Jan 1; 14(1):108-14.

3 Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: Proc KDD. 2013. p. 847–55.

4 Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, Weiner JP. Comparing Population-based Risk-stratification Model Performance Using Demographic, Diagnosis and Medication Data Extracted From Outpatient Electronic Health Records Versus Administrative Claims. *Med Care*. 2017 Aug;55(8):789-796. doi: 10.1097/MLR.0000000000000075

5 J. M. Franklin, C. Gopalakrishnan, A. A. Krumme, K. Singh, J.R. Rogers, C. McKay, N. McEllwee, and N. K. Choudhry. 3/2018. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory. *American Heart Journal*, 197:153-162. doi: 10.1016/j.ahj.2017.09.019

6 Weissman, G. E., & Harhay, M. (2018). Incomplete Comparisons Between the Predictive Power of Data From Administrative Claims and Electronic Health Records. *Medical care*, 56(2), 202. doi:10.1097/MLR.0000000000000848

7 Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2016/, based on November 2018 SEER data submission, posted to the SEER web site, April 2019.

8 Hunter TM, Boytsov NN, Zhang X, Schroeder K, Michaud K, Araujo AB. Prevalence of rheumatoid arthritis in the United States adult population in healthcare claims databases, 2004-2014.*Rheumatol Int*. 2017 Sep;37(9):1551-1557. doi: 10.1007/s00296-017-3726-1

9 Komanduri, S., Jadhao, Y., Guduru, S. S., Cheriyath, P., & Wert, Y. (2017). Prevalence and risk factors of heart failure in the USA: NHANES 2013 - 2014 epidemiological follow-up study. *Journal of community hospital internal medicine perspectives*, 7(1), 15–20. doi:10.1080/20009666.2016.1264696

10 Hofmeister MG, Rosenthal EM, Barker LK, Rosenberg ES, Barranco MA, Hall EW, Edlin BR, Mermin J, Ward JW, Ryerson AB. Estimating Prevalence of Hepatitis C Virus Infection in the United States, 2013-2016.*Hepatology*. 2019 Mar;69(3):1020-1031. doi: 10.1002/hep.30297

11 Centers for Disease Control and Prevention. Nov 2015. Available from: https://www.cdc.gov/ibd/data-statistics.htm

12 Reveille JD, Witter JP, Weisman MH. Prevalence of axial spondylarthritis in the United States: estimates from a cross-sectional survey. Arthritis Care Res (Hoboken). 2012 Jun;64(6):905-10. doi: 10.1002/acr.21621

13 Anne G. Wheaton, Timothy J. Cunningham, Earl S. Ford, MD, and Janet B. Croft., “Employment and activity limitations among adults with chronic obstructive pulmonary disease — United States, 2013,” Morbidity and Mortality Weekly Report (MMWR), 64 (11), pp. 289-295 (March 7, 2015), Centers for Disease Control and Prevention (CDC)

14 Hugh Waters and Marlon Graf. The Cost of Chronic Diseases in the U.S, May 2019.

15 Medical Expenditures Panel Survey(MEPS), Agency for Healthcare Research and Quality, US Department of Health and Human Services, 2008-2012

16 Bui, A. L., Horwich, T. B., & Fonarow, G. C. (2011). Epidemiology and risk profile of heart failure. Nature reviews. *Cardiology*, 8(1), 30–41. doi:10.1038/nrcardio.2010.165

17 Razavi, H., Elkhoury, A. C., Elbasha, E., Estes, C., Pasini, K., Poynard, T., & Kumar, R. (2013). Chronic hepatitis C virus (HCV) disease burden and cost in the United States. *Hepatology* (Baltimore, Md.), 57(6), 2164–2170. doi:10.1002/hep.26218

18 The facts about inflammatory bowel diseases. Crohn’s & Colitis Foundation of America website. http://www.ccfa.org/assets/pdfs/updatedibdfactbook.pdf. Published November 2014. Accessed September 15, 2015

19 Guarascio, A. J., Ray, S. M., Finch, C. K., & Self, T. H. (2013). The clinical and economic burden of chronic obstructive pulmonary disease in the USA. ClinicoEconomics and outcomes research : CEOR, 5, 235–245. doi:10.2147/CEOR.S34321

20 Chen, Q., Jain, N., Ayer, T., Wierda, W. G., Flowers, C. R., O'Brien, S. M., Chhatwal, J. (2017). Economic Burden of Chronic Lymphocytic Leukemia in the Era of Oral Targeted Therapies in the United States. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 35(2), 166–174. doi:10.1200/JCO.2016.68.2856

21 S Gala, A Shah, M Mwamburi. Economic Burden Associated with Chronic Myeloid Leukemia (CML) Treatments in The United States: A Systematic Literature Review. *Value in Health*. November 2016 Volume 19, Issue 7, Page A727. doi: https://doi.org/10.1016/j.jval.2016.09.2179

22 Beth L. Nordstrom, Jason C. Simeone, Karen G. Malley, Kathy H. Fraeman, Zandra Klippel, Mark Durst, John H. Page, and Hairong Xu. Validation of Claims Algorithms for Progression to Metastatic Cancer in Patients with Breast, Non-small Cell Lung, and Colorectal Cancer. *Front Oncol*. 2016; 6: Published online 2016 Feb 1. doi: 10.3389/fonc.2016.00018

23 Savannah L. Bergquist, Gabriel A. Brooks, Nancy L. Keating, Mary Beth Landrum, and Sherri Rose. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. *Proc Mach Learn Res*. 2017 Aug; 68: 25–38.

24 Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2016). Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. *Computational and structural biotechnology journal*, 15, 26–47. doi:10.1016/j.csbj.2016.11.001

25 Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC heart failure*, 6(2), 428–435. doi:10.1002/ehf2.12419

26 Chen JH, Alagappan M, Goldstein MK, Asch SM, and Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform*. 2017 Jun;102:71-79. doi: 10.1016/j.ijmedinf.2017.03.006

27 Min X, Yu B, Wang F. Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD. *Sci Rep*. 2019 Feb 20;9(1):2362. doi: 10.1038/s41598-019-39071-y.

28 Ross S. Kleiman, Paul S. Bennett, Peggy L. Peissig, Richard L. Berg, Zhaobin Kuag, Scott J. Hebbring, Michael D. Caldwell and David Page. High throughput machine learning from electronic health records. Quantitative Methods (q-bio.QM); Machine Learning (cs.LG); Machine Learning (stat.ML). arXiv:1907.01901

29 Tran, T., Luo, W., Phung, D. et al. A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinformatics 15, 425 (2014) doi:10.1186/s12859-014-0425-8

30 Oded Maimon and Lior Rokach. 2010. Data Mining and Knowledge Discovery Handbook (2nd ed.). Springer Publishing Company, Incorporated

31 Black M., Hickey R. (2004) Detecting and Adapting to Concept Drift in Bioinformatics. In: López J.A., Benfenati E., Dubitzky W. (eds) Knowledge Exploration in Life Science Informatics. KELSI 2004. Lecture Notes in Computer Science, vol 3303. Springer, Berlin, Heidelberg

Maladies of Claims Data: Manifestations, Origins, and Cures

JP Tsang, PhD & MBA (INSEAD), President of Bayser Consulting and Igor Rudychev, Head of US Digital, Data, and Innovations, AstraZeneca

Abstract: The data we work with on an everyday basis such as Patient Claims to inform business decisions is incomplete and punctured with holes. Left unattended, the analyst may produce insights that are somewhat off when not completely wrong. That's how we get estimates of new patients starts that are way too high, a targeting list that leaves out important physicians or accounts, estimates of compliance and persistence that come in too low, and market share of our drug that bears little resemblance with reality.

We'll start off with a data analysis story to provide a concrete example of the dangers of ignoring the shortcomings of claims data and taking the findings gleaned from the data at face value. We then follow up with 10 other cases where insights run amok to make sure the lead story is not dismissed as a one-off outlier. We then take up the next 3 sections to describe the different maladies of claims data, the etiology of these maladies, and prescriptions to cure these maladies. To conclude, we go back to the story we started off with and describe how we deployed Machine Learning to solve the problem.

Keywords: Claims, Gaps and Holes, Incomplete Data, Fill in Missing Information, Bayesian Reasoning

It's not the eye that sees but the brain. The eye merely captures photons and produces a messy and incomplete image on the retina. The brain inverts the image, fills in the countless gaps, and curates it. The end result is so good that we are unaware of all the work that the brain does to pull this off. Just the same, data sources merely capture data points, so let's make sure we summon our brain to work its magic.

1. Introduction

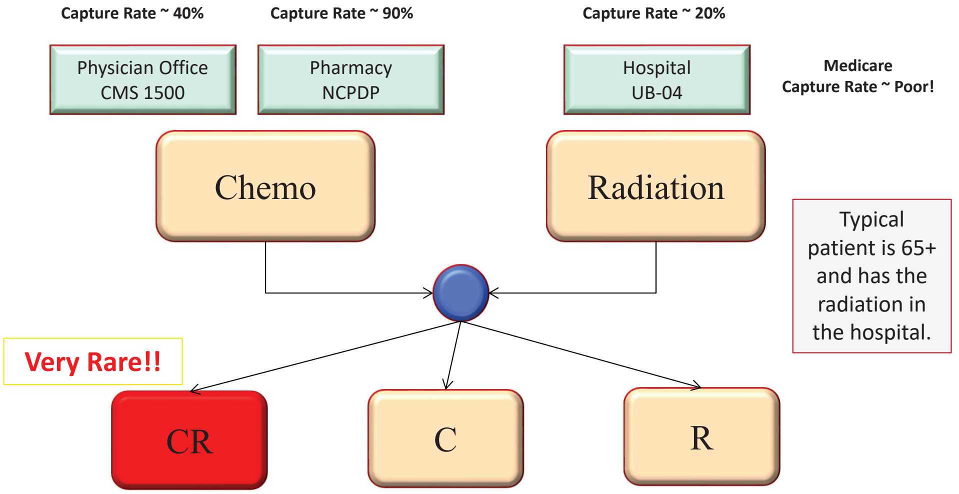
Let's start with a story that will sound familiar to anyone who has been involved in generating insights from Claims data or has been a recipient of such insights.

The ask from Upper Management is simple and direct and it is to shed light on the number of patients that have undergone chemo and radiation, which we'll refer to as CRT therapy. The rationale for the ask is that CRT patients are the ones who are eligible for the drug of interest. The finding will be used for targeting, Long Range Planning (LRP), and ultimately Wall Street.

The data analysts go to work and come back claiming that there are actually not a whole lot of these CRT patients. In other words, the prospects for the drug are very grim. Needless to say Upper Management immediately challenge the analysis. The data analysts stand by their finding and offer two pieces of evidence to back their position. One, they used claims data which is the best that the industry has to offer. Two, they used not one but four different claims data sources and they all provided very similar answers. Then all hell breaks loose.

It turns out the data analysts were wrong. Their blunder was to believe the data because 4 different data sources pointed in the same direction. They should have known better: All syndicated data sources are afflicted with the same setback, so the fact that multiple data sources agree with each other does not mean much.

Figure 1: CRT Patients Are a Rarity in Syndicated Claims Data



Here's what's going on. Radiation is conducted in the hospital and is reimbursed under Medicare Part A when the patient is over 65. Hospitals use a UB-04 claim to invoice Payers for services rendered in the hospital. Chemotherapy is administered in the physician office or dispensed by a pharmacy. Physicians invoice Payers using a CMS-1500 medical claim and Pharmacies an NCPDP pharmacy claim.

Now, it is well known that syndicated data sources do a very poor job capturing UB-04 hospital claims, especially Medicare Part A, where the radiation is reported. By contrast, they do a better job capturing chemo in the physician office (CMS-1500) and an even better job when the chemo is dispensed by a pharmacy (NCPDP). What this means is that the odds of seeing radiation in Claims data is low and the odds of seeing both chemo and radiation lower still (See Figure 1).

That's why the data analysts found so few CRT patients. Actually, many of the CRT patients came across as chemo only patients. Also, since this data capture problem issue is not specific to one particular syndicated data source, it makes sense that all 4 data sources were pointing in the same direction. Indeed, that they agree with each other cannot be used as evidence to support the finding of the analysts.

Unfortunately, this CRT story is not an isolated case. It is more the rule than the exception. Analyses that are subject to data issues range from patient journey to market share and from targeting to measurement of Impact and ROI. See Table 1 for an illustration of how rampant the problem is. It lists the top 10 business metrics where the data analysts would report erroneous findings if they do not take any measure to address the shortcomings of the claims data.

2. Maladies of Claims Data

Claims data is afflicted with 3 types of maladies: Missing information, Wrong information, and Information that we wish were there (See Figure 2). Strictly speaking, the third type is not a malady as the information we wish to see is not part of Claims data.

A. Missing Information

The first type of malady has to do with missing information. This is information that we'd expect to be documented in the Claims but that is not there. Let's go through the various manifestations of this problem.

Table 1: Business Metrics Where Claims Data May Mislead

Type	Business Metric		Issue	Explanation	Source of Problem
Patient Journey	1	New Patient Starts	Patients labeled as new are not	A patient is deemed new when we see a claim for the first time and no claims in the look-back period. Claims prior to the ones we see are not in the data.	Longitudinal Holes
	2	Adherence - Compliance (Medication Possession Ratio) and Persistence (Days on Therapy)	Lower than reality	Not all claims of a patient are captured, leading us to conclude that the patient is less compliant and less persistent than what the patient is in reality.	Longitudinal Holes
	3	Line of Therapy Labeling	Out of whack	This is due to two factors: (1) Missing claims and (2) Misrepresentation of combination therapies as smaller combinations and/or mono therapies.	Longitudinal Holes & Differential Rx/Mx Capture
	4	Regimen Market Share	IV drugs are under-reported relative to oral drugs. Larger combinations are mistaken for smaller ones and/or mono therapies.	IV drugs are missing more frequently than oral drugs leading us to see less of combination therapies that include IV drugs.	Differential Rx/Mx Capture
Market Share	5	Drug Market Share	Underestimate competitive market share	Our SP data captures our drug better than Syndicated claims data does and, as a result, the merged SP/Syndicated data captures our drug better than the competition.	Differential SP/Syndicated Capture Rate
	6	Share by Payer	Anemic Capture of Medicare and Over-representation of Commercial	Two issues: (1) Traditional Medicare is poorly captured because of the way contracting works, and (2) the Data Vendor fails to move Managed Medicare claims back under Medicare where they belong and leave them under Commercial.	Poor Capture of Medicare & Failure to Reclassify Managed Medicare under Medicare
Targeting	7	Targeting and Segmentation	Many physicians and accounts are missing from the target list.	Big swaths of Providers are missing in the data including Kaiser Permanente, VA, DoD, IDNs, etc.	Either too expensive for the data vendor to acquire the data or the data supplier refuses to sell the data to the data vendor.
	8	Capture of Hospital Procedures	Under representation of Radiation, Surgeries, Stem Cell Transplants, and more generally procedures performed in the hospital.	First, the hospital is a poorly captured setting and, as a result, any procedure performed in the hospital. Second, Medicare is a poorly captured Payer.	Poor Capture of Hospital & Poor Capture of Medicare
Indication	9	Patient Profile by Indication	Patients are assigned the wrong indication, resulting at times in a biased and non-representative sample.	Three factors come into play: (1) The coding system does not allow us to zero in on the indication of interest (e.g., PAH), (2) the physician is genuinely mistaken (misdiagnosis), and (3) the physician up-codes to get a higher reimbursement or down-codes to avoid stigma.	Coding System not precise enough & Provider Inputs wrong information
Lab	10	Impact/ROI of interventions to alter Rx Behavior of Physicians when Lab Results are available	Very difficult to assess and measure	Two reasons: (1) Capture of physicians with lab results available for only a sliver of the prescribing physicians, and (2) Lab results come in late, oftentimes after the physician has decided which therapeutic route to follow.	Poor Capture of Lab results & Delay in receiving Lab results

Figure 2: Three Types of Claims Maladies

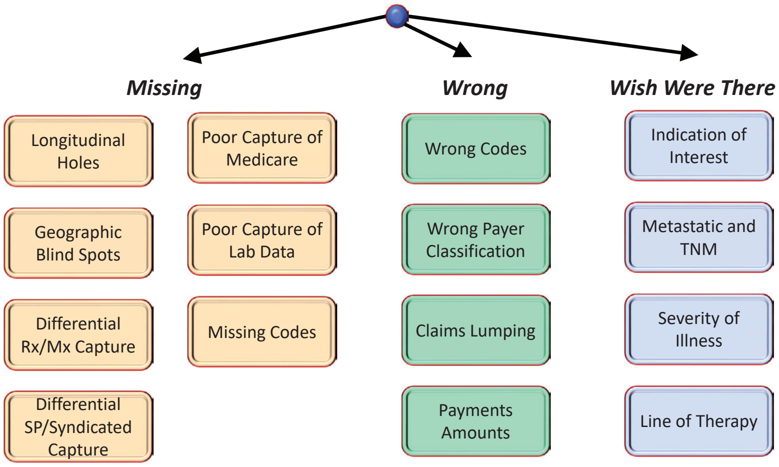
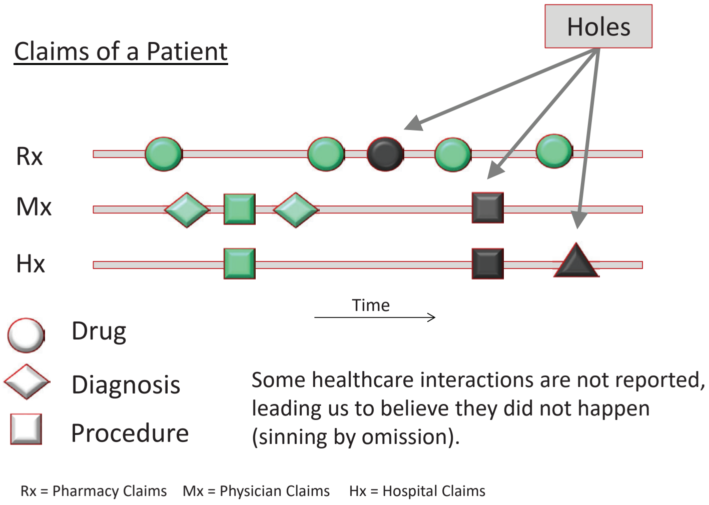


Figure 3: Longitudinal Holes in the Claims Data


Claims of a Patient
Holes
Rx
Mx
Hx
Time
Drug
Diagnosis
Procedure
Some healthcare interactions are not reported, leading us to believe they did not happen (sinning by omission).
Rx = Pharmacy Claims Mx = Physician Claims Hx = Hospital Claims

1. Longitudinal Holes
Claims data is the most valuable commercial data asset out there thanks to its longitudinality. This feature allows us to follow a patient over time and examine all the interactions the patient had with the healthcare system as they unfolded including drugs prescribed, visits to physicians, diagnoses and procedures, labs tests, hospitalizations, and the like.

The number one malady strikes at the heart of its longitudinality. Indeed, there are holes in the data. One or more interactions the patient had with the healthcare system go unreported (see Figure 3). This impacts any analysis that pertains to the healthcare journey of the patient. New patients starts are overestimated as we do not see claims that took place before the ones that we see and they are the ones that signal the start of therapy. Compliance, persistence, adherence, and days on therapy are underestimated as measurements are not made on the full set of claims. For the same reason, where one line of therapy ends and another starts is thrown off kilter.

2. Geographic Blind Spots
A second type of malady has to do with entire geographies going dark (see Figure 4). Examples

Figure 4: Geographic Blind Spots in the Claims Data

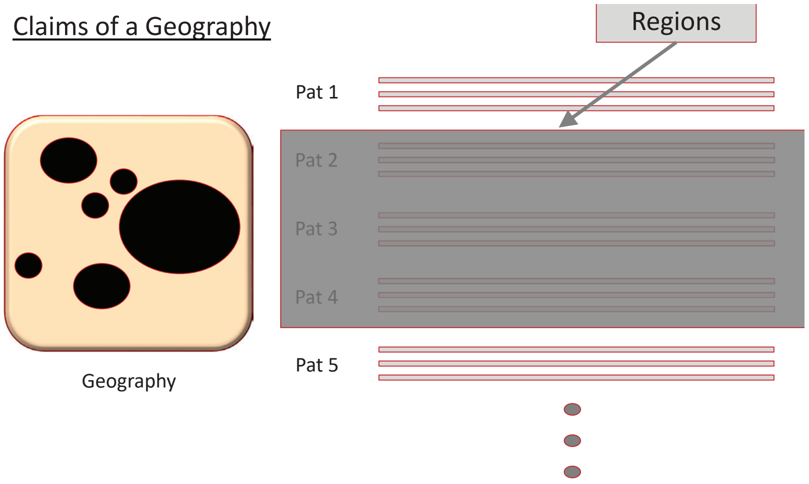
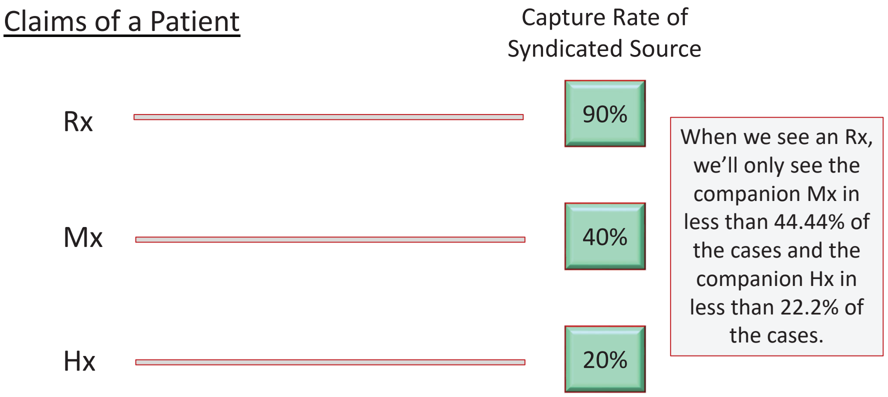


Figure 5: Differential Capture Rate between Rx, Mx, and Hx


Claims of a Patient
Capture Rate of Syndicated Source
Rx
90%
Mx
40%
Hx
20%
When we see an Rx, we'll only see the companion Mx in less than 44.44% of the cases and the companion Hx in less than 22.2% of the cases.

include Kaiser Permanente, VA (Veteran Affairs), DoD (Department of Defense), large IDNs (Integrated Delivery Networks), and the like. Note: this malady is not specific to Claims data.

Impacted analyses include targeting where we leave out important physicians or accounts, incentive compensation as we are not sure if we are giving proper credit to reps for the activity that is taking place in their territories, and market intelligence as we are not privy to what's happening in the dark spots of our data.

3. Differential Rx/Mx Capture

Another malady of Claims data stems from the difference in capture rate between Rx (pharmacy claims - NCPDP), Mx (Medical Claims - CMS-1500), and Hx Hospital Claims - UB-04). Indeed, the capture rate of Hx is much lower than that of Mx, and the capture rate of Mx is much lower than that of Rx (see Figure 5).

This malady impacts analyses that pertain to combination therapies where one therapy is delivered in the hospital (e.g., radiation) and the other in the physician office or by the pharmacy (e.g., chemo) as in the CRT story we recounted earlier. Likewise, combination therapies that straddle Oral and IV may not be portrayed accurately. In Multiple Myeloma, for instance, where patients get RVd (Revlimid, Velcade, and dexamethasone), the data may suggest that a significant portion of the patients do not get RVd but Rd instead.

More generally, the data captures smaller combinations or even mono therapies instead of the full combination therapies, thereby clouding our understanding of what's happening in the marketplace.

Also, when computing market share of drugs in a market that has both oral and IV drugs, the market share of the IV drugs tends to be understated and that of oral drugs overstated.

4. Differential SP/Syndicated Capture
Except in a few cases, our SP data only talks about our drugs (no competitive drugs) and does so extremely well under closed distribution and rather well under open distribution. Add hub data to the SP data and the picture can only get better.

Syndicated Claims data, on the other hand, captures a more partial picture of all the drugs including ours, hence the differential capture of our drug between SP and Syndicated Claims data.

Since we see disproportionately more of our drug than drugs of the competition, we may be tempted to conclude that our market share is higher than what it is and in some cases that we are leading when we are not.

5. Poor Capture of Medicare
The way contracting is conducted between Data Vendors and Data Suppliers is such that Medicare data ends up not well captured in the Syndicated Claims data, a phenomenon that, by the way, has given rise to Data Vendors that specialize in reselling their own flavor of Medicare data from CMS ranging from more recent vintages to better integration with other data assets. This throws off our picture of not only Medicare but of the other Payers as well. A Share by Payer breakout is bound to overestimate Commercial as it underestimates Medicare.

6. Poor Capture of Lab Data
Lab data comes in two flavors: lab order which is the test that the physician orders and lab result or lab value which is the outcome of the test. The lab order appears as a CPT-4 procedure on the CMS-1500 claim the physician sends to the Payer for reimbursement and this takes place shortly after the physician sees the patient. The lab value is only available after the test is done, which happens way after the claim has been sent to the Payer.

The interest in lab data is obvious. Overlaying lab results on top of Syndicated Claims provides a much sharper picture of the patient, allowing us to zero in on those patients that we know for a fact are eligible for our drug.

There are two problems though. First, there is a significant delay before the lab-enriched claims data becomes available which means we may not have the chance to act upon the information. Second, the overlap between lab values and claims may not be big enough, thereby limiting the scope of our interventions.

7. Missing Codes
Procedure codes trigger payments and as a result will always be reported on the claim. Diagnosis codes essentially provide context for the reimbursement. When it is clear that the provider will be reimbursed, the diagnosis codes are almost optional and that's when they are left out. Another instance where codes vanish is when a more serious and urgent diagnosis needs to be conveyed. In this case, the current diagnosis is simply not that important anymore and may be left out.

The mistake here is to confuse absence of diagnosis code with absence of condition. Just because a diagnosis code no longer appears in subsequent claims does not mean the patient is in remission.

J-codes fall under the HCPCS coding system and are used to identify drugs that are administered in the physician office. Several drugs may be assigned the same J-code and a newly launched drug may in some cases have to wait a year or more before it gets a J-code of its own. In the meantime, it gets a catchall temporary J-code. To put it mildly, J-codes are not very clean.

Obviously, this makes it hard to track the activity of IV drugs with the same accuracy with which we track the activity of oral drugs where NDC codes is the de facto standard.

B. Wrong Information

The second type of malady is arguably worse than the first one. Indeed, it's not that the information is missing, it's there and plain wrong. Let's go through the most prominent cases.

1. Wrong Codes

There are several instances where the claim carries the wrong diagnosis code. The most common occurrence is when the physician misdiagnoses the patient and enters the wrong code. In other instances, the physician up-codes the condition so as to get a higher reimbursement. In yet other instances, the physician down-codes the condition to avoid the stigma associated with the condition. It is not uncommon for a psychiatrist to write a diagnosis of bipolar when the physician knows full well it's schizophrenia.

Whatever the rationale behind the wrong code, the implications may be far-reaching. See Sidebar 1 for a compelling example we owe to Marc Duey, the CEO of Prometrics (now part of ConcertoHealth) that shows how up-coding ARDS (Acute Respiratory Distress Syndrome) can wreak havoc.

In other instances, it may be clear that the codes that are reported on the claims cannot be trusted. See Sidebar 2 for a good example that we owe to Vishal Chaudhary, Director of Advanced Analytics with Sanofi.

Sidebar 1: Impact of Up-Coding ARDS for Pneumonia

Instead of coding for ARDS (Acute Respiratory Distress Syndrome), physicians choose Pneumonia as the latter triggers a higher reimbursement. This trips the data analysts in two ways. First, we conclude that there are fewer ARDS and more Pneumonia patients than there are. Second, we infer that ARDS patients are much sicker than they are, judging by their high utilization of ventilation. Indeed, up-coding siphons away a good portion of the less severe cases of ARDS and puts them into Pneumonia, leaving the ARDS patient pool with sicker patients.

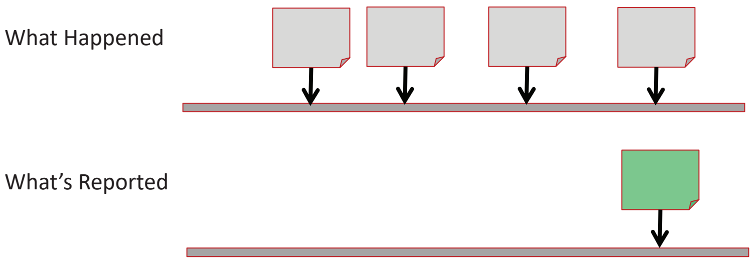
Sidebar 2: Too Many Patients Have Type 1 and Type 2 Diabetes at the Same Time

Claims data indicate that many patients have a diagnosis of both type 1 and type 2 diabetes at the same time. While this is possible, the number of such patients is way too significant to be believable. The fix was developed in conjunction with HEOR and consists of labeling the ambiguous patients as type 1 when the following 2 conditions are met: (1) the patient has more type 1 claims than type 2 claims, and (2) the patient does not take an oral metformin. Ideally, more conditions would have to be satisfied but these are the only 2 that could be deployed in Claims data.

2. Wrong Payers

Managed Medicare is business that Medicare subcontracts to Commercial Payers. As a result, the Claims data comes in to the data vendor through the Commercial route where they need to be moved from Commercial to Medicare. From time to time, the data Vendor omits this

Figure 6: Claims Are Lumped and Reported as One



crucial step and delivers Medicare claims as Commercial claims.

Another issue we came across is Commercial Claims being classified as Cash. It was clear that was happening since the high price tag of the drug puts Cash share squarely in the low single low digit and that was not what the Claims data was saying. Digging into the matter revealed that the physician practices in question were using paper claims, and the poor handwriting also contributed to the misclassification.

3. Lumping of Claims
Instead of sending the claim to the Payer after each patient visit, the Physician holds onto the claims and then sends several of them at the same time, presumably to cut back on the hassle of sending claims (See Figure 6). How do we know that? Well, the dosing of the drug is several times the maximum a human being can take. Not only the patient does not die, the patient comes back the following month for more.

4. Inaccurate Payment Amounts
Take the simple case where payment is split between the Payer and the patient. Even in that case, it may not be clear how much the Payer actually pays as the claim reports amount charged, not amount paid. For that we need the EOB (Explanation of Benefit) on the remit, which may not always be available.

Alternatively, the patient may have purchased supplemental insurance which means that both the primary and secondary payers need to be captured to reconstruct what happened. The manufacturer may offer a coupon or a discount card, and when that's the case, we need to understand if the OOP (out of pocket) includes or excludes the coupon amount (we have seen both cases). Also, a foundation may pitch in to alleviate the financial burden of the patient. If any one of these parties is missing or misrepresented, the payment amount from the Claims is off.

The industry has known this for a long time. That's why anyone that has been around for some time takes Claims data with a pinch of salt when it comes to payments and reimbursements.

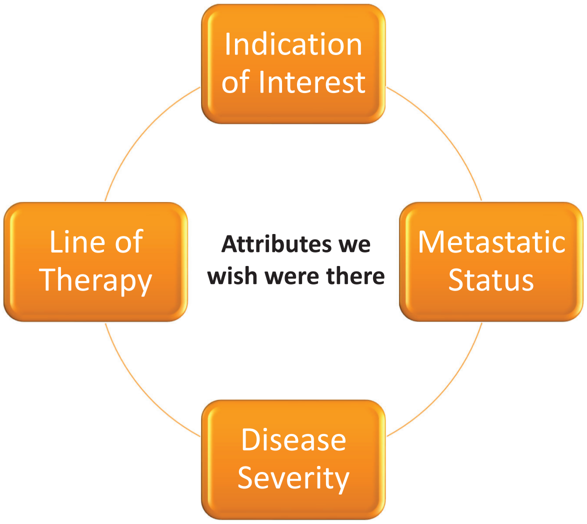
C. Information We Wish Were There

The third type of malady is strictly speaking not one as it pertains to attributes of Claims that we wish were there but have never been there. These attributes were never meant to be in Claims but would make our lives easier if they were.

The most common wished-for attributes are the following (see Figure 7):

1. Indication of Interest - The ICD Coding System may not have a code specifically for the indication of interest. Take PAH (Pulmonary Arterial Hypertension) for instance. There is no diagnosis code for PAH. There is one for PH (Pulmonary Hypertension) but it is much broader than PAH.
2. Metastatic Status - Claims data does not say if the cancer is metastatic or not let

Figure 7: Attributes We Wish Were in Claims



alone describe the size, stage, and spread (TNM) of the tumor.

3. Disease Severity - Claims data provides little information regarding severity of illness of a patient. This has to be inferred with no guarantee of success.
4. Line of Therapy - Claims data does not indicate when a line of therapy ends and when another starts. This has to be inferred.

3. Etiology of Maladies

Why is it that the Claims data from our Syndicated Data vendors have so many issues? It turns out that all the maladies we discussed originate from 3 sources: Contracting, Footprint, and the Capture-Transmission-Delivery process. Let's take a look at each of them.

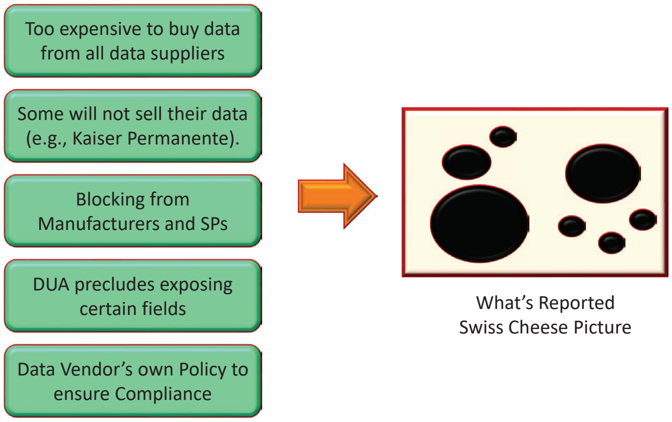
1. Contracting

Data is too expensive for data vendors to purchase all the data sources out there. Instead, they have to be discerning and acquire those data sources that lead to a representative sample. In some cases, money is not even the issue as some will not sell their data and Kaiser Permanente is a prime example. Blocking from Manufacturers and SP's may also punch holes in the data before it makes its way to the Data Vendors. Finally, the DUA (Data Usage Agreement) in place may require the Data Vendor not to expose certain data fields, not to mention the Data Vendor's own policy which may require additional redaction for compliance with HIPAA regulations. See Figure 8.

Data Vendors purchase data from Pharmacies, Payers, Clearinghouses (switches) that in essence move the claims from Providers to Payers, and also directly from Providers such as large physician groups and IDNs (Integrated Delivery Networks). See Figure 9.

First off, the same claim may come in duplicate copies, which the Data Vendor needs to take care of. Next, Manufacturers may ask SP's not to share the activity of their drugs with others. Clearinghouses may be asked the same. Interestingly, they do not always follow through and some of the activity of the drug may show up in the data, a phenomenon known as leakage. The data analyst sees some activity of the drug and, if unaware of the blocking situation, concludes that they are seeing the full activity of the drug. We refer to this as the "illusion of completeness", which needless to say may lead to unfortunate findings. Finally, Payers are not subject to blocking as they need to see the claim to pay the Provider. This means

Figure 8: Holes in the Data Start with Contracting



Too expensive to buy data from all data suppliers

Some will not sell their data (e.g., Kaiser Permanente).

Blocking from Manufacturers and SPs

DUA precludes exposing certain fields

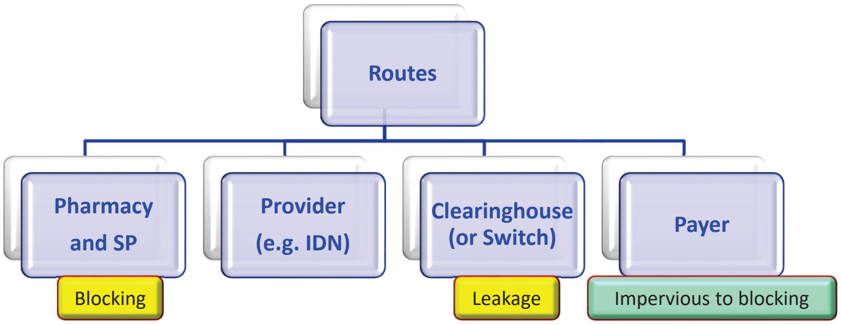
Data Vendor's own Policy to ensure Compliance

What's Reported
Swiss Cheese Picture

The picture is punctuated with holes.

DUA: Data Use Agreement

Figure 9: Data Comes to Data Vendors Through Various Routes



Routes

Pharmacy and SP — Blocking

Provider (e.g. IDN)

Clearinghouse (or Switch) — Leakage

Payer — Impervious to blocking

that Closed Claims datasets such as Optum, Truven, and Pharmetrics Plus are impervious to blocking and as such are good yardsticks to compare against to gauge the extent of leakage.

2. Footprint

Footprint refers to the set of Providers the Data Vendor receives data from. Providers that fall outside of the Footprint are the reason for the blind spots in the data. Now, a patient may see a Provider that is in the Footprint and later on a Provider that is outside of the Footprint. As a result, the corresponding claims "flicker" in the data in that they appear, disappear, and reappear again. See Figure 10.

Below are common reasons for flickering:

1. Patient fills Rx's in Pharmacies close to home and close to work (one is in the footprint but not the other).
2. Patient is admitted in a Hospital that is not in the footprint.
3. Patient changes Payer and Providers of that Payer are not in the footprint.
4. First injection takes place in the physician office (medical benefit) and subsequent ones at home (pharmacy benefit).
5. Patient goes somewhere else during winter (snowbirds).
6. Leakage of Clearinghouse is fickle.
7. Data contracts expire and are not renewed.

Another observation worth pointing out is the asymmetry between Rx and Mx claims. For starters, Rx claims are processed under pharmacy benefit and Mx claims under medical benefit. Differences in reimbursement can be quite sizable and invite creative schemes to have the reimbursement processed under one benefit instead of the other.

Figure 10: Flickering is the Result of Patients Interacting with Providers Inside the Footprint and Providers Outside of the Footprint

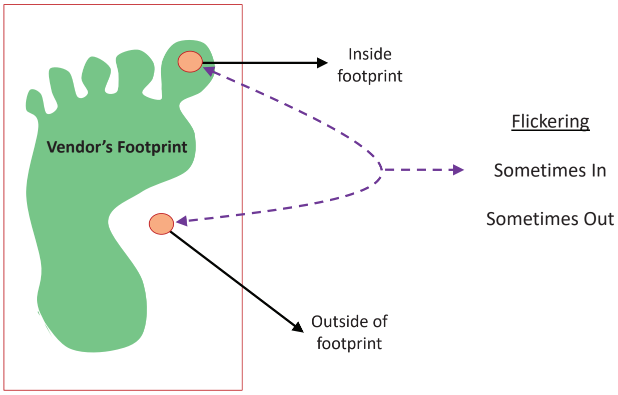
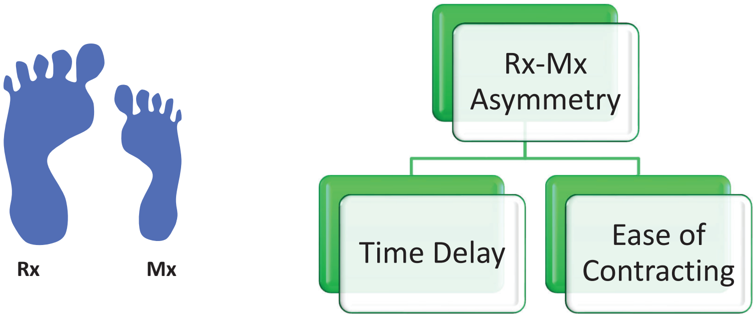


Figure 11: Differences Between Rx and Mx from a Data Standpoint



From a data standpoint, Rx and Mx claims differ in two respects (see Figure 11). First, Rx claims are processed promptly and are available to the Data Vendor in a matter of days. Mx claims, on the other hand, take weeks or even months before they become available to the Data Vendor. Second, Rx suppliers are more concentrated than Mx suppliers, which means that for the same contracting effort, the Data Vendor captures a larger share of the Rx market than that of the Mx market, resulting in the Rx/Mx differential in capture.

3. Capture-Transmission-Delivery

Let's now take a flow of goods perspective and identify the issues that bedevil the data as it moves along the supply chain from the supplier to the receiver.

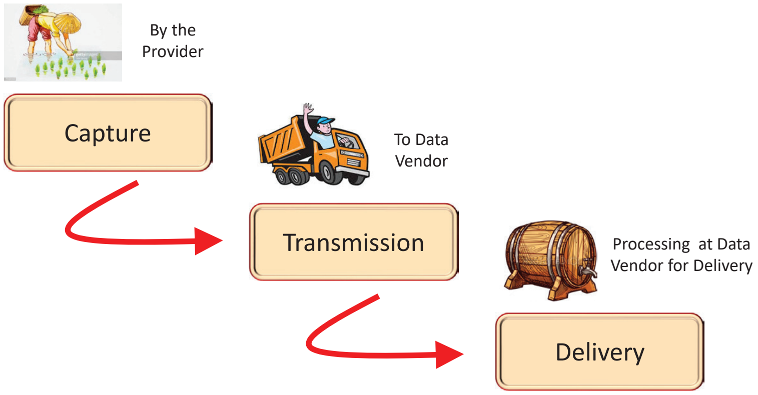
First off, what the Provider is capturing in the system may already be tainted. Here are some examples of what may be happening.

1. Lumping of Claims
2. Wrong ICD diagnosis codes (misdiagnosis, up-code, down code, no code)
3. Fuzzy J-code (not the fault of the provider)
4. Name of the head of practice appears for all transactions of the practice
5. Names on the UB-04 hospital claim may not reflect the actual person doing the work – Operating, Attending, Other, and Other.
6. Units unclear (10 ml or 10 vials)

Next, the data undergoes partial redaction as it gets transmitted from the Provider to the Data Vendor. Below are common examples:

1. Manufacturer and SP Blocking means that big chunks of the data are no longer available
2. Redaction of Payers from the Payer field in the data
3. Leakage from Clearinghouses creates the illusion of completeness (discussed earlier)

Figure 12: The Capture-Transmission-Delivery Data Process



Finally, the data vendor takes possession of the data and creates some additional problems in the process of getting the data ready for delivery. Below are common issues that creep up at this stage.

1. Patient Encryption – An encrypted patient id has to be generated and since PII (Personal Identifiable Information) is spotty at times, the encryption process makes two errors. In the first one, a patient is given 2 ids (e.g., women that get married and change name and address) and in the second one, 2 patients are assigned the same id (relatively easy to spot).
2. Information Redaction - The DUA (Data Usage Agreement) may preclude the release of certain data fields. The Legal Department of the Data Vendor may require further redaction of certain fields to ensure HIPAA compliance.
3. Payer Misclassification - Managed Medicare and Managed Medicaid Claims are inadvertently left under Commercial when they should be reclassified under Medicare or Medicaid. Also, the Vendor makes wrong guesses when the situation is unclear as in the cash share too large example we mentioned earlier.
4. Wrong Profile Information - The Vendor assigns the wrong address, specialty, and COT (Class of Trade) to physicians and accounts.
5. Dollar Amounts - Discounts and chargeback are not applied, resulting in grossly overestimated dollar amounts.

4. Prescriptions for Maladies

Let's take a look at how our vision system works as this will provide context for our discussion on how to address the myriad of issues in claims data (see Figure 13). Indeed, we do not see with the eye but with the brain. The eye merely captures photons.

First off, the image that is captured on the retina is messy and full of gaps and, in many respects, it is like the data we get from the Data Vendor.

For starters the image is inverted. To convince skeptics that was the case, René Descartes in the early 17th century placed a small screen in place of the retina of an excised bull's eye. The audience gasped in awe when they saw that the image on the screen was indeed upside down. Also, part of the image is missing and it is the part that falls on the optic nerve. We all remember the awesome experiment where the marking on the paper magically disappears when we bring the paper close to the eye while the other eye is closed. Color vision is only available in the dead center of the retina where most of the

Figure 13: The Image on the Retina Is Very Messy and We Are Not Aware

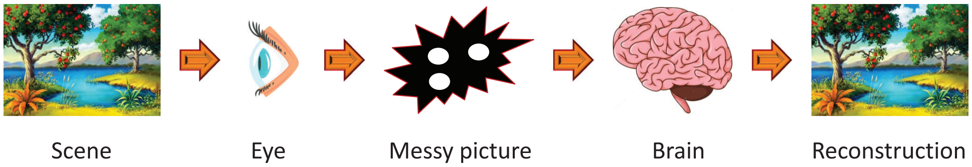
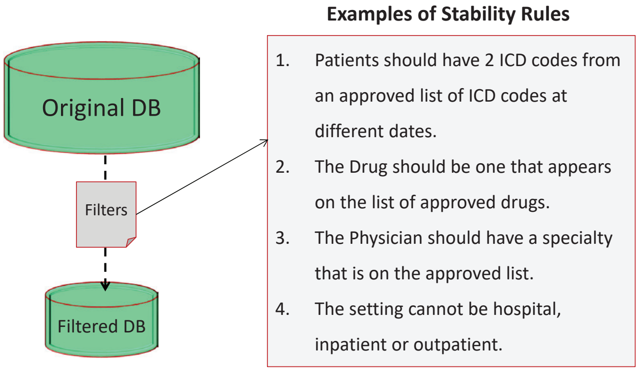


Figure 14: Apply Filters to Remove Noise in the Data



Examples of Stability Rules

1. Patients should have 2 ICD codes from an approved list of ICD codes at different dates.
2. The Drug should be one that appears on the list of approved drugs.
3. The Physician should have a specialty that is on the approved list.
4. The setting cannot be hospital, inpatient or outpatient.

cone receptor cells are concentrated; everywhere else, it's black and white. Finally, what we have is two 2D images, one from each eye, not the 3D picture we always see.

Indeed, it is with this less than ideal picture that the brain goes to work and does its magic. The end product is so good that we have trouble believing that the input is that bad. The lesson here is that just like the seeing brain, we may be able to reconstruct a much better dataset than the one that was handed over to us by the Data Vendor. There are 3 approaches that we use as an industry and, interestingly, they all have a counterpart in optics. Let's review them one by one.

Fix 1 - Apply Stability Rules to Reduce Noise in the Data

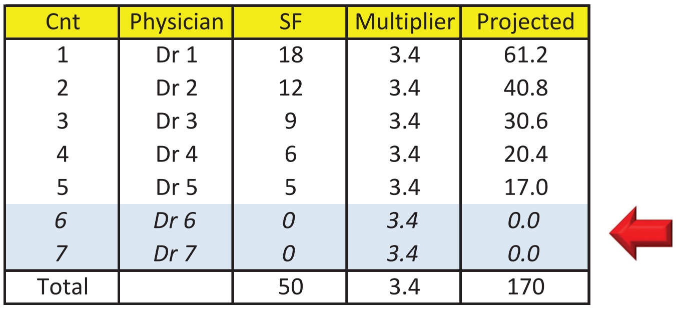
The principle is straightforward: Apply multiple filters to remove noise from the dataset. For instance, throw away patients that have only one claim of the ICD code of interest. Remove physicians if they are of the wrong specialty. While this works well in general, it does have the disadvantage of introducing biases. It is well known that measurements of adherence conducted on filtered datasets overestimate the actual adherence of the population they are supposed to represent.

Analogy from Optics
To get a sharp image, reduce intake. This is the idea behind the pinhole which harks back to the camera obscura (dark room), a device that was known by Aristotle and the Chinese in the 5th century BCE. When reducing light intake, the smaller the hole, the better, up to a point beyond which the image becomes too faint for us to see.

Fix 2 - Use Projection to Make Up for Missing Activity

We know from trusted sources that the sum total of the activity of a given geography is Y but our dataset falls short and only reports X. The idea of projection is to scale up the observed activity by a factor of Y/X. In actuality, other dynamics come into play and the scaling factor that we apply may not be the same for each observation.

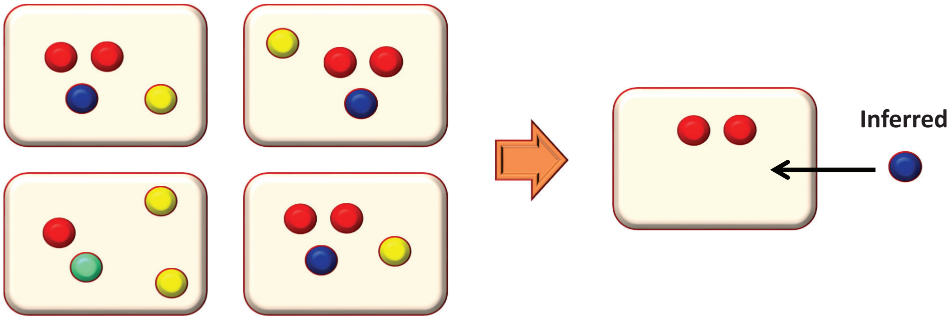
Figure 15: Scale Up the Data Points to Make Up for Missing Activity



Cnt	Physician	SF	Multiplier	Projected
1	Dr 1	18	3.4	61.2
2	Dr 2	12	3.4	40.8
3	Dr 3	9	3.4	30.6
4	Dr 4	6	3.4	20.4
5	Dr 5	5	3.4	17.0
6	*Dr 6*	*0*	*3.4*	*0.0*
7	*Dr 7*	*0*	*3.4*	*0.0*
Total		50	3.4	170

0 multiplied by any constant is still 0.

Figure 16: Pattern Recognition Allow Us to Identify Missing Elements



Take a look at the numerical toy example in Figure 15. The dataset reports 50 units for 5 physicians and 0 units for 2 physicians. Trusted data sources say 170 units, so we apply a scaling factor of 3.4 (170 divided by 50) to each physician and the projected sum total activity aligns perfectly with the trusted data sources.

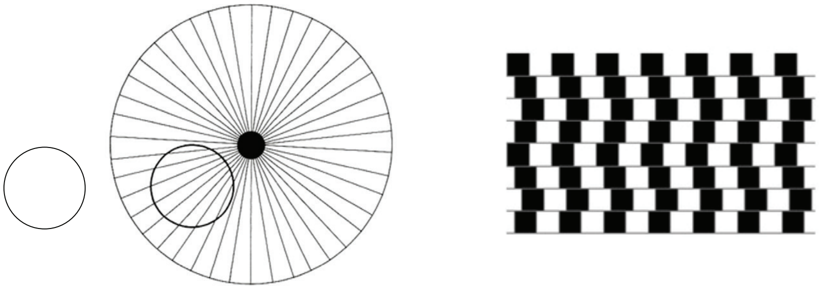
Two observations are in order. First, physicians now have fractional activity. How can a physician write 20.4 Rx's? It's either 20 or 21. By the way, you'll see this happening all the time in datasets such as IQVIA's Xponent and Symphony's IDV (Integrated Data Verse). It's a practice that is meant to make the numbers work out at the aggregate level even if it is at the expense of the granular level. Second, physicians like Dr 6 and Dr. 7 who had 0 activity in the raw data still have 0 activity in the projected data. In some sense, projection has reduced their importance since the gap between Drs 6-7 and Drs 1-5 has widened. By the way, we found a good way to address this problem and it consists of leveraging SVD (Singular Value Decomposition), the same algorithm the Bellkor's Pragmatic Chaos Group used as the centerpiece of their approach to snatch the $1 million Netflix Prize in 2009.

Analogy from Optics
Say you are spear fishing. If you throw the spear right at the fish, you'll miss it. That's because refraction bends the light that comes from under water, creating the impression that everything is further away. You need to adjust your throw by an amount that factors in distance from the fish, level of water relative to your eyes, and refractive index of the water, so as to undo the effect of refraction.

Fix 3 - Use Peripheral Information to Infer Presence of What's Missing

Consider the toy example in Figure 16. Each time we see 2 red dots, we see a blue dot, so when we see 2 red dots and no blue dot, we infer that the blue dot is missing.

Figure 17: Two Optical Illusions That Show We Are Being Fooled



The two small circles are identical!

The horizontal lines are parallel!

This is the type of reasoning we employ to establish that the patient underwent CRT therapy when radiation is missing in the data. If the blue dot is radiation therapy, what are the red dots? Well, an analysis of the claims data shows that the red dots can be drugs such as Paclitaxel, Keytruda or Alimta, diagnoses such as "Encounter for antineoplastic radiation therapy" (Z510 ICD-10), and procedures such as "Radiation planning therapy" (77261-77263 CPT code).

Analogy from Optics
We cannot see a black hole but we can infer its presence thanks to gravitational lensing. Gravity around the black hole is so strong that it bends light that passes near it. If there is a black hole in the line of sight between us and a distant galaxy, we'll see multiple copies of the galaxy and if the alignment is just right a ring of galaxies. By the way, this is how we know that black holes and dark matter exist.

Word of caution: All the fixes we discussed only work to some extent and several issues may persist. This means that we always need to be on the lookout and exercise judgment. See Sidebar 3 for a cautionary tale about seeing what's not in the data.

Even our almighty brain that handles messy images so elegantly can be fooled. In Figure 17, it is hard to believe that the two small circles on the left are identical or that the jagged lines on the right are horizontal and parallel to each other. We handle these situations by having our reasoning override what we see.

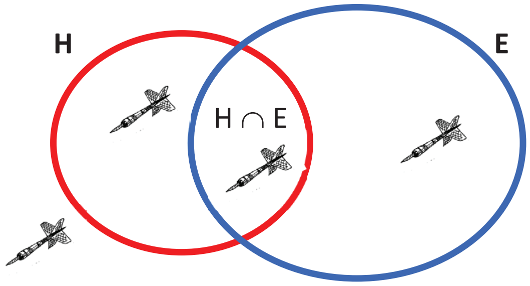
Sidebar 3 - Life on Mars Blunder with Schiaparelli and Lowell

Giovanni Schiaparelli, an Italian astronomer, looked up in his small telescope in the late 1870's in Milan and reported seeing canali on Mars. Canali in Italian refer to either artificial or natural canals and were immediately taken to mean artificial canals, suggesting intelligent life on the red planet. Yes, green little men may indeed be scurrying around! Percival Lowell, a wealthy amateur astronomer from Boston, established the Lowell Observatory at Flagstaff, Arizona to study Mars and spent a good deal of his life mapping the Martian canal structure. Lowell knew a thing or two about astronomy as he predicted the existence of a planet beyond Neptune which turned out to be Pluto. Later on, Schiaparelli saw two sets of the canals running parallel to each other, which he called gemination (as in twins). Actually, his eyes were failing and he was literally seeing double after so many years of straining. Also, he was color blind, which is ironic for a man whose claims depend so much on his eyesight. In truth, no such canals exist on Mars.

<u>5. Machine Learning</u>

Thanks to the explosive growth of Machine Learning in recent years, we now have a whole array of techniques to choose from to fix the

Figure 18: Probability that the Dart Lands in the Overlap of Circles H and E



issues in claims data. We'll focus here on plugging missing data and leave the problem of fixing wrong data for another venue.

It has become clear that the major pushback regarding ML techniques is their black box nature. If it's a neural network that is employed, the only explanation we get, if we can call this an explanation, is the weights that the back-propagation algorithm assigns to the synapses between the neurons. We human beings cannot take a finding just like this and run with it. We need to understand the thought process that led to the conclusion, the considerations that were and were not taken into account, and the caveats associated with the finding. Only then will we do anything with the finding if we so choose.

The good news is not all ML techniques are black boxes. While they do not explain their reasoning as a human would, some allow us to understand how they arrived at the conclusion. We had great success with several of them and will focus here on Bayesian Reasoning. We'll describe how it works and then how we applied it to address the CRT problem we described at the beginning.

A. Bayesian Reasoning

Say you catch a glimpse of a woman with long dark hair at the airport. Later on as you line up for the men's restroom, you notice her again lining up in front of you. Then it hits you: it's a he, not a she. Originally, you believed she was a woman but in light of the restroom evidence, you changed your mind and you are now convinced it's a man with long hair. This is Bayesian reasoning at work.

Let's plug in some numbers to get a better understanding of how Bayesian reasoning works. Let's recap the major facts about the incident.

1. H is the Hypothesis and it is "It's a woman"
2. E is the Evidence and it is "Lined up at men's restroom"
3. p(H) = Probability it's a woman = 99%
4. p(E) = Probability we see a person lining up at the men's restroom = 1%
5. p(H|E) = probability it's a woman given the restroom evidence = ?

At this point, what we need is a way to infer P(H|E). For that, consider Figure 18.

What is the probability of a dart landing in the overlap of H and E? This is p(H ∩ E). It is the probability that the dart is in E given that it is in H when we know it is in H. This is p(E|H). p(H). Since H and E are interchangeable, it is also the probability that the dart is in H given that it is in E when we know it is in E. This is p(H|E). p(E). What we have then is P(H ∩ E) = P(E|H).P(H) = P(H|E).P(E). Putting p(H|E) as the subject of the formula, we have the Bayes formula:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

At this point, we are missing only one thing and it is p(E|H) which is the probability that someone would line up at the men's bathroom given it's a woman. This is extremely rare (women's restroom is broken?) and let's say this happens 1 in 10,000 cases which gives us a probability of .0001.

$$P(H \mid E) = \frac{.0001 * .99}{.01} = .0099$$

In light of the restroom evidence, the probability that it is a woman is less than 1 in 1000 which is the complete opposite of what we believed before, namely, that 99 out of 100 it's a woman.

In practice, it is more convenient to work with odds than with probabilities. If the probability is 50%, the odds are 1. If the probability is 75%, the odds are 3:1 . The odds are simply the ratio of probability to 1 minus probability. The odds counterpart of the Bayes formula is obtained by dividing p(H|E) by p(~H|E) which gives us:

$$O(H \mid E) = \frac{P(E \mid H)}{P(E \mid \sim H)} O(H) = Lik(E \mid H) * O(H)$$

If we assume that the pieces of evidence are independent of each other (this is Naive Bayes), we can then simply multiply all likelihoods with the prior odds to get the posterior odds.

Now let's see how Bayes Reasoning can help us plug missing data. At this point, it does not matter if the missing data was left out from the dataset like Radiation in the CRT therapy example or has never been part of claims data to start with such as when the patient will discontinue therapy or change line of therapy.

Let's assume we are looking at the following:

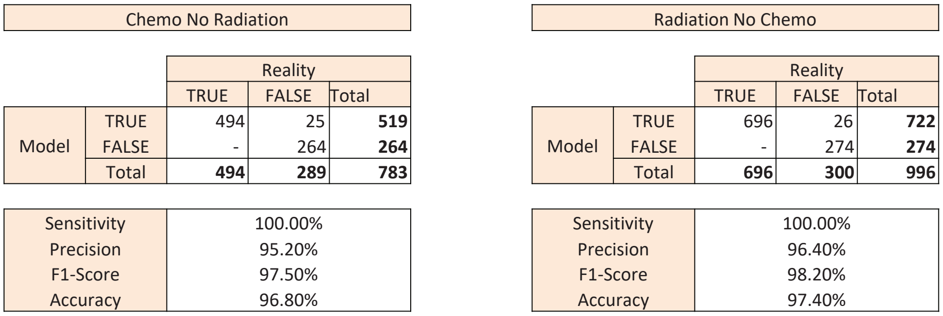
1. H - the missing event did happen but was not captured.
2. ~H - the missing event never happened and as a result is not in the data.
3. O(H) is the odds that a patient taken at random from the dataset will have an event missing and not be captured. This is established using knowledge we have on how well or poorly the event is captured in the claims data.
4. E - evidence that may impact our belief in H or ~H. Examples include drugs, diagnoses, procedures, lab orders, lab results, hospitalizations, referrals, up/down dosing, etc.
5. O(H|E) is what we are after and it is the odds that the missing event did happen when the Evidence is present.

Two pieces are needed to establish O(H|E). They are the numerators and denominators of Lik(E|H). The numerator is the probability of observing the evidence when the missing event did happen but was not captured. That's p(E|H). The denominator is the probability of observing the evidence when the missing event never happened and so is not in the data. That's p(E|~H).

To that end, we construct 2 cohorts of patients, a positive and a negative. In the positive cohort, the missing event did happen but was not captured and in the negative cohort, the missing event never happened. We then go through all the pieces of evidence we can think of and call out those where the frequency of occurrence of the evidence is high in one cohort but low in the other. Note that a piece of evidence that appears as frequently in the positive cohort as in the negative cohort is irrelevant as it does not alter our current belief in H one way or the other.

Now that we have everything we need, we can use the odds formulation of the Bayes formula to compute the posterior odds that the missing event did happen but was not captured. If the odds are greater than 1, we'll say that the missing event happened but was not captured, otherwise that it never happened.

Figure 19: Performance of Bayesian Reasoning Model



Chemo No Radiation

		Reality		
		TRUE	FALSE	Total
Model	TRUE	494	25	**519**
	FALSE	-	264	**264**
	Total	**494**	**289**	**783**

Sensitivity	100.00%
Precision	95.20%
F1-Score	97.50%
Accuracy	96.80%

Radiation No Chemo

		Reality		
		TRUE	FALSE	Total
Model	TRUE	696	26	**722**
	FALSE	-	274	**274**
	Total	**696**	**300**	**996**

Sensitivity	100.00%
Precision	96.40%
F1-Score	98.20%
Accuracy	97.40%

Note on Explanability
The Bayesian reasoning is very transparent and it is clear how the conclusion was reached. Indeed, for any patient, we know what the prior odds of the missing event happening but not captured in the data are and also all the pieces of evidence that were flagged as relevant for that patient along with their respective likelihoods.

B. Using Bayesian Reasoning to Solve the CRT Problem

Problem

The CRT problem, in essence, has to do with the fact that the claims are incomplete and as such prevent us from concluding that a patient underwent Chemo-Radiation therapy (CRT) when Chemo or Radiation is absent in the data. Indeed, absence of evidence is not evidence of absence. The Chemo or Radiation may or may not have happened. Our problem is to resolve the ambiguity by establishing yes or no if the patient actually underwent CRT therapy when Chemo and/or Radiation is missing.

Approach

Claims data carry more than Chemo and Radiation information regarding the patient. They also document diagnoses, procedures, lab tests, lab results, drugs, and hospitalizations the patient underwent. Using the Bayesian approach we discussed, we established which ones of these markers are predictive of the presence or absence of CRT therapy and quantified their likelihoods. Given a patient where either Chemo or Radiation is missing, this approach assigns a probability with which CRT therapy happened.

Performance Results

To evaluate our model, we created 4 groups of patients. Group 1 and 3 are made up of patients that are known to have undergone CRT. For Group 1 (494 patients), we hid the Radiation part and Group 3 (696 patients) the Chemo part. Groups 2 and 4 are made up of patients that have not undergone CRT. Group 2 (289 patients) contains patients that did not undergo Radiation and Group 4 (300 patients) patients that did not undergo Chemo.

Figure 19 displays the result of the performance evaluation. The model predicts with 96.8% accuracy that the patient underwent CRT among those where Radiation is missing and 97.4% accuracy among those where Chemo is missing.

This approach allows us to identify 4 times more patients than what a direct read of the data suggests.

6. Conclusion

Below are the key takeaways.

1. Claims data is full of traps and unless you have a good understanding of its shortcomings, it's very easy to draw the wrong conclusion.
2. The data is guilty until proven innocent. The burden of proof sits squarely on us, the folks that draw insights from the data, although we may feel at times it's clearly the fault of the data vendor. Here's why: When you make a claim that turns out to be bogus, it is you who have to do the explaining, not the data vendor.
3. Be extremely cautious when you uncover some truly remarkable insights. Use common sense and other sources of data to pressure-test your findings. Are you seeing Schiaparelli's canali and concluding the existence of little green men?
4. ML Techniques are a great way to plug in holes in the data. Remember the CRT story. It's clever ML, not more data sources that saved the day!

About the Author

Jean-Patrick Tsang *is the Founder and President of Bayser, a Chicago-based consulting firm dedicated to sales and marketing for pharmaceutical companies. JP is an expert in data strategy and advanced analytics. JP has published 25+ papers, given 80+ talks at conferences, and completed 250+ projects. In a previous life, JP deployed Artificial Intelligence to automate the design of payloads for satellites. JP earned a Ph.D. in Artificial Intelligence from Grenoble University, advised 2 PhD students, and earned an MBA from INSEAD in Fontainebleau, France. He was the recipient of the 2015 PMSA Lifetime Achievement Award.*

Igor Rudychev *is Head, US Digital, Data, and Innovations with Astrazeneca. Igor has 25 years of analytics & insights experience and for the last 18 years he has been specializing in Pharmaceutical Sales & Marketing Analytics, Business Insights, Market Research, Resource Allocation, Promotion Response, Sales Force Optimization, Brand Analytics, Patient-Level Data, and Forecasting. Prior to joining AstraZeneca, Igor was leading Samples Operations, Resource Allocation, Advanced Analytics, and Commercial Assessment & Forecasting at Pfizer. Before Pfizer, Igor was Vice President of Operations with Bayser, a Management Consulting firm in Chicago, where he provided strategic and analytical insights to the top 20 global pharmaceutical companies.*

Igor has been published extensively in trade & peer review journals (including ASCO, WCLC, Journal of Clinical Oncology, Journal of Thoracic Oncology, AMCP, MMM, PDT, DMT, Phys. Lett., Class. & Quant. Grav., and Nuclear Physics B). Igor has an MBA from the University of Chicago Booth School of Business in Analytical Marketing & Analytical Finance and Ph.D. in Theoretical Physics (Superstring Theory & Black Holes) from Texas A&M University. Igor has served PMSA's board for five years, as Treasurer, Professional Development Chair, and Research and Education Chair.

AI Algorithms for Disease Detection: Methodological Decisions for Development of Models Validated Through a Clinical, Analytical, and Commercial Lens

Brian Malpede, Manager; Goksu Dogan, Principal; Scott Moreland, Data Scientist; Rabe'e Cheheltani, Consultant; Brittany Fischer, Associate Consultant; Suyin Lee, Manager; Nadea Leavitt, Principal and US Lead; Orla Doyle, Lead Data Scientist; John Rigg, Senior Principal and Global Lead, IQVIA Predictive Analytics

Abstract: Disease detection driven by artificial intelligence (AI) has demonstrated to be an effective tool for identifying undiagnosed patients with complex common as well as rare diseases. The use of these algorithms is driven by awareness that underdiagnosis leads to a heavy burden for patients and healthcare professionals, and is also a challenge for pharmaceutical companies seeking to expand the patient pool for their medications, whether to power clinical trials or to efficiently target healthcare providers (HCPs). However, despite widespread awareness and usage of this application, methodologies utilized are highly variable and learnings are rarely shared. In addition, the commercial application of models built for pharmaceutical companies is not always considered during model development stages, despite the importance of methodological decisions to the efficient and successful real-time implementation of AI driven diagnostics. In this paper, a cross-functional methodological approach to AI algorithm design for undiagnosed patient detection will be detailed, an approach honed through the development of numerous algorithms applied to a wide-range of diseases, from common to ultra-rare, in diverse therapeutic areas. Methodological and technical considerations will be described that consider relevant aspects of clinical, analytical, and commercial environments to develop an AI solution that is statistically robust, clinically relevant, interpretable, and operationally tenable.

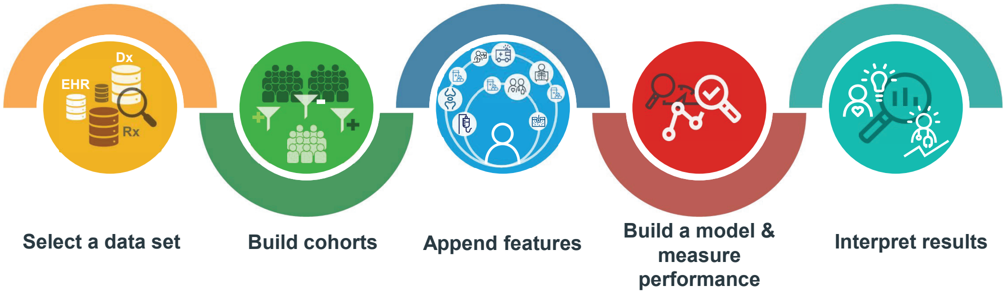
Keywords: Artificial Intelligence, Machine Learning, Predictive Analytics, Disease Detection, Rare Disease, Algorithms

1.0 Introduction

Disease detection algorithms driven by artificial intelligence (AI) have demonstrated to be an effective tool for identifying undiagnosed patients with underdiagnosed, un-coded, and rare diseases. The application of these algorithms is greatly influenced by challenges that patients and healthcare professionals face, as well as those encountered by pharmaceutical companies trying to expand the pool of candidate patients for their medications, whether to power clinical trials or to efficiently target healthcare providers (HCPs).

Despite the popularity and widespread use of disease detection algorithms, the methodology design varies highly across studies and insights into best practices are rarely shared. Developing an effective disease detection algorithm is a multifaceted solution involving technical, clinical, and operational expertise. These capabilities are essential in informing each step of study design, model development, and deployment. Clinical validation and interpretation of the model is equally important to the evaluation and the optimization of advanced AI techniques. Further, the implications on business operations, which are key to the development and real-time implementation of AI driven diagnostics, are often overlooked during model development and deployment phases.

Figure 1: The Five Main Steps in Developing an AI Model Designed for Prediction of Undiagnosed Disease



Source: IQVIA illustration

In this paper, we detail a cross-functional methodological approach to AI algorithm design for undiagnosed patient detection, established over several years and applied to various diseases, ranging from common to ultra-rare. We describe methodological and technical considerations that reflect relevant aspects of clinical, analytical, and commercial environments to develop an AI solution that is statistically sound, clinically relevant, interpretable, and operationally tenable. We will focus on three main areas, including:

1. Application of analytical techniques that drive robust clinical and statistical validation as well as interpretability and insight of AI models
2. Inputs and techniques that foster development of a model that is appropriate and actionable for the desired commercial implementation
3. The outlook for building and utilizing diagnostic algorithms developed with AI

2.0 The Process of Building a Model

The primary and essential elements of building a model to predict diagnosis are consistent across disease states and applications. Details within each step may be variable, but the overall process can be summarized into five main steps, shown in Figure 1.

2.1 Selecting a Dataset and Building Cohorts

Selection of the dataset is a key aspect of AI modeling. The dataset is important for several reasons, including:

- The identification of patients from which the model will learn
- The type and volume of predictors that can be leveraged
- The ongoing business application

Important considerations for selection include a balance of cost, patient coverage, and application (i.e. clinical, commercial, etc.). Adjudicated or non-adjudicated medical claims and electronic health records (EHR) are commonly used or considered datasets. Additional datasets that might be used to supplement modeling include patient registries, lab claims, and consumer data.

The first and most important question to consider is model application and key goals for the modeling effort. For clinical trial recruitment and pharmaceutical marketing, the goals include broad and timely identification of new potential patients and their HCPs. As such, an open claims dataset with robust coverage of patients and HCPs and near real-time updates would be most applicable. Using closed claims for this application is a significant disadvantage given there is an extended lag time in the data. In contrast, for development of a clinical decision tool, the goals might include clinical insight or diagnostic indicators and limited disruption of clinician workflow. For this specific application, an EHR dataset would be more appropriate as it would mimic the environment in which it would be deployed.

Additional datasets, such as lab data, patient registries, and specific consumer data, can supplement modeling. These datasets serve two main purposes, including the identification of known diseased patients (e.g. through disease-specific lab/genetic testing), and profiling of patient subgroups to gain insights into the studied patient population (e.g. through consumer attributes). While all datasets have individual value, researchers should use caution in considering using all datasets for the same model. Cost and complexity for both initial development and ongoing deployment compared to gain in model performance should be evaluated and balanced appropriately.

2.2 Patient Cohort Design

After selecting the appropriate dataset, the next critical element of disease detection modeling is the development of clean, validated patient cohorts, or the sets of patients from which the model will learn to differentiate disease from non-disease. These groups of patients are often referred to as positive cohort (with disease) and negative cohort (without disease).

2.3 Selection of a Positive Cohort

In the simplest form, positive patients can be selected based on defined criteria that is indicative of the disease of interest. For example, if the goal of the model is to predict patients diagnosed with shingles, the selection criteria could be defined as evidence of a claim with the ICD-10 diagnosis code specific to the disease (Herpes Zoster, B02 family of codes). However, selection of positive cohorts may not always be this straightforward, and more complex steps to identify a validated positive cohort may be necessary. Several examples of more complex scenarios are discussed below, including:

A. The use of multiple claims for the disease of interest to increase confidence in the diagnosis
B. The use of "proxies" such as: 1) treatments indicated exclusively for a target disease state or 2) the combination of multiple diagnostic codes that together define a specific disease state
C. A selection period after October 2015 for diseases in which the single ICD-9 code is shared among multiple diseases, where only patients with a definitive and non-shared ICD-10 code are included as positive
D. The use of a supplemental data source such as EHR, lab, or patient registry

Scenario A. The existence of a single occurrence of an ICD-10 in a patient's medical history may not be indicative of the patient truly having the disease, but rather of testing for the disease. In these instances, criteria for a positive patient can be refined to the requirement of at least two instances of the ICD-10 code specific to the disease. This selection can drive a higher confidence of confirmatory diagnosis for positive patients, eliminating patients that may have been tested for disease but not ultimately diagnosed.

Scenario B. In some cases, the disease of interest may have a non-specific ICD code, where the code itself is shared among diseases of the same family or is used for patient populations outside of the disease of interest. For example, if the model is being developed to predict patients with hereditary angioedema (HAE), a challenge that arises is that the ICD-10 code for HAE (D84.1) is shared with other forms of angioedema. In this case, other proxies of a confirmed disease state in a patient's medical history should be considered to select those diagnosed with only the disease of interest, and thus the cleanest sample of positive patients with which to train a model. Once again using HAE as an example, a specific diagnosis may be defined as patients with evidence of treatments indicated exclusively for HAE, or through evidence of a combination of diagnosis codes for HAE (i.e. when a patient has evidence of both the ICD-10 D84.1 and broad ICD-9 277.6, which codes for numerous disorders under "deficiencies of circulating enzymes") along with evidence of non-specific treatments used for management of the disease.

Scenario C. Many rare diseases may have a specific ICD code within version 10, but within the ICD-9 definitions shared diagnostic coding with a broader group of conditions. In these cases, a positive patient selection period can be limited to after October 2015, the beginning of ICD-10 release in the U.S., to help ensure a clean positive cohort. The definition of first instance of diagnosis for such patients, however, should be based on the first observed diagnosis code in their history, whether from the shared ICD-9 or exclusive ICD-10 version, such that the timing of the patient's initial diagnosis is identified with highest confidence.

Scenario D. Finally, a single data source may not always be enough to identify patients of interest. Situations exist in which a group of positive patients cannot be identified solely through diagnosis and treatment coding in claims data, such as when there is simply no ICD code or available proxy. In these cases, addition of other data sources such as EHR, patient registries, or lab results (e.g. genetic testing) may be beneficial. As an example, to identify a disease severity not captured in ICD coding, EHR data can be used to reveal patients with evidence from provider notes of the disease state of interest. These patients can then be linked back to claims data (or other datasets) for model training.

2.4 Selection of a Negative Cohort

At a basic level, patients in the negative cohort can be selected based on the absence of evidence for the disease of interest. In some cases, further filtering of the cohort can eliminate patients who are ineligible to have the disease and thus would not serve as suitable comparators. For example, male patients should be excluded from a negative cohort for a model identifying patients with endometriosis, a disease of the uterus, and elderly patients would be inappropriate to include for a model focused on a pediatric disease.

In some instances, lack of evidence for disease is not necessarily indicative of a patient that is not affected, but rather a result of limitations in data coverage, coding practices, or under-diagnosis. An understanding of estimated prevalence of unknown or unlabeled positives, which could be wrongly labeled as negatives, is helpful in approximating potential impact on model performance. Negative patients are often selected from a random sample of the general population with no evidence of the disease of interest, and thus for modeling in rare and ultra-rare diseases, the existence of false negatives (positive patients that are not identified as such due to the limitations mentioned above) is typically negligible because of the low prevalence of disease. In other instances, however, such as a case of finding patients with common (typically under-coded or under-diagnosed) diseases, the existence of these 'unknown positives' wrongly labeled as negatives could have substantial implications on model development, with the biggest impact on potential incorrect measurement of model performance. Techniques that seek to understand or mitigate the impact of unlabeled positives are discussed in a later section.

Cohort selection is crucial to the modeling process, but researchers should avoid the urge to overclean the cohorts, thus running the risk of introducing bias, reducing sample size, and hindering model efficacy. The introduction of bias is an especially critical consideration in study design. Main sources of bias include changing data coverage, seasonality, market events (diagnostic or procedure coding changes, patient pathway updates, new treatments introduced), and inappropriate selection criteria that lead to an improper positive cohort. Mitigation of these sources of bias reside in the proper selection of cohorts as outlined above as well as in suitable selection of a study time period to be utilized for model training.

3.0 Appending Features (Predictor Design)

Following development of positive and negative cohorts, the next step is to append potential predictors, or in other words, identify the medical history that will be used to train the model. The importance of developing a set of predictors (also called features, model inputs, or variables) cannot be understated. These inputs form the basis for how a model makes its decisions about patient predictions and are thus critical in both driving the predictions themselves and in gaining clinical insight from the model. There are two main approaches to predictor generation: an automated data-driven process, and a hypothesis, or knowledge-driven process. Finding a balance between the classical hypothesis-driven and automated data-driven feature generation is essential for an interpretable and operational model.

3.1 Hypothesis (Knowledge) Driven Features

Hypothesis, or knowledge-driven feature generation, is valuable in that it allows for testing of predictors considered to be clinically important. As such, these predictors are easy to interpret and simple to understand in clinical terms. However, despite their interpretability, these predictors may not capture all relevant aspects of medical history for a specific disease. An example of this type of feature is the roll up of a set of diagnostic ICD-10 coding for abdominal pain. This physical manifestation is captured in several individual ICD-10 codes, each of which defines a specific region of the abdomen as well as the type of pain. A classical knowledge-driven feature would usually contain each of these codes “rolled-up” into one clinical bucket, such that the only information the model sees is *abdominal pain*, and none of the underlying specificity from individual clinical claims codes. Several mapping systems, such as SNOMED[1],exist to align coding with clinically relevant roll-ups, allowing for straightforward assembly of interpretable knowledge-driven clinical features.

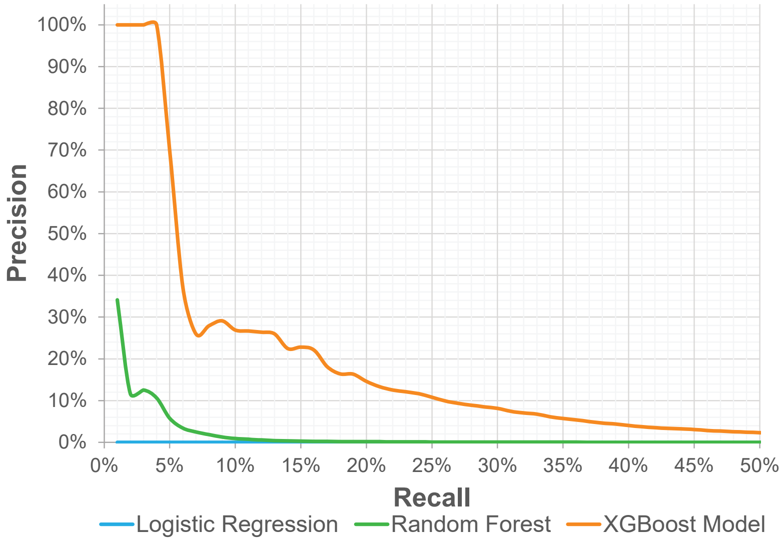
3.2 Data-Driven Features

To leverage additional medical history, an automated data-driven process can be utilized in conjunction with hypothesis-driven features. This process involves selecting features based on the data alone to define inputs for modeling. Leveraging the data to define potential predictors often leads to an initial assessment of thousands of different features. This lengthy list of features can then be narrowed down using a variety of selection techniques to a set of those that are most relevant. Using the data in such a way can reveal previously unknown predictors and is especially useful in disease states where there is limited understanding of the patient journey. While this process is valuable, one caveat with data-driven features is that they are often presented with substantial granularity (such as a single CPT or a single ICD-10 code). This granularity can be helpful but may also result in challenges with clinical interpretation of model decision making.

4.0 Building a Model and Assessing Performance

With positive and negative patient cohorts selected, and a set of features built for input

Figure 2: Comparison of Model Performance for Several Techniques



Comparison of logistic regression, random forest, and XGBoost for a model trained to predict undiagnosed patients with an ultra-rare neuromuscular disorder. XGBoost outperforms the other techniques for detection of this disease. ***Precision***: proportion of positive patients correctly identified by the model; ***Recall***: proportion of positives identified by the model out of all known positives. *Source:* IQVIA case study.

into the modeling process, binary classification is the intuitive solution to develop diagnostic algorithms. Several modeling techniques are typically employed for model training, including logistic regression, random forest, gradient boosting, and neural networks, all of which can work quite well depending on the circumstances. With that said, decision tree algorithms based on gradient boosting (such as XGBoost)[2] are found to work particularly well for the domain of disease detection (See Figure 2).

4.1 Model Validation

Validation of the model's performance is critical for assessing commercial, clinical, or other applicability. This process helps define how effectively a model can identify undiagnosed patients in a real-world commercial setting as well as try to understand a model's complex decision-making process. There are several ways to approach performance measurement, including the precision-recall (PR) curve alluded to above, the F_1 score (equal to the harmonic mean of precision and recall), and the area under the receiver operator curve (AUC) (defined as the integral of the model's true positive rate as a function of its false positive rate).

The F_1 and AUC scores examine model performance as a single number, making them popular choices for performance ranking and hyperparameter optimization. Their definitions, however, average and integrate over competing terms in the confusion matrix – e.g. the true positive rate and false positive rate – rendering them unsuitable for nuanced applications where a certain region of the precision-recall or receiver operating curve is of particular interest, as is most often the case for disease detection applications. In these situations, a more flexible variant of the F_1 score known as the generalized F_β score can be used to quantify model performance, biased to higher or lower recall levels by tuning the value of the parameter β. This value is best suited for instances where there is significant imbalance between the positive and negative patient populations.

4.2 Evaluation of a Precision Recall Curve

While the F_β and AUC measures can be useful for ranking and optimization through measurement of overall performance, they contain only a fraction of the information encoded in the full PR curve. The PR curve is typically the most valuable metric as it provides an intuitive assessment of model performance as well as highly actionable outputs. In addition, the curve allows for an adjustable threshold to suit multiple commercial and clinical deployment initiatives. Identifying a recall threshold at which a patient is identified as high likelihood (with highest levels of precision) versus a patient that is lower likelihood (lower levels of precision) allows for targeted differentiation of predicted patient candidates. Examples of applications include choosing a personal vs. non-personal promotion in a commercial setting or advocating for an expensive diagnostic test versus less costly patient monitoring in a clinical setting.

Reviewing a PR curve is highly useful in understanding a model's real-world application through quantification of potential performance when implemented. However, the PR curve assessment must be used appropriately. Specifically, the curve must be calculated using a representative ratio of positive to negative patients, or in other words, the model should be assessed for performance on a set of patients that reflects the expected real-world population. If the model is provided a 1:1 ratio of positive to negative patients for testing, the number of false positives (patients predicted to have disease that do not actually have the disease) will be grossly underestimated.

In contrast, if the ratio is set according to expected prevalence of disease in the population, the actual expected number of false positives, and therefore a true understanding of real-world application, will be defined. The example PR curve in Figure 3, generated for a model designed to detect an ultra-rare neuromuscular disorder, demonstrates the importance of utilizing a representative ratio for model evaluation. Take, for instance, the point on the curve at 10% recall. When the precision of the model is evaluated on a 1:1 ratio it is nearly 100%. Even at a 1:1,000 ratio the precision is approximately 95%. This precision is deceptively high or artificially inflated as the true precision is 28% when evaluated on a ratio that best approximates real-world disease prevalence.

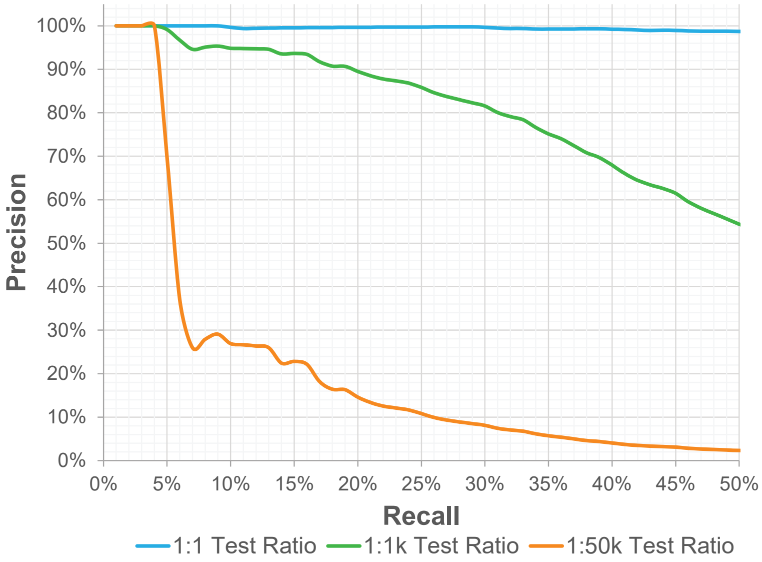
The definition of good performance is thus dependent on the specific modeling exercise and disease of interest. While the actual value of precision may be low (see Figure 3), the relative increase in performance relative to selecting patients at random, a measurement that can be based on disease prevalence, is a more apt way to assess the function of a model.

4.3 Considerations for Model Training with the Expectation of False Negatives

In situations where there is the expectation of an unclean negative cohort, or in other words that the proportion of false negatives (also called unlabeled positives) will be high, an approach known as positive and unlabeled (PU) learning can help.[4] An example of a situation in which this type of AI learning could be appropriate is a disease in which a significant portion of the diagnosed population are un-coded in claims data. These situations may arise due to stigma associated with the disease, or in the absence of specific treatments and thus limited incentive or awareness for HCPs to submit a claim for the disease itself. This issue may impede the model's ability to learn and differentiate positive from negative, and thus adversely affect the patterns and profiles that the model leverages when applied to a real-world dataset for undiagnosed patient identification.

Here the presence of unlabeled/unknown positives in the claims data can be inferred by

Figure 3: PR Curves Adjusted Across Changing Ratio of Positive to Negative Patients



A PR curve should be calculated for a representative ratio of positive to negative patients for diagnostic modeling. If not, the curve, and thus evaluation of model performance, can be misleading and lead to inappropriate assessment of model performance. Each of the PR curves in this image are calculated for the same model at different ratios of positive (diseased) to negative (control/non-diseased) patients. *Source:* IQVIA case study.

comparing the observed incidence of disease to clinical estimates of the disease's published incidence or prevalence. In practice, these quantities can disagree for a number of reasons, including incomplete data capture and fundamental differences between clinical reality and the actions of health care providers with regard to clinical coding and documentation. If the goal of a study is to detect all positive patients, even those which may go undiagnosed in the absence of external intervention, then a model grossly underreports existence of positive patients since it uses the features for those identified as positives as distinguished from those marked as negative. Due to the negative class having a significant number of unlabeled positives, the features that would be used to identify a positive are severely diluted.

One intriguing PU learning method proposed in the literature is to use "spies" to identify clean negative examples in the pool of unlabeled examples. These patients can then be used with the known positives to train a traditional binary classifier.[5] The spies are randomly sampled positive examples that are artificially injected into the pool of unlabeled examples during training. The unlabeled examples (that are known positives) are then modeled as if they were purely negative, and a traditional classifier is trained on the resulting positive and negative examples. Specifically, the largest possible decision threshold *t* is found such that only a small fraction *f* of spies, e.g. *f* = 10%, have a classifier score *smaller* than *t*. It is assumed that the examples with scores less than *t* are mostly clean negatives. The clean negatives are finally combined with the known positives and used to train a second stage traditional classifier with a high purity negative cohort. Once the second stage classifier is trained, it can be used to score previously unseen examples which pull from the same overall distribution of positive and unlabeled examples, with the goal of more effective identification of true positive versus true negative patients.

5.0 Interpreting Model Results

One of the main advantages of AI algorithms is the ability to detect patterns in big data invisible to the human eye from thousands of features generated without a priori hypotheses. However, increasing complexity of modeling approaches comes with reduced interpretability, rendering perception that many models are "black box." Interpretability of model decisions is critical in validating the model's efficacy, or in other words, building confidence that the model is thinking correctly about potentially undiagnosed patients. Several techniques can be employed to help achieve an understanding of model behavior, including predictor importance, relative risks, and SHAP.

A note here is that machine learning algorithms are—at their most basic level—geometric structures that live in multi-dimensional "feature space". Sometimes these structures admit low-dimensional representations that can be easily visualized, allowing one to elucidate the model's decision-making process with relative ease. Quite often, however, low-dimensional representations do not exist or are not readily available.

This latter situation is often the case for rare disease modelling, where patient outcomes depend on non-linear interactions between numerous features. In such cases, one may require a model that is irreducibly complicated and thus difficult to explain/visualize in one or two dimensions. That is to say that while the methods discussed here can provide clarity and confidence in a model's decisions and structure, they cannot necessarily eliminate certain facets of models that may remain too complicated to visualize and understand. Below we discuss the application of several techniques that intend to clarify and validate a model's predictions, with the ultimate goal of building confidence that a specific model is truly suitable for use in a real-world setting.

5.1 Predictor Importance

Predictor importance is typically the first step in understanding a model's decision making. The output ranks predictors by the amount each one contributes to a model's ability to identify potentially undiagnosed patients. This output presents a ranking of predictors, allowing for a straightforward initial glance into the model's processes (see Figure 4). Predictor importance, for decision-tree based models, can be effectively measured with a metric called gain. This metric calculates the relevance of a given predictor to each tree, and thus to each decision/split point, in the model. A higher gain value implies that the predictor is more important to the decision-making process.

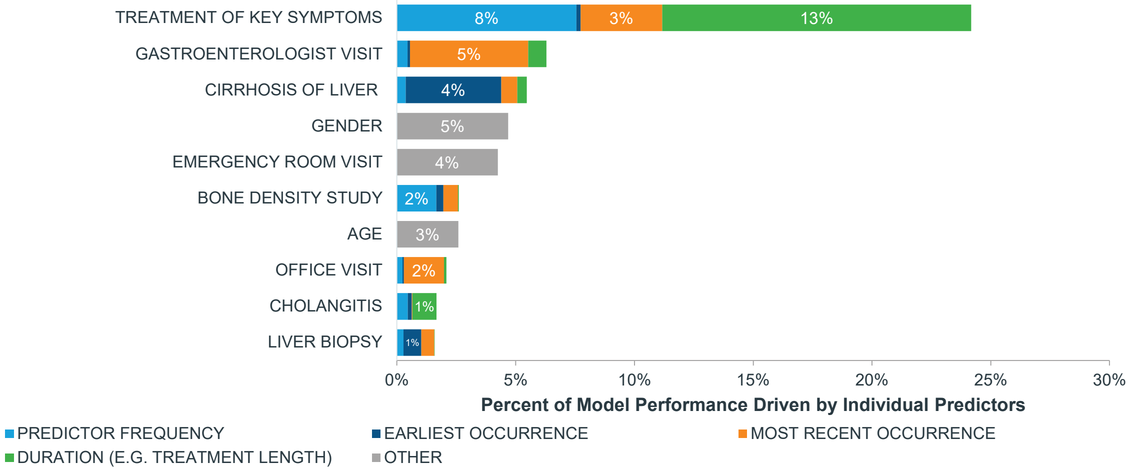
5.2 Evaluating the Magnitude and Direction of Predictor Importance

While predictor importance is valuable, it doesn't provide detail on the magnitude or direction (positive or negative correlation) for a given predictor. For these purposes, relative risk measurements are useful in more detailed quantification of a model's assessment of disease risk associated with specific aspects of each predictor. This metric allows for a calculation of magnitude, or strength, of the risk associated with a given predictor, but also the magnitude associated with a specific facet of the predictor (e.g. frequency, occurrence, timing).

For example, in examining a predictor of a specific disease, such as the occurrence of emergency room visits, relative risk can clarify (see Figure 5):

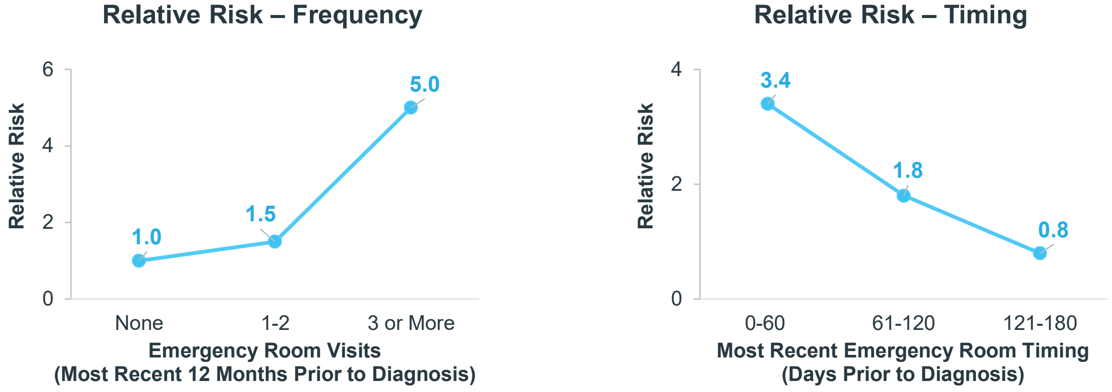
1. The risk associated with the occurrence of a visit
2. Risk associated with how frequently the event occurred prior to diagnosis
3. Risk associated with specific timing (typically focused on the first event and most recent event prior to diagnosis)

Figure 4: Assessment of Predictor Importance



Feature importance as measured by model gain and broken out by predictor type (e.g. frequency, timing, and other – gender/age). The gain measurement adds to one-hundred percent for all predictors included in a diagnostic model. The chart shown here is not exhaustive of all model features, but rather shows illustrative top ten predictors for a gastrointestinal disorder. *Source:* IQVIA case study.

Figure 5: Example of Relative Risk Measurements for Model Validation



Relative risk is defined as the increased risk of diagnosis associated with a specific predictor, such as the frequency and timing of emergency room visits. *Source:* IQVIA methodology; illustrative examples shown here.

In addition to associated risk, the directionality, or positive versus negative impact, of a predictor can be understood. Some predictors may show a simple trend in the positive versus negative direction, whereas others may fluctuate depending on the specific value associated with the predictor itself (e.g. frequency versus timing).

5.3 Additional Patient Level Analysis

To further evaluate patient-level predictions, a technique known as SHAP can determine, for a given patient or patient subtype (i.e. gender, age group, disease etiology or pathway), the specific set of predictors and quantified contribution of each predictor to a model risk score.[6] This method allows for model-driven profiling of individual patients or subgroups and helps clarify the complex and intricate ways in which an AI model derives its risk measurements.

6.0 Rare Disease Case Study – Putting the Methodology into Action

To illustrate the above methodology, a recent study focused on a rare hereditary disorder is summarized below. We describe the overall real-world process of leveraging the above mentioned methodological flow to build a model that seeks to identify potentially undiagnosed patients and provide clinical insight into the pathway to disease.

6.1 Study Background

Patients with the rare condition of interest often present with symptoms that resemble more common chronic illnesses. Due to the rarity of the disease, physicians are not familiar with the diagnosis, and thus it is not top of mind in most cases. These factors make it difficult for patients to be identified and diagnosed, often resulting in delays to proper diagnosis, incorrect treatment, and unnecessary surgical intervention. The goal of the study was to leverage a model to identify HCPs that would benefit from increased awareness of the disorder and understand the pathway to diagnosis, ultimately to accelerate time to diagnosis and appropriate management of the debilitating symptoms associated with the disease. Given that the goals of the study were combined clinical and commercial outreach endeavors, the team selected an open claims data set for the analysis.

6.2 Patient Selection (Positive and Negative Cohorts)

A challenge in this study was that both the ICD-9 and ICD-10 code are shared across multiple conditions. This required specific refinement of positive patients identified in the claims database. Treatments included medication with a label specifically indicated for the disease (disease-specific treatment) and medications that are used across several conditions (disease non-specific treatment). As such, patients were selected into the positive cohort if they had at least one claim for a disease specific treatment. Additionally, patients were selected into the positive cohort if they met the composite criteria of at least one claim for a shared ICD-9 or ICD-10 code as well as at least one claim for the disease non-specific treatment.

For generation of the negative cohort, patients were selected if they did not have any evidence of the disease specific treatment or the combination of diagnosis code and disease non-specific treatment mentioned above. Given the size of the negative cohort and the rarity of the disease (an estimated prevalence of ~1 in 30,000), an evaluation of the size of potentially unrecorded patients (false negatives) in the negative cohort concluded that the machine learning techniques utilized could address a miniscule level of noise expected, eliminating the concern of unknown positives.

6.3 Feature Generation and Model Training

A combination of a data-driven and hypothesis-driven approach was used to generate a comprehensive list of over 300 medical events considered as potential predictors in the model. To fully capture the richness and complexity in the data, metrics including the frequency, sequence and timing of events were generated for each predictor resulting in over 1200 total variables used by the model.

A gradient boosting tree model (XGBoost) was trained using the dataset described above. Model performance was evaluated using a PR curve projected to the prevalence of the disease. In testing, the model successfully identified patients at a precision of 23% at lower recall levels. Comparing this level of performance relative to examining patients at random for the disease, based on the estimated prevalence mentioned above, the model is shown to be highly effective in finding potentially undiagnosed patients (more effective than random by a factor of almost 7,000x). Predictor importance and relative risk analysis confirmed key medical factors in identifying potentially undiagnosed patients with the rare disease. Insights around the importance of the timing of these medical events and the impact on the likelihood that a patient is potentially undiagnosed evaluated in relative risk curves provided guidance on how to design outreach messaging focused upon accelerating diagnosis.

7.0 Conclusion

AI modeling for disease detection has ample opportunity to drive earlier diagnosis for patients in need, and in guiding pharmaceutical companies with highly advanced, targeted diagnostics to help these patients get properly diagnosed and treated earlier in their disease journey. As these algorithms expand in use, applications will widen, and include, for example, timed prediction of diagnosis (i.e. predict a diagnosis a certain amount of time in advance), on-going autonomous learning based on additional newly diagnosed patients (and to account for market changes), and incorporation into EHR systems to predict risk across not just one, but numerous disease states all at once. The use of these models, and advancement in the healthcare space, is undoubtedly valuable, but must be approached with the proper methodological inputs, business considerations, and statistical validation.

About the Authors

The authors of this publication represent the Predictive Analytics practice in IQVIA's Real-World Solutions global group. The team develops innovative solutions to solve challenging healthcare problems based on patient-level data using a variety of advanced statistical and machine learning methods. This development encompasses applications such as physician targeting and risk stratification algorithms aimed at, for example, finding undiagnosed patients or identifying patients suitable for treatment escalation. Our efforts help improve retrospective clinical studies, under-diagnosis of rare diseases, personalised treatment response profiles, disease progression predictions, and clinical decision-support tools. For questions or more information regarding the information in this article, please contact Goksu.Dogan@IQVIA.com or Nadejda.Leavitt@IQVIA.com.

References

1 Snomed ct & other terminologies, classifications & code systems. URL https://www.snomed. 314 org/snomed-ct/sct-worldwide

2 Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of 319 the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 320 KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 321 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785. 322

3 Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc 323 plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), 2015. doi: 324 10.1371/journal.pone.0118432

4 Jessa Bekker, Jesse Davis. Learning From Positive and Unlabeled Data: A Survey. https://arxiv.org/abs/1811.04820

5 Liu B, Lee WS, Yu PS, Li X. In: Proceedings of the Nineteenth International Conference on Machine Learning (ICML): 8-12 July 2002. Sammut, C., Hoffmann, A.G., editor. Vol. 2. Sydney, The University of New South Wales (UNSW); 2002. Partially supervised classification of text documents; pp.387-394

6 S. Lundberg, S. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

Empowering Clinical Decisions Using Machine Learning Prediction of Prognostic Biomarkers for Patient Disease Progression

Ariel Han, Engagement Manager, Commercial Effectiveness, Symphony Health Solutions; Vikram Singh, Principal, Commercial Effectiveness, Symphony Health Solutions; Rick Rosenthal, Senior Principal, Commercial Effectiveness, Symphony Health Solutions; Ewa J. Kleczyk, PhD, Vice President, Symphony Health Solutions

Abstract: As the demand for disease early intervention continues to grow, so does the understanding of precise disease prognosis. Capturing the key moments during the patient journey to properly manage the conditions becomes critical for diseases with multiple stages. In oncologic care, the success in identifying malignant progression and early treatment is highly associated with metastasis free survival and better quality of life. With the emergence of targeted therapies, biomarkers start to play an important role in cancer management.

In a recent case study of prostate cancer, we evaluated the probability of patients likely to observe malignant progression by analyzing events in the patient's history with the use of classification algorithms. This undertaking defines a malignant progression as an elevation of prostate-specific antigen (PSA), qualified on patients with prostate cancer diagnosis and previously treated with androgen deprivation therapy. The approach of predicting biomarker results using machine learning is able to provide physicians with insights on patient prognosis with a satisfying accuracy ahead of time. Through these early therapy interventions, the brand teams will now have many competitive advantages on their tactical and messaging plans.

Keywords: Machine Learning, Prostate Cancer, Metastatic Progression, Clinical Decision Support, Physician Targeting

Background

As the demand for early disease intervention continues to increase, we see a growing interest in predicting precise disease prognosis. Patients with breast cancer, multiple sclerosis, blood cancers or prostate cancer may show signs of progression characterized by symptoms, comorbidities, patient care events, and other diagnoses at various speeds. Patients with specific genetic mutations progress even faster. Identifying key moments during the patient journey and enabling optimal decisions in those moments becomes critical for diseases with multiple stages. In oncologic care, identifying malignancy progression and implementing early treatment are highly associated with metastasis-free survival and better quality of life.[1]

With the emergence of targeted therapies, biomarkers now play an important role in cancer management. Prognostic biomarkers such as Breast Cancer Gene 1 or 2 (BRCA1/2), Prostate Specific Antigen (PSA) for prostate cancer, and Chromosome 17p deletions/ TP53 mutations for chronic lymphocytic leukemia, provide valuable indicators regarding clinical outcomes, cancer recurrence and disease progression. As we further explore the utilization of biomarkers in cancer management, we come to realize that though clinically helpful throughout the patient journey, biomarker tests can be challenging to use routinely in chronic disease management due to high costs and risks associated with constant screening such as infections, stakeholder time and effort, and healthcare system resource constraints. Building accurate predictions of biomarker thresholds and applying them to early disease stages holds the potential to add tremendous value to clinical decision support.

This article presents a case study demonstrating the use of a machine learning model to predict lab results, and its value in clinical support and long-term disease management. It examines the application of a biomarker called PSA (Prostate Specific Antigen), and its clinical significance in predicting time to metastasis for prostate cancer patients. We develop a mechanism to further strengthen the prediction by adding multiple dimensions from other healthcare data sources, and to replicate PSA's predictive power in an environment where PSA is not available to further evaluate prostate cancer progression. This is achieved by analyzing patients' history with the use of classification algorithms.

Case Study Background

Every 17 minutes another American man dies from prostate cancer. Currently there are nearly 3.1 million American men living with the disease – roughly equal to the population of Chicago. Fortunately, high survival rates for prostate cancer continue over time thanks to early intervention and advances in treatment. The overall 10-year survival rate is 98%, and the 15-year survival rate is 96%. However, once the cancer metastasizes to bones, organs, or distant lymph nodes, the 5-year survival rate drops from nearly 100% to 30%.[2] Metastasis to the bone is the main cause of death.[2]

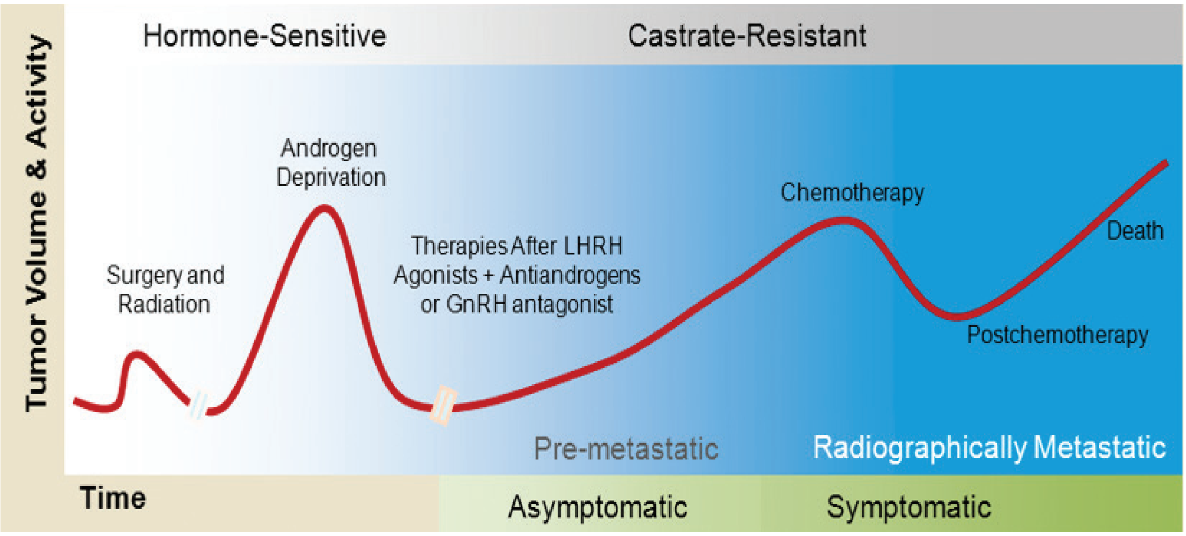
While early treatment with surgery or radiation has achieved success in preventing patients from ultimately dying from prostate cancer, research has shown that the treatment is also associated with significant morbidity such as urinary incontinence, erectile dysfunction, and related infections.[3] As a result, many patients are willing to adopt an active surveillance strategy[4], in which doctors closely monitor cancer progression and consider treatment only when the tumor appears to show signs of growing and spreading. For some patients it takes up to 15 years for the cancer to metastasize. Within the past five years, fewer than 10% of prostate cancer patients were assessed to be in the advanced stage where the cancer spread beyond the prostate gland at the time of initial diagnosis. This may reflect the growing and effective use of PSA testing in early screening.

One of the biggest challenges in prostate cancer management is deciding which patients have clinically important tumors, and intervening early enough to prevent malignant progression. This issue is relevant not only in patients with newly diagnosed tumors, but also in relapsed patients after primary treatment. One method to improve ability to predict prognosis is utilizing PSA doubling time (PSADT), the number of months it takes for PSA to increase two-fold, calculated as:[5]

$$PSADT = \frac{\ln 2 * (date_{test1} - date_{test2})}{\ln \frac{PSA_{test1}}{PSA_{test2}}}$$

Many prostate cancer studies have been trying to examine PSADT as a predictor of cancer prognosis. Studies have shown that patients may have very different disease profiling pre and post radical prostatectomy. Primary clinical studies indicate that among patients treated with radical prostatectomy and ongoing androgen deprivation treatment (ADT), PSADT appears to have greater utility in predicting clinical and systemic progression. After the initial operation, patients with PSADT less than 6 months have significantly higher risk of experiencing metastasis compared to those with PSADT longer than 6 months (Figure 1).[6]

Figure 1: Illustration of Prostate Cancer Progression[7]



However, while playing an important indicating role in metastatic progression, the test itself has limitations:[8]

- False positive results: a prostate cancer patient's PSA level is elevated but the cancer is not actually progressing. Benign conditions such as prostatitis or benign prostatic hyperplasia (BPH) can cause PSA levels to rise.
- False negative results: a prostate cancer patient's PSA level is low (either lower than a previous test, or within a normal range) even though the tumor has been flaring up with corresponding symptoms.

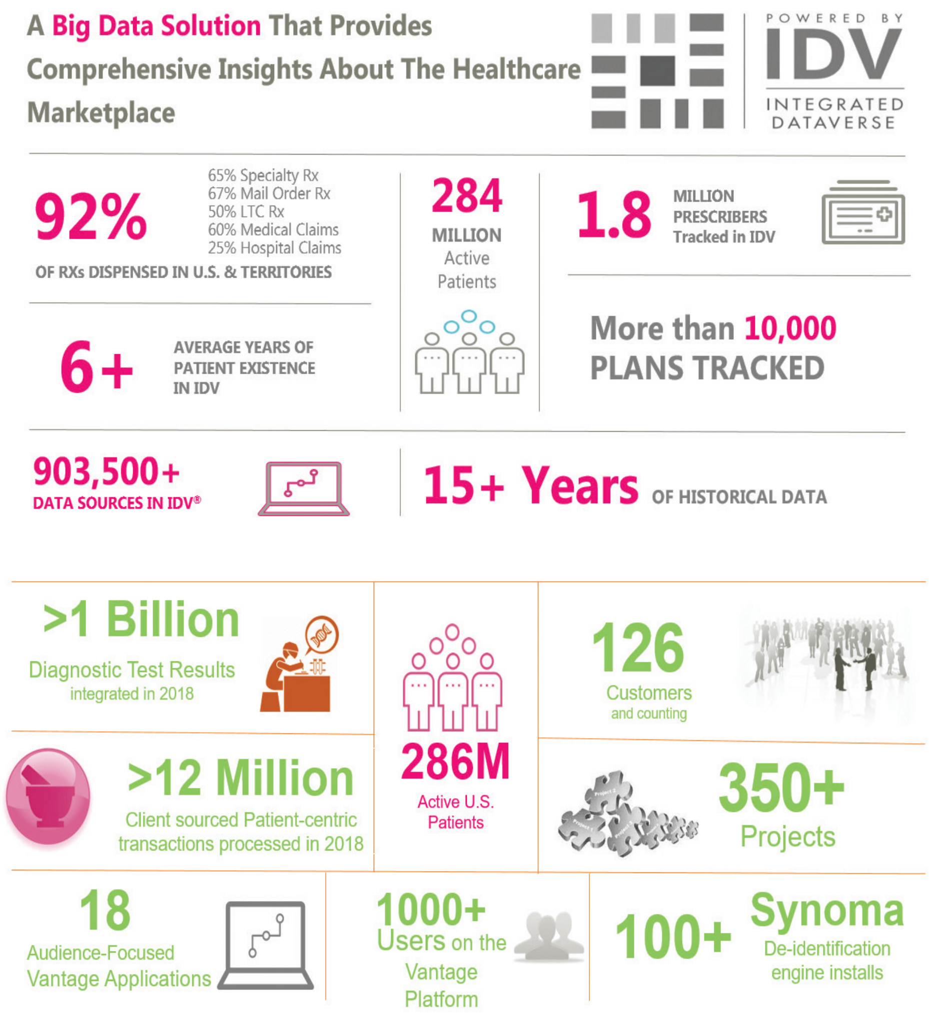
Dataset Overview

In full awareness of the predominant application of PSADT in predicting prostate cancer metastatic progression, our Team was tasked with utilizing a Diagnostic Test dataset to access de-identified PSA test results and incorporate into administrative claims data at the individual patient level.

Claims data usually provides information on the drugs dispensed by pharmacies, procedures performed, diagnoses at office visits, and health insurance plan and copay. Companies that collect these data often have their proprietary algorithms to preserve and merge datasets overtime based on known and universal patient variables such as name, address, and date of birth. As a result, it provides a longitudinal and holistic view of in-patient and out-patient care.[9]

The longitudinal patient database that the Team uses is representative of the US population based on age, gender, and insurance type. The sample is based upon healthcare claims and, through common de-identification process, links patients across time and healthcare settings such as pharmacies, clinics/office practices, and hospitals. This approach to patient linking offers the ability to track patients year-over-year regardless of health insurance or other demographic changes. Consider in 1 year, 30% of patients change payers at the "national" level, 11% of patients potentially change names (e.g., marriage/divorce), and 15% of patients change pharmacies. This patient-linking methodology and stringent requirements in the longitudinal patient database mitigates the loss and miss-assignment of patients.

Figure 2: Overview of Claims and Lab Data

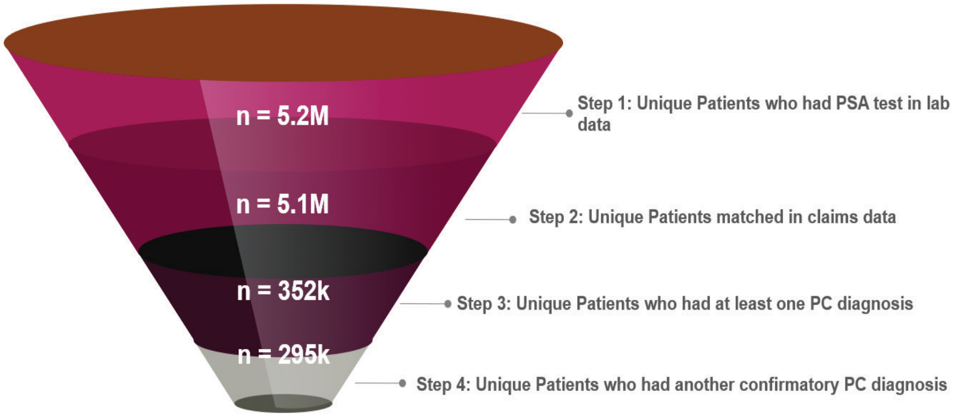

A Big Data Solution That Provides Comprehensive Insights About The Healthcare Marketplace
POWERED BY IDV INTEGRATED DATAVERSE
92%
65% Specialty Rx
67% Mail Order Rx
50% LTC Rx
60% Medical Claims
25% Hospital Claims
OF RXs DISPENSED IN U.S. & TERRITORIES
284 MILLION Active Patients
1.8 MILLION PRESCRIBERS Tracked in IDV
6+ AVERAGE YEARS OF PATIENT EXISTENCE IN IDV
More than 10,000 PLANS TRACKED
903,500+ DATA SOURCES IN IDV®
15+ Years OF HISTORICAL DATA
>1 Billion Diagnostic Test Results integrated in 2018
286M Active U.S. Patients
126 Customers and counting
>12 Million Client sourced Patient-centric transactions processed in 2018
350+ Projects
18 Audience-Focused Vantage Applications
1000+ Users on the Vantage Platform
100+ Synoma De-identification engine installs

Although the Team believes that there are many benefits to using claims data in patient level analysis, the dataset itself might lack complete views of full patient history due to the differential rate of capture across vendors and claims types. Hospital claims usually have a low coverage rate for rendered services and sometimes include inconsistent reporting formats. This impacts the ability to track in-patient care and link other data sources to out-patient treatment after hospitalization. A low capture rate for infused and injectable drug procedures also results in small sample sizes, especially for rare and orphaned diseases.

Based on the Team's knowledge and experience, while the claims data has visibility on lab tests status (ordered and/or executed), it typically does not include the actual test results, which makes tracking disease prognosis challenging. In oncology and immunology, lab results serve as an important indicator that impacts the treatment pathways with great heterogeneity.

To bridge the gap for this initiative, the Team supplements the claims data with PSA laboratory results, and EMR data. For each patient, the Team can access PSA test values & date, claims data (treatment, diagnosis, and

Figure 3: Patient Volume Waterfall



procedure history), and physician information at each interaction point (specialty, place of service) to form a comprehensive patient view of not only prostate cancer but also overall individual healthcare journey. Both claims and test datasets cover around 280 million US patients and the match rate between these two datasets is around 90%.

Figure 3 shows the sample size and match rate between lab data and claims data. In this market, the Team is able to match between lab data and claims data up to 98%. However, not all patents with PSA test are diagnosed with prostate cancer as this is a quite standard and routine monitoring test. The drop from step 2 to step 3 indicates that only 7% of patients with PSA test had a prostate cancer diagnosis. Per research from Cancer.org, 1 out of 9 (11%) men have prostate cancer at some point in their life.[10] This sample is in line with the incidence rate in this market.

The goal is to use classification algorithms to build an accurate prediction mechanism, using a sample group of patients with known PSA values and in-depth claims data. This predictive model was designed to overcome PSA accuracy challenges. It extrapolates from a small sample of known PSA value patients to a broader patient population where PSA values are absent. The model predicts the probability that patients will experience metastasis within a 90-day of treatment window with newer anti-androgens such as enzalutamide, apalutamide and darolutamide – FDA-approved medications that significantly prolong the time to metastasis.[11]

Methodology Overview

The following steps are taken to execute the machine learning analysis:

1. Identify the patient cohort
2. Reduce variable dimension
3. Select the 'Best' model(s)
4. Validate the selected model(s)
5. Create ensembles of 'Best' model(s)

1. Identify the Patient Cohort

Using the in-depth dataset built for this patient universe, the Team's next step involves setting up patient cohorts for the study: a target group and a control group.

As presented in Figure 4, the target group is defined as high-risk patients whose PSADT <= 10 months. Patients have been diagnosed with prostate cancer for at least one year, have gone through at least one of the defined treatment procedures (surgery and/or radiation therapy), and are currently still on androgen deprivation

Figure 4: Patient Cohort of the Study During Prostate Cancer Progression Journey[7]

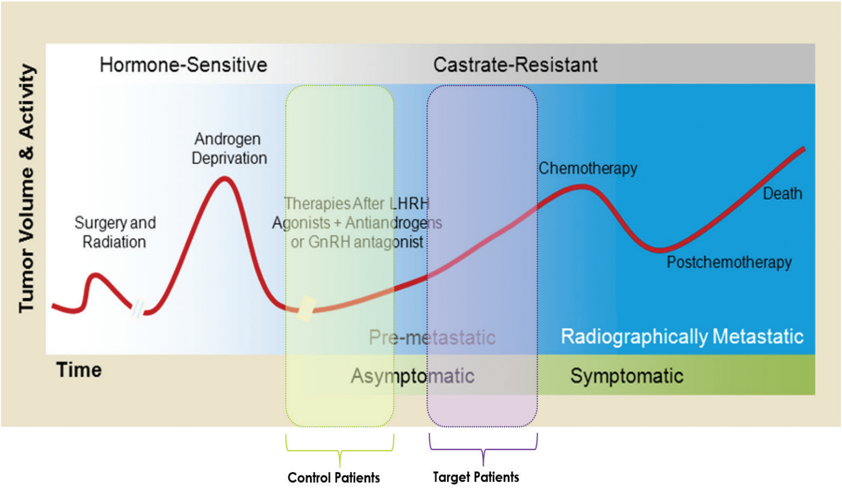


Table 1: Raw Features Category

Raw Variable	Examples	Source
Demography	Age, Ethnicity	Outpatient Office Visit Claims
Patient History	Family history of PC; personal history of other cancers	Outpatient Office Visit Claims
PC Treatment	Anti-androgens, androgen deprivation therapies, radiation therapy, brachytherapy, cryotherapy and prostatectomy	Outpatient Office Visit Claims
Symptoms	tumor flare reactions, prostate cancer related	Outpatient Office Visit Claims
PSA values	date of the test, frequency, actual values	Lab data
Other Diagnostic Tests	CT scan, x-rays, blood tests	Outpatient Procedure Claims
Comorbidities	type 2 diabetes, hypertension, constitutional comorbidities	Outpatient Office Visit Claims
Physician Attributes	specialty (urologist, oncologist), hospital setting	NPI/AMA information

therapy such as luteinizing hormone-releasing hormone (LHRH) agonists and gonadotropin-releasing hormone (GnRH) antagonists. The target group patients have not been diagnosed with any metastatic features, or been on any chemotherapy. The control group is defined as patients who met the requirements from the target group and also had at least two consecutive PSA tests, with PSADT outside the high-risk range. Test and control patient volume ratio is around 1:4.

2. Reduce Variable Dimension

To fully understand the differences between target and control groups, the Team captures around 8000 features and group into categories (see Table 1).

Besides these categories, other similar studies[12] also mention features such as hemoglobin, alkaline phosphatase, alanine transaminase, blood urea nitrogen, creatinine, and prothrombin time. While there are some differences between test and control groups for variables such as prothrombin time and hemoglobin, alkaline phosphatase and alanine transaminase don’t seem to present with significance in our models. This could be explained by the fact that these two substances measure liver functions and will rise to an

Figure 5: ROC Curve

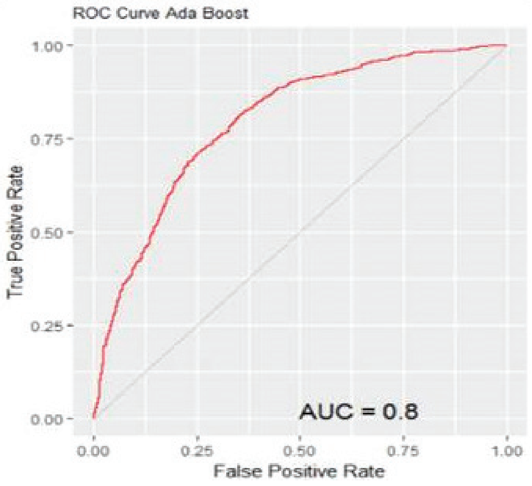


Figure 6: Confusion Matrix

n=2,085	Predicted: No	Predicted: Yes	
Actual: No	TN = 792	FP = 254	1,046
Actual: Yes	FN = 318	TP = 721	1,039
	1,110	975	

abnormal level when prostate cancer has spread out to liver and kidney, which will be an important indicator for post metastasis study.

A couple of dimension reduction methods are run at the same time, such as principal component analysis (PCA), Correlation matrix, and Multi-dimensional scaling (MDS) to validate the level of collinearity and dimensionality within the data. Highly correlated variables are also removed. The goal is to look for consistency of results and make decisions on model variable inclusion. The primary focus is not only just the statistical part of the model interaction but also the practical and business side.

3. Select the Models

The Team explores and employs a number of machine learning models including tree algorithms, Support Vector Machine (SVM), and artificial neural network. The dataset is split in a 70/30 ratio to establish training and validation samples. Further ensemble on tree model from bagging to boosting is also utilized to optimize the modeling approach.

Each model is trained with both a balanced and unbalanced test and control sample.

Two important terms are used to describe model results below: Receiver Operating Characteristic (ROC) curve and confusion matrix.

- A classification model ROC Curve is a plot of the true positive rate against the false positive rate (Figure 5). ROC indicates how well the model can distinguish between the target and control groups (in this case, whether a patient will experience a malignant progression within the next 90 days or not). The closer the ROC curve approaches the upper left corner, the higher the overall accuracy. Figure 5 came from one of the models selected, Adaptive Boosting (ADA Boost). The Area Under the Curve (AUC) measures how well the model performs at predicting the target and control patients.
- A confusion matrix is a table describing classification model performance for a set of control data in which the true values are known (Figure 6). Within the confusion matrix, the team typically evaluates two metrics:
    - Precision: Answers the question “when the model predicts that a

patient is in the target group, how often is it correct?". Precision is calculated as TP / (TP + FP) = 74%, where TP = True Positive, and FP = False Positive.
- Recall: Answers the question "when a patient is in the target group, how often does the model correctly place them in the target group?". It's calculated as TP / (TP + FN) = 70%.

These models end up with strong predictive capabilities compared to other classifiers:

- Tree model ensemble:
  - Bagging trees via Random Forest (RF): In RF, each decision node uses the best among a subset of predictors randomly chosen at that node. This method has been able to resist over-fitting. RF is fairly stable when new dataset is introduced.[13]
  - Boosting methods:
    - ADA Boost: creates a highly accurate classifier by combining many relatively weak and less accurate classifiers.[14] In this setting, ADA Boost has the highest precision across each sample group, making the prediction less sensitive to sample over or under representation.
    - Extreme Gradient Boosting (XGBoost): training is very fast and performs well in unbalanced sample.
  - SVM: works relatively well with a high dimensional data, especially with features at diagnosis, treatment and procedure level. However, because training takes a long time, this classifier is more difficult to tweak and test. It is also sensitive to outliers.
  - Deep learning in Artificial Neural Network: this classifier requires large data sample, which is currently not supported by the augmented claims dataset.

The models returned additional important variables. These included features (diagnoses, treatments, procedures, etc.) contributing the most to predicting malignant progression for a patient:

- Treatment history including hormone therapies such as bicalutamide, GRH analogs such as leuprolide and goserelin; anti-depressants (possibly treating androgen deprivation therapy side effects), steroids, radiotherapy, and overactive bladder treatment.
- Comorbidities such as hypertension, anemia, and benign prostatic hyperplasia.
- Diagnostic tests such as ultrasound and CT scan.

The Team further analyzes the distributions for these features within target and control groups across different time periods: all history, the average of the previous few quarters, the quarter prior to PSADT vs. the average of prior quarters (most recent quarter as a separate variable to see if there was any difference). These "time to event" features are added back to the model as an additional boosting feature to the existing significant variables. Generic product variables are manually removed if they have high importance value.

Figure 7 illustrates the percentage of predictors explained by each category.

4. Validate the Models

In the validation phase, the Team creates a "real world" dataset with a new set of prostate cancer patients outside of the modeling time period and an extended look-forward period of 90 days. Using this longitudinal data, the Team

Figure 7: Predictor Constancy

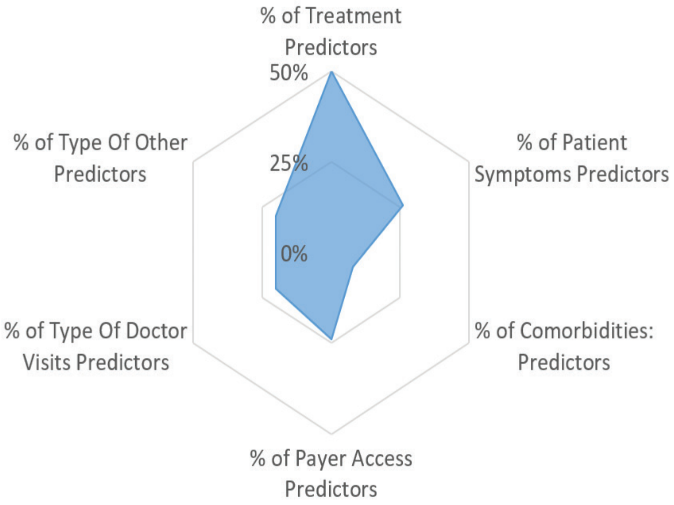
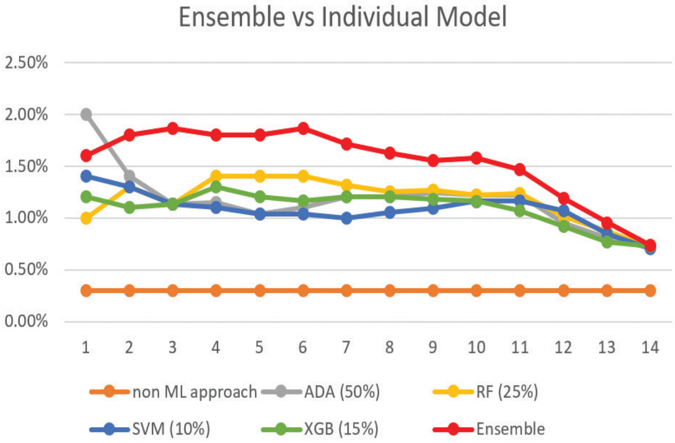


Figure 8: Final Model Performance in Real World Data



is able to see whether the patients experienced malignant progression (such as PSADT<10, initiated chemotherapy, secondary malignancy, etc.). Final ensemble model is applied to this dataset to find the number of true positive patients who qualify for the target group and are also correctly predicted by the model, and calculated precision.

5. Create Ensembles of 'Best' Model(s)

An ensemble of RF, ADA Boost, SVM and XGBoost is created that maximizes precision, by probability score at desirable patient volume (Figure 8). Each model is sensitive to the small prediction size differently as sample might over or under represent certain important variables. The precision curve gets smoother once the patient size goes up. The final ensemble (red dotted line) maximized the precision at each given patient volume point and outperformed each individual model.

The orange dotted line represents precision from a non-machine learning technique to identify patients, such as High Value Targeting (HVT). HVT algorithm is applied where patients are evaluated by a set of variables with a pre-determined threshold such as frequency of office visit, frequency of PSA tests, symptoms prior to metastasis, etc. to identify a list of patients who might fit into the target patient profile. Unlike machine learning that scores each patient based on a combination of conditions, HVT results turn out to be far less accurate. For example, when scoring 3000 patients, ML's precision is 5 times higher than HVT within first 90 days look-forward. In other words, to capture the same number of target patients, HVT needs to reach to a population that is 5 times bigger than ML.

Reaching the Right Physicians at the Right Time

Based on the model propensity scores, the Team is able to generate a list of de-identified patients who are highly likely to have malignant progression within the next 90 days. This type of list can facilitate outreach to the right physicians (those managing their care) to alert them to the risk of disease progression, and the potential need to consider next line treatment options. Prostate cancer at a non-threatening stage is typically treated by urologists up until a point where the tumor metastasizes. Should the tumor metastasize, patients would then be referred to oncologists. At the same time, based on our field research, patients may seek additional opinions from multiple urologists or oncologists.

Reaching out to all patient-linked physicians is not cost effective, as some may no longer see these patients or be integral to their prostate cancer care.

Timing is also crucial. Reaching out to physicians after their patients become metastatic may be too late. However, getting contact too early is not ideal either because patients might not experience or show any symptoms of progression, thus the information loses relevance and value in the treatment decision process.

To overcome these challenges, the Team establishes a set of attribution rules to look at the most recent and most frequently-visited physicians among each patient's oncologists and urologists. When a patient had multiple urologist and oncologist touch points in recent months, our rule was able to distinguish the noise (second opinions and non-current treaters), the outdated (patients who transitioned to another doctor within the same specialty), and progression transition (urologists who had transferred patients to oncologists).

With a list of accurate physicians that patients are currently seeing, and a 90-day window to influence the decision, the machine learning model output is now driving valuable clinical support.

Business Applications

Predicting biomarker results using machine learning shows promise as a tool to assist in both clinical and commercial settings.

- Clinical decision support: besides providing physicians with insights on patient prognosis with satisfying accuracy, this approach is also beneficial for alternative treatment considerations. When prostate cancer patients are no longer responsive to the existing therapies, it might be a good opportunity to enroll in clinical trials before metastasis, which might be a very short time window.
- Commercial applications:
  - Quickly finding patients who are appropriate candidates for therapy has become a desirable goal for life sciences marketers. With earlier identification of appropriate patients, we believe the marketers will have the opportunity to offer stakeholders highly relevant information, education, and support resources precisely at the moment they're most needed.

- With omni-channel marketing widely used, “next best action” has been getting more and more attention as we continue to refine commercial strategies. How the pharmaceutical companies optimize the strategy to properly guide investment for high value physicians in the market has been frequently asked by the brand teams. With properly attributing eligible patients to the true target, the brand team can further customize samples, calls, and other promotional tactics.
- With a 90-day window built in for prognosis prediction, we believe the brand team can now schedule the office visit at a timely manner. Frequency will also be adjusted if physicians currently don’t have any eligible patients and resources can be reallocated.

Next Steps

To refine the model, the Team is planning on taking the following actions as next steps:

- Explore newer advanced algorithms, such as deep learning and neural networks, to further improve precision. Extending the coverage on lab data will also increase the modeling sample size. The Team is working with multiple lab data providers to evaluate the additional lab data for the prostate cancer markets.
- The Team is also considering integrating consumer information into the dataset. With new dimensions added such as occupation, income, education, lifestyle, hobbies, magazine, and/or TV channel subscriptions, and buying behaviors, the Team will be able to further improve predictions. For brand teams, understanding the mosaic of actionable patient segments will help guide direct-to-consumer campaigns to optimal impact. Consumer data may also hold signals that reveal opportunities to communicate with caregivers, such as the spouses and other relatives who often play an important role helping prostate cancer patients with medication, appointments and activities contributing to their quality of life.

The Team has already received positive physician feedback regarding timing and accuracy of the model. Next steps will include measuring real-world effectiveness, and continuing to improve model performance. Our hope is that through improved predictions regarding precise disease prognosis, the Team will help more physicians and patients experience the benefits of early interventions, and that metastasis-free survival and quality of life for prostate cancer patients will continue to improve.

About the Authors

Ariel Han *is an Engagement Manager of Commercial Effectiveness with Symphony Health. She brings 8 years of experience in analytical consulting for pharmaceutical companies. Previously Ariel worked at IQVIA where she led multiple high-profile client engagements in marketing campaign optimization and measurement, TV targeting, and sales force effectiveness. After joining Symphony Health, Ariel has been focusing on machine learning and complex patient journey in the rare disease and oncology areas. Ariel holds a Master of Management in Clinical Informatics degree from Fuqua School of Business, Duke University.*

Vikram Singh *is a Principal of Commercial Effectiveness with Symphony Health. He has more than 15 years of experience leveraging customer insights and big data advanced analytics to help organizations make better sales and marketing decisions. Vikram holds a Master degree in Information Technology from Indian Institute of Technology.*

Rick Rosenthal*, Senior Principal, heads Symphony Health's Commercial Effectiveness practice. In over 25 years in the life sciences industry, he has progressed through sales and marketing leadership roles at Johnson & Johnson, and commercial consulting leadership roles at Cognizant Technology Solutions and Booz Allen Hamilton. Rick is an honors graduate in Economics from Tufts University.*

Ewa J. Kleczyk*, PhD, is a Vice President of Client Analytics with Symphony Health Solutions and an Affiliated Graduate Faculty at the University of Maine. She has expertise in the primary and secondary data analytics, including targeting and compensation, managed market analytics, brand analytics, patient journey, forecasting, personal and non-personal promotional response, etc. She has given presentations in the areas of healthcare at several industry conferences, including DTC National, PMSA, PMRG, and PBIRG. She has published in several journals, including PM360, Journal of Medical Marketing, and PMSA Journal. She has written chapters for multiple books, and has been featured on live radio shows. Ewa has a Doctorate Degree in Economics from Virginia Tech.*

References

1 Smith MR, Mehra M, Nair S, Lawson J, Small EJ. Relationship Between Metastasis-free Survival and Overall Survival in Patients With Nonmetastatic Castration-resistant Prostate Cancer. *Clinical Genitourinary Cancer* [Internet]. 2020 Apr 1;18(2):e180–9. Available from: https://doi.org/10.1016/j.clgc.2019.10.030

2 Facts and statistics [online] 2019 [cites 27 February 2019 available from: https://zerocancer.org/learn/about-prostate-cancer/facts-statistics/

3 Mirza M, Griebling TL, Kazer MW. Erectile Dysfunction and Urinary Incontinence after Prostate Cancer Treatment. *Semin Oncol Nurs* [Internet]. 2011;27(4):278–89. Available from: http://www.sciencedirect.com/science/article/pii/S0749208111000702

4 Pendick D Harvard Health Blog [online] 2013;[cited 27 October 2019 available from: https://www.health.harvard.edu/blog/prostate-cancer-lives-as-it-is-born-slow-growing-and-benign-or-fast-growing-and-dangerous-201308146604

5 Pound CR, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural History of Progression After PSA Elevation Following Radical Prostatectomy. *JAMA* [Internet]. 1999 May 5;281(17):1591–7. Available from: https://doi.org/10.1001/jama.281.17.1591

6 Wieder JA, Belldegrun AS. The Utility of PSA Doubling Time to Monitor Prostate Cancer Recurrence. Mayo Clinic Proceedings [Internet]. 2001 Jun 1;76(6):571–2. Available from: https://doi.org/10.4065/76.6.571

7 Beekharry R. Exploration of a xenograft model of human prostate cancer to predict patient treatment response. (2014). Thesis [online] available from: https://hydra.hull.ac.uk/resources/hull:12623

8 National Cancer Institute Prostate-Specific Antigen (PSA) Test [online] 2017; [cited 27 October 2019] available at https://www.cancer.gov/types/prostate/psa-fact-sheet

9 Kleczyk E, Evans D Leveraging Predictive Analytics to Derive Patient Adherence Drivers. *Journal of the Pharmaceutical Management Science Association*. Spring 2017 p 33-34

10 American Cancer Society Key Statistics for Prostate cancer [online] 2019; [cited 27 October 2019] available at https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html

11 Gomella LG, Singh J, Lallas C, Trabulsi EJ. Hormone therapy in the management of prostate cancer: evidence-based approaches. *Ther Adv Urol*. 2010 Aug;2(4):171-81. Available from: https://journals.sagepub.com/doi/10.1177/1756287210375270

12 SU P-J, Fang Y-A, Chang Y-C, Kuo Y-C, Lin Y-C. Establish a predictive model for high-risk de novo metastatic prostate cancer patients by machine learning. *Journal of Global Oncology* [Internet]. 2019 Oct 7;5(suppl):13. Available from: https://doi.org/10.1200/JGO.2019.5.suppl.13

13 Li W, Kockelman K [online].2019;[cited 27 October 2019] available from: http://docplayer.net/156712279-How-does-machine-learning-compare-to-conventional-econometrics-for-transport-data-sets-a-test-of-ml-vs-mle.html

14 Naresh K The Professionals Point [online].2019;[cited 27 October 2019] available from: http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html

A Novel Interpretable Machine Learning Approach as a Commercial Decision Support Tool

Avgoustinos Filippoupolitis PhD; Michael Kusnetsov PhD; Nicola Lazzarini PhD; Hariklia Eleftherohorinou PhD—Machine Learning & Artificial Intelligence Solutions Global Unit, Real World Solutions, IQVIA

Abstract: Interpretability of a machine learning (ML) model is of high importance, as it enables the users to understand which features contribute to a prediction. As the ML model is no longer seen as a 'black-box', interpretability promotes trust and provides actionable insights into the model's outputs. In this work we present an interpretability approach that goes beyond global feature contribution, and allows the attribution of the relative importance of ML drivers to individual predictions and to population sub-groups. We demonstrate the results of our approach on an ML model based on Gradient Boosting Trees, trained to classify Heart Failure with Preserved Ejection Fraction (HFpEF) patients. We further demonstrate how our approach enables the identification of sub-cohorts for which a feature is important although its global relative importance is low, allowing to identify high-value market segments.

Keywords: Model Interpretability, SHAP Values, Feature Importance, Real World Data, Gradient Boosting Trees

1. Background

Machine Learning (ML) applications show strong potential as commercial decision support tools, with an increasing body of literature demonstrating ML outperforming traditional commercial analytics. As ML adoption in commercial applications increases, interpretation of the results and understanding the patterns identified by the ML model is of paramount importance to translate ML outputs to actions with both improved patient outcomes and commercial impact. That said, most ML interpretation approaches focus on calculating average relative importance of the ML drivers across an entire population and fall short in interpreting targeted sub-populations of high value. This is especially impactful in rare diseases and specialty brands, where the populations are small and diverse, at times hard to specify with medical codes, making it challenging for launch and commercial teams to understand the ML inputs/outputs and develop effective targeted strategies, often resulting in overlooking high value market segments and missed commercial opportunities.

2. **Objectives**

The novel approach presented in this paper takes traditional ML feature attribution techniques one step further and acts as interpretable ML in healthcare that allows the attribution of the relative importance of ML drivers to specific population sub-groups. It builds on previously published work of successfully using ML to identify commercially viable patient populations and now helps understand the patterns of ML to translate them to actions. We compare results with traditional approaches for calculating the global relative importance of ML drivers and we discuss the benefits of our approach in both flexibility and interpretability. As annual health care marketing spending increased from \$17.7 billion in 1997 to \$29.9 billion in 2016, with

Table 1. Characteristics of Dataset

	Patients with HFpEF	Patients without HFpEF
Age (mean)	69.74	64.57
Age (std)	8.46	9.32
Gender (% of male)	45.79%	45.84%
Gender (% of female)	54.21%	54.16%
Count	1,646,563	16,465,630

direct-to-consumer advertising for prescription drugs increasing from $2.1 billion to $9.6 billion during the same period[1] , our novel decision support tool for sub-population identification and ML interpretation can have a strong impact on optimizing resource allocation and increasing revenue for pharmaceutical companies.

3. Data

We employed patient-level data that were extracted from transactional IQVIA US prescription and medical claims between 2010 and 2019. Prescription data are received from pharmacies, and contain information such as product, provider, age, gender, and date of service. Medical claims data are derived from office-based professionals, ambulatory and general health care sites, and include diagnosis and procedure information. The size of the dataset was 18.1 million patients, which is more than 400 times larger than the next largest dataset reported in the literature for Heart Failure patient classification.[2]

4. Methods

In order to construct a dataset for supervised learning, patients diagnosed with Heart Failure with Preserved Ejection Fraction (HFpEF) in the period between 2015 and 2019, are defined as positive, while non-diagnosed are considered negative. We demonstrate results on a binary classifier based on Gradient Boosting Trees[3] for diagnosing HFpEF. This is a complex clinical condition, which is manifested by signs of heart failure, left ventricular diastolic dysfunction, and by a preserved left ventricular systolic function.[4] The predicted percentage of hospitalized heart failure US patients that will have HFpEF by 2020 is 50%.[5] We used features capturing information on demographics, treatments, procedures and symptomatology, including temporal associations between the timing of events. These features were selected based on clinical expert opinion, as potential risk factors for HFpEF. After applying a 1% prevalence filter, the total number of features was 98. Table 1 illustrates the characteristics of our dataset, along with the class ratio. Identification of the best model parameters has been realized using a Bayesian optimization approach.[6] In particular, we use a Tree-Structured Parzen Estimator (TPE) algorithm for hyperparameter space exploration. Traditionally, hyper-parameter selection is based on grid-search, an exhaustive search of a specified subset of hyper-parameter values. Instead of this, the Bayesian optimiser iteratively evaluates subsets of values and automatically identifies the direction towards moving to improve the results. The TPE algorithm has been shown to outperform both grid-search and random search over the configuration space of hyper-parameters. However, the performance of the Bayesian optimisation approach depends on the probability distributions that define the domain of hyper-parameters over which to search.

Figure 1: ROC Curve of the Predictive Model and the Area Under the Curve Value (AUC)

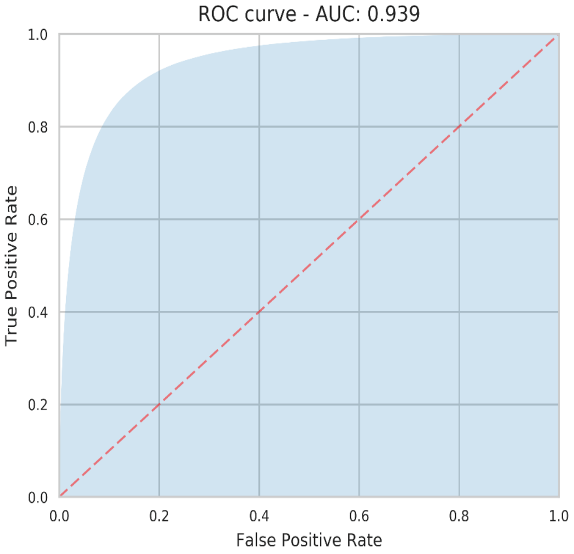
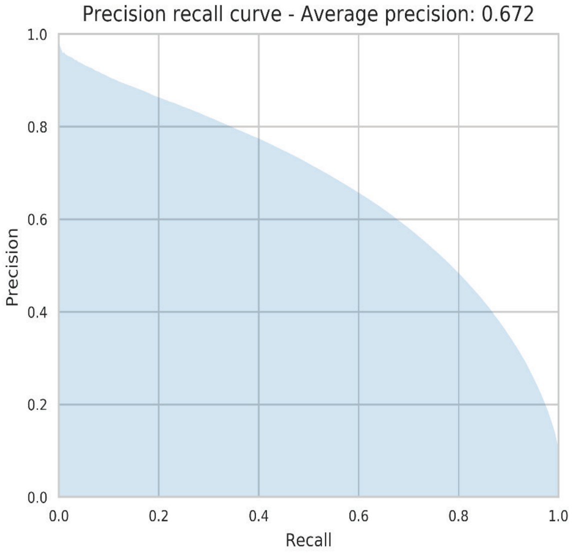


Figure 2: Precision-Recall Curve of the Predictive Model and the Average Precision of the Model



We identified and analyzed key drivers of the trained model using SHapley Additive exPlanations (SHAP) values – a cutting-edge interpretability approach that is based on recent applications of game theory.[7] SHAP values describe how each feature used for modeling contributes to any prediction made by the model. The approach is model-agnostic but is optimized for tree-based models such as Gradient Boosting Trees. SHAP values have two significant advantages over other existing interpretability methodologies. Firstly, it is the only methodology to have rigorous theoretical underpinning. Secondly, it enables a much wider suite of analytic and visualization techniques as we show below.

5. Results

Figure 1 illustrates the Receiver Operator Characteristic (ROC) curve, where we can confirm that our approach performed well in identifying HFpEF patients, with an Area Under the Curve (AUC) value of 0.939. As our dataset is unbalanced with a class ratio of 1:10 between positive (HFpEF) and negative (non-HFpEF) patients (which can be confirmed by Table 1), the Precision – Recall curve is a more suitable metric as it is robust to class imbalance.[8] Figure 2 depicts the Precision Recall curve for our model, where we can observe that the Average Precision Score is 0.672. In particular, the ML model achieved 91% and 86% precision at 10% and 20% recall respectively, in identifying HFpEF patients and their sub-populations, improving by more than 20% on the performance reported in the literature.[9]

To illustrate the utility of our novel interpretability approach, we first apply it to demonstrate the global feature contribution significance variables for the entire population, as illustrated in Figure 3, which illustrates the top fifteen features contributing to the prediction of HFpEF. The value of each feature is the mean absolute SHAP value for each of the features in the test set. This is compatible to the insights produced by traditional ML interpretation approaches, that focus on calculating average relative importance of

Figure 3: The Top 15 Features Contributing to the Predictions; The X-Axis Shows Mean Absolute SHAP Values for Each Feature in the Test Set

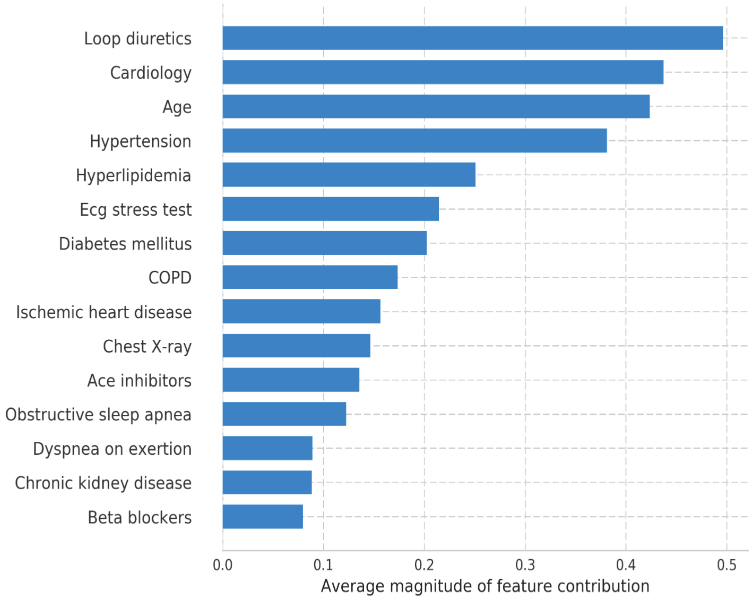
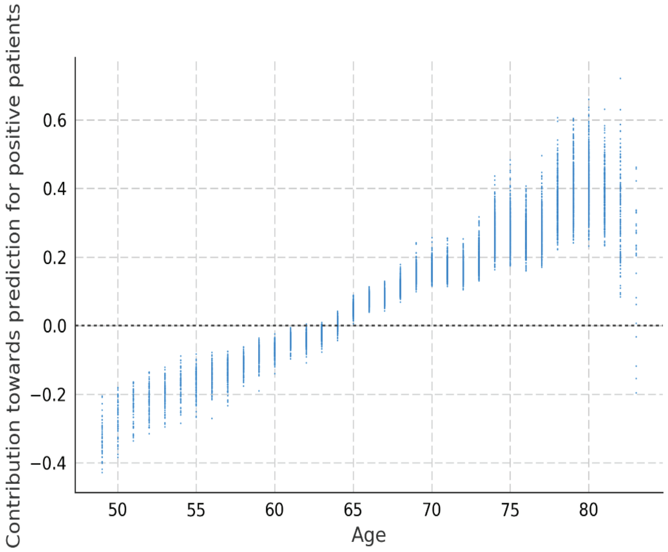


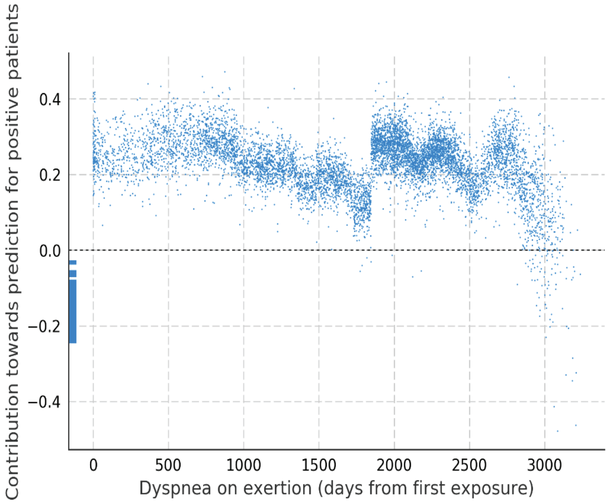
Figure 4: Dependence Plot for Age-Based Population Segmentation Identifies a Patient Sub-Group for Which Age Has High Significance in Predicting Them as Diagnosed, Compared to the Global Population



the ML drivers across an entire population. Beyond global feature importance attribution, our approach can also provide attribution of the relative importance of ML drivers to specific sub-populations. The model features illustrated in Figure 3 are in accordance with prior research and with guidelines[10] recommending control of HFpEF symptoms with diuretics as well as managing comorbidities, including hypertension, because these appear to be the drivers for the inflammation that lies at the root of the condition.

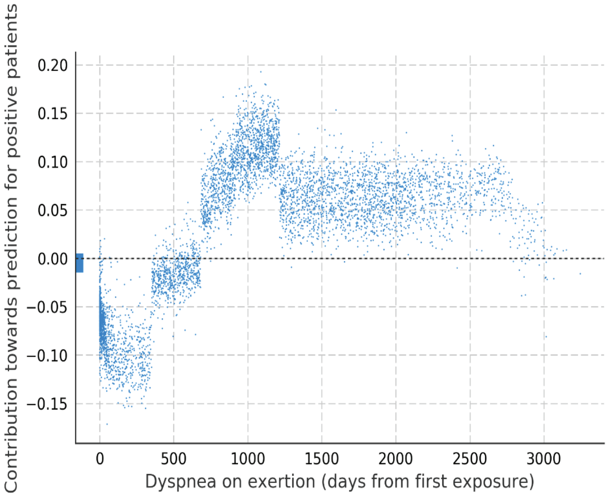
A first example of this capability is depicted in Figure 4, where we illustrate the feature

Figure 5: Dependence Plot for Days from First Occurrence of Hypertension, Indicates that this Feature’s Contribution Is Significant Across All Patients



contribution significance for the Age of the population. Specifically, the x axis depicts the age value of each patient, while the y axis denotes the importance of each value for predicting a patient being diagnosed with the disease. The importance values near the dotted horizontal line (zero-contribution boundary) do not contribute significantly to the diagnosis. The values above the boundary contribute to making a positive diagnosis, while the values below the boundary contribute to making a negative diagnosis. We can confirm that our interpretability approach not only highlights age as a globally important feature, but also identifies a specific group of patients (80 years old) for which age is more crucial in predicting them as being diagnosed with the disease, compared to the rest of the population. This can guide commercial teams in designing appropriate physician education strategies on points of intervention with earlier diagnosis and treatment for highly probable HFpEF patients and then further classification of HFpEF to their sub-populations.

Figure 6: Dependence Plot for First Occurrence of Dyspnea on Exertion, Identifies Patient Sub-Groups for which Dyspnea Has High Significance, Although the Global Relative Importance of Dyspnea is Low



To further demonstrate how our interpretability approach identifies patient sub-groups, Figure 5 illustrates the feature contribution significance for the days since first occurrence of hypertension. As expected, the feature contribution varies for different values of the days since first occurrence; however, we can confirm that the contribution of this feature is significant across all patients, as the importance values are consistently above the zero-contribution boundary. We should also note that the accumulation of negative importance values near the start point of the x axis corresponds to the subgroup of patients that do not have hypertension present in their history, and indicates that the absence of this feature is contributing towards making a negative diagnosis.

Our interpretability approach can also reveal useful insights for features with a low global relative importance. An example is depicted in Figure 6, which illustrates the feature contribution significance for the days since first occurrence of dyspnea on exertion. As we can confirm from Figure 3, this feature has a low global contribution significance. We can also note that for the sub-population that had the first occurrence of dyspnea within the last two years before HFpEF diagnosis, dyspnea was indeed not important in predicting them as diagnosed with the disease. However, for the sub-population that had the first occurrence of dyspnea three years before HFpEF diagnosis, dyspnea was an important factor in predicting them as diagnosed with the disease. These results illustrate how we can combine the attribution of the relative importance of ML drivers to specific population sub-groups, to identify sub-cohorts for whom a specific feature is important even though traditional approaches calculate the global relative importance of the feature to be low. This would allow pharma companies to identify high-value market segments that traditional approaches often fail to target.

6. Business Implications

The enhanced insights enabled by our approach are beneficial for commercial teams, as they enable them to better interpret ML outcomes to identify relevant patients and intervention points across the patient journey for early diagnosis and treatment. They also help design effective physician education strategies and improve the efficiency of marketing strategies. This approach can find wide application across ML commercial uses and can help bridge the gap from ‘black-box’ to ‘glass-box’ ML, as ML becomes increasingly embedded to commercial decision making. The proposed approach can also be applied to the area of rare diseases, considering the limitations related to class balance. In particular, given that the number of patients in this case will be very limited, the size of the positive cohort will be small compared to the negative cohort. However, this presents a challenge for all ML algorithms, as the trained model must achieve a high accuracy value by learning from a highly imbalanced dataset.

One limitation of the presented approach is that its usefulness depends on the accuracy of the underlying trained model. When a model cannot accurately predict whether a patient will be diagnosed with the disease, then the calculation of the global and individual feature contribution will provide insights that are less credible. A second limitation is related to combination of features and the significance of their contribution. Currently, our approach provides attribution of the relative importance of single ML drivers (features) to sub-populations. Extending this to calculation to combinations of features is a direction for future work.

7. Conclusions

We presented a ML interpretation approach that takes ML a step further by helping understand the drivers and patterns behind the model predictions. Our approach was applied to a complex disease and market, such as Heart Failure, with important implications to understanding and acting timely to sub-populations, demonstrating the power to identify the drivers behind predicting population sub-groups. Our approach can identify patient sub-populations of high value, expose the drivers behind the HFpEF diagnosis and highlight patients for which a specific drug is likely to yield improved outcomes.

About the Authors

Avgoustinos Filippoupolitis *leads the Data Science group of IQVIA Machine Learning & Artificial Intelligence Solutions, focusing on research, design and development of scalable machine learning methodologies for a wide range of areas including but not limited to digital disease diagnostics, HCP segmentation and targeting, and decision support algorithms. He is responsible for the design and development of Machine Learning methodologies and algorithms, using Real World Evidence (RWE), patient-level, and commercial datasets. Avgoustinos holds an MEng in Electrical and Computer Engineering from the National Technical University of Athens, and an MSc in Signal Processing from Imperial College London. He obtained his Phd in Emergency Simulation and Decision Support Algorithms at Imperial College London. Before joining IQVIA, Avgoustinos was a Senior Lecturer in Disruptive Technologies at the University of Greenwich.*

Michael Kusnetsov *is a Data Scientist at IQVIA with 10 years of multi-disciplinary experience. He conducts research for the machine learning team, focusing on model performance and interpretability. Michael has worked in the legal and healthcare industries with exposure to the financial and regulatory sectors. He received his PhD in Financial Mathematics from London School of Economics as well as MRes in Financial Computing from University College London and MSc in Mathematical Finance from Imperial College London. Michael is also a non-practicing qualified solicitor.*

Nicola Lazzarini *is a Machine Learning Data Scientist with a computational and bioinformatics background, working at IQVIA since 2018. He conducts research-related and client projects that involve the development of highly predictive and interpretable machine learning models. Nicola has over 7 years of experience in applying machine learning techniques to biomedical and pharma data, both in the academic and industrial fields. He has published various research papers in high quality academic journals and top-tier conferences, including Nature biotechnology and BMC Bioinformatics. His work has been cited by 200+ research papers. Nicola received his PhD in Computer Science from Newcastle University, UK. He also received both a BSc and an MSc in Computer Engineering from the Università of Padova, Italy.*

Hariklia Eleftherohorinou *is the global Head of AI and Machine Learning Solutions on Real World Data in IQVIA. The AI product portfolio spans across several application areas including but not limited to cost effective study design, commercial strategy development, AI-led disease diagnostics and screening, patient identification, Salesforce effectiveness, and AI-driven brand optimization. She is a trusted advisor to clients, guiding and supporting them as they are building their Predictive and AI/ML capabilities, translating predictions and insights into actions and informed decision making following a 'glass-box' approach. Prior to IQVIA, Hariklia was the Data Science and Advanced Analytics Practice Leader at Deloitte UK Consulting. She holds an MEng in Electrical and Computer Engineering from the Aristotle University of Thessaloniki; an MSc in Bioinformatics & Systems Biology and a PhD on Machine Learning In Medicine, from Imperial College London.*

References

1 Schwartz LM, Woloshin S. Medical marketing in the United States, 1997-2016. *Jama*. 2019 Jan 1;321(1):80-96.

2 Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, Dahlström U, O'connor CM, Felker GM, Desai NR. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*. 2018 Apr 12;7(8):e008081.

3 Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems 2017 (pp. 3146-3154).

4 Huis AE, De Man FS, Van Rossum AC, Handoko ML. How to diagnose heart failure with preserved ejection fraction: the value of invasive stress testing. *Netherlands Heart Journal*. 2016 Apr 1;24(4):244-51.

5 Oktay AA, Rich JD, Shah SJ. The emerging epidemic of heart failure with preserved ejection fraction. *Current heart failure reports*. 2013 Dec 1;10(4):401-10.

6 Bergstra J, Yamins D, Cox DD. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in science conference 2013 Jun (pp. 13-20).

7 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in neural information processing systems 2017 (pp. 4765-4774).

8 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3).

9 Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*. 2013 Apr 1;66(4):398-407.

10 Deaton C, Benson J. Time for correct diagnosis and categorisation of heart failure in primary care. *British Journal of General Practice*. 2016;66(652):554 - 555.

xDM Explainable Decision Models

Marc-david Cohen, PhD, Chief Science Officer Emeritus, Aktana Inc. and Pini Ben-Or, MA, MPhil, Chief Science Officer, Aktana Inc.

Abstract: The interest and growth in recommendation engines is accelerating in pharma brand management, marketing and sales operations. Such systems are evolving to manage decisions on how, when, and what to say to physicians to improve sales and physician engagement. To be most effective, systems should integrate brand strategy, business constraints, and models that are predictive of human behavior. While the effect of each of these decision drivers may be individually understandable, the behavior of the composite overall system may be much harder to explain. This is particularly true for systems that rely on machine learning (ML/AI) based analytics. Even if the decision system is not business rule constrained, and solely relies on a single ML/AI model, its decisions will likely need to be understandable to be persuasive to the stakeholders. For example, if a system recommends that a sales rep deliver a particular message to a physician in person, it is likely to be important for the rep to know why the system decided that in order for the rep to gain confidence in the recommendation and actually do it.

Recently there has been significant effort in developing approaches to explaining predictions of individual ML/AI models. This area of research is called explainable AI (xAI). While it is a rapidly expanding area of research, there has been very little focus on explainability of more general decision models or systems which may also include strategy, optimization, business rules, business constraints and multiple predictive models. This paper will start to address this need and propose a broader approach called explainable decision models (xDM).

We will give an overview of the current thinking on xAI which includes structural approaches and model induction approaches and will focus on two model induction ideas currently in vogue—the local and global models as represented by the LIME and SHAP, respectively. This paper will show how these techniques might be applied to decision models and demonstrate the concept using a simplified decision model. It will also discuss some of the more subtle issues regarding what it means for a model to be explainable and will conclude with a discussion of strengths and limitations.

Keywords: Machine Learning (ML), Artificial Intelligence (AI), Explainable Artificial Intelligence (xAI), Explainable Decision Models (xDM), Recommendation Engines, LIME, Shapley

Introduction

The pre-AI world of statistical models had a strong focus on models that are predictive and interpretable; models that are good at predicting the expected value of the target variable and that can be easily interpreted. Here interpretability meant that the researcher could gain a good understanding of the impact of a predictor or group of predictors on the target value. Many experiments and models were and still are carefully designed to be able to distinguish the effects of predictors on the prediction with a high degree of certainty. To aid in this, models were typically parametric, and often linear, with a considerable effort expanded on what transformations of the data were required so that the data supported a well understood parametric model whose parameters could be interpreted to gain insight into the underlying relationship between the predictors and the target or predicted.

As the science and technology for building predictive models evolved, it has resulted in more complex and opaque models. The neural networks, deep neural networks, and ensemble models are some methods that are currently popular. This popularity stems from the improved predictive accuracy they provide when compared to traditional parametric models. The cost of this improved accuracy is the loss in clarity in how the predictors impact predictions and their relationship to the structure of the model. In general, there is a trade-off between the accuracy of the model and the transparency of the model and how it works. If the relationship between the predictors (features) and the target is simple then parametric models can be used and easily understood by analyzing the parameters that describe the relationship between the predictors and the target. However, if the relationship is complex and non-linear then parametric models will not give accurate predictions without complex transformations of the predictors to simplify the relationship before building the parametric model. This in itself is a difficult task.

Understanding the roles that predictors play in complex machine learning models is currently referred to as "explainability." The need was always important and has been a topic of study. For example, it was addressed explicitly with ensemble tree based models like random forests with the concept of an importance family of metrics. These metrics are based on the average change in prediction accuracy or decrease in the Gini importance across the ensemble of trees using an out-of-bag (OOB) sample of the data—data not used in building the trees.[15] However, the need for understanding model behavior has not kept up with the advancement in modeling technologies as evidenced by the recent growth in approaches to explainability.

The topic of explainability in AI/ML models has largely focused on explaining models' classification of pictures and text. When a model has been trained to identify when a picture has a particular element, often using neural networks with numerous layers, the question is asked—how does one understand what is driving that categorization and how can we trust the prediction? Even notions of explainability and the motivations for such are not simple and are varied.[10] There are also important legal aspects to explainability as several authors have noted; with the European Union passing a law, "Right to Explanation" requiring models to have some levels of interpretability.[3, 6]

The focus of this paper is to explore the current thinking and approaches to model explainability and to extend them to the context of decision models. Although much of the technology that's been developed is for models that classify pictures or text, we explore applying this technology to decision models in the domain of marketing analytics where the objective is to identify the best next marketing action for one-to-one personalized marketing.

The next section reviews recent work in explainability. Then, a formal definition of a decision model within the marketing analytics domain is presented. Finally, the approach is illustrated using a real-world example and several different explainability models.

Review

David Gunning of DARPA is largely recognized as coining the term xAI for explainable or interpretable AI. The question at the core of interpretability is whether humans understand a model enough to make accurate predictions about its behavior on unseen instances and whether humans have enough confidence in the model to "believe in it." He breaks down the notion of explainability into several categories:[13, 5]

- Deep Explanation: modified neural nets and deep learning where the nodes are identified with features so that the weights on the various layers illuminate the particulars driving the model.
- Interpretable models: linear models, parametric models, tree models, Bayesian models, and other models where the structure is relatively transparent and can be explicitly understood.
- Model Induction: models of models, examples include: LIME (local interpretable model-agnostic explanations)[13], SHAP (Shapley additive explanations)[11], Anchors[12], CLEAR (counterfactual local explanations via regression)[16], LOCO (leave one covariate out)[9], etc.

As the list illustrates there is a lot of attention being focused on what Gunning calls “Model Induction” methods. These are methods which in effect are models of models. The assumption is that the underlying structure of a complex ML based model is hard or impossible to explain so you build another model “on top of the underlying model” that can more easily be explained. Much has been written about what explainable or interpretable means in this context.[3, 4, 10] Some have suggested that interpretable means that a human can predict what the model will do with unseen instances—can the human predict what the model will do?[12] This topic has been of interest particularly in Europe where model based decisions are being scrutinized for biases and the need for interpretability is being pushed into law. It is interesting to note that issues of bias may not be due to the modeling methods but may be inherent in the data on which the model is built. If there are biases in the data one shouldn’t expect the model to ignore that.

Analytic approaches to this explainability fall into one of two classes, either local or global depending on the type of explanation model. This may be somewhat confusing because each approach may rely on data that is both local and global and the boundary is fuzzy. However, in this context local often means that the explanation applies to a specific prediction of the underlying model, namely at a single point in the space of training or test data. Since many of these approaches were developed for classification of pictures, local means, for this specific picture, identifying the drivers of the specific classification. A broader question is how well does the model generalize to other pictures, and what are the explanations and how can one gain confidence in how well the model would classify other pictures.

As Ribeiro says “Most local approaches provide explanations that describe the local behavior of the model using a linearly weighted combination of the input features. Linear functions can capture relative importance of features in an easy-to-understand manner. However, since these linear explanations are in some way local, it is not clear whether they apply to an unseen instance. In other words, their coverage (region where explanation applies) is unclear. Unclear coverage can lead to low human precision, as users may think an insight from an explanation applies to unseen instances even when it does not. When combined with the arithmetic involved in computing the contribution of the features in linear explanations, the human effort required can be quite high.”[12]

Local

Although some of the techniques present themselves as locally focused and others as globally, the distinction is sometimes somewhat ambiguous. Both the LIME and the Anchors approaches claim to be locally accurate. LIME, which stands for local interpretable model-agnostic explanations, fits a new linear model in a local neighborhood around a given data point.

LIME saves a collection of weighted predictions from the model at sampled instances around the point of interest. Weights are determined by distance to the point of interest. The coefficients of that model hold the explanation. It then uses this new, linear approximation to the underlying model to explain how the more complex model behaves.

The Anchors approach is motivated by the lack of the ability of the LIME linear models to account for interactions effects. It is also a local explanation algorithm. Anchors is motivated by the fact that the univariate aspects of the importance coefficients can lead to some ambiguous attribution of explanations particularly in text mining applications. Anchors looks for a set of features such that if any features not in that set are included the predictions do not change "substantively." Substantively is defined by the expected value of the likelihood of a change in prediction being less than a prescribed amount. The approach is computationally complex since a large space needs to be searched in order to satisfy the Anchors criteria.[12]

Another approach that seeks to improve on the LIME model is called CLEAR. This exploits the use of counterfactuals and also provides to expand the LIME univariate limitations. CLEAR uses the concept of w-counterfactuals to explain a prediction by answering the question of "what if things had been different" with the feature set. Rather than randomly sampling the data and weighting by proximity to the point of interest as LIME does, the CLEAR method is to systematically search the space around the data point of interest and evaluate the model at those points producing counterfactuals to identify classification changes. The points at which this occurs can then be used to build a regression model for explanation, thus improving the fidelity of the explanation around the point in question.

LOCO (leave one covariate out) is an approach for generating metrics that measure variable importance. The metric is based on differences in errors from a complete model or a model built without one of the covariates.[9] The metric can be analyzed in a local manner or a global manner by applying it to every instance in the test data set and then analyzing the distribution of the variable importance metric. The single instance metric is similar to the variable importance measure used in random forests by analyzing the decrease in node purity by changing the order of variable splits.[1]

Global

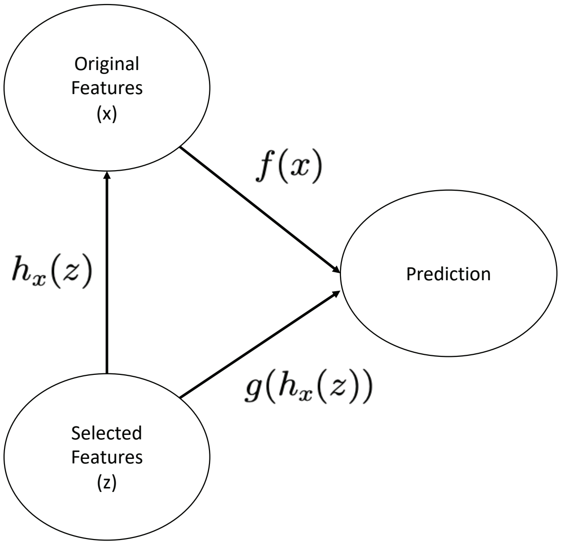
The SHAP (SHapley Additive exPlanations) is presented as a unified framework for interpreting predictions; it assigns each feature an importance value of a particular prediction.[11] In this way it is similar to some of the local approaches described above.

The framework presented by Lundberg, illustrated in Figure 1, is called additive feature attribution methods. Let *f*() be the underlying prediction model to be explained by explanation model *g*(). At a single point of *X* a transformation of the features *x* to *z* by a function $h_x(z)$ such that $f(x) \approx g(h_x(z))$ where $z \in \{0,1\}^M$ and *M* is the dimension of *z* which may be smaller than *x*. Then, the linear additive explanation model is defined as

$$g(z) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i .$$

For LIME, *h*() maps the binary interpretable *z* variables to the original variables *x*. Specific values for φ are found by minimizing a loss function that measures the distance from *f* to *g* with a penalty for complexity—to reduce the number of non-zero φs. Since this additive explanation model is linear, the values of the ϕ_i in the solution easily explain the impact of the features in the underlying model *f*() at the point *x*. Finally, under three regularity

Figure 1: SHAP Framework at a Single Point x



conditions (Local accuracy, Missingness, Consistency) Lundberg derives a fully characterized SHAP solution that is exact for $g()$. See Lundberg[11] for details.

This approach is similar to LOCO in that a new model is built for each predictor leaving the predictor out and then that new model is evaluated at the point of interest (that's the local part) and the difference in the value of the prediction with the prediction from the full model is weighted by the non-zero occurrences for that predictor as shown above.

This review is not meant to be exhaustive but touches on some of the newer and more unique thinking and algorithms. There are many others including partial dependence plots, recursive partitioning, other decision tree methods, and just about any "white box" model.

The Decision Model

A decision model (DM) is a model that recommends a value maximizing action to be taken. These types of models are used when there are decision choices made to maximize value to the firm. The choice of model depends on numerous factors in the decision problem, including the frequency with which the decision needs to be made, the value in making a "good" decision vs the opportunity costs in making a "bad" decision; the information available at the time the decision needs to be made; the reliability of the available information; factors and restrictions associated with implementing the decision; and other issues. There are many different approaches and models used to navigate decision problems based on these factors. The focus here, as discussed above, is on decision models for marketing decisions. These are high frequency decisions that individually have low value but collectively can be of significant value in driving revenue and shaping brand perceptions.

Marketing focused decision models are often associated with recommendation engines. Recommendation engines are designed to maximize the market basket; if you bought this product you are likely to purchase this other product. This is somewhat tangential to the marketing decisions required by pharmaceutical sales operations and brand management. In this domain the market basket is somewhat limited; the competition tends to be well defined; the marketing tools are limited and can be legally restricted; the objectives can also be longer term in the sense of building brand loyalty in addition to increased sales.

This domain specificity and complexity requires a special kind of decision model, one that is flexible and can accommodate multiple inputs: the result of numerous ancillary predictive models; raw data; rules and conditions; and constraints on decision actions. The following section describes a decision model which has this needed generality. This is referred to as the underlying model and it is the object of explanation in this paper. This decision model can be very complex and in fact need not be a stochastic model but is assumed to be a "black box" and complex enough so that its behavior is opaque and requires explanation. In practice decision models are very often stochastic models and rely on the solution of multiple prediction problems for which ML is used.

Note that the description that follows is in reality time dependent but to simplify this exposition we will suppress that dependence. We believe that will not result in a loss of generality.

Y be a target variable. It could be a categorical variable such as whether a physician takes a certain action like open an email or read an online report. Or, it could be a continuous variable like the Rx for a target therapeutic or market share associated with a therapeutic, segment membership, perceptions, or other measures of HCP value.

X be the features that are believed to be predictive of the target. These could be:

demographic data that characterizes the HCP such as age, gender, educational background, segment membership, HCP potential, and such,

patient data that describes the HCP patient population characteristics,

contact history that captures the history of contacts with the HCP, such as number of visits, emails sent, topics of emails, topics of visit conversations, documents provided, webinars attended, conferences attended, and such.

For the non-time-dependent version, each row in *X* is a single HCP. For the time-dependent characterization, data for a single HCP could be replicated in multiple rows and the data in each of these might be for different time points. There are other useful representations of the data which might aid in model development; exploring them would not be of relevance to this discussion.

The typical setting is to find a model $f(X)=Y$ that can be used to predict the target with all the input data as described above. Of course, regardless of the method for fitting the model it will not be perfect so we represent the fitted model by $\widehat{f()}$ so the error associated with the model is $Y-\widehat{f(X)}$.

In this context, an explainable model is one for which humans can understand why $\widehat{f()}$ is predictive, which areas of the domain *X* the predictions are particularly good, where they aren't, and most importantly why we should believe that $\widehat{f()}$ captures the structure in the relation of the variable *X* to *Y*. Without that confidence and insight into the model it is hard to justify its use in making business decisions, particularly to those who are not versed in the details of model building, machine learning, and AI.

In the context of using a predictive model as a decision model we seek to identify some set of variables within *X* as decision variables. These are variables on which we have some control, for example, send an email; send an email about topic *A*; visit and discuss topic *A*; visit and then send an email about topic *A*. Let's denote that subset of variables as *D* so we can re-characterize the predictive problem as to find *f*() such that *f(X,D)=Y*.

The goal of finding *f*() is to use the information contained in it to make decisions on what actions, what specific values of the *D* variables are best, to take to maximize some value *Y*. We can express this as finding the *d*(x)* that maximizes *f*() as

$$d^*(x) \equiv argmax_d f\widehat{(x,d)}$$

In practice all possible choices for *d*() may not be feasible from a business perspective. For example, the maximizing *d*(x)* may be to visit an HCP immediately. While that may be desirable from the perspective of maximizing the value of an action, it may not be feasible because of logistical realities; the medical representative may not be available at that time. Other examples of constraints include

- maintaining a pacing of visits
- coordinating visits with non-face-to-face interactions
- traversing the territory systematically

To capture those realities in the characterization we denote a set of constraints by *C*. The definition of *d*(x)* becomes

$$d^*(x) \equiv argmax_{d\in C}\ f\widehat{(x,d)},$$

where $d\in C$ denotes that the searchable space of *d* values satisfies the constraints.

Another reality encountered in practice are the business rules that the brand management and sales operations teams require the decisions to meet. These can result from various plans and information that may not be captured in the relationship between *(X,D)* and *Y*. For example, there may be a recent introduction of a competitive therapeutic into the marketplace that the brand team wants to proactively address. There are numerous other scenarios for which the brand teams would want to specify actions and decisions derived from the data might not be desirable. Other examples include:

- coordinating interactions with uncontrolled publications and other information
- requiring visits when commercial metrics change in statistically relevant ways
- timing interactions with seasonal commercial drivers
- coordinating messages across therapeutics

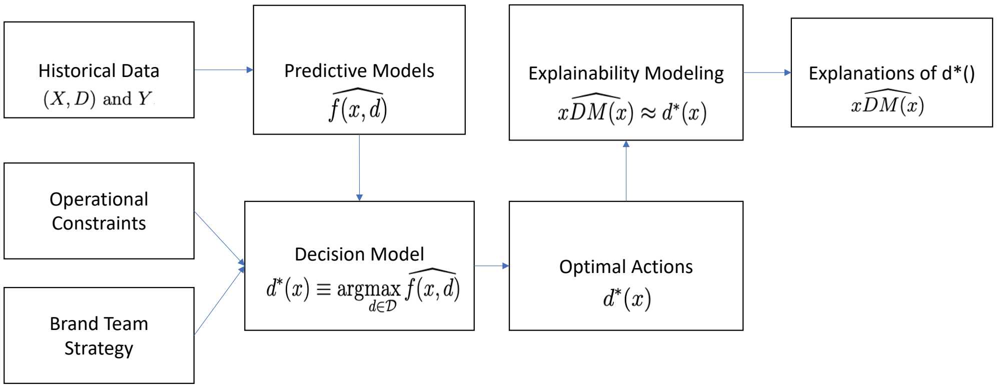
Let *R* denote the set of rules and then represent *D* as the union of constraints and rules, namely $D \equiv C \cup R$.

$$d^*(x) \equiv argmax_{d\in D}\ f\widehat{(x,d)}$$

It is important to note that the strategy of maximization over the *D* variables contains an implied causality associated with them. For the purpose of this paper we are assuming that an experiment can be designed to capture this causality in the data. This could be done by sampling from the population of HCPs (i.e. potential *X* variables) and then randomly assigning actions (i.e. potential *D* variables) to them and execute on those decisions. Of course, there are optimal experimental designs that could be used; that is a topic for another paper.

Finally, although *d*(x)* as presented is based on a single fitted model, in practice the function being optimized could be an algorithm with many components including heuristics, raw data, feature engineered data, and the results of statistical and machine learned models. This generality does not change the approach presented below, in fact, the more opaque the function being optimized by *d*() the more to be gained by employing an explanation model of *d*(x)*.

Figure 2: Schematic of the xDM Components



Explainability

Explanability in the case of a decision model differs from explanation models commonly in practice. This application does not seek to identify the features driving the classification of an instance of a picture or text but seeks to identify the features driving an optimal decision. As for xAI, the business users of decision models may be reluctant to rely on an opaque model that just gives a decision or action. They often seek a deeper understanding of what the model is doing and what areas of the predictor space lead to specific decisions.

The rich literature of xAI presents numerous algorithms and approaches to leverage. Many of these can be applied to understanding and explaining DMs. So far, there have been limited research on explanation algorithms outside classification. One notable example uses a neural network to learn the underlying explanation space produced by a ranking algorithm.[8] We take a similar approach in building a model to learn the underlying explanation space from a decision model (DM). To achieve this we seek an explanation model *xDM*(*x*) that predicts the result of the DM, namely *d*(x)*, and can be characterized simply to give insight into the behavior of the DM. Such an explainability model would be $\widehat{xDM}(x)$ and would satisfy $\widehat{xDM}(z) \approx d^*(z)$ at a specific point *z*.

There are many ways to accomplish this, as is evidenced by the research summarized above. The specific data used to estimate *xDM*() could be used for the explanation model following the LIME approach, or counterfactuals could be used following the lead of the CLEAR approach. Counterfactual data are observations that were not sampled from the original data used to build the model but are data from which predictions of model are calculated. These are typically used for scenario analysis and exploring the implications of using the model for some practical purpose, like "what would the model predict if this happened."

The approach chosen here is to use the counterfactual data that covers the space of *X* or part of that space surrounding, for example, *z*. This flexibility means that the explainability model can be used to gain insight into the broader applicability of the DM to the business objective. The schematic in Figure 2 shows how the various components of the xDM explainability model fit together.

In addition to generalizing the application of explanation models to DMs, a contribution of this approach is the use of counterfactuals as the sample space of the underlying model that is to be explained, in ways similar to the CLEAR approach. This presents the opportunity not only to explain a large part of the decision space but enables the exploration in a very deliberate and controlled way. The cost of this is that if the space of (the counterfactual data of) the DM model i.e. *(X,D)* is very large, the computational costs become large and potentially infeasible to completely explore. However, insight into *(X,D)* and in particular in relation to the constraints $d \in D$ provides potential ways to address the explosion of the decision space.

Example

Rather than use a "toy example" we illustrate the concepts presented above using anonymized data from an overseas client. The business objective is to assign the number of quarterly visits to each facility (doctor's office, clinic, or hospital) that maximizes the sales of each of two therapeutic products. There is strong motivation to reduce costly individual visits, potentially replacing them with group conferences and/or emails and freeing up resources so that more facilities can be served with the same resource overhead. In other words, to maximize sales as a function of visits by identifying where and when it is best to visit. To do this, a decision model (DM) is built that finds the sales optimizing number of visits to each facility. Using the terminology from above the description below shows:

- a prediction model $\widehat{f(x,d)}$ to map features to sales,
- an unconstrained decision model *d*(x)* of visits to facility *f* that has a history of interactions,
- a constrained decision model *d*(x)* of visits to facility *f*,
- two different explanation models to interpret the decision model *d*(x)*

Data

The data are anonymized on contact frequency and sales on 2 products reported on a quarterly basis for 3 years from about 19,000 medical facilities. The total number of records exceeds 475,000. The variables in the data set are summarized below. Each record contains:

facility: a code for one of 11 facility groups indicating the quantile for the sales of the particular facility.

appointment: the number of sales rep's visits to one of the HCPs in facility that was scheduled before the visit. This variable has 6 unique values, 0 to 5 for the number of appointments at a facility in a quarter.

conference: the number of conferences that HCPs within the facility attended. This variable has 2 unique values, 0 and 1 for the number of conferences at a facility in a quarter.

group: the number of group meetings that the HCPs within the facility attended. This variable has 8 unique values, 0 to 7 for the number of group meetings at a facility in a quarter.

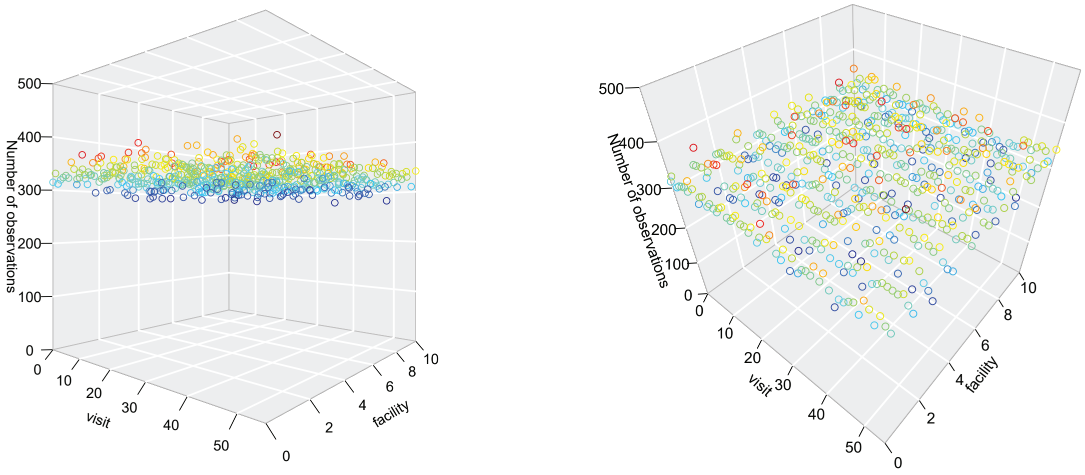
email: the number of emails sent to HCPs within the facility. This variable has 6 unique values, 0 to 5 for the number of emails sent to a facility within a quarter.

visit: the number of unscheduled visits to the HCPs within the facility. This variable has 56 unique values, 0 to 55 for the number of visits to a facility within a quarter.

qtr: the quarter for the data. This variable has 4 unique values.

product: a number indicating the specific therapeutic product which was the focus of the meeting or email. This variable has 2 unique values, 0 and 1 identifying the therapeutic.

Figure 3: Data Coverage



To illustrate the approach, a random sample of 500 facilities was taken from the cleaned data set. All the observations from each of the sampled facilities are included in the analysis. The average sales per facility for each product was calculated and the deciles for the distribution of facility sales means was used to bin each facility into one of the sales deciles. Unfortunately, the data does not include the number of prescribers in each facility, but we use average sales in a facility and facility as a surrogate for the missing information. The target variable to be maximized was the sales difference from the average facility sales.

Figure 3 shows two views of the number of observations visit vs facility which show that there is good coverage of the distribution of observations across these two dimensions. Note that the other dimensions are much more sparsely populated, but since visits is the decision variable of the DM it is important that there be good coverage across the more dense variables in the data. Finally, if a more complete DM was being built, features that were functions of the observed data would be constructed to capture, among other things, the impact of actions over time.

Prediction and Decision Models

The prediction model $\widehat{f(x,d)}$ was built using random forest[*] regression with the target (Y from the earlier description) being the sales deviation from mean group facility sales. All the variables listed above were included in the model which explained 72% of the variation. The variable importance for the predictors is shown in Table 1.

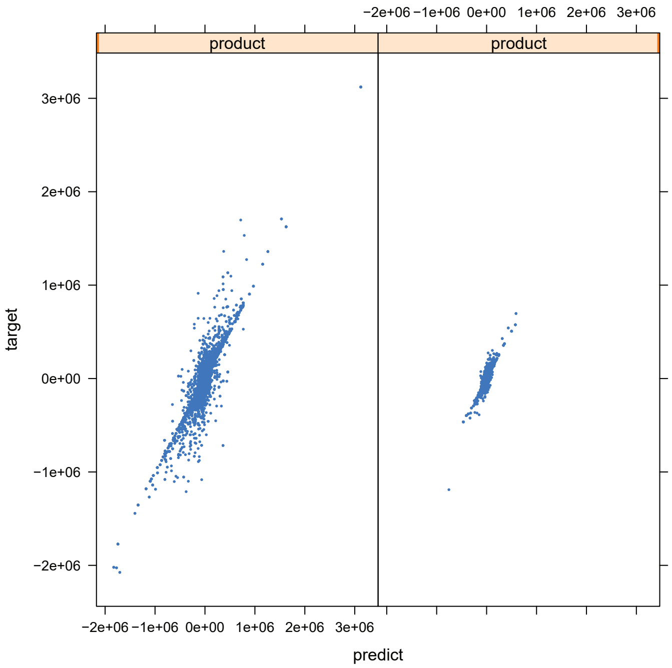
The table shows that facility group, visits, and quarter are the most important predictors. %IncMSE refers to the error of the random forest model and by how much the model would be made worse if that variable was replaced by randomly permuted (destroying correlation with the other variables in the model), whereas IncNodePurity is a measure of homogeneity. It is desirable to have the nodes to be homogeneous each time a node is split. Since the variable *qtr* has the greatest %IncMSE it may be considered as the

[*] The *randomForest package in* R is used.

Table 1: Random Forest Variable Importance

	%IncMSE	IncNodePurity
qtr	133.16188	4.089630e+13
visit	125.76667	1.780477e+14
facility	102.33283	7.961670e+13
email	84.30202	8.880591e+13
appointment	77.56009	1.281616e+14
group	56.38927	5.964828e+13
conference	47.18138	3.337538e+13
product	31.50224	7.460655e+12

Figure 4: Model Fit

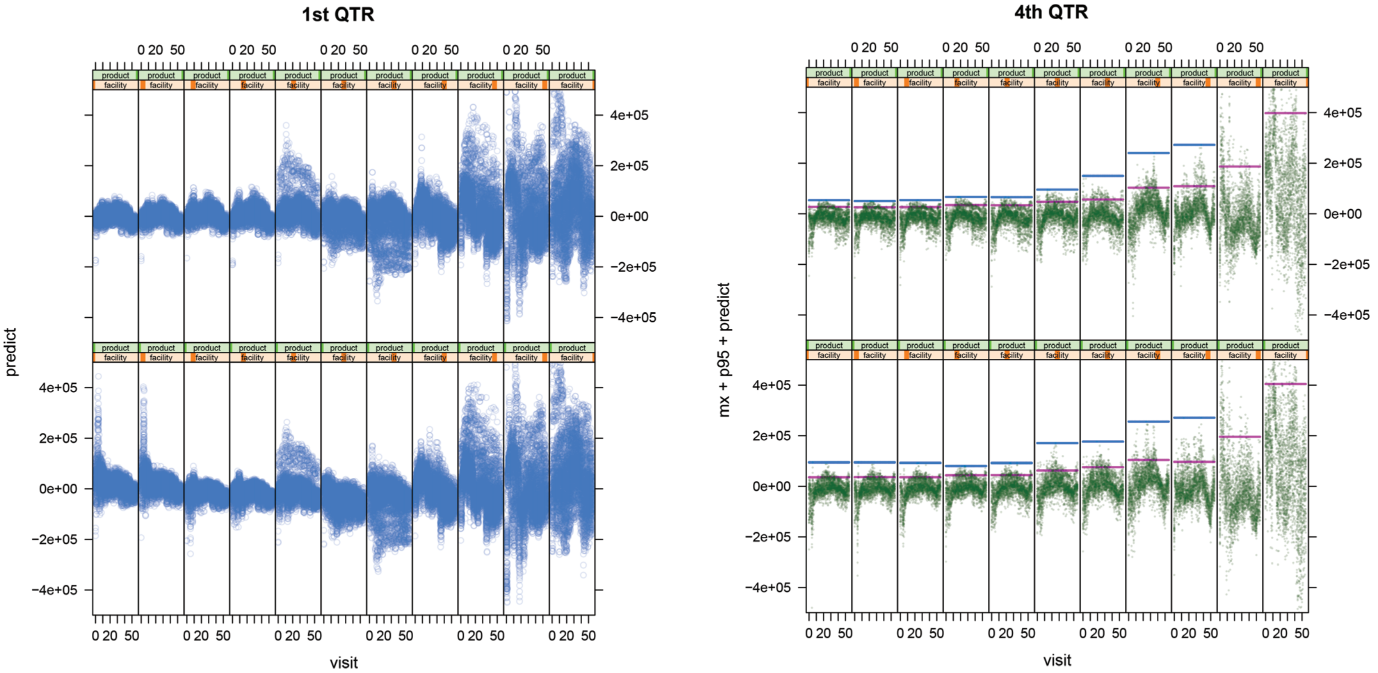


most important predictor, because replacing it with noise would most degrade model performance compared to replacing other variables with noise. The plots in Figure 4 show the model's predicted values vs. the actual target values for each of the products. The plots show a strong diagonal pattern which confirms that the model fit is good.

As described above, the approaches to building an explanation model to evaluate the prediction model either on a sample of the data set used to train or test the model or on a set of counterfactuals. We use counterfactuals to generate 2,838,528 observations that cover the complete space of the predictors. These data will be used to build the DM.

The surface defined by the predictions of $\widehat{f()}$ is an 8-dimensional surface. Since the observations that go into defining the "surface" are from the predictions of a random forest model and not some parametric smooth model, there are discontinuities in the surface as is evident in the plots. Figure 5 shows that surface across 4-dimensions for two quarters. Note that

Figure 5: Prediction Surface



the surface varies across the quarters (only two are shown), across the facilities, and across the products. The first row in each plot is for product 1 and the second is for product 2; as the plots move from left to right the facility decile increases. Some of the variance and fluctuations in the plots are due to the discontinuities of the random forest model and some are due to the hidden variables that are not shown on the plots.

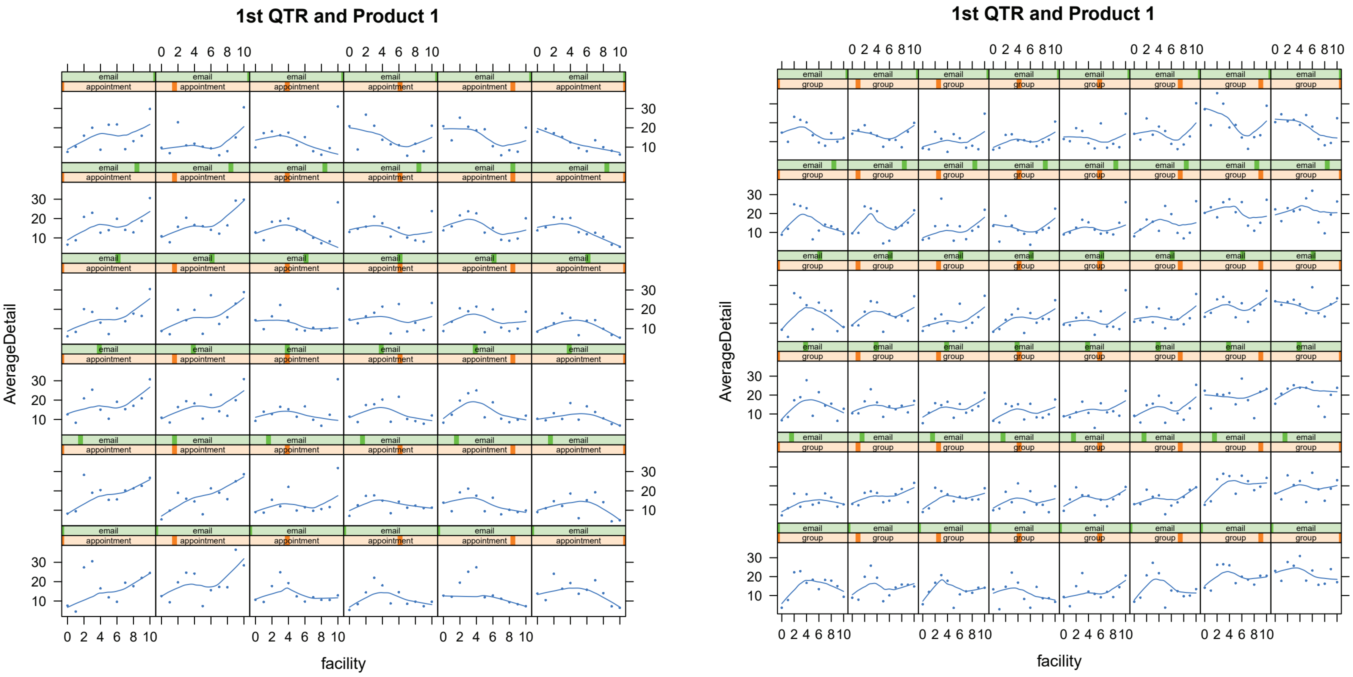
Figure 5 also shows more detail on this prediction surface as well as providing some insight into the DM. Notice that the plot on the right has blue and red lines; these are the maximum and the 95% quantile for the prediction in each of the identified dimensions. The value of visit where the maximum intersects is the value for *d*(x)* for that set of predictors. Since there is variance associated with the prediction model method, the average number of visits where the prediction for those values is above the 95% quantile within that bin of predictors is used as the value for *d*(x)*.

The plots in Figure 6 show the average of the number of visit values above the 95% quantile for several combinations of predictors. They also show a kernel estimated line through those points.[*] Even though the maximum is not used and the average above the 95% quantile is used for convenience, it will be referred to as the maximum.

In the plot on the left this line shows the maximizing number of visits as a function of facility sales size, number of emails sent, and number of appointment details. Appointment details are increasing as the plots move to the right and email sends are increasing as the plots move up the page. The plot shows that the value of visits increases with facility size when there are fewer appointment details but that trend inverts as appointment details grow. One might expect that appointments are more important as facilities grow. It also shows that the impact of email sends is more subtle.

[*] Kernel estimate is from the smoothing option in the *lattice package of* R.

Figure 6: Decision Surface



The plot on the right is similar but focuses on group details in place of appointment details. Group details are increasing in the plots to the right and email sends are increasing as the plots move up the page. Group details are more cost effective since a number of prescribers are simultaneously in a meeting. In general, it shows that more visits are needed as facility size grows. This is somewhat counter-intuitive but may suggest that the HCPs need more explanation in face-to-face visits after group meetings. Since these are views of marginal slices through the decision space, it is difficult to get a complete understanding of the drivers and shape of *d*()*, hence the need for an explanation model.

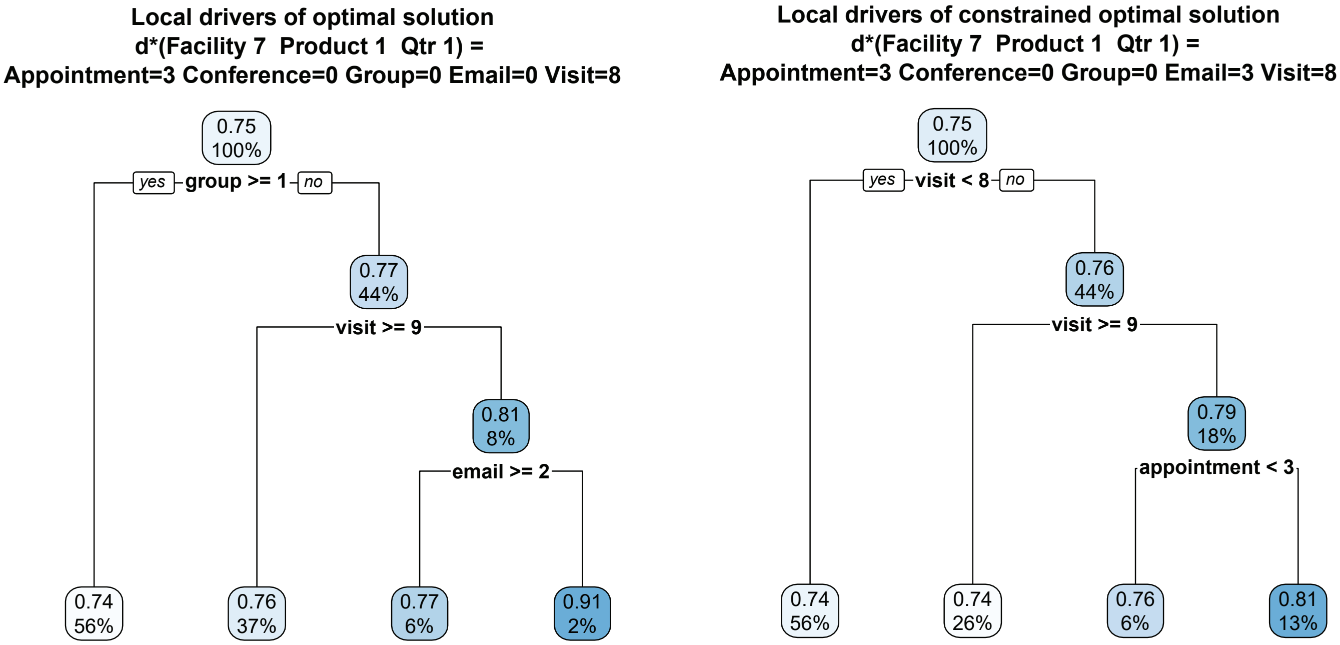
Explanation Model

In explainability approaches like LIME and CLEAR, explainability is obtained by analyzing the behavior of the underlying model at a single point by sampling the underlying model in the space around the point of interest similar to above. In the case of LIME, a linear model is built based on those points as described in the review section above. These approaches do not give multidimensional insights in the sense of identifying the impact of covariates.

This is similar to looking at marginal plots across a handful of dimensions, as discussed above, which gives some insight into the underlying decision model but does not capture all interactions and their relative strengths. Nor does it provide a clear understanding of the relative impact of the different variables on the optimal decision *d*()* as, for example, variable importance in random forest models. The explanations models in xAI, like variable importance, provide insight into what drives predictions of the underlying model. The analogy for DM is to explain what is driving the optimal decision and how varying from that decision impacts the result, namely, what are the importances of the variables involved.

To explore this, let's focus on an instance *d*(z)* where *z* is a particular decision point. There are general questions: what is driving the optimal

Figure 7: Local Drivers of Near Optimum Performance


Local drivers of optimal solution
d*(Facility 7 Product 1 Qtr 1) =
Appointment=3 Conference=0 Group=0 Email=0 Visit=8

Local drivers of constrained optimal solution
d*(Facility 7 Product 1 Qtr 1) =
Appointment=3 Conference=0 Group=0 Email=3 Visit=8

$d^*(z)$ with regard to changes in z, namely with regard to facility, or product, or quarter; the other is when z is fixed (say at z^*) what is driving the particular solution $d^*(z^*)$. As described above, if the model $\widehat{xDM}(z)$ is understandable then because $\widehat{xDM}(z) \approx d^*(z)$ at the point z insight into what variables drive the optimal decision is achieved.

Consider an arbitrary value of z^* as facility 7, product 1, and quarter 1. In a business setting all facility groups, products, and quarters will be of interest and will have sales efforts applied to them, so it makes sense to want to optimize the marketing treatment with that segmentation applied. In fact, one might expect that the sales organization would identify the optimal treatment in each of the facility, product, and quarter segments and then use that result to assign resources to maximize return on effort.

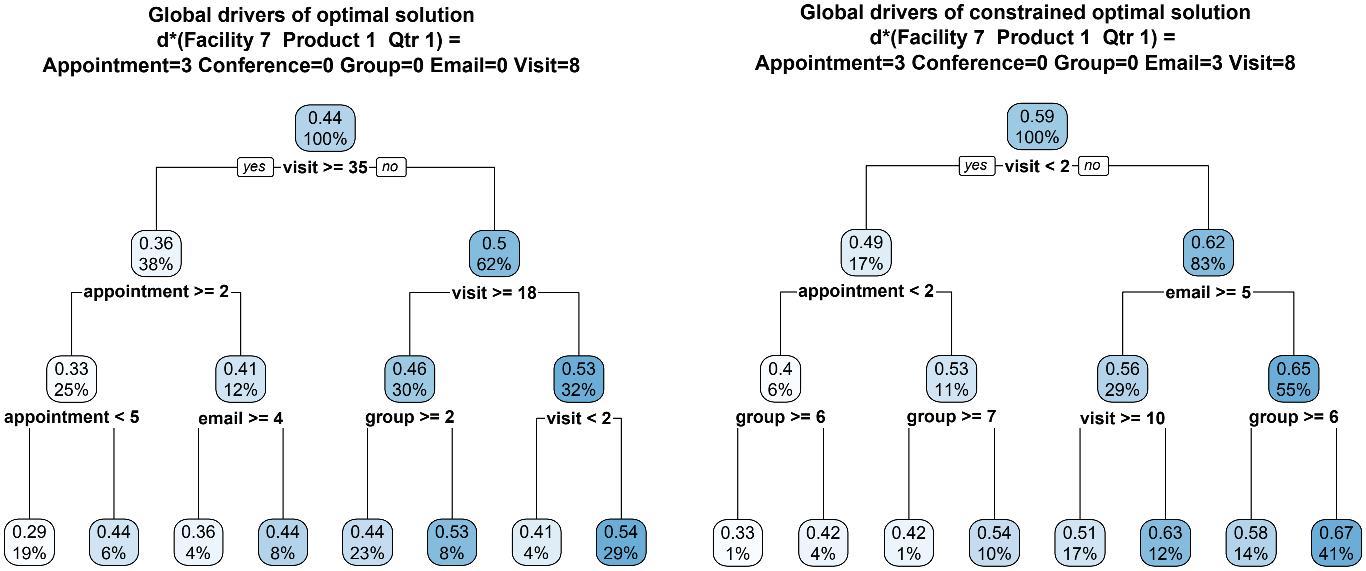
Focusing on this particular z^* segment for the moment, the optimal value for $d^*(z^*)$ calculated across all counterfactuals is 3 appointments, 0 conferences, 0 group meetings, 0 emails, and 8 visits. Note the use of counterfactuals (as with the CLEAR approach) to evaluate $f(\widehat{z^*, d})$ and find $d^*(z^*)$ means that the complete space is available.

Since $d^*(z^*)$ is a point in the solution space based on an optimization of $f(\widehat{z^*, d})$ which is noisy (see Figure 5) exploration of the surface defined by points near $d^*(z^*)$ with a model that can be understood will provide insight into what is driving the optimization, namely what variables impact the solution near the solution? To do this, several choices to fitting $\widehat{xDM}(z)$ are examined including recursive partitioning.[*]

Figure 7 shows two trees fit to predict proximity to $d^*(z^*)$ by using all solutions within 70% of the optimal solution and as target the percent of optimal. The tree on the left is for an unconstrained decision model and the tree on the right is for a constrained model.

[*] The *rpart package in* R is used

Figure 8: Global Drivers of Optimum Performance


Global drivers of optimal solution
d*(Facility 7 Product 1 Qtr 1) =
Appointment=3 Conference=0 Group=0 Email=0 Visit=8

Global drivers of constrained optimal solution
d*(Facility 7 Product 1 Qtr 1) =
Appointment=3 Conference=0 Group=0 Email=3 Visit=8

In the tree on the left, the top node labeled 0.75 and 100% indicates that the average percent of optimal is 75%, and that accounts for all subnodes. The subgroup having *group* value 1 represents 56% of the population and has a mean percent of optimal of 74%. The tree also shows that the optimal solution can have visits of 8 or less and have 0 or 1 email and achieve 91% of optimal. The tree shows which variables are not drivers are not important and not drivers of away from optimality—those not included in the tree. It is important to note that this is a local explanation giving insight into the variables that impact optimality within a neighborhood of the solution *d*(z*)*.

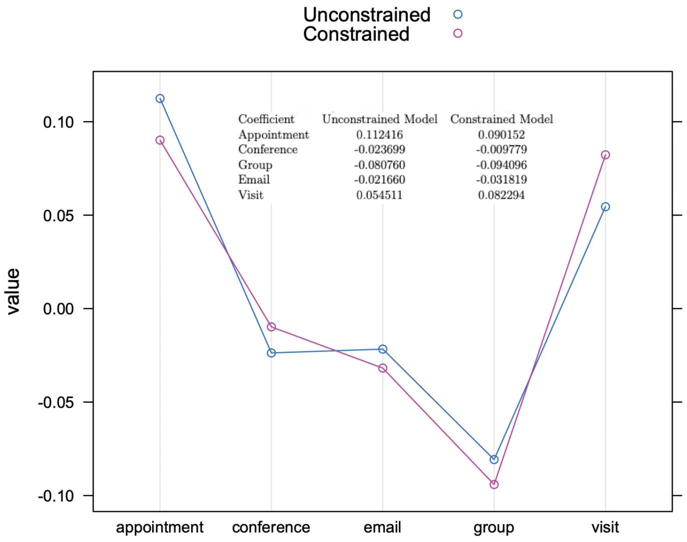
The constrained decision model incorporates a constraint that requires the number of emails sent to be at least 25% of the number of visits. This is a typical type of constraint that sales operations and brand management teams would want to impose. The tree on the right in Figure 7 shows that the *visit* variable is the most consequential in driving to the optimal solution. It is interesting to compare the optimal values of the constrained and unconstrained solutions. The optimal value of the constrained solution is approximately 30% less. This is to be expected since a constrained solution would result in a decreased optimum. This model does not capture the business benefits of the distribution of marketing resources.

Global Explanation

The exploration ideas developed in the previous section focused only on gaining insights into *d*(z*)* in a neighborhood of *z**. Here we use the same approach to gain insight from a "more global" perspective and to find the variables that globally drive optimality at the point of interest.

Figure 8 shows the same approach as used above and shown in Figure 7 but instead of restricting the search space to within 70% of the optimum the entire space of predictions on all the counterfactuals are used in the recursive partitioning algorithm. It is not surprising that the variables' order of splits tracks with the order of importance as shown in 1, namely *visit, email* and *group*. The constrained analysis in the tree on the right shows the impact of the constraint driving emails into the solution. The

Figure 9: LIME Estimates



Coefficient	Unconstrained Model	Constrained Model
Appointment	0.112416	0.090152
Conference	-0.023699	-0.009779
Group	-0.080760	-0.094096
Email	-0.021660	-0.031819
Visit	0.054511	0.082294

branch on the right labeled *email >= 5* contains 83% of the space and accounts for a mean 62% of the constrained optimum value. The sub-branches show there is a tradeoff between *group* visits and *visits* that help the decision model to navigate the email constraint.

LIME Explanation

How do these recursive partitioning based explanations compare to other “more standard” derived explanations, like from the LIME algorithm? An implementation of the LIME algorithm was developed for this example. The standard implementation samples from the test set and then builds a linear model using predictions on the sample points weighted by distance to the point of interest. The coefficients associated with the linear explanation model yield the importance of the predictors for that particular explanation point. The current implementation has been modified to use counterfactuals across the entire space used to evaluate *d*(x)*.

Similar to the standard LIME approach a point is selected (the z^*) for analysis, but then rather than sampling additional points around the point of interest, all counterfactuals within a hypercube with side length 2 are sampled. Note that this example has all integral predictors so a unit hypercube is a natural choice. If some of the predictors were continuous a similar approach could be taken, although a different strategy would be needed to evaluate the DM counterfactuals. The current implementation also weights the observations in the LIME explanation model by $\exp(-w)$ where w is the distance from the points in the hypercube to the point of interest at the center of the hypercube.

A linear model is built with the weighted hypercube values as predictors and where the target of the model is the percent deviation from the optimal value. This is the same target as was used in the recursive partitioning exploration of the space. The plot in Figure 9 show the coefficients for two estimated models: one for the constrained model and one for the unconstrained model. The predictors are along the horizontal axis and the coefficients on the vertical axis. The table shows the values. The r^2 for the unconstrained model is .97 and the constrained model is .98. For each model the variables *appointment, conference*, and *visit* were highly significant and the constrained model the variable *email* is significant at the .05 level. These results generally agree with the observations above. The difference is that with the LIME model the multivariate impact of the predictors as explainers in the decision model is not available.

Conclusion

The importance of machine learning continues to grow and become more central to decision making in marketing analytics. With that increased reliance comes the importance of understanding how the complex algorithms do their work and why the results should be trusted. This paper has reviewed the leading approaches to explainability and the trends and ideas that these approaches are using. It presented a general characterization of a decision model, one that can be used to make optimal decisions in the domain of marketing analytics and one-to-one marketing and personalization of recommendations. The model relies on the results of other predictive models as well as business constraints and requirements. It illustrates the use of explainability models to gain insight into how a specific decision model is behaving using real-world data. There are many unanswered questions; of interest for further study would be to compare the efficacy and accuracy of different explainability models on decision models. For example,

- suppose that one of the recursive partitioning models were used in place of the DM that is being explained, how degraded would the resulting performance be?
- what are appropriate measures of the reliability of an explanation model?
- how well would an explanation model work if the underlying decision model was on a much larger scale than the example?

Perhaps it shouldn't be too surprising that the world has progressed to a point that in order to gain trust and confidence in the result of algorithms that have high predictive accuracy, models that explain models are needed. But it is an interesting development.

Acknowledgments
Thanks to an anonymous reviewer for many helpful comments.

About the Authors

***Marc-david Cohen**, PhD, Chief Science Officer Emeritus, Aktana, is an experienced business leader with a background in operations research and statistics. At Aktana he participates in 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.*

***Pinchas Ben-Or**, MA, MPhil, Chief Science Officer, Aktana, is an experienced technology leader who oversees artificial intelligence (AI) and analytic innovation at Aktana. Previously he served as Global Head of Analytics at Actimize where he helped transform the company from reliance on rule systems and expert models to deploying fully agile machine-learning-based models for financial crime detection. Pini has a BSc in Physics, Mathematics, and Philosophy from The Hebrew University in Jerusalem, and a MA and MPhil in Philosophy from Columbia University in NY, where his research areas were Decision Theory, AI, and Philosophy of Physics.*

References

1 Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

2 Budzik, J., “Four Approaches to Explaining AI and Machine Learning,” https://www.kdnuggets.com/2018/12/four-approaches-ai-machine-learning.html, 2018.

3 Gilpin, Leilani H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., “Explaining Explanations: An Overview of Interpretability of Machine Learning,” arXiv:1806.00069 [cs.AI], The 5th IEEE International conference on Data Science and Advanced Analytics (DSAA 2018). Feb 3 2019.

4 Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F., “A Survey Of Methods For Explaining Black Box Models,” arXiv:1802.01933v3 [cs.CY], Jun 21, 2018.

5 Gunning, D., “Explainable Artificial Intelligence (XAI),” www.darpa.mil.

6 Hall, P., “Guidelines for Responsible and Human-Centered Use of Explainable Machine Learning,” arXiv:1906.03533v1 [stat.ML], Jun 8, 2019.

7 Hall, P., Gill, N., An Introduction to Machine Learning Interpretability, An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI. O’Reilly Media, Inc., 2018.

8 Hoeve, M., Odijk, D., Schuth, A., Rijke, M., “Faithfully Explaining Rankings in a News Recommender System,” arXiv:1805.05447v1 [cs.AI], May 14, 2018.

9 Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R., Wasserman, L., “Distribution-Free Predictive Inference For Regression,” arXiv:1604.04173v1 [stat.ME], Apr 14, 2016.

10 Lipton, Z. “The Mythos of Model Interpretability,” arXiv:1606.03490v3 [cs.LG], Mar 6, 2017.

11 Lundbrerg, S., Lee, S., “A Unified Approach to Interpreting Model Predictions,” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA.

12 Ribeiro, M., Singh, S., Guestrin, C.,”Anchors: High-Precision Model-Agnostic Explanations,” Association for the Advancement of Artificial Intelligence (www.aaai.org). 2018.

13 Ribeiro, M.T., Singh, S., and Guestrin, C., “Why Should I Trust You?” Explaining the Predictions of Any Classifier,” CHI 2016 Workshop on Human Centered Machine Learning. (arXiv:1602.04938v1 [cs.LG] 16 Feb 2016.

14 Strobl, C., Boulesteix, Kneib, T., Augustin, T., Zeileis, T., “Conditional variable importance for random forests,” BMC Bioinformatics, 9(307), 2008. URL http://www.biomedcentral.com/1471- 2105/9/307.

15 Strobl, C.,Boulesteix, A.,Zeileis, A., Hothorn, T., “Bias in random forest variable importance measures: Illustrations, sources and a solution,” BMC Bioinformatics, 2007, 8:25.

16 White, A., Garcez, A., “Measurable Counterfactual Local Explanations for Any Classifier”. arXiv:1908.03020v1 [cs.AI], Aug 8, 2019.

AI Plus Real-World Data for Early Prediction of Disease Progression and Operationalized Precision Targeting

Brian Malpede, Manager; Stephanie Roy, Principal; Patrick Long, Data Scientist; Emin Ozkan, Data Scientist; Ryan Hopson, Consultant; Nadea Leavitt, Principal and US Lead; Orla Doyle, Lead Data Scientist; John Rigg, Senior Principal and Global Lead, IQVIA Predictive Analytics

Abstract: Artificial intelligence (AI) offers a highly effective solution to classic problems in marketing later line therapies, particularly for niche patient populations and in increasingly competitive markets: how do companies reach the right healthcare providers (HCPs) at the right time to impact treatment decisions? What message should they deliver? How are these messages properly deployed to the field and monitored for ongoing performance? There is a compelling impetus for pharmaceutical companies to find ways to more accurately and efficiently time outreach activity to HCPs treating patients who are most suitable for a specific therapy. AI can leverage large scale datasets to find the right patient at the right time to achieve advanced predictive targeting of therapy escalation through timed prediction of disease progression. This paper focuses on methodological inputs for AI models that seek to predict these events in the therapy journey, including: 1) defining and building the dataset, 2) setting and understanding the window for predictive timing, and 3) effectively assessing model performance prior to commercial deployment.

Keywords: Artificial Intelligence, Machine Learning, Predictive Analytics, Disease Progression, Treatment Journey, Precision Targeting

1.0 Introduction

1.1 Predicting Disease Progression with Artificial Intelligence

Treatment decisions such as new therapy initiations or therapy escalation in the event of disease progression are complex, multifaceted, and may occur over a narrow window of time. Effective messaging with a view towards impacting treatment decisions thus requires a highly targeted approach, particularly for products intended to treat niche patient populations or indicated for later line therapy. Increasingly, pharmaceutical companies face the challenge of not only reaching the appropriate health care provider (HCP), but more importantly, of delivering the right message at the right time in a patient's disease progression to most efficiently operationalize outreach.

Medical and prescription claims are most commonly used for HCP targeting due to patient coverage, cost, and the commercial applicability of linking a patient to a treating HCP. However, claims data can be noisy and complex, making predictions of disease progression (typically evidenced by therapy initiation or escalation) a difficult task. Artificial intelligence (AI) provides the technical foundation to effectively mitigate challenges associated with predictions of disease progression including:

- Medical history data in the form of open claims, typically used for commercial targeting endeavors, can be noisy and fragmented – AI can handle such data and help to maximize the value of targeting operations through deployment at scale and without requiring highly specific biomarkers that are indicative of disease progression.
- Therapy decisions are often driven by complex and interrelated factors in a patient's history as well as by their HCP's unique preferences – AI can address both patient and HCP characteristics and navigate the complex clinical factors that influence progression to the next line of therapy.
- Data used for predictions is likely to have a lag (the time between the actual data event and collection/availability) – AI models can account for this lag by predicting an event in advance (e.g. 30 days prior to the target event) providing the necessary time for salesforce mobilization.
- Identification of target patients may focus on a niche group suitable for a specific medication – AI can selectively mine data to focus on a broad patient population or a narrow, highly specific subpopulation when determining who is most likely to benefit from a medication.

AI is a well-suited tool to identify patients who may experience disease progression and enables the precision needed for targeted outreach. Due to our growing ability to process ever larger and more complex data sources, AI models can now use the presence and interplay of clinical events to predict which patients are likely to progress as well as use the timing of these events to inform when progression may occur. The ability to integrate the timing of events also allows insights from AI to be proactive rather than reactive, resulting in targeted and timely interventions. This paper will discuss key methodologies critical to success in disease progression predictions and will speak to the future of AI applications in the predictive precision targeting space.

2.0 Methodological Design for Timed Predictions with AI

2.1 Defining Potential Predictors of Disease Progression

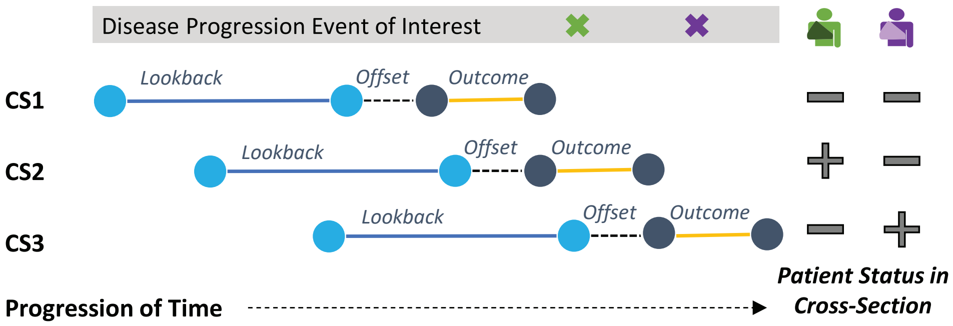
AI algorithms 'learn by example' whereby a model is rewarded for finding patterns in patient medical data that distinguish a target patient from a large universe of non-target patients. When applied to prediction, AI algorithms analyze data as a collection of predictors (or features) that may guide a model in its decision making. A first step in model development is defining the clinical features that may be used for model training. This process may be accomplished in two ways: 1) through a hypothesis/knowledge-driven approach where domain experts manually curate potentially relevant clinical features and 2) through an automated data-driven approach where features are algorithmically extracted from patient data without human interference. Knowledge-driven features ensure that domain knowledge is leveraged for feature identification whereas the data-driven approach provides the opportunity to identify previously unknown clinical factors that may prove indicative of disease progression or informative of potential therapy escalation. A hybrid of domain-driven and data-driven predictors further ensures that the potential for insights from the data is maximized.

2.2 Data Transformation for AI Predictions

Conventional Approach

The well-designed use of patient medical data is essential to the success of AI for disease progression prediction. A conventional approach for data extraction in predicting progression is to take a single snapshot of patient medical history preceding an outcome

Figure 1: Multiple Snapshot Cross-Sectional Approach to Data Extraction



The cross-sectional approach defines patients as positive or negative based on multiple snapshots of medical history. Thus, a patient can be negative (absence of disease progression) in earlier snapshots and transition to positive (experience disease progression event) in a later snapshot. **CS** = cross-section; + indicates that the disease progression event has occurred for the patient, designating them as positive in the cross-section; **-** indicates that the patient has not experienced the disease progression event and is thus negative in the cross-section. *Source:* IQVIA methodology.

of interest, e.g. patient history prior to the initiation of a new therapy. This set of data is then used for model development. Within the dataset itself, two groups of patients are identified: a positive group that experiences this event of interest (disease progression) and a negative group that does not experience the event.

The strategy of using a single snapshot presents several challenges. For one, an outcome of interest may occur multiple times over a patient's medical journey, such as a patient who initiates multiple biologics for the treatment of rheumatoid arthritis over the course of a single year due to adverse reactions or poor treatment response. This situation raises the question of which event date to use as an anchor point for prediction of progression. Additionally, it can be unclear as to which snapshot of medical history is the most appropriate for a negative patient, as they do not have an event of interest on which to anchor, in contrast to positive patients that experience defined progression.

This conventional approach consolidates patient histories into a single longitudinal snapshot at a single point in time, and therefore only allows the model to learn about this timeframe in the data. There is thus no flexibility to view a patient across multiple points in their journey. In addition, the dependence on an event date of interest to guide data extraction may lead to sampling biases when target events occur with greater or lesser frequency due to temporal factors like seasonality or variations in data coverage. As a result, models trained using this single snapshot strategy may generalize poorly, meaning they will not perform well when deployed commercially on new, previously unseen patient data.

Advanced Cross-Sectional Approach

A promising alternative to the single-snapshot approach seeks to overcome the above challenges while leveraging the full breadth of patient medical data over time. This approach sub-divides patient histories into a series of time-bound rolling snapshots defined as cross-sections (see Figure 1 and key definitions in Table 1). Each cross-section includes a defined lookback window usually on the order of one or two years from which patient medical history is extracted to create features for model training, as well as a narrow forward-looking

Table 1. Key Definitions Utilized

Concept	Definition
Lookback	Medical history used to train the AI model
Outcome	Clinical event defining disease progression, often the initiation or escalation of therapy, that identifies a patient as positive
Outcome Window	Time period over which progression is predicted to occur
Offset	Time period prior to the outcome window that can be incorporated to accommodate lags in data collection and operational needs (i.e. the model can predict an event X amount of time in advance, where X is defined by the offset)
Cross Section	Snapshot of patient medical history defined over a set time period
Precision	Proportion of patients correctly predicted with disease progression, defined as: $\frac{\textit{True Positive Patients}}{\textit{True Positive + False Positive}}$

prediction window to detect an outcome of interest (disease progression). This outcome window defines the time period over which disease progression is predicted. For example, if the window is three months the model will be trained to predict patients who will transition over a three-month time horizon. Successive cross-sections are shifted by a given interval (e.g. monthly increments) to form a final dataset containing multiple cross-sections of data defined by iterative timeframes of medical history.

A key advantage of this strategy is that it captures multiple snapshots of the patient journey wherein patients are labeled according to their current therapeutic status (within the specific snapshot of time), such as drug initiation versus no initiation. Within this definition, a patient may be seen by the model as a negative patient during an earlier snapshot of data, and subsequently seen as a positive in a later snapshot once the patient has initiated the medication of interest. This allows AI algorithms to learn from more varied and comprehensive representations of patient history regarding disease progression. It also helps overcome challenges arising from small sample sizes, which is often the case for niche products, products that have recently launched, or products with narrowly defined market segments, since the number of patient instances used for model training scales in relative proportion to the number of cross-sections. In other words, patients are used more than once for model training, amplifying the signal obtained from each individual.

There are a number of benefits of this multi-snapshot approach to patient data extraction related to customization that best suits a given clinical scenario and commercial application:

- The prediction window can be designed such that patient predictions are valid for a given period of time (e.g. a 3-month window of opportunity).
- The "offset" period prior to the prediction window can be incorporated to accommodate lags in data collection, data processing, or the mobilization of clinical or commercial resources (e.g. a 1-month time period prior to operationalizing AI predictions in the field).
- This method provides the opportunity to train on multiple cross-sections, allowing for better monitoring of indicators of

model drift (reduced performance due to market or other changes in the data), and thus for mitigation of temporal biases in patient sampling due to seasonality, fluctuations in data coverage over time, or shifts in market dynamics such as a new therapy launch or changing treatment pathways.

This approach to patient data extraction also enables model validation strategies that evaluate model performance exclusively on "future" data. Specifically, a model can be trained on the bulk of historical medical history snapshots and validated only on the most recent snapshots to produce representative indicators of model performance after real-world commercial deployment. Within this framework, the model is essentially being evaluated on data it has not seen from a future time period, as it is trained exclusively on earlier snapshots of data.

Despite its many advantages, one should exercise care when using this multi-snapshot strategy for several reasons:

- First, as always when using longitudinal data steps must be taken to prevent instances of "data leakage" (i.e. scenarios where the model is mistakenly exposed to future patient data during training since the model will not have access to this patient information at deployment).
- Second, as this approach takes full advantage of historical data one must be vigilant to shifting market dynamics that could cause the model to overfit to obsolete market trends and so should select the study time period appropriately.
- Finally, choosing the duration of a snapshot outcome window requires balancing the needs of a desired model use case with the ability to accurately predict over narrower or broader time horizons. Finding the optimal outcome window may require some experimentation and adaptation based on a given therapeutic area and corresponding market dynamics.

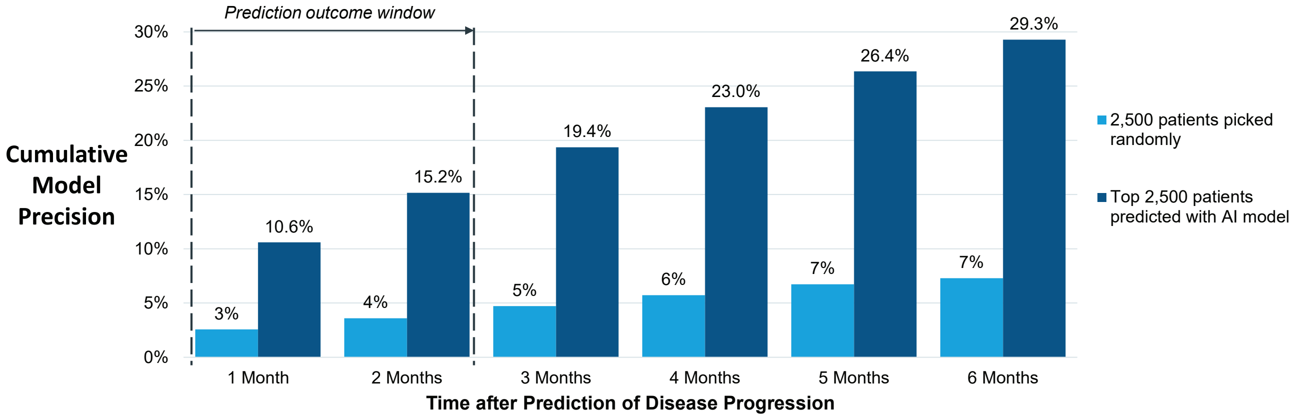
3.0 Methodological Design for Model Validation and Performance Assessment

Deploying AI using longitudinal patient data is most beneficial when frameworks for commercial implementation are carefully considered in the evaluation of model performance. Proper assessment of model performance prior to deployment is essential to understanding model utility in a real-world setting. Below, two main elements of model evaluation prior to deployment are discussed. The first assesses the implications of an extended outcome window on model performance and the second introduces precision measures based on potential target HCPs. A process for AI-driven predictive messaging in real-world deployment is also outlined.

3.1 Case Study Data and Modeling Overview

A recent example of a model developed in the autoimmune space illustrates the cross-sectional approach. This model was trained to predict disease progression among patients with an autoimmune disorder using an open claims dataset. The patient sample included ~25,000 patients who switched to a later line medication (the positive cohort) and ~340,000 patients with a diagnosis of the disorder but no evidence of switch from a first to later line therapy (the negative cohort). Features were derived from both a data-driven approach (based on claim prevalence in the positive and negative populations) and a knowledge-driven approach (claims were selected under the guidance of domain experts). Features were engineered as the recency (first and last date) and frequency of claims during each cross-section. An Extreme Gradient Boosting model (XGBoost) was trained to predict disease progression.

Figure 2: Model Precision with an Extended Outcome Window



Model precision in identifying patients with disease progression with an autoimmune condition is increased approximately three-fold (~10% to ~30%) as the outcome window is extended.

Hyperparameter tuning and cross-validation was performed to maximize model robustness and predictive performance.

3.2 Model Performance Measurement with an Extended Outcome Window

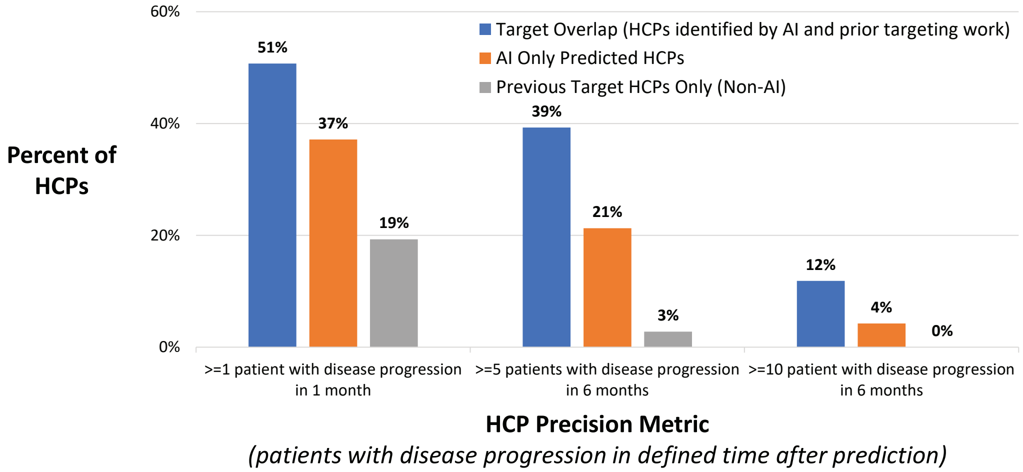
The outcome window defines the period of time over which the model predicts a disease progression event of interest. AI models may be trained with narrow outcome windows in order to find patients for whom disease progression is imminent. This approach is reasonable as patient history close to an event of interest often proves predictive, such as a rare procedure that is frequently performed shortly before the initiation of a new medication. However, a narrow outcome window to predict a patient transition is not always ideal as an overly narrow window may rely on a signal to noise ratio that does not exist in reality.

Our autoimmune case study illustrates this idea. The model was initially trained to predict disease progression within a narrow outcome window of two months. For this clinical population there proved to be insufficient signal in patient medical history to effectively make such high-resolution predictions and consequently the precision of this model was lower than desired (15.2% precision for the top 2,500 patients predicted by the model). However, assessing performance this way underestimates the usefulness of this model because a substantial percentage of false positive patients within the constrained time window ultimately experience disease progression when the outcome window is relaxed to six months beyond the time of prediction (29.3% precision for the same set of 2,500 patients – see Figure 2). This precision value is only truly meaningful when compared to a baseline of performance in the absence of the AI model. In the case of the model defined here, AI-driven precision was four to five times better than selecting patients at random for disease progression, a substantial increase over baseline.

3.3 Measuring Model Precision at the HCP-Level

Model precision is often measured at the patient level as outlined above, where the performance is quantified by how many patients the model identifies correctly. However, for commercial targeting purposes, it is not just the effective prediction of patient events that is important but also of HCP-related events. Or, in other

Figure 3: Model Precision Measured at the HCP-Level



HCP-level precision demonstrates the efficacy of the model in identifying HCPs who see patients with a disease progression outcome of interest, which is defined here as a transition from an initial treatment to later line therapy with a biologic in the autoimmune space.

words, predicting if an HCP will transition a patient to a new medication is just as important if not more important towards the validation of a model's performance and applicability in a real-world commercial setting.

Looking back at the same model developed for disease progression predictions in the autoimmune space mentioned above, the measurement of HCP-level precision is constructed by defining the number of predicted patients that are linked to a potential target HCP and identifying HCPs that recognize disease progression in one or more of these predicted patients and transition them to a medication of interest. Constructed this way, an HCP-level performance assessment demonstrates that the AI model can achieve high levels of precision (see Figure 3). For this specific model, the view of HCP-precision defines the proportion of potential target HCPs that are expected to actually transition a predicted patient to a specific biologic (new treatment) in a predicted time window. For commercial application, one could consider that the set of HCPs identified by AI would be targeted with 37% precision, meaning that roughly two out of five targeted HCPs would be expected to act on the disease progression and transition a predicted patient in a real-world setting within one month of the prediction.

3.4 Leveraging AI for Predictive Messaging

Once individual predictive features have been identified, it is critical to ensure these insights are both interpretable and actionable. Knowing that treatment with six individual therapies is predictive of disease progression may be informative; however, this level of granularity may not be valuable when making messaging decisions. A view of messaging that maintains the insights but simplifies the content can be produced by collecting individual related predictors into broad domains (see Table 2). For example, while the model may identify a recent rheumatologist visit, duration of a specific treatment, and recent visit for a comorbid disease as important individual features in predicting disease progression, it

Table 2: Predictive Messaging Informed by an AI Model

Domain	Potential Predictors Included
Disease Activity	• Recency of rheumatology interaction • Recency of visit for comorbid disease • Duration of maintenance therapy
Reduced the Risk of Event	• Rheumatologist visit frequency • Recency of high or mid-dose steroids
Reduced Steroid Dose	• Proportion of days covered with steroid treatment in most recent six months
Opportunity for brand use as an earlier-line therapy	• Naïve to specific treatment

Source: Illustrative examples driven by studies in the autoimmune space.

may be more useful to craft a single message related to all three labeled as "disease activity." These messages can then be presented to target HCPs with predicted patients that fit a certain profile of disease progression.

3.5 On-going Algorithm Re-optimization

While a model is trained using data collected over a limited historical period, it may benefit from re-optimization after real-world deployment. Through this process, an established predictive model may be retrained with the addition of new and more recent data, including additional positive samples as more patients experience disease progression and transition to later line medications. With on-going optimization, the model can also account for potential market shifts to ensure that predictive performance remains stable during deployment.

Two main options can be implemented for ongoing model re-optimization, including:

1. Refreshing with newer data (including updated positive patients)
2. Refreshing with data collected in the field (i.e. call and response data)

For the first option, on a routine basis the model can be updated with additional data that is collected between the initial predictions and subsequent rounds of predictions as patients are newly identified as positive (e.g. undergo disease progression after model deployment). For the second option, an additional model may be developed to track previously predicted positive patients and their corresponding HCPs to understand how many predicted patients ultimately experience disease progression.

3.6 On-going Optimization of Messaging

The refresh approaches above provide the opportunity to tune both the number of aggregated messages as well as the granularity of messaging based on modelling results and the commercial or clinical application. For example, a model that derives the bulk of its predictive power from a small subset of features may only need two or three aggregate categories to effectively summarize outputs for HCP outreach. On the other hand, a salesforce large enough to call on a broad range of HCPs or to maintain frequent contact with a select group of HCPs may require increased detail such as the types of drugs HCPs prescribe to relevant patients at either the class or individual product level, the number of patients seen by the HCP who have certain symptoms or comorbidities, or the types of other specialists seen by patients tied to that HCP.

The messaging can also be adjusted over time either based on feedback from the field force or the results of a newly trained model (refreshed based on re-optimization element 1) mentioned above). For example, if a certain treatment is traditionally used for alleviation of symptoms associated with a disorder, it may initially be highly predictive of disease. But over time, newer treatments may become more prominent, and as such the older treatment may no longer be as relevant or predictive. In this case, we would expect this feature to be de-emphasized in a re-optimized model and the messaging to be adjusted accordingly. This approach specifically allows for adjustment of HCP messaging strategy to reflect changes in features driving model predictions.

4.0 Business Implications of this Methodology

Overall, there are several key benefits to the approach outlined above and its translation from model development to real-world application. These benefits include:

- The model is trained on a more representative dataset for a better estimate of performance.
- This method provides a better ground truth for actual model performance once applied in a commercial or clinical setting.
- The model is well suited for iterative, ongoing refreshes as it is designed on a rolling set of data that is set up to be optimized over time.
- A model can be trained on older data, and then tested on a "future" dataset as the cross sections are rolling in time and can be split temporally.

Certain challenges with this technique do exist, such as:

- Rapid expansion of the amount of data used to train the model and thus increasing computational power required for analyses (expansion occurs because a patient's medical history can be utilized at multiple times, amplifying the amount of times this patient is present in the dataset)
- In less prevalent disease states or when targeting a niche patient population, the positive sample in an individual cross section can be low, rendering it more challenging for the model to identify a positive signal to differentiate patients of interest.

Conclusion

Predicting disease progression over a pre-defined time window can be quite powerful towards informing pharma commercial strategies. These predictions are not an easy task, as progression is informed by many clinical events, treatment journeys are complex, and HCP preference can play an important role in therapy changes. AI can be successful in not simply predicting these events but also in informing messaging specific to disease progression. However, the effective use of AI is not easy and requires clinical, methodological, and AI expertise. In addition, enhancement of the technique is possible, and upcoming improvements include comparing different algorithms, such as neural networks (e.g. deep learning) relative to prevailing tree-based algorithms (e.g. XGboost) as well as experiments on the types of clinical features that may be engineered to support predictive modeling.

About the Authors

The authors of this publication represent the Predictive Analytics practice in IQVIA's Real-World Solutions global group. The team develops innovative solutions to solve challenging healthcare problems based on patient-level data using a variety of advanced statistical and machine learning methods. This development encompasses applications such as physician targeting and risk stratification algorithms aimed at, for example, finding undiagnosed patients or identifying patients suitable for treatment escalation. Our efforts help improve retrospective clinical studies, under-diagnosis of rare diseases, personalised treatment response profiles, disease progression predictions, and clinical decision-support tools. For questions or more information regarding the information in this article, please contact Stephanie.Roy@IQVIA.com or Nadejda.Leavitt@IQVIA.com.

A Method for Physician Segmentation and Brand Activation Prediction Using Claims Data

Mert Sahin, Chief Marketing Officer, Imaging, GE Healthcare and Ashish Patel, Co-founder, CareSet Systems

Abstract: Pharmaceutical manufacturers use several key performance indicators (KPIs) to assess the performance of their products. One such KPI, which has become the standard for assessing prescription write-and-fill events for pharmaceutical benefits, is total new prescriptions (NRx). It uses real-time data and has become a part of established data feeds from companies that provide weekly physician-level NRx information to their clients (e.g., brand, regional sales, and field teams of pharmaceutical manufacturers). However, the NRx KPI is limited in measuring the performance of intravenous, subcutaneous, injectable, and other biologic drugs, which require final-action administrative claims instead of retail prescriptions. In this article, a new Patel-Sahin Proxy (PSP) score is presented to approximate the NRx KPI for biologic drugs by using procedure and diagnosis codes found in administrative claims. The PSP score was applied to the Centers for Medicare and Medicaid Services (CMS) Part B claims data to produce prescriber profiles and make sales predictions. The results showed that the PSP score produced a more actionable representation of provider segments and made better sales predictions compared to the NRx KPI.

Keywords: Medicare, Claims Data, Commercial Analytics, Brand Marketing, Utilization, Activation, Biologics, Buy and Bill, Part B, Physician Segmentation, NPI, HCP

Introduction

The marketing methods used by pharmaceutical manufacturers to select physicians have shifted from the use of quarterly prescription and claims data to the use of real-time data from large data providers.[1] To track the number of new patients a physician has educated and activated with a specific brand or treatment, the most relevant metric is the number of new prescriptions per physician, also known as total new prescriptions (NRx).[2] Marketing and sales teams may receive weekly updates on brand activation NRx. The teams can also pay to obtain real-time notifications once a physician writes a prescription. These new written prescriptions (NWRx)[3] enable pharmaceutical manufacturers to reliably measure and predict the success of their products.

However, most biologic drugs, along with equipment used for diagnosis and treatment, do not necessarily rely on written prescriptions. In those cases, manufacturers have to be innovative in their methods for measuring success. One such way to do that is to more efficiently explore the vast amount of new physician-level data and technology.[4]

GE Healthcare is one pharmaceutical manufacturer to have contextualized this challenge. Among a diverse offering of healthcare technologies and services, GE produces several injectable contrasting products used in medical imaging to diagnose several cardiovascular and neurological degenerative disorders. These injectables are "buy and bill" and collecting data on products obtained through a hospital or specialty pharmacy is difficult. The concept of "buy and bill" is when an institution, hospital, or medical group purchases and warehouses units of medication, and bills for each unit or dose after it is administered to a patient.[7]

In this article, a new method of physician segmentation and brand activation success prediction is reported. The initial intent was to estimate brand utilization by a physician over the next two years. The method was created, executed, and compared against the current method of reporting by two cooperating teams: (1) marketing and sales branches of GE Healthcare, and (2) the data scientists and engineers of a leading Medicare data company. The new method used quarterly claims data from Medicare Part B. The qualities and success of the new model were compared to the performance of the current method to answer the question: Can product utilization be accurately predicted using medical benefit claims?

Methodology

GE Healthcare is launching a range of new biologic drugs and is interested in developing innovative methods to increase access and adoption in the adult population. The partner organization followed a strategy built on Medicare claims used to help payors, accountable care organizations (ACOs), and hospital systems to develop robust physician networks. Regional physician recruitment teams helped develop several models to identify and prioritize physician-based estimates of growth, potential, risk appetite, and network impact. When combined, these parameters could predict the arrival of newly diagnosed and treatment-naive patients.

Focusing on networks and physician utilization, analysis of Medicare data can identify care continuity gaps with therapeutic area-specific perspectives highlighting the patient's journey through the healthcare system. Leveraging that approach, GE received the relevant expertise in Medicare Part B claims data, and were empowered to collaborate within their teams. The Medicare data provided the total visibility of claims, with no missing markets, providers, or organizations since any provider accepting Medicare, the largest payer in the US, was visible. Additionally, patient identity and privacy were protected by following the CMS cell size suppression policy, which sets minimum thresholds of 11 for the display of any CMS data (e.g. admissions, discharges, patients, services, etc).[5]

Several datasets were used to find and identify physicians. The evaluation relied upon the innovative use of healthcare administrative claims (Medicare fee-for-service (FFS) Part B Research Identifiable Files (RIF) claims with a 100% population sample, which contains the full FFS census). The observation period was 24 months. Physicians were identified by the National Provider Identifier (NPI). NPI is a Health Insurance Portability and Accountability Act (HIPAA) Administrative Simplification Standard used for administrative and financial transactions. Nurse practitioners, physician assistants and pharmacists that may administer treatments are also identifiable by NPI. In addition to identifying physicians, the NPI also identifies hospitals, organizations and group practices. The number of established or new patients added to the practice, and the number of diagnoses related to the proposed product was calculated based on each physician NPI.

In contrast to Medicare Part D, which covers pharmaceutical benefits for typically oral medications obtained through retail pharmacy, Medicare Part B covers Institutional Outpatient and Carrier Outpatient setting medical benefits, meaning that all claims in this category reimburse for procedures, such as evaluations, labs, radiology, and other same-day services, and prescriptions, such as those medications that are injectable, intravenous, and infused. When the outpatient benefit is administered in an institution or hospital setting, it is billed using a UB-04/CMS-1450 claim form and is found in the Institutional Outpatient claims file. All non-institutional claims, such as those from private practices and free-standing settings are billed in Carrier Outpatient claims using a CMS-1500 claim form.[6]

CMS provides qualified researchers with quarterly RIF data that includes procedure and diagnosis codes, place of service, and payment details, including patient out-of-pocket, third party payment amount, Medicare reimbursement, and provider charge data.

Healthcare Common Procedure Coding System (HCPCS) procedure codes (Px) appearing in Part B claims describe the supplies and services a physician provided; as such, the descriptions provide insight into the patient experience in that physician's office. The most common Px reimburse physician office visits for patient evaluation and management (E&M). Most importantly, there is a code-level distinction that describes if the patient is "new" to the physician or if the patient has an "established" relationship with the physician. For example, 99201 procedure code is used for a "new" patient and 99211 for an "established patient". Therefore, the evaluation of the procedure codes across a physician's Part B claims allows the team to calculate the percentage of new patients. This is done by dividing the number of new patients by the total number of patients and multiplying the figure by 100. This practice growth measurement is called the "% of New Patients" for the physician and is a critical component of the model.

From the pharmaceutical manufacturers' experience, if a practice is no longer billing office visits for new patients and using only the established patient E&M codes, then that practice may be likely to close over time, or they may not be open to treating new patients. There is an assumption that those physicians are less likely to adopt new treatment tools and are unlikely to be ideal recipients of marketing efforts. The proposed model focused on identifying physicians who are growing their practices, recording new indications, managing newly diagnosed patients, and, therefore, thought to be open to pharmaceutical innovations.

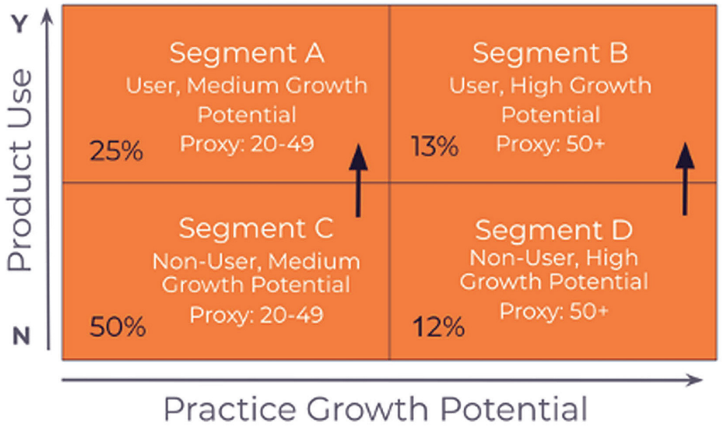
Most biologic drugs are identified with an HCPCS, common procedure terminology (CPT), or national drug code (NDC) on the claim. These claims allow for objective utilization measurements of drugs and related products. It also allows for subsequent identification of physicians who administer medications and procedures, medical billing groups, and hospital systems. Diseases are classified in Medicare claims using the 10th revision of the International Classification of Diseases and Related Health Problems, Tenth Revision, Clinical Modification (ICD-10-CM) codes. These ICD-10 codes are embedded in claims and can be used to count the number of unique patients diagnosed or managed by physician NPI and by organization NPI. Based on this information, physicians can be segmented to identify those who would find greater utility from novel treatments and diagnostic procedures to manage and evaluate patients. The PSP model included a set of ICD-10 codes related to all the diseases targeted by GE's diagnostic offerings.

GE's biologic assists in the early diagnosis of a degenerative disorder, which can dramatically improve outcomes. Finding physicians treating the symptoms of the disease allows for faster diagnosis and efficient marketing efforts, allowing more relevant product-related information to be made available to these physicians. Based on this objective, the model was defined to predict the number of patients a physician (or account) may have in the future by combining factors like indicated populations, product utilization, and a growth factor measured from E&M codes.

Figure 1: PSP Score Formula

$$\textit{Patel—Sahin Proxy Score} = \underset{\textit{Indication 1} \ \cup \ \textit{Indication 2}}{\underset{\uparrow}{\textit{Number of Indicated Patients (n)}}} \ x \ \underset{\frac{\textit{New Patients}}{\textit{All Patients Seen}} \ X \ 100}{\underset{\uparrow}{\textit{\% New Patients (c)}}}$$

Figure 2: Segmentation Table with Physician Allocations Based on a Comparison of Product Use and PSP Scores ≥ 20


Y
Product Use
N
Segment A
User, Medium Growth Potential
25%
Proxy: 20-49
Segment B
User, High Growth Potential
13%
Proxy: 50+
Segment C
Non-User, Medium Growth Potential
50%
Proxy: 20-49
Segment D
Non-User, High Growth Potential
12%
Proxy: 50+
Practice Growth Potential

The intent was to estimate brand utilization by a physician over the upcoming two years. First, we determined a physician's total "Number of Indicated Patients" (n) by counting the unique patients who had claims with ICD-10 codes related to a specific product in question. Then, we counted the occurrence of 99201-related CPT codes to represent "new" patients and the occurrence of 99211-related CPT codes to represent "established" patients. The ratio of new vs established patients multiplied by 100 provides the "% New Patients" (*c*). The PSP score was derived by multiplying % New Patients (*c*) and Number of Indicated Patients (n), which was used to predict brand activation and future utilization. PSP score is defined in the formula in Figure 1.

In order to segment the qualified physicians for marketing action, an easily interpreted model was used. Figure 2 is a 2x2 segment map that divides physicians based on whether they already utilized GE's products (plotted on the y-axis) and used the PSP score to group physicians as high- or medium-growth physician practices (plotted on the x-axis). After that, marketing teams were able to efficiently target providers who were more likely to be receptive to new products.

Results

With previous segmentation efforts, the NRx model identified approximately 8,000 physicians, while the PSP model initially found approximately 80,000 potential providers with positive PSP scores. The goal was to interact with 6,000 physicians by optimizing the marketing field force resources GE had at hand, so a minimum PSP score threshold of 20 was chosen. Physicians with a PSP score of 20 or more were chosen for the segmentation effort and the rest were set aside for this exercise (Table 1) to revisit in the future. With the threshold filter applied, the model predicted approximately 5,600 physicians would have 20 more treatment/diagnostic naive patients

Table 1

Model	NRx	PSP no Filter	PSP Filtered
Filter	All providers (no filter)	All providers (no filter)	Providers with PSP scores ≥ 20
Grand Total	**~8,000**	**~80,000**	**~5,600**

over the next two years. Each of these 5,600 physicians was considered to have the greatest fit and highest probability of response because they met two critical criteria: (1) they treated the diseases of interest, and (2) showed high growth rates of their practice. Furthermore, 13% of physicians displayed high growth rates, with PSP scores of 50+, and did not utilize GE's products, representing immediate opportunities to increase awareness and access.

There were immediate results with the PSP model at the center of the marketing effort design. Figure 2 illustrates the distribution of 5,600 physicians into segments for next-step actions. Pending the segment under which a physician falls, they will have an individualized value proposition. For example, approximately 12% of physicians were found with high growth and no evidence of GE product utilization. This audience represented a huge opportunity for personal and non-personal communication activity.

The GE marketing team received new PSP data in the fall and began immediate communications with each physician with segment-specific and customized messaging. Within a few quarters, there was a noticeable 60%-70% increase in product utilization and revenue. The following four quarters showed 100% revenue growth overall. From GE's perspective, this increase was attributable to the ability to reach the correct audience for prescribing therapeutics and referring patients for diagnostics.

Discussion

This method contrasts the NRx method in several ways. First, the older method assessed prescriptions as an indication of active disease management, whereas the PSP score relied on the presence of E&M billing codes for indicated patients to accomplish the same. The older method exposed how often a physician activates a treatment-naive patient, whereas the PSP score predicted the number of patients with similar indications from the physician's entire patient panel. The superiority of the PSP score was based on the physician coverage of Medicare, the fit of Part B claims for biologics utilization, and the positive changes in the rate of sales associated with switching from the old NRx model to the new PSP model.

The formula (Figure 1) provides a proxy based on Medicare data that is akin to, but not the same as, NRx data from large data providers. While the PSP score combines the number of indicated patients the physician treats and practice growth rate, the novelty stems from the growth factor (c), which can be added to any measurement, including NRx, to create a new prediction. To that end, c is referred to as the Medicare coefficient and can be "subscripted" to deliver physician and practice-level growth rates for subsets of patients (i.e., c_1, c_2, c_3, etc.).

This new physician segmentation protocol was rolled out across the portfolio of products in GE's biologics business. The resulting model helped in understanding new market dynamics, procedure flow, and emerging key opinion leaders (KOLs). PSP scores can be updated quarterly allowing consistent recalibration of marketing resources based on each therapeutic area's quarter and annual trends. Additionally, it is possible to identify which procedures are increasing or decreasing, and which physicians are shifting treatments and diagnostics away from the inpatient setting towards the institution outpatient and ambulatory clinic setting.

The PSP score enabled the ability to visualize KOLs through the lens of utilization and practice growth. It was possible to identify impactful physicians in their regions and to facilitate high-quality discussions between them. These physicians can partner with pharmaceutical manufacturers to highlight the importance of the product, diagnosis, and the impact on patient's lives.

Conclusion

This manuscript discusses a new KPI used by brand marketing teams to assess where the market is going, rather than its current state. A novel Patel-Sahin Proxy (PSP) Score was derived using Medicare Part B claims data (with a quarterly lag) to produce prescriber profiles and make sales predictions to help marketing professionals allocate resources. The new method showed a 100% increase in sales for a brand compared to the older method using NRx data from prescription claims vendors. The proposed PSP scoring approach was reliable, easy to communicate, could easily track results quarterly, and helped the pharmaceutical marketing teams to predict new opportunities with maximum efficiency.

About the Authors

Mert Sahin *is a marketing and commercial executive with 14+ years experience in leading growth, innovation, and transformation drawn from extensive international experience in Life Sciences, Medical Devices, and Pharmaceutical industries. During this initiative, Mert served as Senior Director of Regional Marketing, and later as the Chief Marketing Officer of GE Healthcare, Imaging US/CAN. He has leadership experience on regional and global teams in matrix organizations with a focus on new product launches, branding and positioning, marketing to various customer types (payer, HCP, consumer), KOL management, market research, commercial development, sales force sizing, training, strategic planning, and implementation.*

Ashish Patel *is co-founder of CareSet Systems, a Healthcare Data Science company with the first access to 100% Medicare Part A, B, C and D claims data. He is currently working to decode Medicare claims data for Life Science companies, helping analyze provider referrals and building efficient sales teams. Ashish's expertise from Payors, Providers and Accountable Care Organizations (ACOs) help Pharma support the under-studied Medicare population. An entrepreneur and healthcare data transparency advocate, Ashish also founded the DocGraph Journal, bringing together the Healthcare Data Science community along with Politico and ProPublica to publish data sets for scientific advancement.*

References

1 Mortimer E. LexisNexis Risk Solutions. Using near real-time medical claims to target physicians with newly diagnosed patients. Pharmaceutical Commerce website. https://pharmaceuticalcommerce.com/information-technology/using-near-real-time-medical-claims-target-physicians-newly-diagnosed-patients/

2 Syndicated Loading Definitions. (n.d.). Available at: https://docs.oracle.com/cd/E12102_01/books/AnyInstAdm784/AnyInstAdmLifeSciences4.html

3 Lurker N. (2014, November 15). Real-Time Response. Available from: http://www.pharmexec.com/real-time-response

4 Sohrabi B, Vanani I R, Nikaein N, Kakavand S. A predictive analytics of physicians prescription and pharmacies sales correlation using data mining. *International Journal of Pharmaceutical and Healthcare Marketing 2019*; 13: 346-63. doi:10.1108/ijphm-11-2017-0066

5 CMS Cell Size Suppression Policy. [Internet] Research Data Assistance Center; 2017 [cited 2019 July 12]. Available from: https://www.resdac.org/articles/cms-cell-size-suppression-policy

6 Medicare Claims Processing Manual. (n.d.). Available from: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/clm104c26.pdf

7 Fein AJ. Follow the Vial: The Buy-and-Bill System for Distribution and Reimbursement of Provider-Administered Outpatient Drugs. [Internet] *Drug Channels*; 2016 [cited 2019 Nov 7]. Available from: https://www.drugchannels.net/2016/10/follow-vial-buy-and-bill-system-for.html

8 Hammer R. Top 6 Reasons Why Physicians Join ACOs. [Internet] ReferralMD; [cited 2019 May 07]. Available from: https://getreferralmd.com/2015/07/top-6-reasons-why-physicians-join-acos/