

SPRING 2024 | IN THIS ISSUE:

Identifying the Right Key Opinion Leaders in Pharma Using Machine Learning

Physician Engagement Optimization

Improve Customer Experience and Omnichannel Effectiveness through Customer Journey Analytics

Optimizing Launch Excellence: An AI-driven Framework for Precision Engagement

Transfer learning approach to enhanced patient classification using Real World Data

Navigating Uncertainty: Evaluating Risks to Enhance Drug Sales Forecast

Healthcare Provider (HCP) Behavior Assessment

Data X Ray – A Novel Tool that Assesses Data Sources Using ChatGPT

Unleashing the Power of Rep Notes



Pmsa

PHARMACEUTICAL MANAGEMENT
SCIENCE ASSOCIATION

Table of Contents

Identifying the Right Key Opinion Leaders in Pharma Using Machine Learning: Creating a Data-Driven Process for Impactful Outcomes	5
<i>Nikhil Jain, Executive Director – Analytics & Data Science, ProcDNA Analytics; Shubham Agrawal, Engagement Lead – Analytics & Data Science, ProcDNA Analytics; Tejaswini Salunkhe, Data Scientist, ProcDNA Analytics</i>	
Physician Engagement Optimization: Reinforcement Learning-based Omni-Channel GenAI approach for Maximizing Email Open Rates and embracing Representative preferences to target HCPs	17
<i>Ashish Gupta, Senior Data Scientist, Omnichannel Analytics, Genpact</i>	
Improve Customer Experience and Omnichannel Effectiveness through Customer Journey Analytics	24
<i>Jingfen Zhu, Ph.D., Chief Science Officer, Genpact; Rakesh Sukumar, Senior Manager, Genpact; Ankit Majumder, Data Scientist, Genpact</i>	
Optimizing Launch Excellence: An AI-driven Framework for Precision Engagement	34
<i>Christel Chehoud, Johnson & Johnson; Xiaoyang Meng, Johnson & Johnson; Manish Kumar, ZS; Kaushik Chafekar, ZS; Mehul Singh, ZS</i>	
Transfer learning approach to enhanced patient classification using Real World Data	43
<i>By Aniruddha Mitra, Atharv Sharma, Arkojyoti Das, Arrvind Sunder, Srinivas Chilukuri, Shubhendu Trivedi</i>	
Navigating Uncertainty: Evaluating Risks to Enhance Drug Sales Forecast	59
<i>Rajat Parmar, Manager, Viscadia; Rishabh Bawa, Consultant, Viscadia; Himanshu Khandelwal, Associate Consultant, Viscadia; Varun Singh, Associate, Viscadia; Sakshi Devi, Associate, Viscadia</i>	
Healthcare Provider (HCP) Behavior Assessment: Identifying latent subgroups of HCPs and Salesforce eSales Aid Impact Analysis	74
<i>Sachin Ramesh, Manager, Genpact; Karthick Karuppusamy, Assistant Vice President, Genpact</i>	
Data X Ray – A Novel Tool that Assesses Data Sources Using ChatGPT	90
<i>JP Tsang, PhD and MBA (INSEAD), President of Bayser Consulting; Maylis Larroque, Senior Consultant, Bayser Consulting</i>	
Unleashing the Power of Rep Notes: Extracting Actionable Insights through NLP driven Analysis with LLMs and Generative AI	103
<i>Arvind Balaji Gunasekaran, Principal Consultant – Data Science, Trinity Lifesciences; Nagasubramanya Nabha, Vice President – Data Science, Trinity Lifesciences</i>	
Effectively predict patient discontinuation with AI and opportunities for Rx switching	113
<i>Ryan Sun, Stuart Kamin</i>	

*Official Publication of the Pharmaceutical
Management Science Association (PMSA)*

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

The key points in achieving this mission are:

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

Please submit correspondence to:

**Pharmaceutical Management Science
Association**

1024 Capital Center Drive, Suite 205
Frankfort, KY 40601
info@pmsa.org
(877) 279-3422

Executive Director:

Stephanie Czuhajewski, CAE
sczuhajewski@pmsa.org

PMSA Board of Directors

Igor Rudychev
President

Nuray Yurt, Merck
Vice President

Nathan Wang, Janssen Pharmaceuticals
Professional Development Chair

Aditya Arabolu, Pfizer
Research and Education Chair

Mehul Shah, Bausch Health
Marketing Chair

Tatiana Sorokina, Novartis
Digital Engagement Chair

Jing Jin, AstraZeneca
Program Committee Lead

Srihari Jaganathan, UCB
Program Committee

Nadia Tantsyura, Boehringer Ingelheim
Global Summit Chair

Vishal Chaudhary, Amgen
Executive Advisory Council

PMSA Journal: Spotlighting Analytics Research

PMSA is pleased to announce the 2024 Journal of the Pharmaceutical Management Science Association (PMSA), the official research publication of PMSA

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*:

Simon Fitall, Tudor Health
Ewa Kleczyk , Target RWE
Sudhakar Mandapati. SRI
Ashish Patel, CareSet
Igor Rudychev, Horizon Therapeutics
Mehul Shah, Bausch Health
Tatinana Sorokina, Novartis
JP Tsang, Bayser
Devesh Verma, Axtria

Editor: Aditya Arabolu, Pfizer

Identifying the Right Key Opinion Leaders in Pharma Using Machine Learning: Creating a Data-Driven Process for Impactful Outcomes

Nikhil Jain, Executive Director – Analytics & Data Science, ProcDNA Analytics; Shubham Agrawal, Engagement Lead – Analytics & Data Science, ProcDNA Analytics; Tejaswini Salunkhe, Data Scientist, ProcDNA Analytics

Abstract: Key Opinion Leaders (KOLs) are experts who provide valuable insights into the unmet needs and preferences of the target patient market as well as have an influential impact on the decision-making process of other healthcare professionals. They are often involved in clinical trials, research publications, educational outreach, thought leadership, guideline development, and policy-making initiatives. KOLs offer immediate and actionable insight that can significantly impact brand awareness and inform strategic campaign planning. Thus, identifying KOLs is crucial for pharmaceutical and biotechnology companies, as the KOLs can provide valuable insights that these companies can leverage to design and execute effective business strategies across R&D and commercial.

However, identifying KOLs is not a straightforward task, as it requires collecting and analyzing data from multiple sources such as clinical trials, research papers, media articles, patient claims, and other healthcare databases. Therefore, a comprehensive and systematic approach is needed to integrate and process the data from various sources and identify the KOLs that match the specific business objectives and criteria.

In this paper, we propose a methodological and standardized approach to identify KOLs in any therapeutic area using machine learning algorithms leveraging data from various sources. We have described the main steps and challenges involved in the process, as well as the possible approaches that can be followed to achieve the desired business outcomes.

Keywords: Machine Learning, Key Opinion Leader, KOL, Real-World data, Feature Engineering, Digital Opinion Leader

1. Background:

1.1 Importance of KOLs:

KOLs are subject matter experts who are integral in driving medical breakthroughs. They create awareness about a drug's Mechanism of Action and lend credibility to products in the healthcare community. They have a wealth of knowledge in their respective fields due to extensive experience in managing diverse patient cases, adept at diagnosing multiple

diseases, contributing to key medical research, and thus, holding significant influence over their peers and patients. They act as a strong bridge connecting patient needs with medical R&D efforts.

Thus, pharma companies are always keen to engage with KOLs to position their drug(s) in the market through proven drug characteristics like newer and effective therapies, better efficacy, fewer side effects, ease of use, etc. which can lead to better adoption and positive

patient outcomes. By extension, this also helps pharma companies commercially in driving up sales and revenue. For example, as per published data in industry articles¹, companies that had MSLS engage with KOLs before product launch to educate them about the latest healthcare developments, had 1.5 times greater treatment adoption across the opinion leaders' healthcare organizations in the first six months of product launch.

1.2 Challenges with the existing approaches:

iii. Lack of a data-driven technique:

Currently, there is an over-reliance on internal teams across medical, sales, marketing, etc. to gather inputs on opinion leaders in the target market with limited usage of various available data points.

iii. Lack of standardization across

diverse sources: With the increase in number of data sources related to research, drug development, social and general media, etc., collating information is a significant challenge and requires:

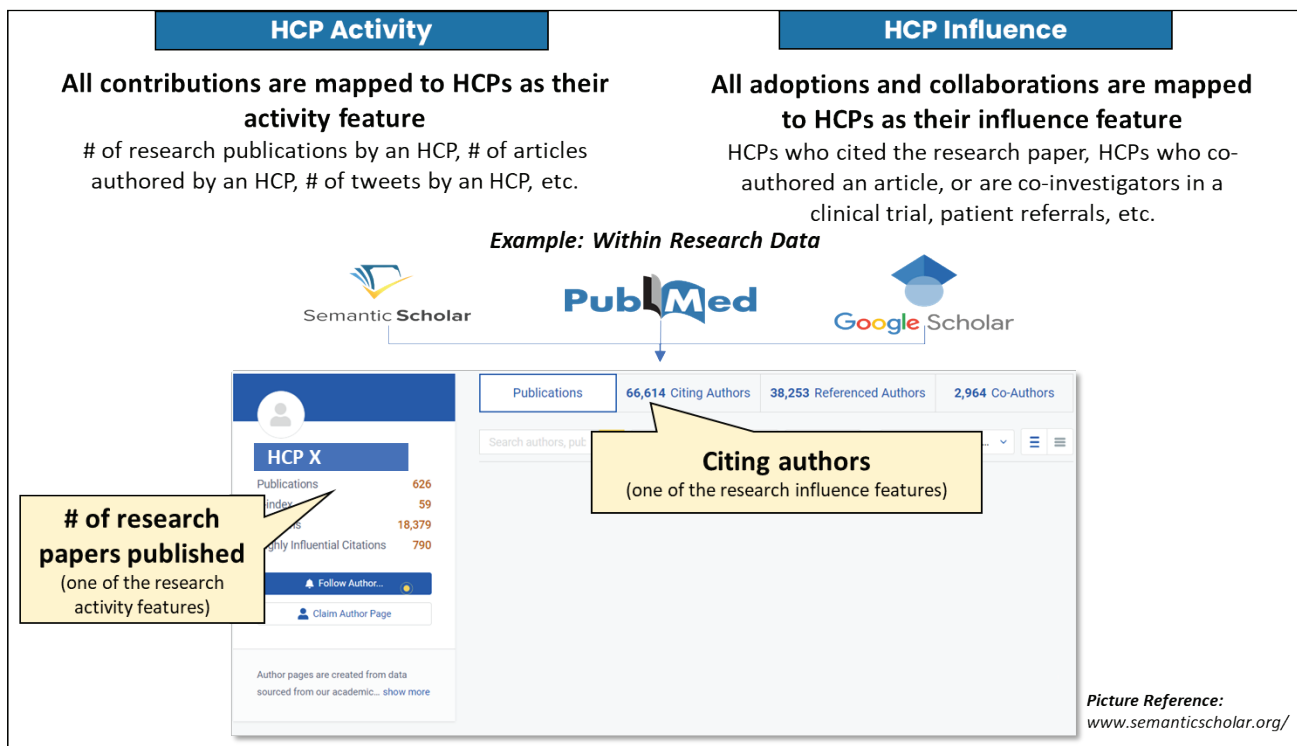
a. Mapping data from different sources to a single HCP using de-anonymization techniques and

b. Filtering through the noise to find the relevant data.

iii. Lack of robust approach to quantify

opinion leaders' impact: Once the required data is gathered, a standard approach needs to be deployed that can measure contributions and sphere of influence for each opinion leader in and across all data sources.

Figure 1: We consider both HCP Activity and HCP Influence metrics for determining KOL status



2. Methodology:

Our KOL identification approach is motivated by the fact that HCPs who have significant contributions as well as high influence in the healthcare community can be considered Key Opinion Leaders. These contributions could include research, publications, innovations, or any other activity that demonstrates expertise. Their level of influence is determined through the adoption of their work and collaborations with other HCPs, suggesting that their work is recognized and valued by their peers. This has been briefly summarized using an example in Figure 1.

All contributions are mapped to HCPs as their activity metric, and all adoptions as well as

collaborations are mapped as their influence metric. These metrics are used for feature engineering of each data category to be input into the machine learning models.

3. Data Preparation

3.1 Identifying Data Categories & Sources:

KOLs are known for expressing their opinions, thought leadership, and research work through scientific publications, national/international conferences, various media sources, etc. They also actively participate in various clinical trials and hold executive positions in professional associations. Thus, identifying the right data categories and then gathering relevant data

Table 1: Data categories considered for KOL identification

S.No.	Data Category	Description of Data Category
1	Research	Scientific research conducted by HCPs
2	Clinical Trial	Clinical trials across all phases of study undertaken by HCPs as principal investigators/ or supporting roles
3	General Media	Articles and opinions written by HCPs that are published in leading healthcare journals and different media outlets
4	Social Media	Tweets (X) or posts (LinkedIn, other platforms) that are written, shared, or liked by HCPs
5	Payment	Payments received by HCPs for research work, speaker programs, consulting services, and other engagements (as part of the <i>Sunshine Act</i>)
6	Thought Leadership	Key positions held by HCPs in various professional associations including top healthcare societies, conferences, congresses, guideline committees, etc.
7	Patient Treatment	Patient claims and patient referrals from/to other HCPs

from the appropriate sources is an important prerequisite for our identification exercise. For our exercise, we have identified and gathered data from multiple sources for seven broad data categories (Table 1) as shown below:

As we further conducted a deep dive into each data category to identify the source(s), we assessed their data on its Availability, Accessibility, and Stability i.e. selecting data repositories that provide:

- i. Availability of comprehensive data coverage across HCPs
- iii. Accessibility to relevant data fields for modeling
- iii. Stability of data format through historical feeds

3.2 Mapping consolidated data to HCPs and indication:

One of the challenges with data gathering is that HCP(s) and Indication(s) might not be explicitly specified within a data source but embedded in accompanying textual information. Thus, a core process for the KOL identification exercise is to map each data information to relevant HCP(s) and Indication(s). Multiple machine learning algorithms are available to conduct this exercise.

For example, to address indication mapping, we employed the BioBERT model, a specialized language representation model. Initially pre-trained on English Wikipedia and BooksCorpus, this model was further fine-tuned using biomedical corpora from PubMed and PMC, enabling effective mining of biomedical text.

3.3 Defining the HCP Cohort:

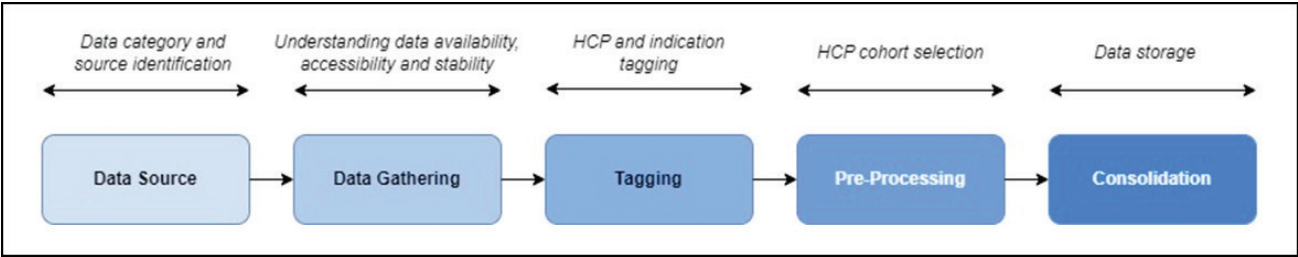
As the next step of data preparation, we filtered the data for the appropriate HCP cohort. A few of the selection criteria were:

- i. **HCP Specialty:** To understand the broader outlook on HCP’s specialization and medical interest.
- ii. **Activity Timeframe:** To identify if the HCP(s) have been active in the last few years (e.g. to consider HCPs in the cohort if they have contributed across any data category in the last 2 years)

3.4 Data Preparation Summary:

We identified and gathered data from different relevant sources across each data category and applied HCP/indication mapping, and pre-processing steps to store mapped data of relevant HCP cohorts in a standard format. This was further utilized by the machine learning models to identify KOLs (Figure 2).

Figure 2: Data Preparation flow for each data category



4. Machine Learning Modeling:

4.1 Feature Engineering:

We looked at HCP Activity and Influence features for KOL identification.

i. HCP Activity Features: For each HCP within each data category, we selected activity features that were used in the Machine Learning model. A reference list of features (not exhaustive) per category is listed below in Table 2.

Table 2: Features considered for HCP Activity across data categories

Data Category <i>(source example)</i>	List of Features <i>(not exhaus- tive)</i>	Description
Research (PubMed etc.)	Publications	Number of publications done by the HCP
	Citations	Number of publications in which the HCP is cited
Clinical Trials (Clinicaltrials.gov)	Studies	Number of studies an HCP is involved in (either as one of the PIs or as a participating HCP)
	Sponsors	Number of unique sponsors for clinical trials
General Media (OncLive, Healio, JAMA, etc.)	Articles	Number of articles authored by the HCP
	Mentions	Number of articles in which the HCP is mentioned
Social Media (Twitter [X], LinkedIn, etc.)	Tweets	Number of tweets tweeted by the HCP
	Retweets	Number of retweets of the HCP
Payments (openpaymentsdata.cms.gov)	Total Payment	Sum of payments received by the HCP
	Institutes	Number of unique institutes from which the HCP has received payment
	Drugs	Number of unique drugs for which the HCP has received payment
Thought Leadership (NCI, AAN, AAD, FDA, etc.)	Type	Type of leadership entity corresponding to the HCP (e.g., Scientific Body, Conference, Congress, Symposium, and Journal)
	Position	Position(s) held by the HCP within the leadership entity (e.g., Board Member, Committee Member, Speaker, etc.)
Patient Treatment (Secondary claims data etc.)	Treatments	Number of indication-specific prescriptions/procedures conducted by the HCP
	Diagnosis	Number of indication-specific diagnoses done by the HCP

ii. HCP Influence Features: To quantify HCP influence, we used network graph analysis within each data category. These help us understand:

- Direct influence on associated HCPs
- Efficient exchange/flow of information between clusters of HCPs through these HCPs

Graphs are mathematical structures used to model many types of relationships and processes in information systems. They are primarily composed of:

- Nodes: which represent system entities (e.g. each HCP is a single node in our case)
- Edges: which represent relationship(s) between entities. (i.e. HCPs connectivity with other HCPs)

Specifically, to identify the important nodes in any graph, we use the concept of centrality. Centrality measures the degree to which a node is central to the network, i.e., it gives us an idea about the number of direct connections a node has or how many other nodes it serves as a link. There exist four centrality measures, each capturing different aspects of a node's significance within a network. These are:

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality
- Eigen Vector Centrality

In the context of our exercise, we selected and employed the following centrality measures to create influence features for each HCP within each data category: Degree Centrality and Betweenness Centrality (refer to Table 3 for their definition and significance).

Figure 3: Centrality Measures² – a visual representation of nodes in a network

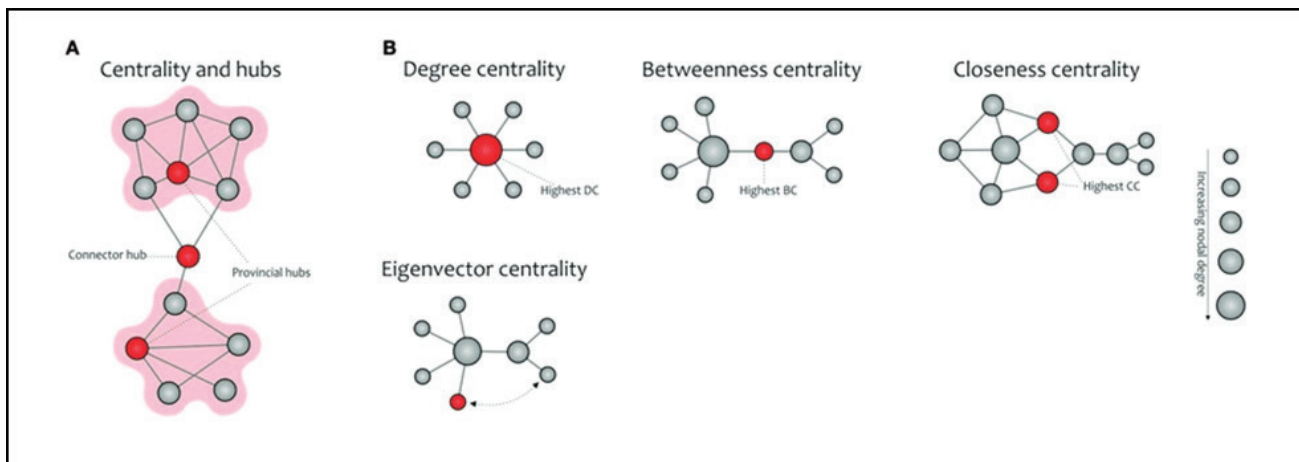


Table 3: Centrality Measures – Importance in KOL identification

List of Features	Description	Significance
Degree Centrality	The degree index denotes the number of connections an HCP may have in a network.	It tells us how many direct “one hop” connections each HCP has to other HCPs. A higher degree implies a faster flow of information within a cluster
Betweenness Centrality	The betweenness index tells us which nodes are ‘bridges’ between other node connections in the network.	It shows which HCPs are ‘bridges’ between clusters of HCPs within a network. Higher centrality implies higher impact on the flow of information within the network

iii. Normalization & Standardization:

After the features have been created for activity and influence, we converted these into ordinal scores with values ranging between 0 and 1. These scores are obtained by normalizing and standardizing the observations of the above features. These scores are then utilized by the Machine Learning model that is elaborated in the next section.

[Normalizing: If x denotes the observation of the feature, then the normalized observation would be $\frac{x - \mu}{\sigma}$ where μ and σ denote the mean and standard deviation of the observations of the feature. Normalizing converts the distribution of observations to one with 0 mean and unit standard deviation.

Standardizing: If x denotes the observation of the feature, then the standardized observation would be $\frac{x - \min}{\max - \min}$ where $\min$ and $\max$ denote the maximum and minimum observation corresponding to the feature. Note that standardized scores lie between 0 and 1.]

4.2 Machine Learning Models Overview:

As we worked with unlabeled data, we utilized unsupervised machine learning algorithms for our KOL identification exercise. The models were filtered which offered anomaly detection techniques since KOLs are HCPs who are outliers due to their significant contributions and influence compared to other HCPs within each data category.

We compared two different types of unsupervised models - the Isolation Forest Model (iForest) and the Gaussian Mixture Model (GMM). The iForest and GMM are both anomaly detection techniques but use different approaches for learning and identifying anomalies. By applying these models separately to each data category’s final output, we attempted to filter Key Opinion Leaders within each data category to assess the performance of both models.

- iForest:** To effectively discover abnormalities, the iForest model makes use of recursive partitioning and random splitting. Anomaly scores are determined using the average path lengths of those anomalies.
- GMM:** This is a probabilistic model that assumes a mixture of Gaussian distributions and is used to generate data. For each Gaussian component, it applies the Expectation-Maximization (EM) algorithm to iteratively estimate parameters and assign probabilities to data points.

4.3 Machine Learning Models Evaluation:

The two models classified the HCPs into two categories:

- Category A (an anomaly) represents

individuals identified as Key Opinion Leaders (KOLs).

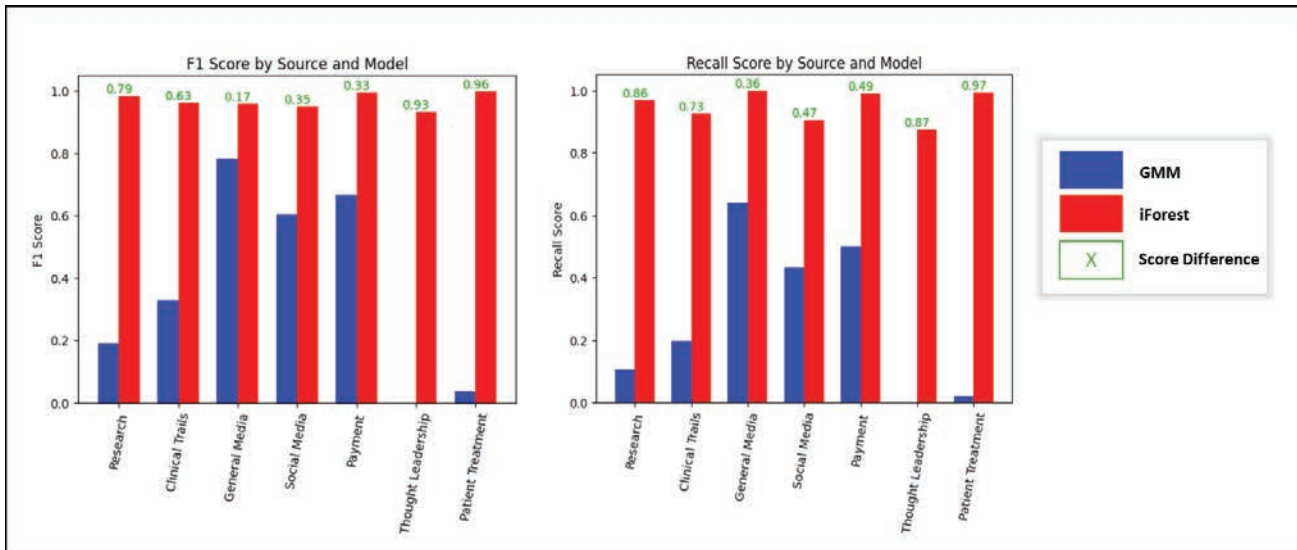
- ii. Category B (not an anomaly) represents individuals who were not classified as Key Opinion Leaders (KOLs).

The performance of the models is evaluated using 'F1 score' and 'recall' across various sources. The F1 score provides a balance between precision and recall, while recall focuses on identifying true anomalies. Visualization of F1 scores and Recall distributions (Figure 4) across sources reveals that iForest has a high precision-recall balance.

iForest Advantages:

- i. **High F1 Scores:** iForest consistently achieved high F1 scores, emphasizing its precision-recall balance.
- ii. **High Recall:** iForest exceeded in identifying anomalies, ensuring a larger proportion of true anomalies are detected.
- iii. **Consistency Across Categories:** iForest has consistently high score across categories demonstrating robustness.

Figure 4: Comparison of F1 Score (left) and Recall (right) for GMM and iForest models



4.4 Machine Learning Model Selection:

Based on the above-observed model performance metrics, we selected and utilized the iForest model for our KOL identification exercise. Its notable qualities, such as a consistently higher recall rate, a well-balanced precision-recall performance, and proven resilience across a variety of data categories, constitute the basis of selection.

4.5 Working of Isolation Forest:

i. **Isolation Tree Construction:**

Multiple isolation trees are constructed independently. Each tree is built by randomly selecting a subset of features. The recursive partitioning process continues until each data point is isolated in its leaf node. The randomness in feature selection contributes to the diversity of trees.

ii. **Path Length Calculation:** For each data point, the path length within each isolation tree is computed. The path length is the count of edges (splits) traversed from the root of the tree to the leaf node containing the data point. Anomalies are expected to have shorter paths due to their isolation.

iii. **Anomaly Score Calculation:** The anomaly score for each data point of each source is determined as the average path length across all isolation trees. Anomalies,

being isolated with shorter paths, will have lower average path lengths. Therefore, instances with lower scores are considered more likely to be anomalies.

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where:

- $s(x, n)$ is the anomaly score for data point x ,
 - $E(h(x))$ is the average path length for x across all trees,
 - $c(n)$ is the average path length for an unsuccessful search in a binary tree of n points.
- iv. **Normalization:** To facilitate better interpretation and comparison of anomaly scores, they are normalized to a standard range, commonly $[0, 1]$.
- v. **Thresholding:** Instances with anomaly scores above the threshold are considered as KOLs, while those below are regarded as non-KOLs.

Figures 5 and 6 illustrate the implementation and intuition of the iForest Model respectively.

Figure 5: Implementation of Isolation Forest Model³

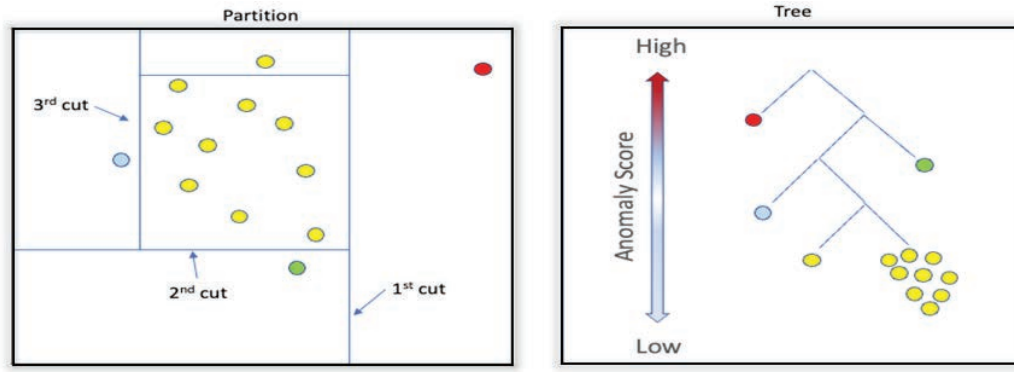
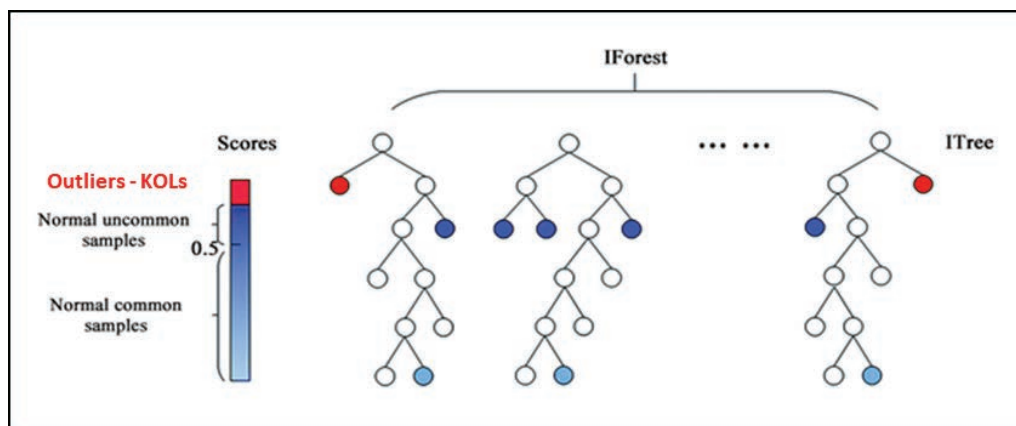


Figure 6: Intuition of Isolation Forest Model⁴



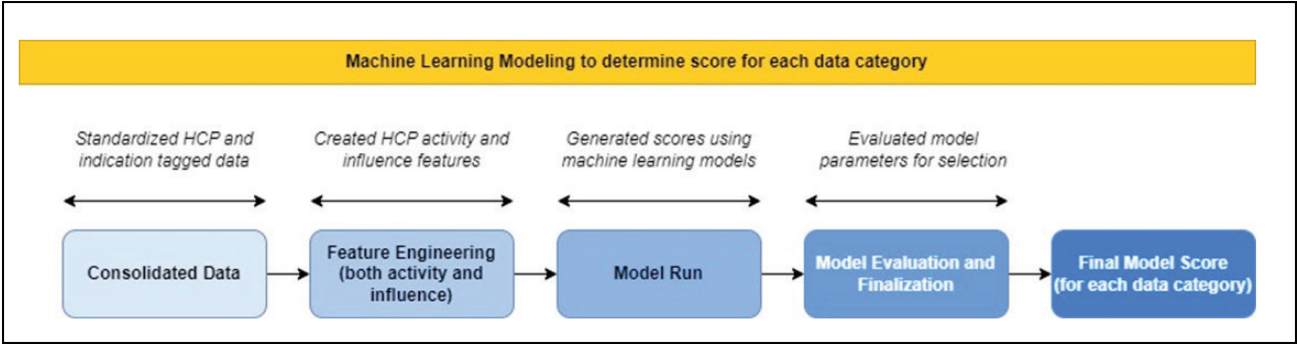
4.6 Machine Learning Model Score Generation:

Normalized features for each data category were used as inputs in the machine learning models to generate scores for each HCP to identify the Key Opinion Leaders (KOLs). The generated scores also help us to identify the relative importance/ranking of each KOL within each data category.

4.7 Machine Learning Modeling Summary:

For our modeling exercise, we created **HCP Activity Features** and **HCP Influence Features** across each data category. These category-specific features were then standardized/normalized and ingested in two different machine learning models (iForest and GMM) to derive relative scores for each HCP within each data category. Through comparison of the modeling performance parameters, we also established that iForest proved to be an effective modeling technique for KOL identification and was used to generate scores for each data category (Figure 7).

Figure 7: Summary of machine learning modeling

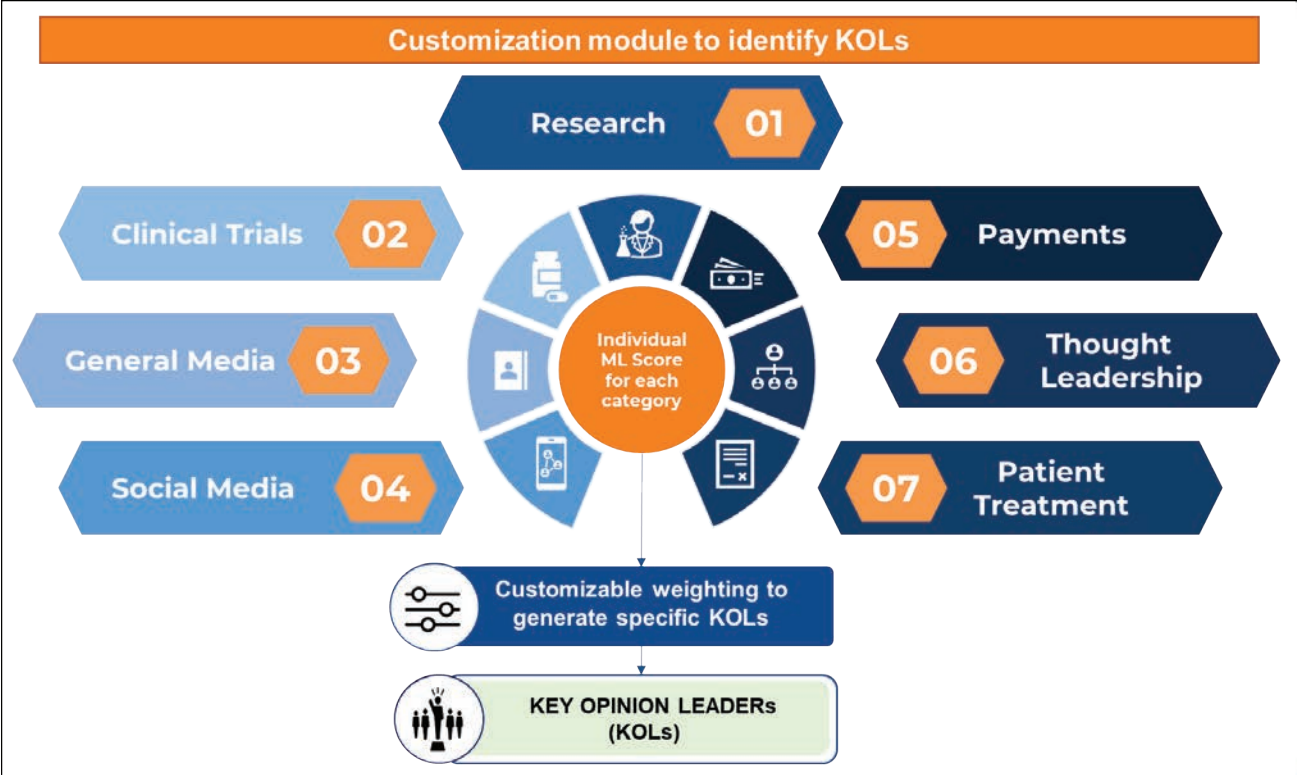


5. Customization as per business objective:

To cater to the evolving business needs, we developed a custom scoring mechanism that provides variable weights to scores of each data category before generating the final score (Figure 8). This customization feature empowers the user to modulate the scores’

importance and generate specific KOLs that are relevant to business requirements. For example, if the business objective is to increase engagement towards digital KOLs, a higher weight could be given to social and general media data categories relative to other data categories to generate a KOL list which is a reflection of Digital Opinion Leaders (DOLs).

Figure 8: Final KOL score calculation based on customization



6. Conclusion and Future Scope:

In this paper, we have presented a detailed methodology to identify KOLs using Real-world datasets and advanced machine-learning techniques. We also talked about an additional customization module that we have developed on top of this to cater to evolving business needs. Our approach could be used to identify KOLs across any geographic and therapeutic area.

The outcomes of our current modeling exercise can be further processed to create additional summaries that can add value to the end user for generating insights. One such domain of interest could be defining the “geographical archetypes” of KOLs. This new metric could potentially quantify the sphere of influence of each KOL within a defined country or geography. Some of the key features that could be considered are – the distance between KOLs and HCPs interacting/connected with them, the number of times they interacted with each other, and the number of sources they have interacted through. Using these features, HCPs could be grouped into three clusters, such as National, Regional, and Local, by using unsupervised machine learning techniques like clustering algorithms.

About the Authors

Nikhil Jain is an Executive Director with ProcDNA where he leads the Analytics & Data Science Excellence Center, focusing on solving challenging healthcare problems based on a variety of datasets using advanced analytics and machine learning methods. He has 12+ years of experience working with the Pharmaceutical, Bio-Tech, and Medical Devices industries. He has multifaceted expertise in areas including product launch preparedness, sales force effectiveness, commercial development, training, and strategic planning. He received his MBA from the Indian School of Business.

Shubham Agrawal and **Tejaswani Salunkhe** represent the Analytics & Data Science Excellence Center in ProcDNA. The team is an expert in designing and delivering AI/ML-enabled innovative healthcare solutions such as KOL identification and activation, clinical trial design and efficacy assessment, and prediction problems like patient events, next-best-action (NBA), and early adopters.

References:

1. Nick Paul Taylor. Pre-launch KOL talks linked to 1.5-time boost to drug uptake. 2023; Available from: <https://www.fiercepharma.com/marketing/pre-launch-kol-talks-linked-15-time-boost-drug-uptake>
2. Farzad V Farahani, Waldemar Karwowski, Nichole R Lighthall. Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review. 2019; Figure 5. Basic concept of network centralities. Available from: <https://www.frontiersin.org/articles/10.3389/fnins.2019.00585/full>
3. Chris Kuo. Handbook of Anomaly Detection with Python Outlier Detection — (4) Isolation Forest. 2019; Figure (A): iTree. Available from: <https://towardsdatascience.com/use-the-isolated-forest-with-pyod-3818eea68f08>
4. Yupeng Xu, Hao Dong, Mingzhu Zhou, Jun Xing, Xiaohui Li, Jian Yu. Improved Isolation Forest Algorithm for Anomaly Test Data Detection. 2021; Figure 1. Normal and abnormal points in iTrees. Available from: <https://www.scirp.org/journal/paperinformation?paperid=111345>

Physician Engagement Optimization: Reinforcement Learning-based Omni-Channel GenAI approach for Maximizing Email Open Rates and embracing Representative preferences to target HCPs

Ashish Gupta, Senior Data Scientist, Omnichannel Analytics, Genpact

Abstract: For an effective Omni-channel Marketing campaign, the four phases of Content creation, Campaign execution, Campaign analysis and Representative feedback should be a continuous iterative cycle with the right balance of choosing Channels, relevant Content and Time of delivery.

This paper presents a Reinforcement Learning (RL)-based Omni-channel Generative AI approach designed to enhance open rates of emails sent to physicians. Leveraging historical engagement data and employing continuous learning, the proposed system utilizes GenAI to dynamically create and select the most promising subject lines and preview texts (first visible points on a physicians' device) using exploration and exploitation technique in real time, thereby increasing the probability of physician engagement.

Moreover, this paper presents a study on embracing representative preferences to improve the effectiveness of suggested calls for physicians. In existing system, these calls are suggested for a set of physicians via Omnichannel Orchestration engine every week and are kept flexible for representatives by having dismissal options.

But each representative also has their own preferences to visit a set of HCPs together that are based on either distance, prescriber type, specialty type, prescription-based segments, bonding with physicians or other factors. These set of HCPs visited together should be considered as feedback to engine which can be obtained by analyzing representative past activities.

Keywords: Omnichannel optimization, maximize email open rates, Representative feedback analysis, improving physician engagement, Generative AI, Reinforcement Learning

I. INTRODUCTION

To optimize the HCP journey, the ideal flow of touchpoints should start with low-cost channels (Representative Emails / App Alerts / 3rd Party Email) that can grab physicians' attention on content and are flexible with their reading time followed by calls / events and then move to follow-up emails. One of the metrics mostly

used to evaluate emails is CTR (Click-through rate) with an industry standard of nearly 2%. But another metric "Open rate" has greater impact on physician engagement. It will not matter how exceptional the quality of content is until the physician opens the email.

Viewing from frame of physician's device, the first three things visible to them are

sender's name, subject line, and preview text. Increasing the open rates can only be achieved by optimizing the subject line and preview text. Physicians preferred type of emails are modelled over historical content engagement data (such as Informational, Action oriented, Question-Based, Personalized, Urgent or time-sensitive, Announcement or Update, Invitation or event related). Based on the type of emails, Gen AI generate multiple versions of the subject line and summarize the email body to get different versions of preview text.

Using Reinforcement Learning - Thompson Sampling Bandit, the system is able to deliver the most engaged version to a maximum number of physicians by utilizing an explore & exploit strategy in real time.. The system ensures that there is minimal loss of touchpoint usage over non-engaging content. Physician Survey and their feedback/questions becomes a source for representatives to prepare for calls on relevant content with them. Moreover, prompt responses from representatives to similar physicians will build physicians' trust in the brand.

From another standpoint, a study conducted for a client aimed to enhance the effectiveness of the Orchestration engine calls by proposing the inclusion of a preferred set of healthcare professionals (HCPs) for each representative within the same week. As per data, representatives were doing calls on their own before being suggested by Orchestration engine and dismissing them. The study comprises the following flow to validate representatives' wisdom.

- i. To what extent do representatives exceed the suggested number of calls? How does the call sequence of HCPs differ between self-execution and orchestrated suggestions over a 13-week period?

- ii. Do those reps who are not following the suggestions have a positive impact on TRx?
- iii. What can we learn from the suggestion-execution pattern to improve future suggestions?
- iv. Are there any set of HCPs called out together by reps within week?
- v. Is the completion rate better for the grouped HCPs?
- vi. Why certain HCPs were called out together?

II. METHODS

A. Effectiveness of Reinforcement Learning (RL) and Generative AI Approach

Below process illustrates flow of Gen AI content creation, its approval, best content delivery based on Reinforcement and feedback for continuous Campaign Execution & Analysis.

- i. Campaign consists of multiple programs and different vendors. They are planned to execute during a fixed period with start and end date.
- ii. To continuously refine the content and strategy, the fixed period of the program is divided into sets. To quickly improve the strategy, the duration of the starting sets should be shorter, while the later ones can have a longer gap. The iteration formula used to calculate 'hours' is shown below. (where days are fixed duration of program, trial sets are based on past study and 'x' initiates from 1 to total trial sets),

$$\text{Days} \frac{x}{\text{Trial Sets}}$$

- iii. Based on physicians' preferred types of emails, which are modelled over historical content engagement data (such as Informational, Action oriented, Question-Based, Personalized, Urgent or time-sensitive, Announcement or Update, Invitation or event related), Gen AI generate multiple versions of the subject line and summarize the email body to get different versions of preview text.

For instance, GenAI generated three versions of subject and preview text with the following business subject line and physicians email preference:

"Subject: GPO Pricing for {Drug class} to Support your Medication Budget."

Version a (Urgent or time-sensitive):

Subject Line: Exclusive GPO Pricing for our {Drug class} - Optimize your Medication Budget.

Preview Text: Save on diabetes management with our affordable {Drug class}.

Version b (Informational):

Subject Line: Enhance Patient Care with Budget - Friendly {Drug class} & GPO Pricing.

Preview Text: Discover how our {Drug class} with GPO pricing can benefit your patients.

Version c (Invitation or event related):

Subject Line: Partnership Opportunity: Access Cost-Effective {Drug class} with GPO Pricing.

Preview Text: Schedule a call to explore exclusive GPO pricing options.

- iv. Subject lines and preview texts go through an email content approval flow for HPRC approval (if needed).
- v. To achieve the best overall email open rate percentage for the available three versions of a program, the version with a higher rate should be triggered more frequently to physicians and version with lower rate should be triggered occasionally. The Thompson sampling bandit method, a type of reinforcement learning, is used to achieve the goal.
- vi. For instance, there are a total of 1000 trials (distinct HCPs) to whom email must be sent among three versions of subject and preview text. Initially, there will be wider curve for each content based on its current Open Rate.
- vii. To choose content for an HCP in an explore-exploit manner, a sample 'open rate' will be chosen from the curve of the three versions, and which ever has the highest rate will be assigned to physicians set. (Refer visualization in https://en.wikipedia.org/wiki/Thompson_sampling)
- viii. As the number of trials (email to physicians) increase, the open rate curve of each version becomes narrower. So, after one point of time the higher open rate content will always be chosen. Technically, with increase in the number of trials, the beta distribution narrows towards the true success rate (probability of opening).
- ix. Additionally, two parallel system detects any anomaly through open rate analysis (sudden spike or drop in opening) and

Subscriber behaviour monitoring (sudden unsubscribing change than normal rate) to avoid any major negative impact.

- x. To complete the loop, customer survey in email, post event survey, physician queries, Star rating in email are taken as physician feedback to promote or pause any content for the next trial set.
- xi. These survey responses and questions from the physicians are summarized by GenAI to identify the trending queries and shared with same therapeutic area representatives before their HCP calls or visits. This helps similar representatives to be well prepared for the queries that were asked by other physicians via emails to other representatives.

B. Improvement with Representative Preferences

Under the study, client data showed that representatives were making physician calls on their own before being suggested by the Omnichannel Orchestration engine, which resulted in a majority of dismissals. In this approach, a suggestion matrix and representative execution matrix of the HCPs sequence (A sequence of the HCPs that were called by the rep like Face-to-Face / Web based / Phone) within 13 weeks period were created using dimensions (last visited HCP Id, current visited HCP Id, days difference between these visits). The approach was optimized over the easiest way of creating all HCPs combination resulting in 0.03% of 65 billion sequences. It was achieved by Sparsity Reduction, HCPs Pair order independent assumption, Space/Memory Minimization, Precise Frequency calculation (i.e., HCPs Pair Repetition).

Among various similarity algorithms, cosine similarity was found to be an ideal fit and was used to evaluate the similarity between the suggestion and execution matrix.

To validate the wisdom of representatives, it was important to analyze the positive impact on TRx by those representatives who were not following the suggestions.

Reps were divided into two dominant groups: Grow and Decline, using hyperparameter-tuned decision tree algorithms, regression, clustering, SVD (Singular Value Decomposition), and quartile binning techniques, as shown in the results section.

The analysis suggested better performance by representatives who took actions on their own, resulting in an understanding of the set of preferred HCPs called out together by them in the same week. Certain conditions and association rules (FP-Growth) were applied over the past execution data to create list of HCPs group preferred by each representative.

This led to another question: whether the completion rate was better for the grouped HCPs (or rules). To generalize the performance of grouped HCPs, a single performance metric (named as the association ratio) was created for each representative's group, as shown in Figure 1.

After calculating completion rates, the question arises as to what is common among these HCPs in a group. Four major factors are hypothesized that influence a representative's preferences.

These factors include the minimum distance between the HCPs, prescription-based segments (Tier 1, Tier 2, Tier 3, Tier 4), prescriber types (Physician, Physician Assistant, Nurse Practitioner, and others), and specialty types (Therapeutic area specialist, Family Medicine, General Practice, Internal Medicine, and others). While the actual distribution was available in the data, multinomial probability was used to obtain the expected distribution of all HCPs belonging to the same category (such as the same segment), as shown in Figure 2.

Figure 1

Association Ratio for the rules are based on conditional probability concept.

$$\begin{aligned}
 &\text{Performance of suggestions (Association Ratio)} = \frac{\text{Completion rate of HCPs when grouped together}}{\text{Completion rate of HCPs when they were NOT grouped together}} \\
 &\text{Suppose, we have a rule [HCP_1, HCP_2]} = \frac{P(\text{HCP_1} \cap \text{HCP_2})}{\sim P(\text{HCP_1} \cap \text{HCP_2})} \\
 &\text{Here, in } \sim P(\text{HCP_1} \cap \text{HCP_2}), \text{ both HCP_1 and HCP_2 will be independent of each other} \\
 &\quad \frac{P(\text{HCP_1} \cap \text{HCP_2})}{P(\text{HCP_1}) * P(\text{HCP_2})} \\
 &\quad \text{In an ideal week, if both are not together, then one will be completed and other dismissed}
 \end{aligned}$$

If the association ratio is significantly greater than 1, it indicates that completion rates of HCPs grouped together are higher.
 If the association ratio is significantly less than 1, it indicates that completion rates of HCPs not grouped together are higher.

Figure 2

If HCPs visited by Rep 101 fall into different prescription-based tiers,

Segment	Tier 1	Tier 2	Tier 3	Tier 4
HCPs Distribution	40%	25%	21%	14%

Suppose for a group [HCP1, HCP2, HCP3], the probability of all HCPs in a group that belongs to same segment can be calculated,

Using Multinomial Probability distribution,

$$P_r = \frac{n!}{(n_1!)(n_2!)\dots(n_x!)} P_1^{n_1} P_2^{n_2} \dots P_x^{n_x};$$

$$\begin{aligned}
 \text{Probability of all HCPs in same segment} &= P(\text{Tier 1} = 3, \text{Tier 2} = 0, \text{Tier 3} = 0, \text{Tier 4} = 0) + P(\text{Tier 1} = 0, \text{Tier 2} = 3, \text{Tier 3} = 0, \text{Tier 4} = 0) \\
 &+ P(\text{Tier 1} = 0, \text{Tier 2} = 0, \text{Tier 3} = 3, \text{Tier 4} = 0) + P(\text{Tier 1} = 0, \text{Tier 2} = 0, \text{Tier 3} = 0, \text{Tier 4} = 3) \\
 &= \frac{3!}{3!} (0.4)^3 + \frac{3!}{3!} (0.25)^3 + \frac{3!}{3!} (0.21)^3 + \frac{3!}{3!} (0.14)^3 \\
 &= 0.09 \text{ or } 9\%
 \end{aligned}$$

III. RESULTS & DISCUSSIONS

After conducting the study, observations from data show rep's behavior deviates from suggestions.

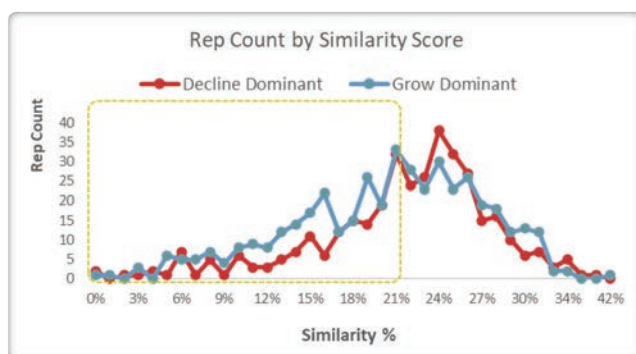
i. Are Reps following suggestions?

Representative execution was no more than 40% similar to suggestions through

the Orchestration engine. This was due to fact that reps are doing more calls, and their HCPs sequence are inconsistent with suggestions. For instance, Orchestration engine recommended to visit HCP A in current week and HCP B in the next week, whereas the representative visited HCP A & B in the same week and dismiss next week suggestion of HCP B. Representatives' behavior differs from suggestions especially with low-writing HCPs.

ii. Are reps' approach better than suggestions?

The reps' approach outperforms the suggestions for the 13-week period (which didn't have calling behavior volatility typically observed during holiday months). Below chart shows that the count of reps who dominate were higher where similarity% between rep execution pattern and suggestion pattern was lower.



iii. Do representatives call a group of HCPs within 7 days?

Yes, most representatives called a group of HCPs together in the 13-week study period. Over 10,000 groups were found for all representatives during the process.

iv. Is completion rate better when HCPs are grouped together?

The data indicates when HCPs are suggested together based on the patterns identified, the completion rate is moderately better.

v. Which factor influences a representative preference?

The data shows that distance, prescription-based segments, prescriber types and specialties all seem to lie behind the HCPs grouping as a representative individual preference.

(vi) Additionally, data showed that not all dismissals were truly made by the representatives which stated the need to train them on choosing the relevant dismissal options.

IV. CONCLUSIONS

In conclusion, the process of choosing content for healthcare professionals (HCPs) in an explore-exploit manner proves to be effective in maximizing open rates. By selecting a sample open rate from the curve of three content versions, the version with the highest rate is assigned to the physicians' set. As the number of trials (physicians) increases, the curve narrows, resulting in the higher open rate content consistently being chosen.

To complete the feedback loop, customer & post event surveys in emails, physician queries, and email ratings are collected as feedback from physicians and used to promote or pause specific content for the next trial set. Additionally, GenAI summarizes the survey responses and physician queries to identify trending topics and shares this information with representatives in the same therapeutic area. This enables representatives to be well-prepared for queries asked by other physicians via emails to other reps, improving overall engagement and satisfaction.

In summary, this comprehensive approach of content selection, feedback collection, and information sharing ensures continuous improvement and optimization of email campaigns, resulting in enhanced physician engagement and response rates.

Based on the results and discussions, it can be concluded that the behavior of representatives deviates from the suggestions provided by the Orchestration engine. The representatives were found to be inconsistent in their

execution pattern and often disregarded the recommended sequencing of calls to healthcare professionals. However, their approach outperformed the suggestions during the 13-week study period. The data also revealed that representatives tend to call a group of healthcare professionals within a 7-day period,

and when HCPs are suggested together, the completion rate improves moderately. Multiple factors together such as distance, prescription-based segments, prescriber type, and specialty were found to influence the representatives' preference in grouping HCPs.

ACKNOWLEDGMENT

I would like to extend my heartfelt appreciation to my reviewers who took the time to review and provide valuable feedback on the paper. Your insightful suggestions and edits have greatly enhanced the quality and clarity of my writing.

ABOUT THE AUTHOR

Ashish Gupta, Senior Data Scientist, Genpact, has over 7 years of experience in Pharma, Consumer Goods & IT industries. With strong belief in taking proactive initiatives, he has actively engaged in solving real-world problems and driving positive outcomes through strategic problem-solving. He has worked on cross function projects involving Marketing, Finance, Sales, Supply Chain & IT.

Improve Customer Experience and Omnichannel Effectiveness through Customer Journey Analytics

Jingfen Zhu, Ph.D., Chief Science Officer, Genpact; Rakesh Sukumar, Senior Manager, Genpact; Ankit Majumder, Data Scientist, Genpact

Abstract: Customer experience is a relatively nascent topic in pharmaceutical industry, and it has rarely been leveraged in omnichannel strategy development. Traditionally, marketers have used tools like Marketing Mix Models and impact analysis to optimize marketing spends at a strategic level. In more recent times, marketing teams are adopting tools like Next Best Action (NBA) models to determine the immediate next touchpoints at a tactical level. However, marketers are now realizing the potential of managing customer experiences as end-to-end omnichannel journeys and the impact that can have on their customer retention and revenue growth. In this paper we propose an analytics framework that leverages past customer interactions data to construct customer journey pathways customized to HCP's awareness and experience with brand. It has been developed and tested using marketing and prescriptions data of a global pharmaceutical manufacturer for its US market. The study's findings were synthesized into a summary of successful and unsuccessful tactics and pathways by stage, as well as recommended pathways for further testing and adoption.

Keywords: Omnichannel, Marketing Mix Models, Impact Analysis, Next Best Action, Customer Journey Analysis, Customer Experiences, Segmentation, Community Detection, Structural Equation Modeling.

Introduction

Pharmaceutical industry has historically relied heavily on sales representatives for promotion of their products. Sales reps often have face-to-face (F2F) interactions with Health Care Professionals (HCPs), providing them with product information, educational resources, coupons, and product samples. This method, though effective, is time consuming, costly, and limited by geographical constraints. Moreover, HCPs are increasingly opting for fewer F2F interactions as these interactions take time and compete directly with their already demanding work schedules. The Covid pandemic has only helped to hasten this trend. Social distancing measures have compelled HCPs to reduce their in-person meetings, forcing pharma companies to focus more on their digital channels.

Digital platforms offer several advantages over traditional channels. They enable companies to reach a larger audience, provide personalized content, and engage with customers in real-time. They also offer valuable data that can be used to gain insights into customer behavior and preferences, thereby allowing for more targeted marketing strategies. However, in the digital space, pharmaceutical companies are compelled to compete for HCP's attention with all commercial entities from within and outside the pharmaceutical industry. Therefore, pharma companies must provide timely, relevant, and personalized content to convey their key messages and engage with HCPs online. In summary, as avenues of brand interactions multiply and competition intensifies in the

pharma industry, offering an exceptional customer experience across all touchpoints can be the decisive factor that separates market leaders from other players.

Customer Journey Analysis (CJA) is the key to managing customer experiences as end-to-end journeys. CJA is a comprehensive approach to understanding the entire process a customer goes through while interacting with a company or brand across all touchpoints, channels, and devices. CJA uses advanced analytics techniques to identify key drivers of

customer behavior, predict customer needs, and optimize engagement strategies. It provides a holistic view of the customer journey, enabling pharma companies to identify key touchpoints that drive conversion and retention. CJA also helps pharma companies to personalize their engagement strategies, deliver relevant content and offers to individual customers at the right time through the right channel. With CJA, pharma companies can optimize their marketing spend, improve customer satisfaction, and drive revenue growth.

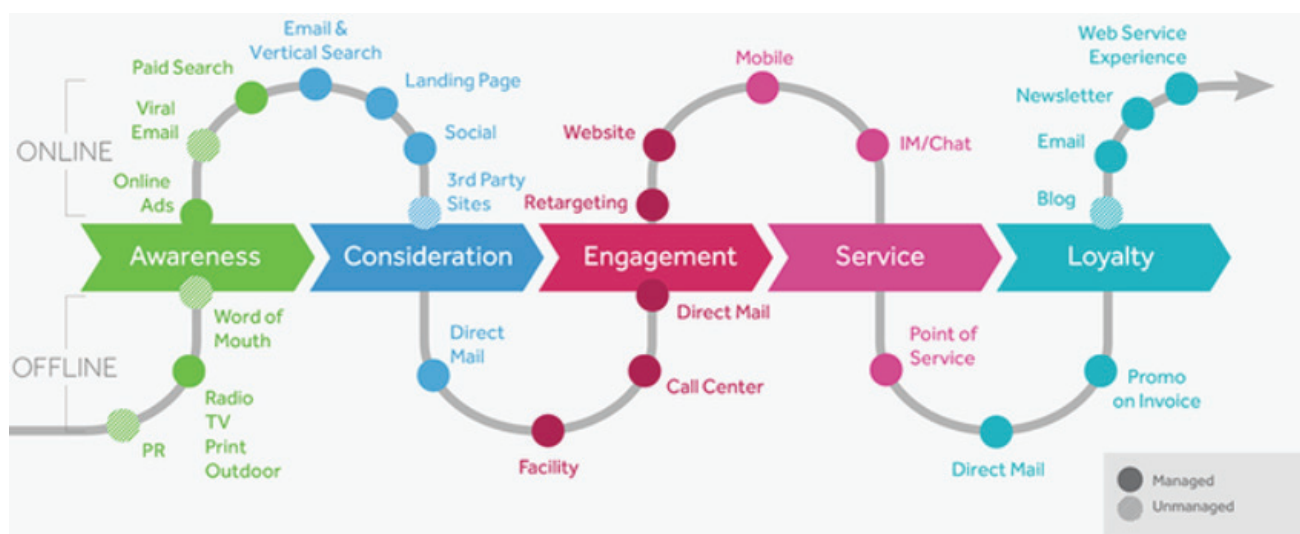


Figure 1: A simple illustration of a customer journey

1.1 Comparison Study of CJA with Marketing Mix Modeling

- CJA evaluates omnichannel marketing efforts from an evolutionary perspective with custom success metrics defined for different stages of the customer journey. Marketing Mix Modeling is applied at a strategic level to determine the required budget allocation for different marketing channels to meet the sales forecast.
- CJA helps identify high value actions along customer journey. Marketing Mix Modeling has one objective (usually sales or prescription) and determines how different marketing channels are contributing to this objective.
- CJA can analyze your marketing data at both content and channel levels while Marketing Mix Modeling is typically done at channel level.

- CJA throws light on campaigns that help HCPs get started or enhance knowledge about a product whereas Marketing Mix Modeling only highlights campaigns with a direct sales impact.

1.2 Comparison of CJA with Next Best Action Analysis

- Customer Journey Analysis strives to embark HCPs on journeys that have been successful in the past. Next Best Action (NBA) is about responding to an immediate trigger such as signing up for an online conference, following up after a sales rep visit or responding to a sales trend break etc.
- CJA helps identify high value actions along the customer journey. These high value actions indicate an increasing level of awareness and experience with the brand. NBA provides a list of trigger points and recommended actions. The trigger points could be positive, such as a request for a Rep visit, or negative such as a falling number of weekly prescriptions.
- Insights from CJA have a wider frame of reference as the insights are valid as long as the customer stays on the same journey stage. NBA recommendations are more tactical and short-lived in nature.
- Customer journey analysis provides insights and recommendations that help inform strategic decision-making. It helps organizations understand customer pain points, preferences, and behavior patterns to make informed decisions about marketing campaigns and customer service improvements. Next best action analysis, on the other hand, focuses on operational decision-making in real-time. It guides frontline employees or automated systems

to take the most appropriate action in the present moment to optimize the customer experience.

Next section provides a high-level roadmap for a customer journey analysis project. Section 3 provides a detailed description of the project steps and their implementation in the case study.

Customer Journey Analysis Project Roadmap

As the first step, identify who your customers are, and what channels are in scope for the project. While these questions may appear trivial at first glance, they require careful consideration of various intricacies. For example, Pharma industry has multiple stakeholders such as patients, payers, HCPs, and administrators who can affect the final buying decision. Similarly, though HCP websites are an important channel for disseminating product information, most of the web traffic data comes with the visitors unidentified and cannot be used directly in a customer journey modeling framework.

Second, the key stages that the customers experience along their end-to-end journey must be identified. For instance, one may define a customer journey as: Awareness > Consideration > Engagement > Loyalty > Advocacy. Segmentation techniques can be used to map your customer to one of these journey stages as demonstrated in the following case study.

Third, success metrics needs to be identified for each stage of the journey. While total number of TRx or NBRx as metrics would directly align with the pharma company's business objectives, they cannot be used as success criteria for customers in the early stages of their customer journey. Metrics that represent an HCP's level of awareness about the product, or their willingness to engage with marketing would be

more suitable for early stages. These metrics must be determined in a data driven way and after discussion and alignment with brand team.

The last step is to generate insights on the main marketing factors that drive the success metrics at each stage of the customer journey. This step assesses performance of marketing efforts at different stages of customer journey. Furthermore, by analyzing the relationships between the marketing drivers, it is possible to develop customer journey pathways that yield the best outcomes at every stage. Marketers should leverage the insights from CJA to customize their marketing campaigns effectively.

CJA Case Study

This case study illustrates how Customer Journey Analysis was applied to a global pharmaceutical client for their US market. The study utilized marketing outreach and engagement data from both traditional and digital channels, in addition to prescription data for the brand and category. Marketing outreach involves the pharmaceutical company proactively reaching out to HCPs with marketing messages, such as promotional emails or digital display campaigns. HCP engagement, on the other hand, signifies the HCP's interaction with the marketing message, such as opening an email or clicking on a digital ad. In cases where the distinction between outreach and engagement is ambiguous, such as a sales representative visit or a sample drop, the data was considered as

both outreach and engagement. This paper uses the terms “HCP engagement” and “HCP interaction” interchangeably to denote the same concept. The study's findings were used to create recommendations for effective tactics and pathways at each stage of the customer journey. The subsequent sections provide a comprehensive explanation of the data processing and analysis conducted for CJA.

3.1 Data

Prescription data and marketing data from multiple channels were combined into a unified dataset, creating a time series record of prescriptions, marketing outreaches, and marketing engagements experienced by all HCPs during the analysis period. Records were classified as either a marketing outreach or an engagement or both. Marketing records were also tagged with high-level topics such as “Efficacy related”, “Key Opinion Leader (KOL) Videos”, “Insurance related” etc. based on the content of the material delivered.

3.2 HCPs to Journey Stages with Clustering

The consolidated marketing and sales data were used to segment HCPs into customer journey stages. Multiple clustering algorithms such as K-Means, GMM, and DBSCAN were evaluated and the result with the closest proximity to the predefined journey was chosen. The segmentation results were reviewed with the brand marketing team, and the insights provided by the team were used to create the final mapping of HCP clusters to journey stages.



Figure 2: Customer Journey Stages

3.3 Identifying success metrics for journey stages using Community Detection

The time series data for each HCP are essentially a chain of time ordered interactions, starting with the first engagement that the HCP had during the analysis period and ending with the last engagement. Individual HCP chains were aggregated to form a global network of marketing engagements and prescriptions. The nodes in this network represented the distinct marketing engagements or prescription. The edges of the network represented the movement of HCPs from the source node to the target node.

Community detection techniques were then applied on the network. Community detection provides a means to mapping marketing engagements to customer journey stages. Community detection algorithms analyze

network structures to identify groups of nodes with dense connections. There are several algorithms available for community detection such as Louvain, Leiden, Label propagation etc. The reasoning behind applying community detection for mapping marketing engagements to customer journey stages can be illustrated with a simple example. If Dr. John is an HCP in the awareness stage of the customer journey, the marketing contents he would be engage with would be those relevant for his awareness level and experience with the brand. Hence, it can be argued that the marketing contents that HCPs engage with are related to their respective stages in the customer journey diagram. Consequently, when the marketing engagements data are aggregated to form a global network as explained above, nodes in the network data related to the same journey stage would have a lot of connections between them and could be grouped together using the community detection algorithm.

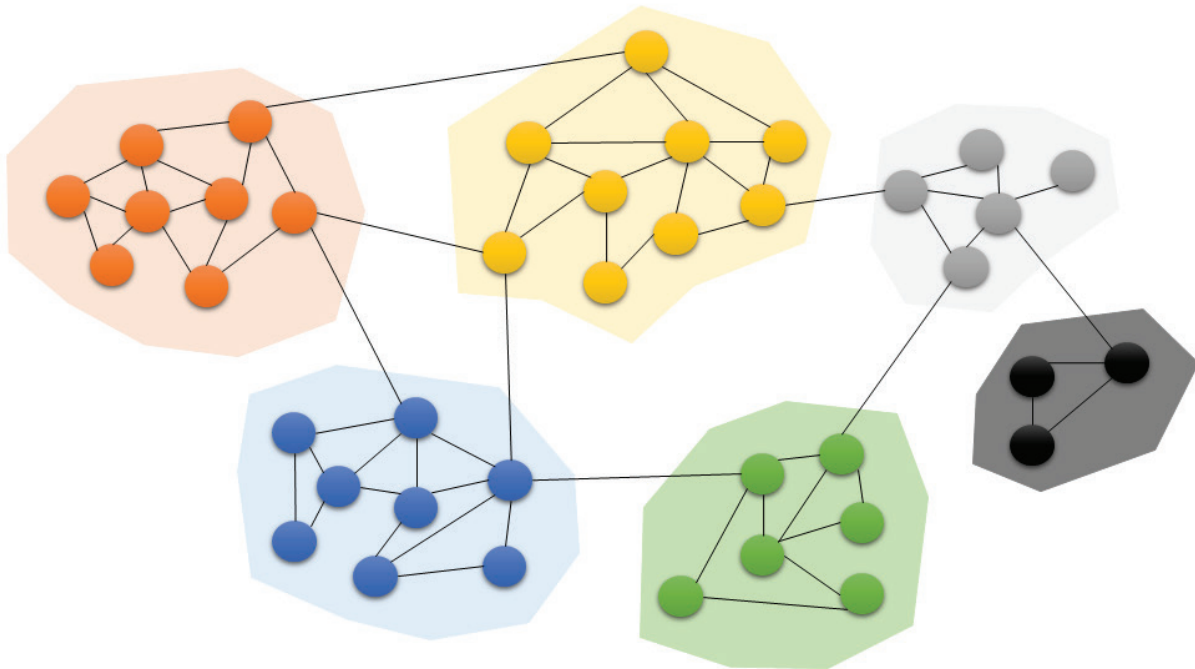


Figure 3:An Illustration of Community Detection Algorithm

HCPs were scored based on the intensity of interactions with different communities and these intensity scores were used to map communities of marketing engagements with customer journey stages. Table 2 shows the relative intensity of customer engagement across journey stages for different communities. More intense interactions correspond to darker shades of green.

The mapping of marketing engagements to customer journey stages were used for a brainstorming session with the brand marketing team to identify appropriate success criteria for each journey stage. For example, engagement with marketing contents related to Insurance Reimbursements and Patient Support Programs were selected as an appropriate success metric for “Engagement stage” and NBRx was chosen as a success metric for “Loyalty stage”.

Marketing Engagement Communities	Customer Journey Stages				
	Awareness	Consideration	Engagement	Loyalty	Advocacy
<u>Community 1</u> Channels: Veeva Interactions, Speaker Events Attendee Topics: Indication, Real World Evidence					
<u>Community 2</u> Channels: Digital Display, Mass Email Campaigns Topics: Efficacy, Key Opinion Leader Videos					
<u>Community 3</u> Channels: Veeva Emails, Digital Display, Mass Email Campaigns Topics: Patient Support Programs, Insurance Related					
<u>Community 4</u> Channels: Veeva ESales , NBRx Topics: Real World Evidence, Key Opinion Leader Videos					
<u>Community 5</u> Channels: TRx, Speaker Events Speaker Topics: Indications, Real World Evidence					

Table 2: An Illustrative Diagram of Varying Customer Engagements Across Journey Stages

3.4 Identifying marketing factors driving success metrics

Once success metrics were chosen for each journey stage, regression models were developed to identify marketing variables

driving the desired outcomes. Confounding variables such as seasonality of the disease area were added to the dataset. The data were split into test and training sets and XGBoost models were fit to minimize the test set’s Mean Squared Error (MSE). SHAP (SHapley Additive

exPlanations) values were used to identify main marketing factors driving success metrics for each stage. SHAP is a game theoretic approach to explain the output of any machine learning model. It assigns each feature an importance

value for every prediction in the dataset. See figure 4 and 5 below for importance ranking of marketing factors at Engagement and Loyalty stages.

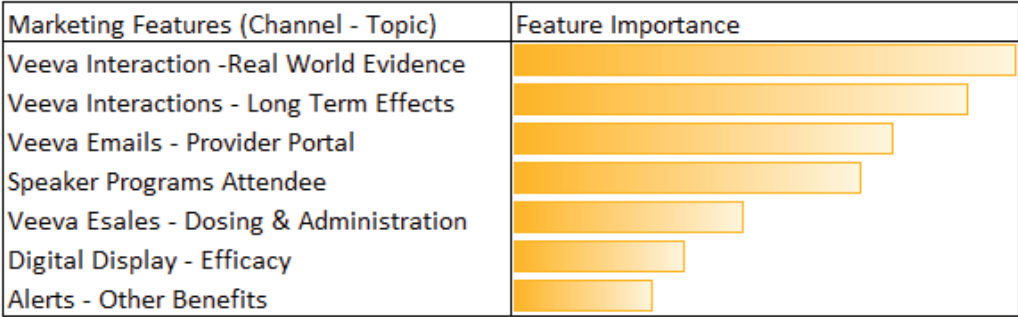


Figure 4: Feature Importance for the Engagement Stage

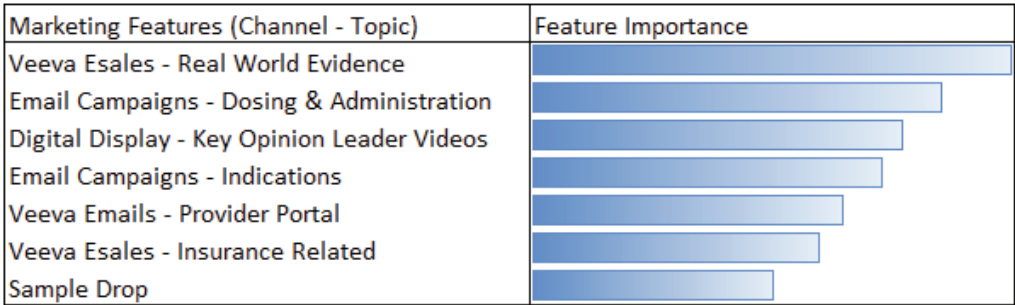


Figure 5: Feature Importance for Loyalty Stage

3.5 Constructing Customer Journeys from significant marketing drivers

Path models were fit to construct customer journey pathways from significant marketing drivers identified from the previous step. Path Modeling is a variant of Structural Equation Modeling used to evaluate relationships between variables in a causal model. It allows to specify a hypothesized causal model which includes a set of variables and their hypothesized relationships and estimates the strength and significance of these relationships by fitting the model to the observed data using regression analysis. It can estimate direct and

indirect effects of variables on an outcome variable, considering multiple pathways and potential mediating or moderating variables.

Figure 6 displays the recommended pathway for Loyalty stage. NBRx, the success metric for the stage is displayed in a green box and all marketing drivers are in blue. Marketing drivers are shown with their corresponding channel and topic of interaction. All marketing drivers have a direct effect on NBRx, thus omitted from the diagram for simplicity. Path analysis also provides path coefficients to indicate the strength and direction of relationships between different variables.

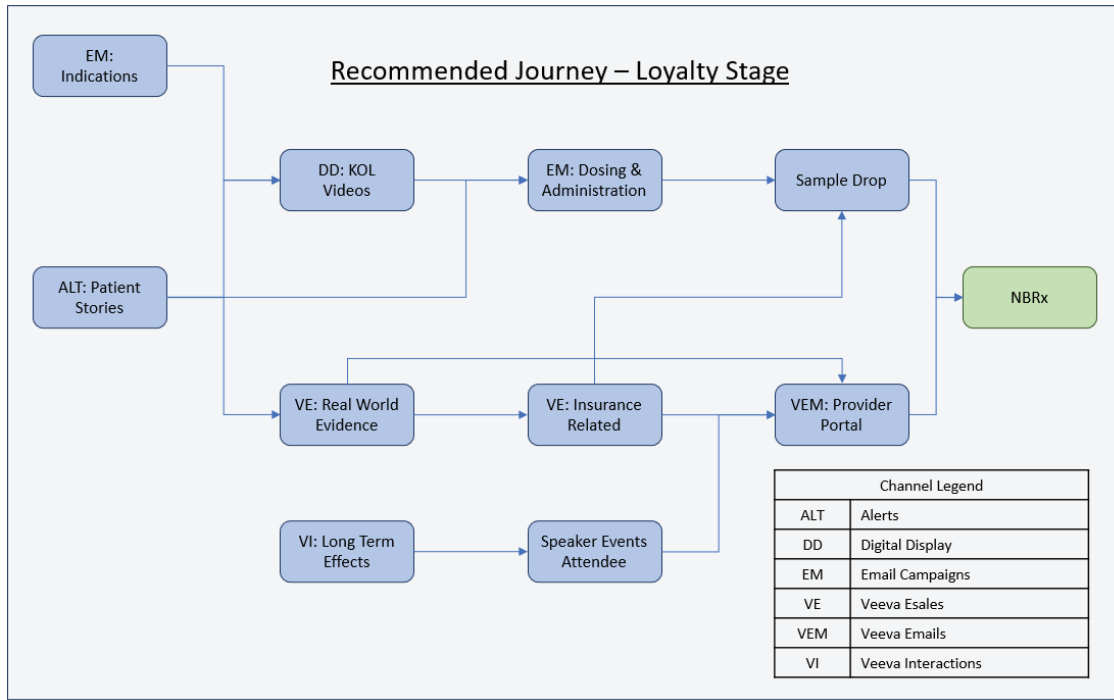


Figure 6: Recommended Customer Journey for Loyalty Stage

Results

Probability of success was calculated for the recommended pathways at each stage. It is defined as the probability of observing the successful outcome after an HCP has experienced the recommended sequence of touchpoints. The time series data of HCP interactions and prescriptions were cut at

each NBRx to create HCP Path Segments and counted to compute the probability values. For example, if “E1 > E2 > E3 > Brand NBRx” is a recommended pathway where E1, E2 and E3 represents different marketing touchpoints, then the probability of success is defined as

$$Prob(E1 > E2 > E3 > NBRx) = \frac{\text{No. of HCP Path Segments with } E1 > E2 > E3 > NBRx}{\text{No. of HCP Path Segments with } E1 > E2 > E3}$$

The recommended pathways were found to have high probability of success at each stage with varying lengths of sequences. Table 3 shows the probability of success for top recommended pathways at the Loyalty stage. Sample Drop is the most significant single marketing interaction predictive of NBRx. The path diagram also highlights the importance

of maintaining and promoting the HCP portal (Provider Portal) as many pathways at this stage pass through it. Although some of the widely promoted messages do not have a significant direct impact on sales (NBRx or TRx), they are instrumental at early stages in increasing awareness and driving HCPs forward in their customer journey.

Length of Sequences	Journey	Outcome Probability
1	Sample Drop	63%
1	Veeva Emails - Provider Portal	54%
1	Veeva Esales - Real World Evidence	51%
1	Email Campaigns - Dosing & Administration	45%
1	Digital Display - Key Opinion Leader Video	39%
2	Veeva Esales - Insurance Related > Veeva Emails - Provider Portal	71%
2	Email Campaigns - Dosing & Administration > Sample Drop	68%
2	Veeva Esales - Insurance Related > Sample Drop	64%
2	Speaker Events Attendee > Veeva Emails - Provider Portal	54%
3	Veeva Esales - Real World Evidence > Veeva Esales - Insurance Related > Sample Drop	72%
3	Veeva Esales - Real World Evidence > Veeva Esales - Insurance Related > Veeva Emails - Provider Portal	65%
3	Veeva Esales - Long Term Data > Speaker Events Attendee > Veeva Emails - Provider Portal	61%
3	Digital Display - KOL Videos > Email Campaigns - Dosing & Administration > Sample Drop	56%

Table 3: Outcome probabilities for top customer journeys from Loyalty Stage

Conclusion

In this paper, we presented an analytical framework that leverages past customer interactions and prescription data to develop stage specific pathways of marketing touchpoints that deliver best outcomes.

The proposed framework offers a thorough approach from identifying journey stages to finally constructing recommended pathway. It is generalizable to different drugs and different therapeutic areas provided there is a good volume of marketing interactions and prescriptions data. Insights gained from

the CJA can be utilized to create marketing campaigns tailored to HCPs to ensure they have the necessary information to effectively serve the patient community. New rules can be designed to improve the NBA engine. Regular updates to the findings of the CJA are essential to adapt to the ever-changing business situations and meet the evolving needs of customers.

References

Lundberg, S.M., Erion, G., Chen, H. et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2, 56–67 (2020). <https://doi.org/10.1038/s42256-019-0138-9>

Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System. <https://doi.org/10.48550/arXiv.1603.02754>

Lemon K., Verhoef P.; Understanding Customer Experience Throughout the Customer Journey; *Journal of Marketing*; 2016.

Zhang J., Fei J., Song X., Feng J.; An Improved Louvain Algorithm for Community Detection; *Mathematical Problems in Engineering*, Hindawi; 2021.

Hastie T., Tibshirani R., Friedman JH.; *The Elements of Statistical Learning*. 2nd ed. Berlin: Springer Series in Statistics; 2009.

Breiman L., Friedman J., Stone CJ., Olshen RA.; *Classification and Regression Trees*. Taylor & Francis; 1984.

Fan Y., Chen J., Shirkey G., Jhon R., R. Wu S., Park H., Shao C.; Applications of structural equation modeling (SEM) in ecological studies: Updated review; *Ecological Process*; 2016.

Biography of Authors

Jingfen Zhu works as Chief Science Officer at Genpact providing thought leadership advisory and innovation guidance for Fortune500 clients, as well as mentoring, lecturing, and speaking at key conferences. She has worked in the Data Science field for over 18 years in

Life Science, Healthcare, Finance and CPG. Throughout her career as both an executive and practitioner, she has developed and guided building of numerous state-of-the-art analytical solutions which have been productionized and generated tremendous financial outcome. The team she coached participated in the 2018 Data Hackathon held by KeyBank and ranks top 3 in Client Centricity solutions among 72 competing teams and 300 contestants. The first-phase deployment of personalized customer retention strategy from this hackathon solution resulted in a signed-off savings of \$38 million. She received Ph.D. in Applied Statistics and M.S. in Management Science from The Pennsylvania State University - University Park. She enjoys baking and tasting gourmet coffee in her spare time.

Rakesh Sukumar works as a Senior Manager at Genpact's Life Science Analytics vertical. He has more than 11 years of IT experience and more than 7 years of data science domain experience. Throughout his professional experience, he has delivered analytics projects for Fortune 500 companies in the Life Sciences, CPG, and IT Service Management industries, specifically in the domains of marketing, supply chain, and operations. He has done MBA with Finance & IT as majors from Indian Institute of Management Kashipur and B.tech in Electronics Engineering from College of Engineering Thiruvananthapuram.

Ankit Majumder works as a Data Scientist at Genpact's Life Sciences Analytics vertical. He has research project experience in applied and bio statistics. He has done MSc. in Statistics from Presidency University, Kolkata and BSc. in Statistics from Bidhannagar College, WBSU. He is a dedicated professional committed to advancing the boundaries of data science in the pursuit of knowledge and innovation.

Optimizing Launch Excellence: An AI-driven Framework for Precision Engagement

Christel Chehoud, Johnson & Johnson; Xiaoyang Meng, Johnson & Johnson; Manish Kumar, ZS; Kaushik Chafekar, ZS; Mehul Singh, ZS

Abstract: AI/ML methods have gained significant adoption in commercial operations across pharmaceutical companies, particularly for multi-channel, omni-channel, and Next Best Action (NBA) customer engagement strategies. However, the lack of robust customer-level data, particularly historical sales data, for use at launch poses a challenge for training these AI/ML methods with traditional ML-driven NBA models built on at least 6-12 months post-launch data. It is also widely recognized that a successful launch plays a key role in determining the future reach of a brand¹. This creates an opportunity for an improved NBA design for launch that adapts to the uniqueness of a brand and serves the unmet needs of customers from the day of launch. Fortunately, a substantial amount of primary market research (PMR) is typically conducted pre-launch and available for use. In this initiative, we sought to translate PMR-based insights, combined with customer preferences derived from product usage data, to build a precision targeting capability for products at launch, particularly in the rare disease space. This approach offers incremental benefits with improved timeliness, increased precision for marketing, and better-informed launch readiness.

Keywords: Precision Engagement, Launch Excellence, Real World Data, Patient Enrichment, Primary Market Research, Next Best Action

Background: Precision Engagement at Launch

Navigating the intricate process of identifying and reaching the right physicians to introduce a new product launch is challenging. Relying solely on historical potential may prove insufficient in estimating patient unmet needs given the increasing complexity of identifying the ‘right patients’ for treatments with precision medicine and targeted therapies. Yet, initiating personalized interactions during this crucial phase can result in an enhanced HCP experience and create a ripple effect. The first six months of a brand’s performance serve as a key indicator of future success and are pivotal in determining a brand’s trajectory². Recognizing

and prioritizing early adopters during this phase can ensure the activation of network effects. Moreover, as the time between competitive launches shortens in the dynamic healthcare landscape³, effective product launches, tailored to resonate with HCP preferences, particularly early adopters, become imperative for sustained success.

Approach & Implementation

Overview: Recognizing the substantial data gap at launch, the conventional machine learning based NBA approach falls short. Our approach for launch is to optimize our reach at a customer level by decomposing the problem into sub-metrics broadly categorized as quantifying

patients' unmet needs and HCP's likelihood of usage of a launch product. We address the likelihood of usage by leveraging insights from aggregated, cohort-level PMR. The likelihood

of usage can be further split into two distinct aspects: the customer's receptivity towards the product and the effectiveness of commercial engagement.

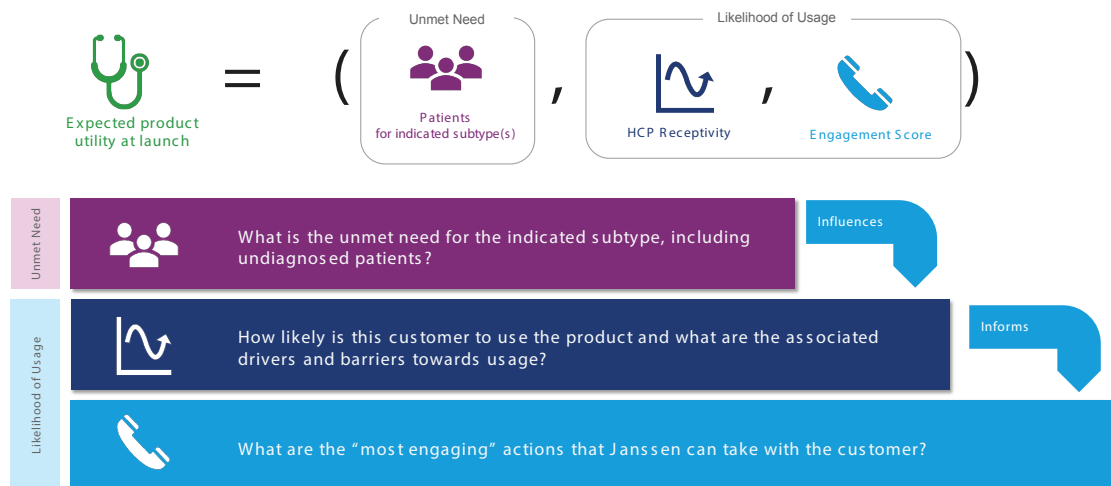


Figure 1: Precision Engagement at Launch Framework

Patient Unmet Need at Launch

Quantifying the patients' unmet needs for an indicated disease, including undiagnosed patients, is crucial. Patient prediction models are now ubiquitous in their attempts to do this. With the growing emphasis on targeted therapies and precision medicine, it has become increasingly important for medical and commercial organizations to be more sophisticated in understanding patient needs and strategically prioritizing efforts. Models can be built pre-launch to estimate undiagnosed and/or misdiagnosed patients. During the launch phase, estimating opportunities within patient subtypes aids in positioning products effectively, especially in cases where access to therapy is contingent on companion diagnostics or prior treatments or large deviations from

historical products are expected, as in the case of a first-in-class launch. Moreover, it is noteworthy that the quantification of unmet needs often falls below epidemiological estimations. The integration of a machine learning based enrichment model could offer a solution⁴. By creating probabilistic models, we gain insights into the true patient opportunity and establish a comprehensive view on the patient funnel.

Our initial exploratory analysis was conducted on real-world claims datasets to assess the gaps in patient capture produced by using a traditional, business rules-based approach. The process highlighted the need for data enrichment due to the significantly lower prevalence of patient sub-types of interest when compared to epidemiological estimates

and competitor product utilization. An ML-based enrichment model was implemented to uncover additional patients to get a more precise estimate of true patient opportunity. The models were validated using hold-out sets, and performance was assessed at different thresholds to determine the optimal patient enrichment model. The most important features that helped identify the additional patients were compared with the literature and reviewed thoroughly with experts to validate the model's clinical reliability.

HCP Receptivity at Launch

While there may be multiple factors contributing to customer base receptivity at launch, prior research suggests that 80% of brand performance is attributable to the level of unmet need, product efficacy, and product novelty⁵. Recognizing the distinctiveness of unmet need and product profile, it is essential to

access them separately to avoid introducing bias through a single analog proxy for receptivity. Estimating and understanding the receptivity of customers towards the launch product enables more tailored engagement strategies for addressing different customer needs.

Fortunately, base receptivity is also captured via PMR for a sample target list using the actual target product profile. While this information is based on a sample of physicians, it helps us learn which attributes matter while making treatment decisions for the patient sub-type. Other indicators like account affiliation of the customer, historical early adoption behavior, experience in treating patient sub-types, experience with drug classes, etc. are also created using Real World Data. Leveraging all this information as an 'algorithm', we can predict the receptivity towards the launch product at the customer level.

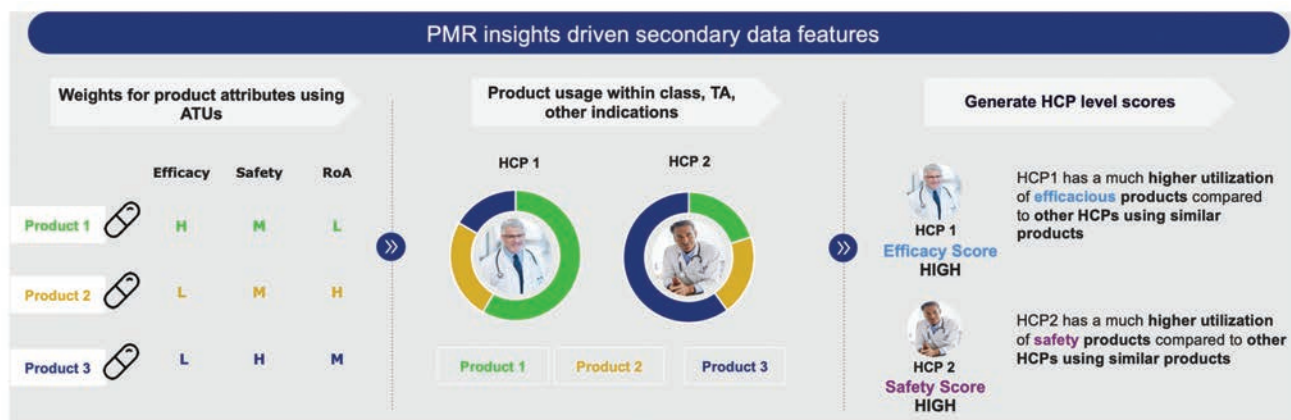


Figure 2: Example features derived through the combination of primary and secondary data

To achieve accurate estimations of base receptivity at a customer level, we integrated (and not projected) PMR insights with customer-level attributes derived from secondary data sources. Analyzing prescribing behavior across therapeutic areas allowed us to estimate attributes such as “Safety Focus”, “Efficacy Focus”, or “Mode of Administration Focus” by assessing deviations in product usage from the average and whether the skew is consistently observed towards products that are differentiated on these attributes. For example, if PMR shows that PRODUCT_A is perceived as much safer than PRODUCT_B, we used secondary data to convert HCPs’ use of PRODUCT_A and PRODUCT_B into estimates of their “Safety Focus”. An HCP prescribing 90% PRODUCT_A and 10% PRODUCT_B would be deemed more “Safety Focused” than an HCP prescribing 10% of PRODUCT_A and 90% PRODUCT_B. These customer-level attributes derived from claims-based data sources uncovered insights into the customers’ unmet needs and attitudes that aided in understanding the ‘why’ behind treatment decisions and uncovering cross-therapy area behavior preferences. Armed with these HCP-level scores on safety, efficacy, dosing, etc., we scored customers using an algorithm that employs PMR-based features to estimate receptivity. This approach of leveraging PMR-based features for HCP receptivity was tested by incorporating it into an existing ‘in-market’ NBA model and assessing lift in performance.

Engagement Score

Building on the insights from unmet needs and receptivity, our next step involved determining the optimal intervention strategy for each customer. This decision is guided by a meticulous analysis of customers’ channel preferences. The engagement score relies on historical interaction data, both internally generated and through industry-syndicated reports. Given that this score is based on a ‘likelihood to engage’ metric based on historical engagement data, we had robust information about most of our customers on how they prefer to consume information about new products in the therapeutic area to foster a stronger connection with them.

The effectiveness of prior interactions with the customers across channels for the indication of interest and other similar disease areas was estimated using internally available historical engagements. This was layered with industry benchmarks on access through external datasets. These scores were then combined into an engagement score that quantified the relationship and ‘potential to engage’ with customers across channels. The validation of this approach occurs in two ways – one based on time series prediction for other products that we have access information for, as well as based on territory feedback closer to launch to ensure launch efforts are prioritized for the most accessible customers given the opportunity cost. We predicted rates of engagement for the first 1, 3, and 6 months of launch, and this was used to ensure the customers with the most unmet needs, who are most receptive to new therapies, and who are likely to engage with the prioritized channel are prioritized.

Channel Affinity, and past engagement history, length of relationship can be used to coordinate between Personal and Digital touchpoints

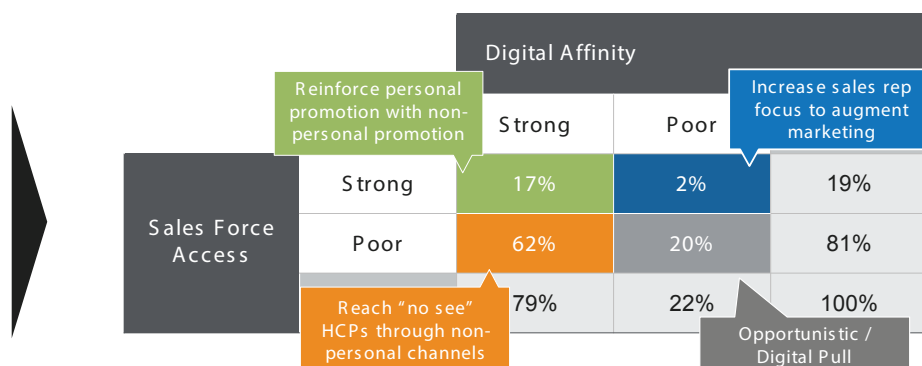


Figure 3: Meeting the customer where they are by incorporating Sales Force Access & Digital Affinity experience data

Launch Precision Engagement Plan

The commercialization strategy for a product launch is intricately defined at the customer level through our three-component approach (Figure 1). Given the unique nature of each component, it becomes imperative to normalize and weigh each score appropriately, recognizing the distinct 'unit' of measurement for each before combining them. The assignment of weights to individual components for every customer subtype is contingent upon brand marketing strategy, influenced by the competitive and dynamic market landscape.

For instance, an industry-wide, unaddressed unmet need might be highly correlated with the number of eligible patients, and on the contrary, customer receptivity towards the launching therapy may be more predominant in a relatively competitive market. A weighting strategy should align with the brand launch strategy using a data-driven approach or business-rule approach as the interplay of these scores is crucial in accessing the relative commercial potential of each customer. Equation 1 illustrates the general form of our three-component approach:

$$S_{composite} = w_1 \times S_{patient} + w_2 \times S_{receptivity} + w_3 \times S_{engagement}$$

Equation 1

Where $S_{composite}$ is the normalized patient unmet need score, $S_{receptivity}$ is the normalized HCP receptivity score, $S_{engagement}$ is the normalized engagement score, W_n is the associated weight on each component, respectively. $S_{composite}$ is the combined composite score which represents the relative commercial potential of each customer for the launch product.

The resultant composite score acts as the objective function for the development of a precision engagement plan for launch. This plan is crafted through the utilization of the NBA optimization engine, which is tasked with maximizing the sum of combined scores of customers at a representative's territory. Operating within the framework of predefined business rules, the optimization engine considers factors such as representatives' capacities, physician or account access constraints, routing considerations, and other pertinent criteria. The meticulous process ensures that the engagement plan is not only data-driven but also aligned with the practical realities of the market, ultimately enhancing the precision and effectiveness of our launch strategy.

Testing

Launch is a team game. Since we started building capabilities of modeling for launch, our three components have been used for several other downstream tasks. Launch is typically enabled by a matrix team, and the ability to have integrated data about the customer that can serve as a single source of truth for downstream planning significantly contributed to ensuring connectivity between teams supporting different commercial operations.

Validation through Simulated Experiment

The combination of three pillars maximizes a "utility" function at launch to ensure we can reach the most available customers. The true measure of success here will be improved productivity of our efforts, i.e. proportion of early adopters among engaged customers, and the proportion of engaged customers among targeted customers. Before launch, we validate the launch engagement plans with the sales force, based on their past experiences. While the true impact can only be validated post-launch, the framework was tested by simulating a launch scenario for an already 'in-market' product. Patient unmet need, HCP receptivity, and engagement score were developed using PMR, real-world data, and analog products - all available prior to launch for a current 'in-market' product. The model was successfully able to identify 84% of HCPs who were true early adopters in Year 1 of launch in this pre-launch simulation. Optimizing engagement strategies armed with this information could have yielded a potential 10% lift in Year 1 sales compared to traditional strategies. These features also significantly improved the performance of 'in-market' engagement models.

In addition to model-based validations, the opportunity and receptivity segments are also being used and validated in attitudinal segmentations to further understand drivers of usage and product positioning. Some of our hypotheses around early use and perceived unmet needs should correlate directly with average responses from certain segments, and this also helps us pre-define schemas for the projection of the research from a sample to the overall target population.

Launch Execution

At launch, we want to ensure that we prioritize the HCPs who have the most patients with the unmet need and who are most likely to engage as well. These are our objective and loss functions, respectively, while we build the launch “call plan”. With inputs from attitudinal segmentation, we can also prioritize the right content to engage the customers. Digital outreach can also be planned based on SF Access and digital preference for resource planning.

Beyond Launch

Implementation at Launch and Beyond In-Market (NBA)

We acknowledge that while each component has undergone rigorous testing individually, an inherent assumption is that all these components will apply to the product and its unique context, for which no secondary data is available. We anticipate variance due to the evolving market landscape, individual customer-level differences, applicability of product analogs, and a myriad of other things. As we observe deviations from model expectations, it is imperative to adapt the underlying models to track and revise erroneous predictions. Particularly in the early launch phase, early signals are invaluable to the models. However, the challenge lies between discerning information and noise and leveraging these data effectively to inform better models. An active learning approach holds promise in addressing this hurdle during early launch. By systematically testing for variance with early signals and modifying the underlying drivers of the model accordingly, we can iteratively improve model accuracy and adaptability. As more data becomes available, the weights of models will gradually shift from these ‘priors’ to being entirely evidence-driven,

aligning with the standard practice for most supervised learning NBA models.

Beyond Targeting at Launch

While the overall capability aims to inform launch resourcing by channel, the individual pillars also help inform other launch enablement activities. A few notable examples are listed below:

- Estimating medical unmet needs and prioritizing customers with behavioral gaps for medical education – the data enrichment model helped identify where there are gaps in testing based on epidemiology estimates, and how we can educate HCPs on using precision therapies to improve patient outcomes. This capability was very relevant pre-launch for the medical teams to understand where to focus.
- Supporting PMR panel builds using clustering-based methods – Using HCP unmet need, and receptivity-based features, we also identified customer types using unsupervised methods. This helped stratify the target population into segments, which were then used to build representative research panels.
- Informing Data acquisition strategies for launch - through a supervised behavior prediction model, which was tested on analogs, we could identify what attributes are most predictive of customer behavior, and what information would provide a competitive advantage at launch. We were able to check for missing-ness in key features to recommend which data sets should be prioritized for data acquisition.

We constantly learn about how the capability and underlying features are supporting other

teams such as attitudinal segment projections from a limited sample to the target population.

Conclusion

Our approach presents a significant leap forward in optimizing product launches. By providing incremental insights across multiple levels, it addresses critical aspects such as expedited customer prioritization, refined marketing strategy at a customer level, and enhanced launch preparedness based on informed decision-making. Yet, it is essential

to acknowledge notable opportunity areas that merit attention, including data limitations across different disease areas and geographies, the considerable expenses associated with real-world validation methods compared to simulations, and the ongoing challenge of reconciling gaps between stated and actual preferences. Despite these hurdles, the potential advantages of this approach underscore its pivotal role in equipping the pharmaceutical industry with AI-driven launches that are optimized for success.

References

1. Ford J, Fezza T, Elsner N, Arora A. Key factors to improve drug launches: Why drug launches miss market expectations and what to do about it. Available from: <https://www2.deloitte.com/us/en/insights/industry/life-sciences/successful-drug-launch-strategy.html> [Accessed February 27th 2023]
2. Rickwood S, Gores M, Karanis Y, Alexandra A. Launch Excellence VI: Launch Excellence in a disrupted world: challenges and opportunity. IQVIA White Paper, 2019.
3. Albrecht B, Ascher J, Menu P, Peters M, Stiehl L. Launches in oncology: The elements of success. McKinsey, August 2 2018. Available from: <https://www.mckinsey.com/industries/life-sciences/our-insights/launches-in-oncology-the-elements-of-success#/>
4. Sunder A, Sharma A, Halder P. Enhancing Patient Classification and Staging in RWD Using Machine Learning. Pharmaceutical Management Science Association. Spring 2023.
5. Walter S, Gurnani K, Mukhopadhyay S. What actually drives drug launch success? Available from: <https://www.zs.com/insights/what-actually-drives-drug-launch-success> [Accessed February 27th 2023]

About the Authors

Christel Chehoud is a Senior Director within the Global Commercial Strategy Organization at Johnson & Johnson Innovative Medicine. Christel focuses on the advancement and acceleration of global go-to-market strategies for product pipeline by developing end-to-end data science solutions for a wide range of applications including healthcare provider customer targeting, optimized medical engagement, forecasting and real-world-evidence studies. Previously, within R&D, she led a team of data scientists delivering advanced analytics solutions for patient-level prediction models, clinical trial optimization, and biomarker identification. Christel holds a Ph.D. from the University of Pennsylvania and an A.B. in Molecular Biology and Computer Science from Princeton University.

Xiaoyang Meng is a Senior Data Scientist within the Commercial Data Science, North America Technology Group at Johnson & Johnson Innovative Medicine. In his role, Xiaoyang focuses on leveraging predictive modeling with machine learning and advanced algorithms to serve patients through omnichannel marketing, global pricing strategies, and real-world evidence studies. Xiaoyang holds a Ph.D. in Environmental Engineering and an M.S. in Computer Science and Engineering from Georgia Institute of Technology.

Manish Kumar is a Data Science Consultant at ZS's Pune office where he focuses on the development and expansion of AI models within Commercial Data Science and Patient-Level Healthcare Modelling across North America and Europe. Prior to this role, Manish contributions were pivotal within R&D use cases, focusing on optimizing site selection for clinical trials. He holds a Bachelor of Engineering in Information Technology from Pune University.

Kaushik Chafekar is a member of the Real-World Evidence practice and a consultant within ZS's India office. In his position, Kaushik is primarily responsible for using Advanced Analytics & Data Science techniques leveraging Real-World data to inform strategic choices for Pharmaceutical Manufacturers. With experience in precision engagement for providers, patient identification, survival modelling, comparative effectiveness and various real-world evidence studies, Kaushik has built and implemented solutions across the product lifecycle.

Mehul Singh is an Associate Principal with ZS' Princeton office and a leader in ZS' AI practice. Mehul has designed and delivered several AI-based products in APAC, EU, and North America to inform commercial strategy decisions for Pharmaceutical Manufacturers and Health Plans. Mehul's primary areas of interest are data driven resource allocation and precision engagement for providers, patients, and payers. Prior to joining ZS, Mehul worked on AI applications in Consumer-Packaged Goods and Energy Industries.

Transfer learning approach to enhanced patient classification using Real World Data

By Aniruddha Mitra, Atharv Sharma, Arkojyoti Das, Arrvind Sunder, Srinivas Chilukuri, Shubhendu Trivedi

Abstract: Real-world data (RWD), provides a comprehensive repository of information encapsulating patients' authentic engagements within the healthcare ecosystem. This forms a robust foundation for unraveling intricate patient journeys, diagnostic and treatment sequences, provider behavior and preferences, payer dynamics and pivotal macro healthcare metrics such as treatment rates, market shares, persistence, and compliance. Notably, pharmaceutical companies harness RWD for a variety of commercial (e.g., targeting, messaging, segmentation, brand performance tracking, etc.), medical (e.g., KOL identification, evidence generation, etc.) and other use cases (e.g., health economics and outcomes research (HEOR) studies, etc.). As a result, RWD has emerged as a cornerstone in orchestrating pharmaceutical strategies.

Within the realm of RWD, claims data—derived from administrative claims—stands out due to its unique systematic patient coverage and inclusion of healthcare provider (HCP) information, positioning it as the primary dataset for informing commercial strategies. However, despite the inherent advantages of claims data, certain limitations impede its full utilization. For instance, identifying specific patient sub-cohorts, such as a certain biomarker positive (say, Epidermal Growth Factor Receptor (EGFR) positive patients) non-small cell lung cancer (NSCLC) patients, can be challenging as detailed information around lab results/ clinical notes (e.g., biomarker status, histological status, severity, etc.) is often not captured.

To overcome such limitations, an intuitive solution involves heading towards richer datasets like electronic medical records (EMRs) or electronic health records (EHRs), which encompass minute details through HCP notes, lab results and other channels. While this approach presents its own set of challenges, the primary drawback lies in the sacrifice of patient pool captured. Moreover, EMR data often lacks HCP information crucial for commercial use cases.

This article proposes the design of an AI-based solution aiming to bridge the gap between disconnected sources—claims data and EMR data. The objective is to **enrich claims data with additional information from other data-sources with rich information** like EMR, broadening its scope while preserving its scale. Such enrichment promises to unlock new dimensions in the commercial applications of claims data for pharmaceutical companies, offering a more holistic perspective across various use cases.

Background and challenges

In the dynamic landscape of pharmaceutical advancements, an increasing number of companies are embracing precision medicine^{1, 2}—a paradigm that promises therapies to the unique characteristics of a specific subset of patients, ultimately improving both short

and long-term outcomes. This strategic shift has led to the introduction of numerous targeted therapies designed for specific patient subtypes. However, extracting meaningful insights from RWD sources, particularly claims data, to delineate patient journeys in this era of precision medicine poses a significant challenge.

As pharmaceutical launches become more nuanced, identifying distinct patient types from claims data has grown increasingly complex. Challenges include identifying the target biomarker positive patients, categorizing tumor stages such as I vs II vs IIIA vs IIIB vs IV and so on, or identifying niche biomarkers such as EGFR in lung cancer or HER2 in breast cancer. The ability to derive precise patient cohorts directly from claims datasets is compromised, impacting the quality of insights for both medical and commercial functions within the pharma organization.

The 2023 edition of the Pharma Management Science Applications (PMSA) journal featured a seminal paper titled “Enhancing Patient Classification and Staging in RWD Using Machine Learning.”³ This paper shed light on the concept of data enrichment and its application, assuming the availability of a gold standard patient cohort identification. However, as pharmaceutical launches continue to unveil therapies targeting niche biomarkers, the derivation of gold standard patient cohorts from claims datasets becomes intricate.

With the rapid evolution of AI and machine-learning (ML), newer algorithms and approaches have emerged, offering solutions to address the evolving challenges in precision medicine data analysis⁴. This paper proposes the application of transfer learning—a concept designed to harness similarities and dissimilarities across diverse data sources. By developing ML models, this approach aims to enable robust predictions even in the absence of gold standard patient cohorts.

The potential applications of transfer learning in this context are vast, including the enhancement of patient classification within claims datasets and the deployment of algorithms across EHRs. By leveraging transfer learning, this paper seeks to overcome

the limitations associated with the evolving landscape of precision medicine, opening avenues for more accurate and insightful analyses in the pharmaceutical domain.

Mitigating data gaps for optimal RWD utilization

The challenge of addressing such data gaps in real-world claims data necessitates a departure from conventional business-rules-based patient identification strategies.

To illustrate this challenge, consider a pharmaceutical manufacturer planning to launch a precision medicine tailored for the treatment of patients—an example we will maintain throughout the paper for simplification. Historically, identifying any given biomarker, say EGFR, positive patients within a cohort of NSCLC patients has been executed through methods often grounded in common business-rule-based approaches. These approaches anchor around proxy information available in claims data, such as specific drug consumption within a defined treatment window and are influenced by predefined notions. Despite meticulous boundary conditions, rule-based strategies inherently carry biases, favoring certain patient subsets and potentially misdirecting downstream strategic decisions. Another drawback of using proxy-based business rules is that they are always retrospective in nature, hindering any prospective insights and actions.

The imperative is to adopt a solution that transcends dependency on predefined rules, instead, one that autonomously learns from the data and adapts to its nuances. This necessitates a move towards data-based solutions capable of extracting rich information from non-claim data sources and seamlessly integrating this knowledge with claims data.

However, taking a traditional ML-based approach poses its own set of challenges, particularly when the objective involves learning signals from one data source and applying it on another data source. Traditional ML models face risks when applied in scenarios where the inherent data distributions of information-rich specialized data sources—such as EMRs or EHRs—and claims data significantly differ. The reasons for these disparities are multifaceted and warrant exploration in subsequent discussions.

In the subsequent sections, we delve into the complexities associated with ML-based patient classification in the context of divergent data distributions, exploring potential avenues to bridge the gap and optimize RWD utilization for robust and unbiased patient insights.

Data differences and the limitations of traditional ML: A visual perspective

The success of any ML solution depends on the fundamental condition that the data on which it operates to produce its opinion should be ‘similar’ to the data it encountered during training. This condition is particularly evident when attempting to construct an ML model to decipher the pattern to identify patients within EMR data, where the target biomarker labels (Positive/Negative) are available, and subsequently leveraging the learned pattern to classify patients in claims data.

In the case study section, we will investigate a specific scenario, but first, let’s discuss the differences in data across claims and EMR sources that pose challenges for traditional ML applications and necessitate novel solutions.

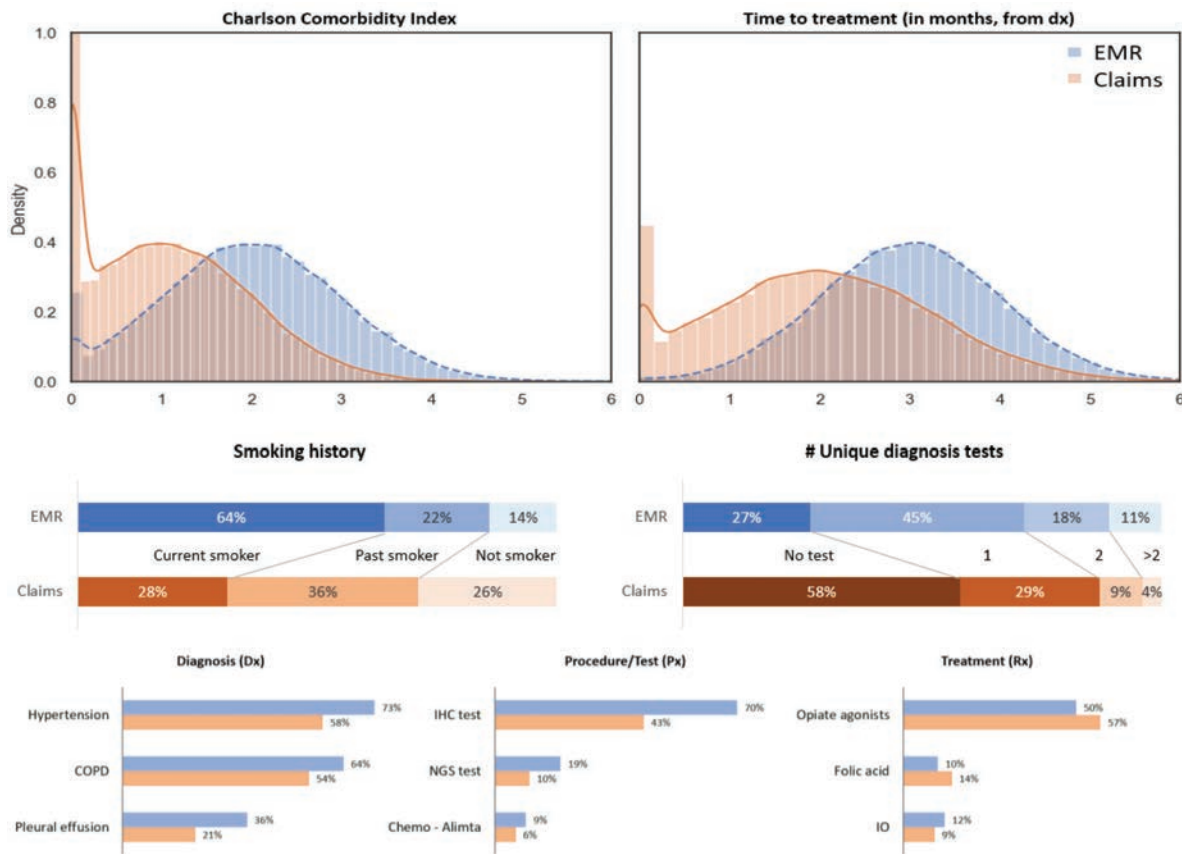


Figure 1: Data differences between EMR & claims data

As depicted in Figure 1, the inherent differences in data distribution are significant and we are able to highlight a few key metrics:

- The median time taken by patients to start treatment in EMR data is approximately three months post-diagnosis, whereas in claims data, the same metric is almost 50% lower.
- In EMR data, over 75% of patients have a documented history of smoking, while in claims data, less than 60% have such a history.

There are multiple reasons for these differences, with the main one being that specialized EMR data often only covers a certain subset of accounts and patients so is not fully representative of the entire population.

If an ML solution existed that could successfully decipher the pattern of the biomarker positive patients from a patient pool in EMR data and accurately identify unseen patients as biomarker-positive/negative coming from same or similar EMR data—there wouldn't be any assurance that the same solution would yield similar performance in data which is 'different' in terms of inherent data distributions.

This prompted a need for a ML solution that acknowledges gaps between the source and target data sources which is designed to transfer information from EMRs to claims data by bridging the gap between data distribution across the two sources.

Transfer-learning: A practical overview

Consider a practical scenario where 200 cameras that are programmed to count cars are scattered across various locations in Manhattan. Each camera is exposed to variations in

calibration, perspective, lighting conditions and more. The challenge is to count the number of cars in the images captured from these different cameras. Acquiring labeled data for all these cameras is expensive, making it impractical to gather labeled images for each camera individually.

The concept of transfer-learning becomes instrumental in this scenario. In a realistic setup, labeled images would be collected for a subset of cameras, and a generalized counting system could be trained across all cameras, regardless of each camera's capabilities.

Domain adaptation: Bridging the gap between source and target domains

Domain adaptation specifically addresses situations where labeled data is available only from the training distribution, known as the source domain. While unlabeled data is derived from the testing distribution, known as the target domain. The complexity arises from the fact that the source domain may differ significantly from the target domain. In the context of the vehicle counting example, different cameras capturing images present distinct pixel distributions due to the pre-established variations.

The goal of a domain adaptation algorithm is to enable the model to generalize effectively to the target domain without the need to have access to labeled samples. This becomes particularly crucial in real-world applications, where practical constraints make it challenging to obtain labeled data that spans the entire spectrum of variations. In the subsequent sections, we investigate the intricacies of domain adaptation algorithms and their practical applications in diverse domains, shedding light on their efficacy in scenarios where labeled data is limited but generalization across domains is paramount.

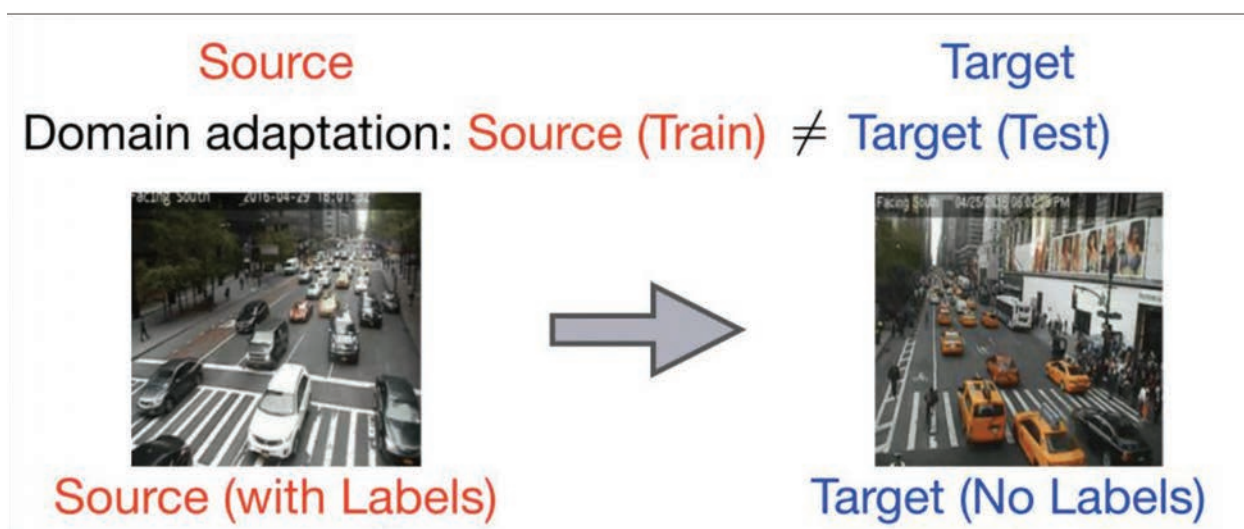


Figure 2: Domain adaption in practice: Source and target domains⁵

Transfer-learning framework: Unifying data perspectives for enhanced claims data enrichment

While the challenge of enriching claims data from smaller yet more information-rich datasets may not directly align with the vehicle counting scenario, a strategic shift in perspective allows us to identify a common thread present in both problems. By introducing specific nomenclatures, we define the **source** as smaller but richer datasets—such as EHR—and the **target** as large real-world claims data lacking detailed information.

To illustrate the application of the proposed solution, we anchor our discussion with a simple example: Identifying a target biomarker (say, EGFR) positive patient within a NSCLC patient cohort in claims data. In data language, this involves classifying the target biomarker patients from their counterparts in claims data. Notably, claims data does not contain labels indicating whether a patient is positive for the

biomarker or not. To tackle this, we need to learn from the EHR datasets where such labels exist and are considered gold standard. The learned classification must then be applied to claims data to accurately identify potential target biomarker positive patients.

However, this presents a challenge – in the car example, where an AI model is trained in a domain with black and white vehicles as source data, using the same model may not yield accurate results in the target domain where vehicles exhibit more colors. Similarly, even if we discern a pattern for identifying the target biomarker positive patients in the source EHR domain with commendable accuracy, applying this learning directly to claims data for the target biomarker positive patient identification may fall short. An AI model's utility is constrained to the extent that the testing data distribution aligns with the training data distribution from which it learned the pattern.

This underlines the need for a solution that facilitates learning across domains, transcending inherent differences in data distribution and ensuring the applicability of the acquired knowledge in diverse datasets. In a later section, we delve into the intricacies of developing a transfer-learning framework tailored to overcome this challenge, ultimately enhancing the accuracy and efficacy of claims data enrichment.

Validation mechanism in transfer-learning: An adversarial approach

Before investigating the intricacies of constructing a solution, it is important to establish a standardized scale of measure for validation. In our context, where patient identification in real-world claims data is the focal point, we acknowledge the existence of a limited number of patients in claims data labeled by business rules. While these rule-based labels may lack representativeness for potentially identifiable cohorts, we can leverage them to assess the overlap with patients identified by ML models.

However, recognizing the inherent bias in business-rule-based identification, we must not solely rely on it for validation. The need arises to develop a mechanism that provides a robust way for validating results obtained through various transfer-learning settings.

At this stage, it is relevant to reflect on two fundamental questions:

- **Why is a ‘validation dataset’ essential in any ML setup?**

The validation dataset acts as a navigational guide, ensuring that the model not only learns from the training data but also generalizes well to unseen data. It serves as a checkpoint to prevent the model

from overfitting to the training data and, instead, encourages learning of inherent patterns applicable to diverse datasets.

- **And what basic qualities define a dataset as a validation dataset?**

The validation dataset serves as a representative sample of the data on which the model is intended to make predictions. It should closely mirror the actual operational space, allowing the model to gauge its effectiveness in a real-world context. Additionally, it must contain the true labels, enabling the model to calibrate itself by comparing its predictions with the actual outcomes. This dual role ensures the validation dataset’s significance in refining and validating the model’s predictive capabilities.

In our case, a conflict has arisen: Gold-standard labels—such as target biomarker positive/negative tags—are available in some data sources—such as EMR—but they inherently differ from claims data-distribution, as we have seen in Figure-1.

To resolve this, an adversarial ML approach is adopted:

- Stratify the source data into k-folds.
- For each of the k folds:
 - » Utilize the remaining *k-1 folds of the source data and the entire target data to build an ML model classifying source patients from target patients.*
 - » Pass the hold-out kth fold source patients through the trained model.
 - » Identify the patients that the model ‘mistakenly’ labels as target data and tag

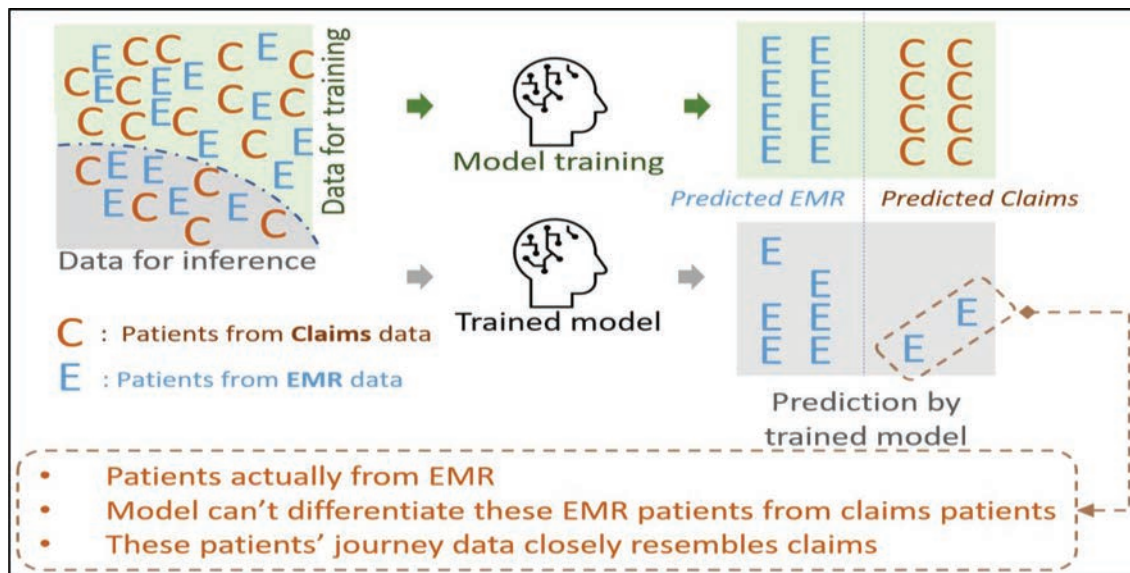


Figure 3: Adversarial ML based proxy validation dataset design

them as ‘target-like-source’ patients from the k th fold.

- Aggregate all the ‘target-like-source’ patients.

This subset has unique advantages:

- It contains gold-standard patient identification labels.
- It closely resembles target data.

It is crucial to note that the definition of a ‘mistake’ in patient identification significantly influences the resemblance and size of the ‘target-like-source’ patient set. Consequently, sufficient care must be exercised when creating this validation set.

This simulated validation set stands out from all patient identification models developed, providing a robust means of assessing the effectiveness of transfer-learning across different scenarios.

Algorithms for transfer-learning: Navigating interpretability and computational efficiency

In the landscape of transfer-learning across domains, a multitude of algorithms exists, each touting its efficacy in different scenarios. However, these algorithms exhibit significant variability in terms of interpretability and computational considerations. Our objective is to test these algorithms, from the simplest to the most recent, with a careful assessment of their potential success and an evaluation of their implications.

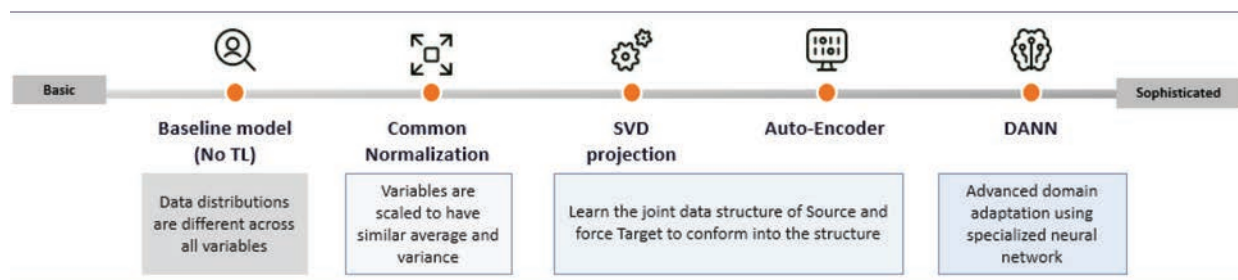


Figure 4: ML solution spectrum

Baseline Model: Establishing the Foundation

Before analyzing the implementation of the straightforward domain-adaptable transfer-learning model, we aim to address the necessity for establishing a baseline performance. This means constructing a model without any domain adaptation components, serving as a reference point for assessing the effectiveness of subsequent models, specifically tailored to facilitate transfer learning within the pharmaceutical domain.

We will construct a simple classifier designed to identify the target biomarker positive patients within the source EMR domain. Subsequently, we will use the same model to discern potential target biomarker positive patients within claims data.

Source and target data alignment:

1. Common normalization—harmonizing data distributions

Given the observed substantial differences in data distribution between source and target domains, addressing this dissimilarity requires careful consideration of various statistical metrics. One of the simplest approaches involves evaluating the proximity of two distributions, with a key focus on mean and variance for individual variables across

datasets. To achieve this, a straightforward, yet effective method involves mean-shifting and scale-normalizing the source and target data independently. This ensures that each variable maintains consistent mean and variance across both datasets, establishing a foundational step toward harmonizing the data distributions.

2. SVD projection—aligning target data with source data distribution

A sophisticated linear algebra-based statistical method serves as a technique for capturing the inherent data distribution in a lower-dimensional latent space, effectively condensing information during the process. First, we perform Singular value decomposition (SVD) on the information-rich source data within a common feature-space shared with the target data. The target data is then projected onto the learned SVD-space, aligning it with the distribution of source data, akin to casting into a mold designed for the source data's inherent patterns. This alignment facilitates the learning of patient attributes and patterns from the SVD-transformed source data, making this knowledge leverageable in the target domain within the source-SVD-transformed space. While this process may incur a minor, unavoidable loss of information from the target data, the overall goal is to synchronize data distributions and enable effective transfer of insights across distinct domains.

3. Auto-encoder (AE) based data alignment— extending beyond linearity

While SVD proves effective as a data alignment technique, it operates under the constraint of ‘linear’ transformation. To transcend this limitation and achieve data alignment beyond linearity, deep learning (DL) based distribution learning techniques like AE come into play. Like SVD, AE condenses rich information in a lower-dimensional space, learning the data distribution in the process. In the AE example, the model is initially trained solely on the source data. Subsequently, the target data is passed through the source-data-trained-AE, aligning it with the structure of the source data. This process aims to harmonize data distributions, extending the philosophy of effective transfer of insights across distinct domains beyond linear transformations.

In the three above-mentioned cases, whether through common normalization, SVD or AE, the overarching objective remains consistent: To mitigate the inherent differences in data distribution between source and target data while preserving the essential knowledge that distinguishes patients based on their biomarker status. The alignment of data distributions allows for comparison between the two datasets, enabling a model trained on the transformed source data to be leveraged effectively on the source-aligned-target data. This process

facilitates patient identification across domains by ensuring that the model’s learned patterns are applicable and transferable, nurturing a coherent approach to insights in distinct data landscapes.

Deep adversarial neural network: Integrating data transformation and patient identification

In the previous cases, the execution was designed in two distinct, independent steps. Initially, both the source and target datasets experienced separate transformations to ensure proximity in data distribution. Subsequently, a model was constructed on the transformed source data, and patient identification in the target space was achieved using the transformed target data. It is important that while the initial data transformation brings the data closer by design, it may not be inherently tied to the primary objective of patient identification in the target space during transformation. This highlights the scope of advantage of integrating both steps, where they work in tandem with each other—an advantage offered by deep adversarial neural network (DANN)⁶. This means DANN facilitates a unified approach, bridging the gap between source and target domains by jointly optimizing the transformation and identification processes, thereby enhancing the effectiveness of patient identification in the target space.

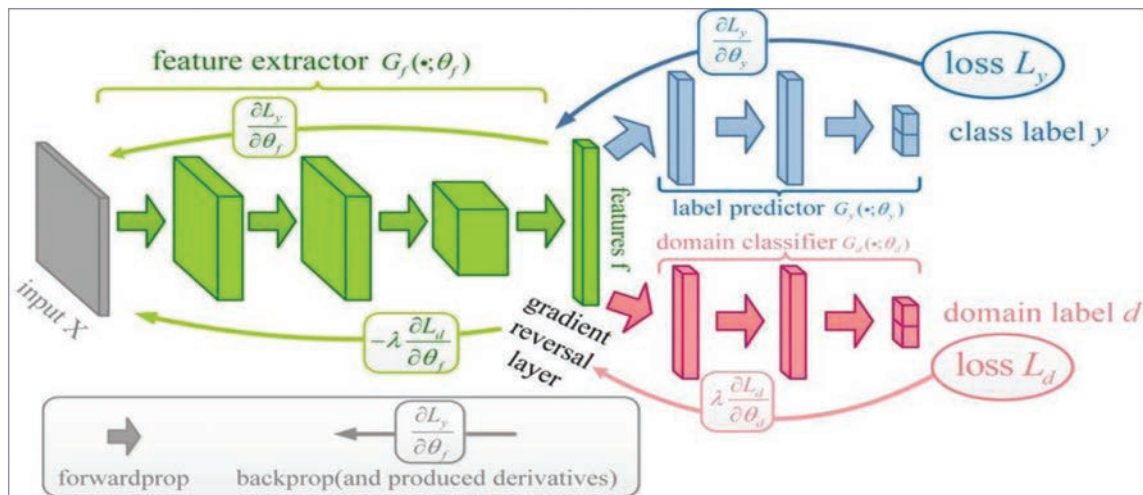


Figure 5: DANN architecture

The original DANN architecture may look complicated, but the underlying design philosophy is simple and intelligent. The detailed architecture can be split into three different but simultaneously working units:

- **Feature extractor:** Applies a common transformation to both source-EMR and target-claims data, ensuring practically undistinguishable source and target patients, and easier differentiation between target biomarker positive vs negative patients in the transformed space.
- **Label predictor patient identifier:** Functions as a ‘classifier’ to identify the intended target biomarker positive patients, similar to the second part in previous solutions.
- **Domain classifier:** Serves as a litmus test to check if source-and-target patients are distinguishable in the common transferred space and provide feedback to the feature extractor on the distinguishability of source-versus-target, suggesting adjustments.

It is worth noting:

- The feature extractor and domain classifier execute the task of transforming the source and target to a space where they become indistinguishable and in proximity.
- The label predictor closely resembles the second-level model to learn the pattern of the target biomarker positive patients in the transformed source space, and it applies the same to the transformed-target space for the target biomarker positive patient identification.

The unique advantage lies in the simultaneous implementation of seemingly two steps, all

tied to a common goal of making the two data sources distinguishable and identifying source the target biomarker positive patients correctly. Once this goal is achieved, the same process is applied to identify potential target biomarker positive patients in the target space.

Ensemble model: Enhancing decision consistency

While each model exhibits unique advantages in terms of efficiency, interpretability, and computational considerations, there often exists a subset of cases where models may not unanimously agree. In binary scenarios like distinguishing between a biomarker positive and negative patient, when two models disagree, one of them is inherently correct. Recognizing that each model carries certain biases—for instance, say, an SVD projection-based model may lean towards tagging patients with folic acid usage as positive, while an AE-based model may be biased towards patients with a low smoking history—our focus is not on pinpointing individual biases but rather on extracting a collective decision from each model.

To achieve this, we employ an ensemble approach, similar to forming a medical board of expert doctors. Each model represents a unique perspective based on its learning and biases. Such as situations where individual doctors may not unanimously agree on a patient’s condition, we establish a rule within the medical board framework, ensuring that the collective opinion of the ensemble always supersedes that of any individual model. In ML terms, this is referred to as an ensemble model. While we will not delve into the technical intricacies of ensemble methods, our primary emphasis remains on their role in complementing and enhancing decision consistency within the broader context of transfer-learning.

Enriching NSCLC claims data with a target biomarker tagging from EMR data sources:

In this section, we will explore the details of the example we have been referencing throughout this paper.

Objective:

The central aim of this case study is to perform a classification of patients within the NSCLC cohort, specifically distinguishing them with respect to a given biomarker as positive or negative. This classification task will be accomplished by harnessing information derived from EMR data sources. In the EMR dataset, patients are already labeled with their corresponding biomarker status. The overarching goal is to showcase the effectiveness of a novel ML solution in accurately categorizing NSCLC patients within the claims dataset, where such initial biomarker information is absent. This process involves training an ML model on the labeled EMR data and transferring the acquired insights to proficiently classify patients within the claims dataset, thereby demonstrating the practical utility of this solution in the realm of pharmaceutical applications.

Data:

- **Claims data:** We possess longitudinal claims data for an identified NSCLC patient cohort. It is essential to emphasize that, in this claims data, biomarker status of the patients is unknown.
- **EMR data:** On the contrasting side, we have EMR data for another identified NSCLC patient cohort. In this dataset, patients are tagged with their biomarker status, thus for the target biomarker, we can have positive/negative labels.

Feasibility assessment:

This crucial step involves conducting exploratory data analyses (EDAs) across both the ‘source’ and ‘target’ datasets to evaluate the feasibility of robustly training a transfer-learning model based on the information available in these two distinct data sources. The assessment encompasses various dimensions, including:

- **Event overlap:** Examination of the number of clinical events—diagnosis, treatment, and procedure—present in both data sources. Ensuring that overlapping events exhibit a sufficiently high capture rate across both datasets.
- **Distribution of events:**
 - » Statistical analysis of the distribution of common events, encompassing metrics such as the median length of patient journeys in both the ‘source’ and ‘target’ datasets.
 - » Evaluation of percentile distribution of lengths, coverage, and other relevant parameters.
- **Count of labels in ‘source’:** It is not ensured to have target biomarker positive/negative tagging of all NSCLC patients even in EMR data. However, the presence of this label is crucial for the intended information transfer. Therefore, ensuring a ‘sufficient’ count of labels becomes imperative for robust model development.

This comprehensive evaluation aims to determine the viability of training an effective transfer-learning model by evaluating the compatibility of information in datasets. The success of the subsequent steps in the transfer-learning process relies heavily on the insights gained from this feasibility assessment.

Patient cohort creation and data preparation:

In this phase, the objective is to identify a suitable set of patients that can be considered representative of the entire NSCLC population. Simultaneously, this patient cohort should exhibit the expected presence of treatment events, substantiated by secondary analysis to facilitate robust downstream numerical analysis. This essential step aligns with standard practices in clinical patient-level modeling analysis.

The process involves collaborative efforts with business and domain experts to formulate business rules for creating patient cohorts in both EMR and claims data sources independently. These business rules are meticulously designed to ensure the selection of patients who not only accurately represent the broader NSCLC population but also possess the requisite treatment events, laying the groundwork for meaningful and reliable numerical analyses in subsequent stages of the study. Going forward, all analyses performed will be based on the defined patient cohorts.

Data difference:

In Figure-1, we have illustrated the disparity in data distribution between Electronic Medical Records (EMR) and claims data. It is this divergence that underlines the necessity for

transfer learning to navigate and address the dissimilarity between the two datasets.

Feature for modeling:

We will refrain from investigating the intricate details of the rationale behind feature engineering, as it is a standard procedure inherent to various processes and not the focal point of this article. Nevertheless, it is noteworthy that our engineered feature encompasses the patients' treatment journey within a stipulated period and that the process of feature engineering from both EMRs and claims data is consistent. Crucially, it is essential to highlight that the inherent data differences observed between the source and target datasets in their raw spaces become evident in the engineered feature space. This underlines the imperative need for transfer-learning to effectively bridge the gap and facilitate accurate modeling across distinct domains.

Model performance evaluation: Unveiling transfer-learning insights

In our quest to assess model performance, the concept of a dedicated proxy validation set was proposed, denoted as 'target-like-source' data. For consistency in our modeling setup, we maintain the same data structure across all stages for ease of comparison.

Data	Domain	Patients	Tx Journey	Target biomarker label
Training	EMR	{Source patients} – {Target like source patients}	✓	✓
Validation	EMR	{Target like source patients}	✓	✓
Prediction	Claims	{All patients in claims data}	✓	?

Figure 6: Model set-up

All reported results are based on the validation dataset & one targeted biomarker. The following

table summarizes the model performance metrics:

Sl.	Model	Precision	Recall	F1-score	AUC
1	Baseline	85%	73%	78%	78%
2	Normalization	91%	88%	89%	84%
3	SVD	93%	87%	90%	89%
4	AE	94%	88%	91%	89%
5	DANN	92%	89%	90%	88%
6	Ensemble	96%	93%	95%	96%

Figure 7: Performance metrics

DANN emerges as a top performer, in the majority of the performance metrics. However, the SVD projection-based model closely follows DANN with nearly negligible computational and time requirements. As anticipated, the ensemble model surpasses each model, demonstrating superior results even on a marginal scale. This underlines the effectiveness of transfer-learning and ensemble techniques in enhancing overall model performance.

Validation of results:

While the numerical model performance metrics exhibit robust performance in the designed validation dataset, the actual assessment of the model's effectiveness in

the real prediction space, such as claims data, remains a critical aspect. Therefore, we adopt a descriptive route to interpret how the model distinguishes potential target biomarker positive patients from the rest in claims data and whether this contrast aligns with the benchmarked known biomarker positive/negative classification in EMR data. This approach not only serves as a validation of the model results but also opens a window for the model to undergo scrutiny by subject matter experts (SMEs) before its implementation in real-world business use cases. The descriptive analysis adds a layer of assurance and comprehensibility to the model's performance in practical scenarios.

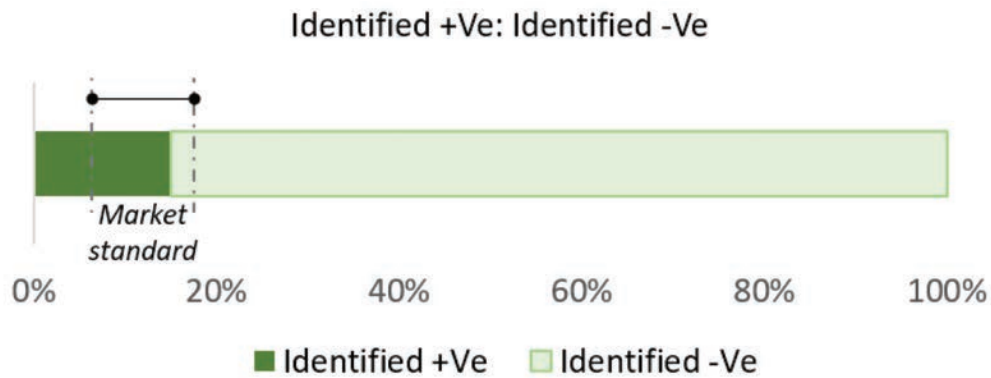
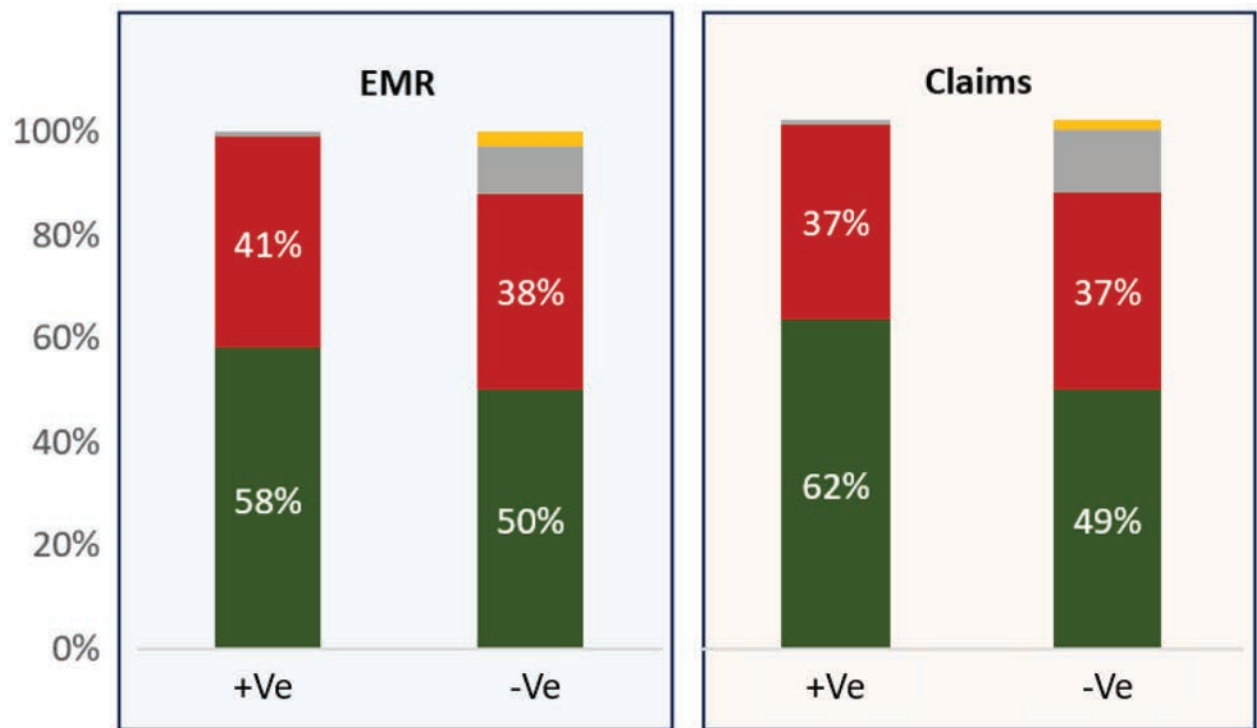


Figure 8: Business validation of the identified patients

The initial and fundamental validation lies in the proportion of patients identified as target biomarker positive. Based on secondary benchmarks, approximately 15% of NSCC are expected to have a positive biomarker status. Remarkably, our solution identifies around

17% of patients as biomarker positive, aligning closely with the expected ratio. This encourages us to advance to the next step, where we delve into the characteristics of the identified 17% positive patients.



Given the distinct distributions of the two datasets, a direct comparison between known EMR target biomarker positive patients and their counterparts in claims data is not feasible. However, a strategic approach involves examining the directional shift in patient characteristics (e.g., drug utilization) from positive to negative. Despite potential dissimilarity in specific patient characteristics between identified patients and EMR patients due to inherent distribution differences, the expectation is that the overall directional shift from positive to negative remains consistent.

As evident from the plotted data, in the known positive/negative split patients, positive patients exhibit a higher tendency to consume VEGF compared to their counterparts. While the exact drug share percentages may not necessarily align closely in identified patients-split from the claim space, the observed trend of reduced VEGF usage in negative compared to positive is consistent in the identified patients. This pattern persists across various dimensions of patient characteristics.

Conclusion:

The application of transfer-learning in healthcare is transformative, particularly in enhancing patient classification within real-world claims databases. This advancement significantly contributes to a multitude of commercial activities, such as precise targeting, segmentation, forecasting, opportunity and latent demand assessment and care gap identification. Beyond these commercial applications, transfer-learning opens avenues for deploying algorithms on EMRs to proactively share patient risk scores with physicians. These transfer-learning models, adaptable to diverse data sources, can be fine-tuned based on the target data source. As the model undergoes training with data from various sources, its accuracy improves, allowing for enhanced scalability and generalizability. Pharmaceutical companies can leverage these powerful models to deploy EHRs and proactively engage with physicians, facilitating proactive interventions through patient service programs and ensuring patients access the right treatment.

Future scope:

While transfer-learning has demonstrated its potential in confidently classifying patients within a dataset using patterns learned from another dataset, there is ongoing research to further improve the scalability and generalizability of this approach. Two key dimensions are currently under exploration:

- **Explore synthetic data generation:** The application of generative AI is being investigated to enrich representation within the 'Source' dataset. This initiative aims to create synthetic data that closely mimics real-world scenarios, fostering a more robust validation process.
- **Federated learning:** Another avenue of research involves federated learning, a methodology enabling cross-learning across multiple data sources simultaneously. This approach prioritizes data privacy and security, addressing the challenges associated with centralized data processing. The ongoing endeavors in these dimensions signify a commitment to pushing the boundaries of transfer learning, aiming to make it even more effective and applicable in diverse healthcare scenarios.

About the Authors:

Aniruddha Mitra: Aniruddha is a data science consultant at ZS, spearheading efforts in life science AI, specifically focused on patient journey analytics using RWD. With over eight years of experience, Aniruddha has a diverse background, having implemented various data-based solutions across sectors like manufacturing, fintech and banking.

Atharv Sharma: Atharv is an advanced data science manager with ZS based out of Philadelphia and is a core member of ZS's RWD and insights venture. Atharv has over nine years of experience in applying ML solutions in healthcare, and primarily helping the commercial pharma industry realize the enhanced impact of leveraging RWD advanced analytics and ML-based techniques.

Arkojyoti Das: Arko is an advanced data science manager based out of India, leading ZS's product development of patient clustering and patient identification capabilities. Arko has over seven years of expertise in AI development, deep Learning applications in healthcare and specializes in RWD and insights.

Arrvind Sunder: Arrvind is a principal from ZS's Evanston office and has been with ZS for 16 years. Arrvind is a leader in ZS's RWD and insights venture where he focuses his time on enabling clients to realize competitive differentiation through data, digital and AI.

Srinivas Chilukuri: Srinivas is the leader of ZS's AI center of excellence in New York, driving innovation in frontier AI. With expertise in automated ML, natural language processing and predictive analytics, Srinivas has authored thought leadership articles and regularly speaks at conferences.

Shubhendu Trivedi Shubhendu is a research scientist with ZS Associates based out of the New York office attached with the AI center of excellence. Prior to ZS, Shubhendu was a research scientist at the MIT CSAIL, working on problems in drug discovery and molecular dynamics.

References:

1. The National Human Genome Research Institute (NHGRI): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7397223/>
2. The Personalized Medicine Coalition: <https://www.personalizedmedicinecoalition.org/>
3. Enhancing Patient Classification and Staging in RWD Using Machine Learning - https://www.pmsa.org/_resources/documents/journal/Journal-of-PMSA-Spring-2023.pdf (p26-36)
4. AI/ML in Precision Medicine: A Look Beyond the Hype: <https://link.springer.com/article/10.1007/s43441-023-00541-1>
5. Unsupervised Vehicle Counting via Multiple Camera Domain Adaptation: <https://arxiv.org/pdf/2004.09251.pdf>
6. Domain-Adversarial Training of Neural Networks, <https://arxiv.org/pdf/1505.07818.pdf>

Navigating Uncertainty: Evaluating Risks to Enhance Drug Sales Forecast

Rajat Parmar, Manager, Viscadia; Rishabh Bawa, Consultant, Viscadia; Himanshu Khandelwal, Associate Consultant, Viscadia; Varun Singh, Associate, Viscadia; Sakshi Devi, Associate, Viscadia

Abstract: Projection of sales of new drugs has always been complex. It is influenced by a dynamic environment characterized by intricate market trends, evolving regulatory landscapes, regional dynamics, and diverse patient demographics. Ensuring precise forecasts is a top priority for manufacturers, however hitting bullseye is generally a tall order. Forecasters are often left dealing with force-fitted structures, imperfect assumptions, and a lack of clarity on how the forecast would be used. Each forecast is different and needs to be treated as such. For example, if doing a regional product forecast in Europe, what pricing assumptions should we consider? Should we simply use the price assumptions from the first launched country, or decide on country-specific regulations that might affect understanding? The output in turn would be used differently too - it would have varied uses such as revenue planning, resource allocation, setting targets for field force, compensation plans, partnering, and R&D. These multi-dimensional factors generate risks and uncertainties in forecasts, creating a need for a comprehensive method to enhance the precision of drug sales prediction.

Risk and uncertainty in a forecast point to variability and lack of precision – these can be minimized by showing a range of possible outcomes or by scenario planning, instead of just showing the most probabilistic future. Running Sensitivities and what-if scenarios are also often considered an effective way to mitigate risks. Although these methods do curb the uncertainty in some way, they don't always help answer questions on probabilistic scenarios like 'How certain are you that we will achieve our sales forecast?' In this context, Monte Carlo Simulation emerges as a powerful tool that can map scenarios and sensitivities with a high probability of an outcome to happen and give a statistically confident range of forecasts.

This paper explores the application of the Monte Carlo method to model complex market dynamics, account for multifaceted uncertainties, and generate a spectrum of potential sales scenarios. We would present a case on how the outcomes from the Monte Carlo method helped a pharmaceutical drug manufacturer in exploring the probabilities of various sales trajectories by utilizing historical sales data, parameter estimation, and scenario planning. The resulting distribution of forecasted sales outcomes equipped the stakeholders with a comprehensive understanding of potential risks and opportunities. With its ability to accommodate different drivers and barriers of sales and unforeseen disruptions, the Monte Carlo method has the potential to emerge as a vital approach for pharmaceutical companies seeking robust and adaptable drug sales forecasts in an increasingly dynamic industry landscape. Furthermore, the paper delves into future directions, envisioning technological advancements and the integration of Monte Carlo simulation with artificial intelligence and machine learning.

Keywords: Pharmaceutical Forecasting, Monte Carlo Simulation, Drug Sales, Uncertainty Management, Risk Mitigation, Supply Chain Optimization, Monte Carlo Advancements

1. Background

Forecasting assumes a pivotal role in shaping strategic decisions within the pharmaceutical industry, conferring a significant competitive advantage through the anticipation of market trends and optimal resource allocation. This paper meticulously explores the essential nature of forecasting, underscoring its significance in fostering confidence in predictions specific to the pharmaceutical sector. The inherent high uncertainty in market parameters within the healthcare industry poses a recurrent challenge of fluctuating forecasts, influencing firm strategies and demand estimates.

To enhance forecast accuracy, we employ Monte Carlo techniques—a methodology deeply rooted in finance, cyber security, quantum computing, and water forecasts. The integration of Monte Carlo techniques proves indispensable in mitigating this variability, providing a nuanced risk assessment essential for effective forecasting. Furthermore, our investigation extends to the integration of Machine Learning and Artificial Intelligence, enhancing the sophistication and historical refinement of the Monte Carlo processes applied. This comprehensive approach aims to address the unique challenges of forecasting in the pharmaceutical landscape, ensuring more reliable and precise predictions for informed and strategic decision-making that align with the dynamic nature of the industry.

2. What is Forecasting and Why it is Important

2.1 Introduction

Forecasting in the pharmaceutical industry involves predicting future demand and market trends for pharmaceutical products. This intricate process relies on a comprehensive

analysis of historical data, market research, and consideration of key factors such as disease area landscape, addressable population, unmet needs, competitive environment, and patient journey. The primary objective of forecasting in the pharmaceutical sector is to empower companies to make informed decisions related to production, inventory management, pricing, and marketing strategies. Accurate demand forecasting enables pharmaceutical companies to optimize production processes, maintain adequate stock levels, and strategically position themselves to maximize market share and profitability.

Furthermore, forecasting in the pharmaceutical industry aids companies in anticipating shifts in healthcare needs, enabling them to align research and development efforts accordingly. This encompasses considerations of the patient journey, including treatment pathways, adherence rates, and patient preferences. Additionally, it comprises considerations related to regulatory approvals and market access. Pharmaceutical companies must anticipate the timing and outcome of regulatory reviews for new drug candidates or modifications to existing products. These approvals significantly impact product availability and market penetration, making them critical factors in forecasting.

Another critical area for forecasting within the pharmaceutical Industry is market access. Market access involves understanding the reimbursement landscape, payer preferences, and healthcare policies that influence access to pharmaceutical products. Forecasting in this context involves assessing how reimbursement decisions and formulary placements may affect product uptake and market share. Moreover, forecasting may involve evaluating the potential impact of health technology assessments (HTAs) and comparative effectiveness research (CER) on product adoption and pricing strategies.

Overall, forecasting in the pharmaceutical industry is a comprehensive and dynamic process that involves considering a wide range of factors, including regulatory approvals, market access, international markets, risk assessment, and scenario planning. By integrating these aspects into their forecasting strategies, pharmaceutical companies can make informed decisions and effectively navigate the complexities of the healthcare landscape¹.

2.2 Importance of Forecasting in the Pharmaceutical Industry

Forecasting plays a pivotal role in the pharmaceutical industry across various domains, driving strategic decision-making and fostering operational efficiency. Firstly, in research and development planning, accurate predictions of future market demands enable pharmaceutical companies to effectively allocate resources and prioritize projects, ensuring alignment with emerging therapeutic needs. This foresight not only optimizes R&D efforts but also enhances the potential for successful product development and market acceptance. Considering the enormous investment required for drug development in terms of both time and resources, it becomes essential for pharmaceutical companies to analyse how the drug will perform after launch, as well as various factors that may influence its performance. By projecting future market demand and potential revenue streams, companies can assess whether the expected returns justify the significant investment needed for drug development, as well as calculate the Return on Investment through different scenarios, enabling more informed decision-making regarding resource allocation and project prioritization. Moreover, forecasting extends beyond R&D, influencing resource allocation in other functions such as marketing. By anticipating market trends,

companies can allocate resources efficiently across sales force, advertising costs, and other marketing initiatives. Additionally, forecasts serve as a basis for setting targets and allocating incentives to the workforce, aligning individual and team efforts with organizational goals and market projections.

Moreover, precise forecasting facilitates supply chain optimization by minimizing inventory costs, reducing wastage, and ensuring seamless operations. Major pharmaceutical players like Pfizer and Johnson & Johnson have exemplified this through reported reductions in inventory holding costs and improved stock availability, achieved through advanced forecasting methodologies. These benefits extend beyond industry giants, as smaller firms leverage partnerships with analytics providers to implement agile and cost-effective inventory management solutions.

Additionally, forecasting aids in regulatory compliance by predicting timelines for approval processes, thereby facilitating timely product launches while maximizing market exclusivity. Lastly, forecasting serves as a crucial tool for risk mitigation, allowing companies to identify and prepare for potential market disruptions, ultimately enhancing resilience and adaptability in navigating the dynamic pharmaceutical landscape. By integrating robust forecasting methodologies across these key areas, pharmaceutical companies can drive innovation, streamline operations, and mitigate risks effectively in an increasingly competitive market environment.

2.3 Current Situation of Forecasting within the Pharmaceutical Industry

The current situation of forecasting within the pharmaceutical industry reveals several challenges contributing to inaccurate

forecasts. These challenges stem from the unique characteristics of pharmaceutical products and the complexities of the market dynamics. One such challenge is accurately predicting the timing and impact of regulatory changes on market demand. Pharmaceutical companies must anticipate how changes in healthcare policies will affect patient access to medications, healthcare provider prescribing practices, and payer reimbursement decisions. However, the complexity and uncertainty surrounding healthcare policy development and implementation make it challenging to forecast the precise impact on product demand accurately.

Moreover, benchmarking studies indicate that forecasting in the pharmaceutical industry still lack the sophistication required to capture the intricacies of market dynamics accurately. As a result, forecasting errors can be substantial, with error rates sometimes reaching around 40%². Additionally, despite the abundance of potentially useful data for more accurate forecasting, such as promotion activities and their effects on sales, the industry faces limitations in data utilization. Challenges such as different data formats, lack of integration tools, data collection times, and data actuality hinder the effective leveraging of available data for improving forecasting accuracy. Overall, these factors contribute to the current situation of inaccurate forecasts within the pharmaceutical industry, highlighting the need for more advanced methodologies and tools to enhance forecasting capabilities.

2.4 Strategies for Improved Forecasting

As discussed in section 2.2, forecasting plays a vital role in the pharmaceutical industry to optimize resource allocation, research and development planning, and risk mitigation, as well as in dealing with regulatory compliance. Hence, it becomes cardinal to ensure precision

in forecasts. Inaccuracies in the forecasts can lead to stockouts or excess inventory, affecting drug availability and operational costs. Timely and precise predictions ensure optimal resource allocation, enabling pharmaceutical companies to navigate regulatory requirements, minimize production disruptions, and meet patient needs efficiently. The use of strategies like sensitivity analysis, scenario planning, and methods like Monte Carlo simulations can play a significant role in enhancing the accuracy and reliability of forecasts.

Sensitivity analysis involves systematically varying key input parameters or assumptions in a forecasting model to assess their impact on the output. By testing the sensitivity of the forecast to changes in variables such as pricing, market demand, or regulatory factors, pharmaceutical companies can better understand the potential range of outcomes and identify areas of vulnerability or uncertainty. For example, in forecasting the demand for a new drug, sensitivity analysis can assess how variations in factors such as pricing strategy, competitor actions, or changes in regulatory policies may affect sales projections. This proactive approach allows companies to develop contingency plans and make informed decisions to mitigate risks and optimize resource allocation in response to changing market conditions.

Monte Carlo simulations are a more advanced tool, which involves using random sampling and statistical modelling to predict the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. In pharmaceuticals, this can be applied to model the success rates of clinical trials or to forecast market dynamics. By running thousands or even millions of simulations, each with randomly selected variables, Monte Carlo simulations provide a distribution of possible outcomes, offering

insights into the likelihood and risks of various scenarios.

By integrating these methods, pharmaceutical companies can enhance the accuracy and reliability of their forecasts. Confidence intervals provide a realistic range of outcomes, scenario planning prepares for multiple future possibilities, and Monte Carlo simulations offer a probabilistic assessment of complex, uncertain processes. Collectively, these methods enable more informed decision-making, leading to efficient resource allocation, successful navigation of regulatory landscapes, and ultimately better patient outcomes.

3. Monte Carlo Simulation

3.1 Monte Carlo Simulation Overview

Monte Carlo simulation, named after the renowned gambling hub of Monaco, is a powerful methodology with diverse applications. Originating in the 1940s at Los Alamos National Laboratory, it gained significance during the Manhattan Project.³ The simulation heavily relies on probability, reflecting its association with casinos that leverage probability for long-term gain. It allows for a dynamic exploration of potential outcomes by simulating the impact of various uncertain variables, mirroring the randomness and unpredictability inherent in real-world scenarios.

Drug forecasting, with its myriad uncertainties encompassing clinical trials, market dynamics, and regulatory changes, can benefit significantly from Monte Carlo simulation. After running the simulation, the forecasting model generates a probability distribution of potential outcomes depending upon the problem statement, such as net present value, market share, and revenue projections. This distribution allows decision-makers to assess the likelihood of meeting

specific targets or facing challenges.

For example, the simulation might suggest that there is a 60% chance of achieving a certain revenue target within a specified period. Additionally, it provides insights into the factors contributing most to uncertainty, helping stakeholders focus on areas that may require more attention or risk mitigation strategies.

3.2 Benefits of Monte Carlo Simulation in Drug Forecasting

3.2.1 Improved Decision Making

In the landscape of drug forecasting, the speed of decision-making processes is one of the most critical metrics. Monte Carlo simulation significantly expedites decision-making by providing probabilistic insights and facilitating scenario analysis. Traditional forecasting methods often rely on deterministic models, which may oversimplify complex situations. Monte Carlo simulation, on the other hand, embraces the probabilistic nature of drug development and market dynamics⁴.

Using these findings, the company can make informed decisions to align with expectations and meet the required demand. Instead of relying on a single forecast, the Monte Carlo simulation empowers decision-makers to consider a spectrum of possibilities, enhancing their ability to adapt strategies in response to evolving circumstances. This increased agility is particularly crucial in the dynamic and uncertain landscape of drug development and market dynamics.

3.2.2 Risk Mitigation

Monte Carlo simulation is a useful tool for identifying and mitigating risks associated with inaccurate forecasts in drug development. Traditional forecasting methods may not

be effective in capturing the complexity and variability inherent in drug development and market conditions, which can lead to inaccurate predictions and unforeseen challenges. By using Monte Carlo simulation, pharmaceutical companies can systematically explore the potential risks associated with different variables and their combinations.⁵ This proactive approach to risk identification allows for timely interventions and adjustments to strategies, ultimately mitigating the impact of uncertainties on drug development projects.

3.3 Challenges and Considerations

3.3.1 Data Quality and Availability

One of the key factors in developing reliable Monte Carlo simulations is the availability and quality of input data. Inaccurate or incomplete data can significantly compromise the accuracy and effectiveness of forecasting models. To overcome these challenges, it is necessary to adopt robust data collection methodologies and ensure that input data is both accurate and complete. This may involve collaboration with various stakeholders, such as healthcare providers, clinical researchers, and data analysts. Additionally, utilizing advanced data cleaning techniques and validation processes can help improve the overall quality of input data⁶.

To address these challenges, the following strategies may be implemented:

1. **Data Cleansing and Validation:** Implement rigorous data cleansing processes to identify and correct inaccuracies, inconsistencies, and outliers. Validation checks are essential to ensure that the available data is reliable.
2. **Data Augmentation:** Enhance the dataset by incorporating additional relevant

information through data augmentation techniques. This could involve leveraging external data sources, real-world evidence, or expert opinions.

3. **Continuous Monitoring and Updating:** Establish a system for continuous monitoring of data quality. Regular updates and improvements should be made to adapt to changes in the pharmaceutical landscape, ensuring the model remains relevant.

3.3.2 Model Complexity

The balance between model simplicity and accuracy is a perpetual challenge in Monte Carlo simulation for drug forecasting. While complex models may capture a more comprehensive range of variables, they often introduce challenges related to the computational burden, interpretability, and the risk of overfitting. Striking the right balance is crucial to ensure that the model remains practical and interpretable while still capturing the essential dynamics of the system.

Considerations for managing model complexity include:

1. **Parameter Sensitivity Analysis:** Perform a comprehensive analysis to determine the most influential variables. The model can be simplified without compromising accuracy by focusing on these factors.
2. **Model Validation:** Thoroughly validate the model by using historical data and real-world scenarios. This process helps assess predictive accuracy and identifies areas where simplification can occur without sacrificing reliability.
3. **User-Friendly Interfaces:** Design interfaces that allow stakeholders to

interact with the model without delving into intricate complexities. Visualization tools and dashboards aid in conveying insights derived from the model in a comprehensible manner.

3.3.3 Missing Data

Dealing with missing data is a significant challenge in Monte Carlo simulation, particularly in drug forecasting. Such missing information can arise due to incomplete clinical trials, patient dropouts, or data collection errors. To handle this problem, various strategies can be implemented, including imputation techniques like mean imputation or multiple imputation, and careful consideration of the assumptions underlying these methods. Incorporating advanced statistical methods such as Bayesian approaches can provide a more robust framework for handling missing data by explicitly modeling uncertainty. It is crucial to ensure that the chosen imputation strategies align with the specific characteristics of the pharmaceutical data, which requires collaborative efforts between statisticians and subject matter experts⁷.

4. Case Study

There are several case studies available to understand the practical applications and benefits of Monte Carlo simulation. This section provides a detailed analysis of one such case that demonstrates the tangible benefits of using Monte Carlo simulation.

4.1 Objective

Accurately estimate the future demand volume for the pharmaceutical company's drug and forecast the corresponding net revenue. This analysis will facilitate strategic production planning for the manufacturing team, ensuring alignment with the company's projections for

the next few years.

4.2 Methodology

To determine the expected demand accurately, a comprehensive understanding of the drug, its associated disease context, and relevant factors is crucial. The specified drug in the case study is commonly prescribed for a prevalent disease and hence involves multiple influencing factors. Consequently, a single forecast proved insufficient for estimating the demand. The requirement for different forecasts, each considering varied values for the variables, led to adopt the Monte Carlo Simulation approach. This section delves into the several key steps involved in Monte Carlo simulation to effectively model different scenarios and determine the range of possible outcomes for drug forecasting⁸.

Step 1: Identify Key Variables:

The first step in a Monte Carlo Simulation is to define the key variables involved in the drug forecasting model. These variables could include parameters such as market demand, pricing strategies, and competitive landscape. Two factors are considered when selecting variables for this simulation: sensitivity and confidence. The grid below shows that variables with high sensitivity and low confidence are the most suitable for the Monte Carlo simulation. Variables with high sensitivity and high confidence, low sensitivity and low confidence, can also be used but their inclusion may impact the sample size. Variables with low sensitivity and high confidence should not be included.

For the given case study, variables such as the total patient pool that will utilize the drug, fill rate sensitivity, persistence, peak market share, payer access and packs per patient per year were considered.

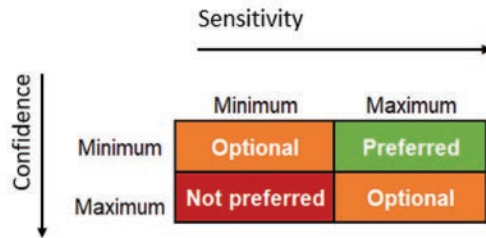


Fig 1: Types of variables suitable for Monte Carlo simulation

Step 2: Define Variable Uncertainty:

After defining the variables, the next step is to assign uncertainty through probability distributions to these variables. This allows for the incorporation of uncertainty into the model, as real-world scenarios often involve varying degrees of unpredictability. If the user has an understanding of the input's behaviour, this step is relatively simple. However, in case of insufficient information, the user may opt to initially utilize a uniform distribution, and as more data is gathered, a more suitable distribution can be employed.

It is very important to understand the type of probability distribution that would be leveraged for each input metric since it directly impacts the sample size and completion time required for simulation. A probability distribution is a mathematical function that helps to generate

random input samples, by providing the range of possible inputs with their associated probabilities. In the context of Monte Carlo simulations, the following types of probability distributions can be employed to model different uncertainties.

1. Normal Distribution: The normal distribution, also known as the Gaussian distribution, is a symmetrical probability distribution that is commonly used in statistics. It is called a bell curve because its visual representation takes the shape of a bell, with a peak at the mean and a downward-sloping curve on each side.

The data that is closer to the mean is more likely to occur than the data that is farther away from it. A normal distribution is used when the uncertainty associated with a variable is low. e.g. GTN of the drug.

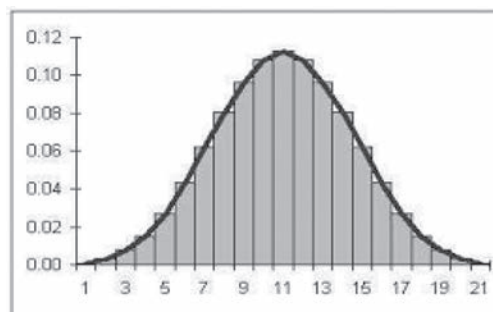


Fig 2: Normal Distribution

The normal distribution can be calculated using the formula below. It only requires the mean (μ) and standard deviation (σ) values to be known.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where:

x= the value of the variable or data being analysed

f(x) =the probability function.

μ = the mean

σ = the standard deviation

2. Uniform Distribution: The uniform distribution is a type of probability distribution that assigns an equal probability to all possible outcomes. It is often referred to as the “rectangular distribution” due to its appearance on probability distribution plots. The probability density remains consistent throughout the distribution, resulting in a flat and even curve. The uniform distribution is typically used when there is a high level of uncertainty associated with a variable, such as the potential impact of a competitor’s product launch on our share of the market.

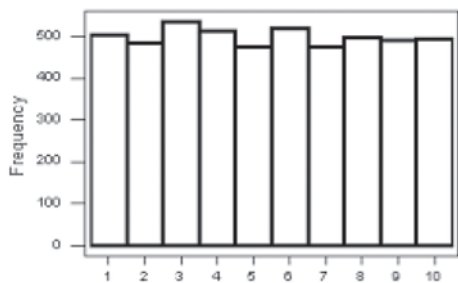


Fig 3: Uniform Distribution

3. Discrete Distribution: This type of distribution is used wherein multiple possible outcomes have different probabilities of occurrence. This is used to model the likelihood of discrete events. e.g. a product undergoing clinical trials can come out to be more / less / equally efficacious in comparison to the competitor drug, leading to three discrete uptake scenarios. Assuming that the graph above represents the market share of a drug, there are three possible scenarios that could occur. The first scenario is where the market share is 30% with a probability of 20%. The most likely event is where the market share is 35%, with a probability of 60%, and similarly, there is a 20% chance of the market share being 40%.

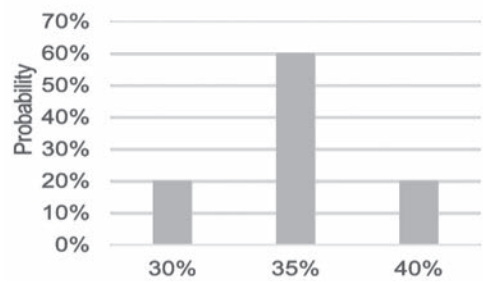


Fig 4: Discrete Distribution

The table below provides a comprehensive list of all the metrics that were considered to estimate the required volume along with the probability distribution used.

All the above-mentioned factors form the variables in the Monte Carlo Simulation, which would fluctuate resulting in multiple scenarios. Since the uncertainty attached to these variables is high, uniform distribution was used for most of the inputs.

INPUT METRICS	PROBABILITY DISTRIBUTION	BASE SCENARIO	MC RANGE		CONFIDENCE LEVEL (FOR NORMAL DISTRIBUTION)
			LOW	HIGH	
Prevalence / Incidence	Normal	1.6M / 233K	1.4M / 207K	1.8M / 266K	95%
% Return of LTFU patients	Uniform	20%	10%	30%	NA
% YoY growth of PD treatment rate	Uniform	3.50%	2.50%	4.50%	NA
Number of episodes (using split across patient segments)	Normal	7.3	5.5 (-25%)	9.2 (+25%)	80%
Peak adoption	Normal	42% / 49%	-0.05	0.05	95%
Persistence inputs	Uniform	-	-40% relative	+20% relative	NA
Payer access (% payers giving Tier 2 access)	Discrete	50%	30%	70%	NA
% Patients with coupons	Normal	80%	70%	90%	90%
Fill rate and OOP mapping	Uniform	-	-20% relative	+20% relative	NA
% Patients consuming 2 dose per episode	Normal	50%	30%	70%	80%

Table 1: Input variables considered in the case study along with the distribution type

Step 3: Generate Random Samples:

Once the variables and their distributions are defined, the simulation involves generating a large number of random samples from these distributions. This step allows for the creation of diverse scenarios that cover a wide range of potential outcomes. The sample size, denoted as 'n,' is a critical factor. The number of samples directly impacts the time required for the simulation, with larger sample sizes reducing the standard deviation of the variable. However, the relationship between the number of samples and the increase in time is not linear, necessitating careful consideration when determining the appropriate number of simulations for the desired accuracy.⁹

The number of iterations required can be increased or decreased based on the number of input variables being considered, and the type of uncertainty being introduced. Therefore, steps 1 and 2 i.e., identifying the right variables and their distribution are very crucial in Monte Carlo simulation.

Below given formula was used to determine the number of iterations:

$$\text{Number of iterations} = ((Z_c * S_x) / \text{Error})^2$$

where Z_c depicts the confidence interval, S_x is the standard deviation of the data, and the error represents the permissible error.

Another way is to plot a curve with the incremental number of samples and the point where the graph stabilizes is the optimal number of iterations. For this case, a confidence interval of 98% and an error of 1.25% were considered, resulting in the requirement of 2000 iterations.

Step 4: Run the Simulations:

The generated random samples are then used to run the simulations. By inputting these samples into the model, the simulation calculates the corresponding outcomes based on the defined variables and their distributions.

4.3 Interpreting the Results

After running a substantial number of simulations, the results are analysed to determine the range of possible outcomes and their associated probabilities. This analysis provides valuable insights into the potential sales and demand for drugs under various circumstances. Different types of visual representations emphasize various aspects of an analysis. The choice of visual representation is based on the problem statement and can significantly impact the analysis of the results. To provide multiple perspectives, both the

Bayesian prediction cone and Cumulative probability curve were leveraged to visualize the simulation results.

A cumulative probability curve, also known as a cumulative distribution function (CDF), is a right-continuous, monotone increasing function that uniquely identifies every probability distribution supported, whether discrete, continuous, or “mixed.” It is often represented graphically as an S-shaped curve, and it can be used to calculate various probabilities, such as the probability of a random variable being below a certain value or within a specific range.

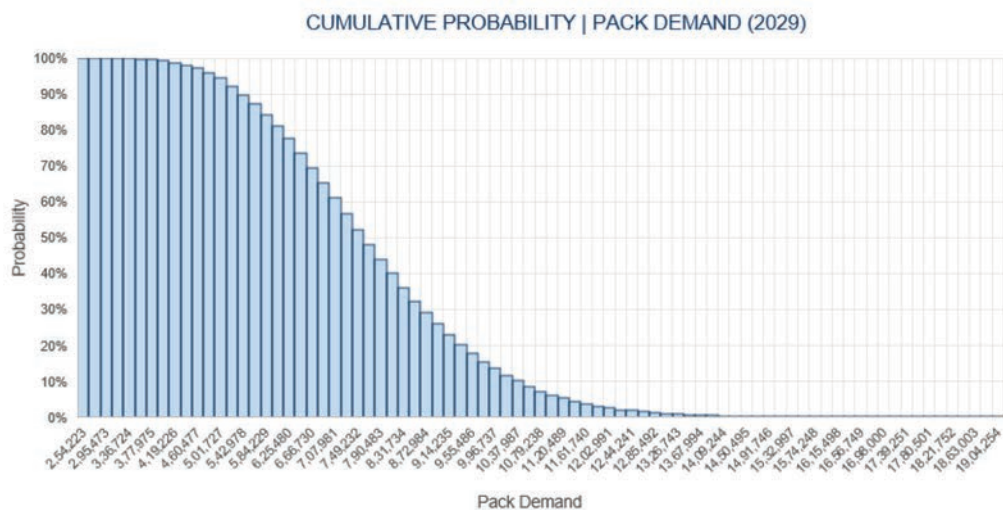


Fig 6: Probability of pack demand achieving certain values using Cumulative Probability Curve

The graph above displays the probability of pack demand achieving specific values:

- 100% probability that the pack demand is at least 254K
- 80% probability that the pack demand is at least 611K

- 40% probability that the pack demand is at least 811K

- 0% probability that the pack demand is at least 1.9M

This data can be used to determine future goals and requirements based on likelihood. Another representation of the same can be observed

in the Bayesian prediction cone, where the different forecast scenarios are highlighted as well. The Bayesian charts represent data as a cone-shaped distribution, providing a more comprehensive understanding of the range of

possibilities and associated probabilities. The graph below shows a range of possible outcomes and highlights the probability of each outcome falling within specific bands.

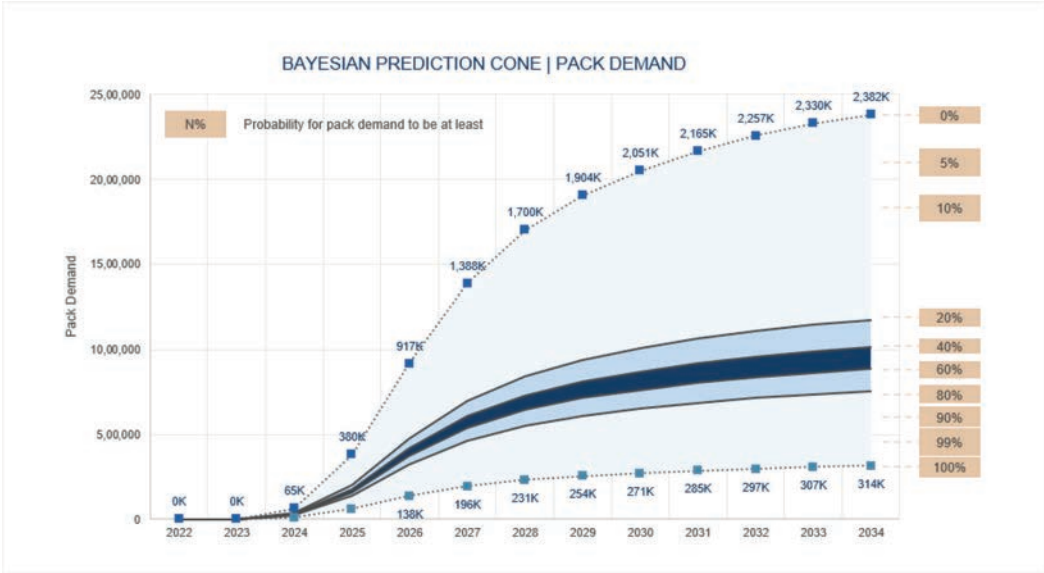


Fig 7: Probability distribution of pack demand using Bayesian prediction cone (High and Low Scenario)

The graph shown above provides a detailed overview of the probability of achieving certain pack demand in the upcoming years, presenting both high and low scenarios. It is almost impossible to achieve a pack demand of 1.9M, whereas it is certain that 254K packs will be required in the year 2029. It can also

be observed that there is 20% possibility that the demand lies within the dark blue band. In comparison, the below graph provides the probability of the demand highlighting the base scenario. The probability of achieving the base forecast as per the analysis is around 26%.

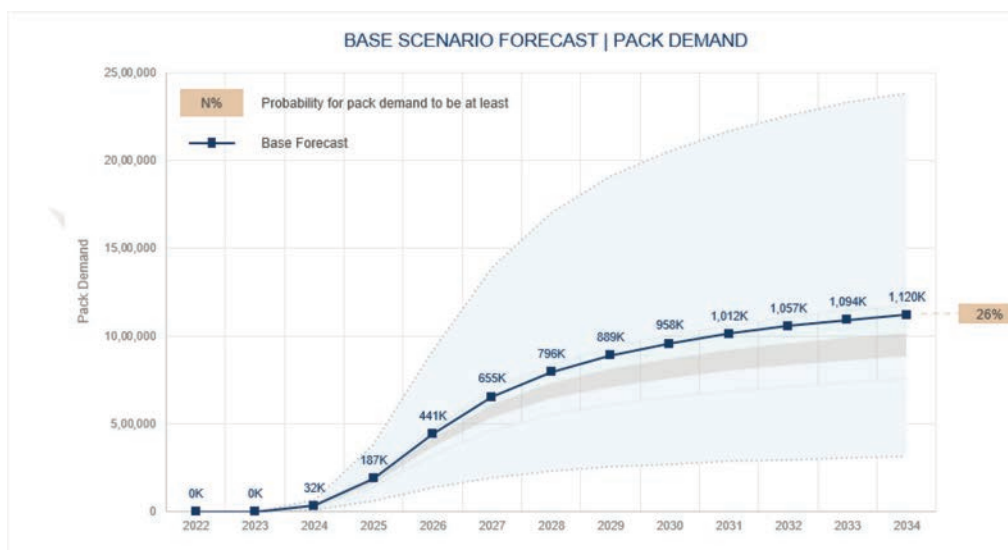


Fig 8: Probability distribution of pack demand using Bayesian prediction cone (Base Scenario)

The above-mentioned graphs can be recreated for the net revenue of the company in the upcoming years. These graphs can be very useful for the leadership team to set goals and targets as they provide the probability of achieving the target. This information can be used to make more informed strategic decisions, allocate resources effectively, and develop risk management strategies to navigate a rapidly changing market.

5. Future Directions and Recommendations

5.1 Advancements in Technology

Monte Carlo simulation stands as a cornerstone in drug forecasting, continuously evolving to meet the dynamic challenges of drug forecasting. However, it has various limitations that hinder its effectiveness, such as high computational demands, which require substantial resources due to the need for numerous iterations to generate reliable forecasts. Convergence issues further complicate matters, particularly when

modelling complex pharmaceutical systems with high dimensionality. Additionally, the method's inability to incorporate exact solutions means it provides only approximate outcomes rather than precise pharmaceutical relationships. Lastly, it can sometimes be difficult to Interpret these results and require careful analysis and communication, especially to non-experts in the pharmaceutical field.

The integration of emerging technologies can not only tackle these limitations but also Increase the effectiveness of traditional Monte Carlo Simulations, propelling drug forecasting to new heights of effectiveness. The use of High-Performance Computing (HPC) emerges as pivotal in this regard. HPC Involves the use of powerful computers and advanced algorithms to process vast amounts of data quickly. It leverages parallel computing using GPUs, such that computational tasks are distributed among multiple processing units or nodes simultaneously. This reduces the time required for computations significantly, improving the efficiency and scalability of Monte Carlo simulations¹⁰. Further, by harnessing

the capabilities of distributed computing frameworks like Apache Spark and Hadoop in addition to HPC, Monte Carlo simulations can overcome current limitations, paving the way for real-time insights and predictive analytics in drug forecasting.

These capabilities can be further enhanced, using advanced statistical techniques such as bootstrapping and Latin hypercube sampling. These techniques leverage probability, to augment the precision of Monte Carlo simulations, as well as decrease computations needed while maintaining relatively high accuracy. Bootstrapping generates multiple datasets from the original data, which can further be analysed, improving the reliability of forecasts. Latin hypercube sampling involves dividing the range of each input variable into equally probable intervals, while ensuring that each interval is sampled exactly once, efficiently exploring the input space, and providing a more representative sample compared to traditional random sampling methods. By synthesizing these methodologies with surrogate modelling approaches, researchers can navigate the complexities of drug forecasting with heightened efficiency and accuracy. Surrogate models, by their very nature, offer simplified approximations of intricate systems, allowing for expedited simulations without compromising accuracy.

Lastly, advanced technologies such as Digital Twins can provide a personalized and dynamic representation of patients or clinical trial cohorts and allow for highly tailored Monte Carlo simulations in drug forecasting. By simulating virtual populations with individual characteristics and disease profiles, digital twins can enable more accurate predictions of drug responses and treatment outcomes across diverse patient demographics. This personalized approach could be pivotal in enhancing the precision and effectiveness of drug forecasting

models, ultimately leading to better-informed decision-making in pharmaceutical development and commercialization.

5.2 Integration with AI and Machine Learning

As discussed in 7.1, Monte Carlo simulations serve as an invaluable tool, yet it faces challenges in accurately predicting pharmaceutical outcomes due to the inherent complexities of the industry. Traditional forecasting methods leveraging Monte Carlo simulations often struggle to capture the multifaceted nature of drug efficacy, safety, and market demand, leading to uncertainties that can hinder decision-making processes. The integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques holds immense potential in overcoming these challenges. By leveraging AI and ML algorithms to analyse vast datasets and uncover intricate patterns, Monte Carlo simulations can be enhanced to provide more accurate, reliable, and actionable forecasts in pharmaceutical development and commercialization.

One of the ways AI and ML can augment drug forecasts using Monte Carlo simulations is by harnessing vast amounts of historical data from clinical trials, real-world evidence, and diverse sources. This enables the development of predictive models that can capture the intricate relationships inherent in pharmaceutical outcomes. Such integration not only facilitates a deeper understanding of drug efficacy, safety profiles, and market dynamics but also addresses challenges associated with parameter estimation and optimization within Monte Carlo simulations.

Moreover, AI and ML techniques can address the challenges associated with parameter estimation and optimization in Monte Carlo simulations. Traditional methods may struggle

with complex models or limited data, leading to uncertainties in forecasting outcomes. However, AI and ML algorithms, such as Bayesian inference, excel at optimizing model parameters efficiently, and adapting, and refining estimates over time¹¹. This iterative approach enhances the accuracy and reliability of pharmaceutical forecasts, empowering companies to anticipate market trends and regulatory outcomes with greater confidence.

In summary, AI and ML techniques enrich Monte Carlo simulations in pharmaceutical forecasting by providing advanced predictive capabilities, optimizing model parameters, integrating diverse datasets, and modelling complex systems. As these technologies continue to evolve, their integration into forecasting workflows promises to drive innovation and efficiency in drug development and commercialization strategies, ultimately benefiting patients and healthcare systems worldwide.

References

1. A. G. Cook, *Forecasting for the pharmaceutical industry: models for new product and in-market forecasting and how to use them*, CRC Press, 2016.
2. G. Merkurjeva, A. Valberga and A. Smirnov, "Demand forecasting in pharmaceutical supply chains: A case study," *ICTE in Transportation and Logistics*, p. 8, 2018.
3. J. Mun, *Modeling risk: Applying Monte Carlo simulation, real options analysis, forecasting, and optimization techniques*. Vol. 347, John Wiley & Sons, 2006.
4. M. Palmblad, "Captario- The Perks of Using Monte Carlo Simulations in Drug Development," [Online]. Available: <https://www.captario.com/post/the-perks-of-using-monte-carlo-simulations-in-drug-development>.
5. J. Lampkin and D. G. Loosschilder, "MARKET FORECASTING – Monte Carlo-Based Forecasting: How to Deal With Uncertainty," *Drug Development & Delivery*, vol. 13, no. 04, May 2013.
6. "Six Sigma - Monte Carlo Simulation: Uses in Industrial applications," 16 August 2018. [Online]. Available: <https://sixsigmacertify.com/monte-carlo-simulation/>.
7. J. A. a. C. M. B. Sokolowski, *Modeling and simulation fundamentals: theoretical underpinnings and practical domains*, John Wiley & Sons, 2010.
8. W. Kenton, M. James and S. Kvilhaug, "Investopedia - Monte Carlo Simulation: History, How it Works, and 4 Key Steps," 02 November 2023. [Online]. Available: <https://www.investopedia.com/terms/m/montecarlosimulation.asp#toc-monte-carlo-simulation-results-explained>.
9. "Medium- Forecaster's Toolbox: How to Perform Monte Carlo Simulations," 13 June 2018. [Online]. Available: <https://medium.com/@melissa.lin/forecasters-toolbox-how-to-perform-monte-carlo-simulations-8bda16d2a35e>.
10. X. Tian, K. Benkrid and X. Gu, "High Performance Monte-Carlo Based Option Pricing on FPGAs," *International Association of Engineers - Engineering Letters*, 2008.
11. J. Rocca, "Bayesian inference problem, MCMC and variational inference," *Towards Data Science*, 2019.

Healthcare Provider (HCP) Behavior Assessment: Identifying latent subgroups of HCPs and Salesforce eSales Aid Impact Analysis

Sachin Ramesh, Manager, Genpact; Karthick Karuppusamy, Assistant Vice President, Genpact

Abstract: Cluster analysis and related subgrouping methodologies have gained prominence in the realm of market segmentation, particularly in categorizing individuals based on behavioral patterns. In the pharmaceutical industry, continuous engagement with Healthcare Professionals (HCPs) is critical for sales success. Subgrouping techniques, when applied to prescriber data, unveil latent characteristics within HCP groups, offering profound insights into intrinsic data trends. Traditionally, segmentation of HCPs relies predominantly on prescribing records, categorizing them based on their organizational value. However, in the context of escalating competition and diverse HCP behaviors, this may be inadequate. A comprehensive understanding of HCP behaviors, including their prescription habits and underlying engagement stimuli & responses, is essential for refining marketing strategies. Incorporating highly targeted communication strategies, particularly digital detailing by salesforce teams, has shown to enhance engagement and drive sales. This paper presents three case studies, each employing a blend of data-driven and business-oriented techniques to delineate latent HCP groups, thereby assessing their individual behaviors and attitudes towards marketing interactions. For each identified subgroup, the impact of digital calls is quantified, alongside an analysis of preferred messaging strategies, their sales impact, and a comparison between current and optimal frequencies of communication. Furthermore, the paper explores how this integrated approach can be adapted in omnichannel marketing strategies. The findings are particularly beneficial for brand marketing and commercial/omnichannel analytics teams, offering strategic insights and metrics for enhancing customer engagement.

Keywords: HCP Subgrouping, Digital Detailing, Individualized Experiences, Focussed targeting, Data & Business driven, Omnichannel

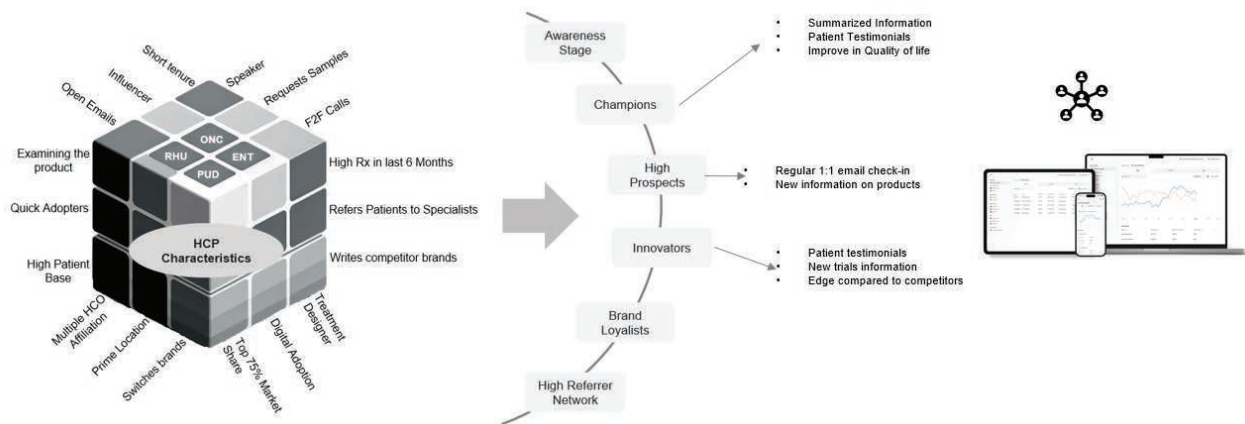


Figure 1: Illustrates the objective of the paper; on the left handside we have multidimensional cube representing HCP characteristics, through which subgroups can be derived for focused digital marketing

Introduction & Background

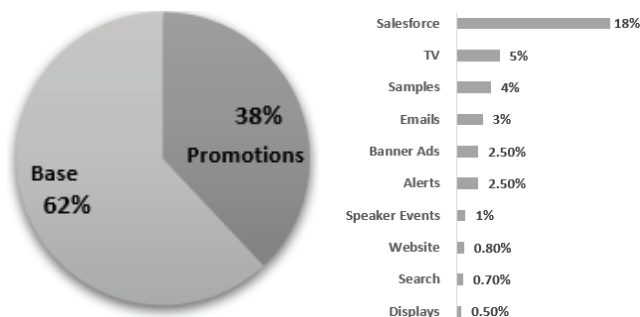


Figure 2: Illustrates typical Marketing Mix Model results

Pharma Salesforce

Empirical evidence highlights salesforce calls as critical marketing drivers for pharmaceutical sales [Figure 2]. Specialized sales representatives connect manufacturers and HCPs and serve as primary conduit for disseminating key drug information that transcends beyond mere sales to influence HCP education and patient outcomes. Despite its impact, challenges include the need for digital skills and data driven adaptation to evolving healthcare trends. With the tendency towards omnichannel marketing, strategic and operational proficiency is crucial.

Digital Detailing

The pharmaceutical salesforce's growing adoption of digital detailing represents a significant industry shift. Utilizing tools like salesforce automation (SFA), e-detailing, and digital marketing, facilitates enhanced data analysis and fosters personalized customer engagement strategies. Digital detailing's Closed Loop Marketing (CLM) gathers healthcare professionals' feedback during digital interactions, enabling more tailored marketing efforts. Digital detailing offers numerous advantages, including improved engagement,

flexibility, cost-effectiveness, and adaptability, significantly transforming the pharmaceutical salesforce landscape.

Contents in Digital Detailing

(eSales aid, eDetailing & Digital detailing are used Interchangeably throughout the paper) Pharmaceutical digital presentations, spanning from PDFs to interactive videos, leverage CRM platforms for their creation and dissemination. Displayed on devices such as iPads, they facilitate dynamic resource navigation, employing storyflows to accentuate critical data, including drug efficacy, patient testimonials, and competitive insights, thereby augmenting engagement and effectiveness.

Why Targeted / Focused Detailing is important?

In the pharmaceutical sector, targeted marketing is crucial for reaching specific audience segments effectively, enhancing brand visibility, patient care, and resource efficiency. It enables precise communication with healthcare market segments via customized messaging, bolstering the relevance and impact of marketing efforts and patient outcomes. The imperative for targeted marketing emerges from the need to meet diverse patient & HCP requirements and adapt to changing technological advancements. Thus, targeted marketing is vital for pharmaceutical companies, offering precision, cost-effectiveness, improved health outcomes, stronger stakeholder relations, market adaptability, and enhanced data use.

What is the need for subgrouping?

Subgrouping is a more refined approach compared to broader segmentation, targets specific groups within larger segments, addressing the diverse needs of HCPs and

patient populations. This method offers strategic advantages by enabling personalized communication and enhancing the overall impact of marketing initiatives. Subgrouping’s effectiveness lies in its ability to allocate marketing resources more efficiently, tailoring messages to unique group characteristics, and increasing engagement and response rates. Additionally, it leverages data-driven insights, providing a deeper understanding of market dynamics and consumer behavior, thus informing, and enhancing future marketing strategies. In essence, subgrouping in pharmaceutical marketing is a strategic necessity for success in the dynamic pharmaceutical industry. It ensures precision and maximizes the effectiveness of marketing efforts.

Case Studies:

Numerous innovative methods, technologies, and techniques exist to address the focused targeting and behavior assessment problem.

Utilizing these tools, one can generate a myriad of segments. However, it is crucial to

acknowledge that misuse or failure to apply business context may lead to erroneous conclusions. While subgrouping effectively addresses data underutilization, it is essential to recognize the risk of excessively dissecting data to identify all conceivable microsegments. The selection of appropriately sized groups, which offer actionable business insights while considering the balance between effort and cost, is pivotal in effective implementation and operationalization. Approaches discussed in the paper ensures that the segmentation process is not only data-driven but also strategically aligned with business objectives, ultimately enhancing the value and applicability of the derived insights.

The 3 subgrouping case studies discussed in this paper aims to adopt both data driven and business driven approaches. All three studies follow similar workflow expect for the step 1 (as shown in the *Figure [3]*). Data used in each of the case studies resembles different therapeutic areas (same data is not used for all three studies).

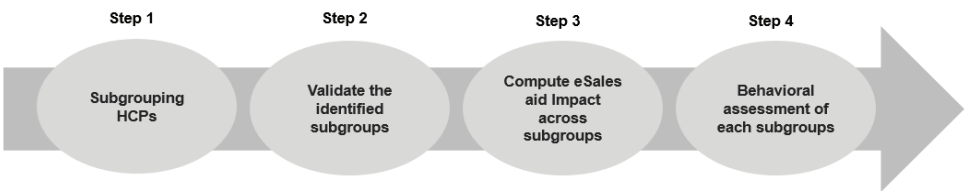


Figure 3: Workflow followed for behavior assessment in each of the outlaid case study

Primary Key		Independent Variables										Dependent Variable			
Dataset	HCP	Month	eSales aid calls	Non eSales aid Calls	emails	MSG 1	MSG 2	MSG 1, MSG2	MCM	DTC	Samples	Vouchers	Coupons	Competitor Sales	Sales
	123	Apr													
	456	Apr													
	789	Apr													
	123	May													
	456	May													
	789	May													

Figure 4: Illustration of final data for esales aid impact analysis

Step 1: Subgrouping HCPs:

Each case study will discuss broadly on different methods used to subgroup HCPs

Step 2: Validating the identified subgroups:

The Kruskal-Wallis H Test is a non-parametric method used for comparing three or more independent samples of equal or different sample sizes. It extends the Mann-Whitney U test when there are more than two groups. This test is particularly useful when the assumptions of one-way ANOVA are not met

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is a method for evaluating the quality of a clustering. It's calculated by the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters. Higher values of the Calinski-Harabasz Index indicate better clustering.

Forecasting the average sales for each group with an intention to view if there is any change in average sales in future assuming all other promotions constant. Example: If the average sales for one subgroup is hardly changing then it can be said that the group does not require much of promotional engagement for continued sales compared to others.

Step 3: eSales aid Impact assessment:

The analysis follows below

a. Data Preparation:

- HCP universe consists of those IDs which either have a Veeva interaction or a claim
- Data is compiled at HCP x Month level (square matrix)

- eSales calls and non eSales calls constitutes the variables of interest for which impact is assessed.
- Data filters are applied for selling teams, segments, specialty codes etc.

For confounding effects and as control variables, other promotions, competitor sales, samples, emails are used. Data structure will look as shown in *Figure [4]*.

b. Transformations:

- Lag effects are considered for the dependent variables, adstock rate is determined using co-relation analysis
- Various nonlinear transformations are used depending on data points
- Data is scaled and normalized.

c. Modelling:

- Regression techniques are used to derive attribution towards sales, multiple iterations are run to stabilize the model and minimize the error.

d. How to interpret the impact from attribution?

- If esales calls are attributing to 18% of Total sales (10,000 sales for example), sales due to eSales calls = $18\% * 10,000 = 1,800$
- Impact is defined as the ratio of sales due to and Calls, (5000 calls for ex.) Impact = $1,800/5,000 = 0.36$
- Impact is drilled down to content level at various granularities providing the comprehensive view as shown in the *Figure [5]*

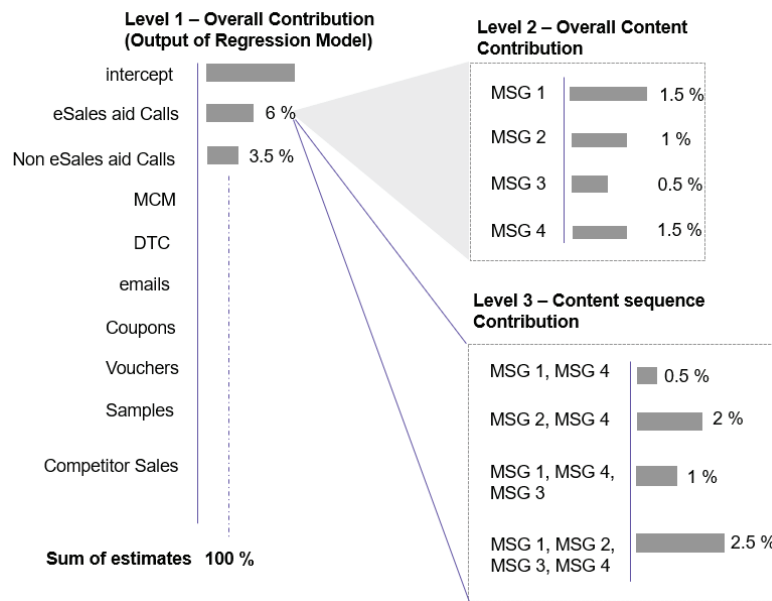


Figure 5: Illustrates the attribution breakdown by contents

Behavioral Assessment

This step involves analyzing the results to differentiate the individual characteristics of identified subgroups especially on eSales aid messages along with insights and summarization. Also, the analysis attempts to recommend Ideal frequency of messages i.e., Optimal number of calls in which a message must be exposed to the HCP for each subgroup, and is computed using below simulation approach:

- Consider the secondary regression model (Level 2 as shown in *Figure [5]*)

$$y = \beta_1 * \text{MSG1} + \beta_2 * \text{MSG2} + \beta_3 * \text{MSG3} + \beta_4 * \text{DTC} \dots + c$$

- To compute ideal frequency for MSG1, considering all other independent variables as constant (no change), sales is simulated for each value of MSG1, the value for MSG 1 here will be number of calls (in the increments of 1) in which this message was detailed in a call.

- Inflection Point is identified by comparing results from below two methods:
 - The second order difference for the values computed in above step is calculated. The diminishing sales (inflection point) is the call value where the second order difference does not change, or change is very minimal (< 5% compared to previous value)
 - A point where second order derivative changes is computed using smoothing functions. For this exercise, Savitzky-Golay signal filter function is used for smoothing, first order and second order derivatives are calculated and value at which second order derivative changes the sign is calculated and mapped for call frequency.

Case Study 1: Behavioral Subgrouping using Data Driven Approach

In this analytical endeavor, considerable emphasis is placed on HCP interactions with various marketing touchpoints. HCPs

demonstrate varying degrees of engagement and responsiveness to diverse marketing stimuli, influenced by multiple factors. Brand marketing teams frequently reevaluate their strategic goals regarding what information about the product should be communicated to HCPs. They then devise tactics and campaigns incorporating a range of product-related information. This includes, but is not limited to, product efficacy across different age groups, recent clinical trial outcomes, patient onboarding processes, insurance and payment support options, access to patient gateways, and drug characteristics such as side effects, symptom reduction etc.

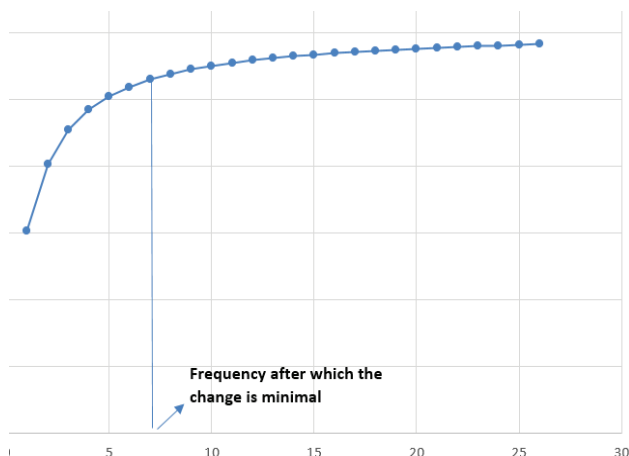


Figure 6: Illustrates response curve from simulation described above

This plethora of product information is disseminated through various channels, including emails, alerts, banner ads, digital displays, web links, direct salesforce engagement, speaker events, and television advertisements. Each HCP may resonate more effectively with certain types of delivery channels, influencing their response patterns. Considering the multifaceted nature of these promotional efforts (encompassing both channels and content) and analyzing the resultant behaviors poses a significant big data challenge. Addressing this

challenge requires innovative approaches. In response to this complexity, a methodological approach combining Cluster Analysis with Graph Analysis (specifically Community Detection) is employed. This dual-analysis technique aims to identify prominent groups of HCPs based on distinct combinations of channels and topics. By leveraging this combined analytical framework, it is possible to gain deeper insights into HCP behaviors and preferences, thereby informing more effective and targeted marketing strategies within the pharmaceutical domain.

Step 1: Subgroup Identification

Community Detection

Community detection in network analysis is a method used to identify groups within a network, where nodes are more densely connected to each other than to nodes in other groups. This technique is key in understanding the structure and organization of complex networks, which can range from social networks to biological systems. By identifying these communities, insights into the underlying relationships and interactions within the network can be derived, which helps to reveal patterns, hierarchies, and influential nodes. In this paper, community detection is used to find prominent groups among channel & content interactions for the HCPs.

Community Detection Data Prep Steps

1. Identify and map the cross-channel topics and contents. This distinct mapping of channel and topic will act as initial node of the graph. These nodes are individual touchpoints / HCP engagement/ stimuli.
 - a. Example: If the channel is Digital salesforce Call, and the topic discussed is about drug efficacy, the Node would be Digital_Detailing_Efficacy (N1)

2. For all the HCPs in the analysis period, Sort the mapped nodes by time/date of their occurrence in the data.
 - a. Example, if an HCP was exposed to Digital_Detailing_Efficacy (N1) in Jan and Email_Patient_Testimonial (N2) in Feb, the order of occurrence for HCP will be $N1 \rightarrow N2$.

HCP ID	Source	Target
123	N1	N3
123	N3	N2

3. In this step for each user ID, create 2 node (Source Node and Target Node) path.

Source_Node	Target_Node	Weight
N1	N3	200
N2	N1	3500

- a. Example, if an HCP has $N1 \rightarrow N3 \rightarrow N2$ sequence of touchpoints, the result of this step would look like
4. In this step, Final data is prepared in a format that can be utilized by any graph analysis algorithm, the format would look like. Here the weight refers to the frequency (count of occurrence) from source node to target node

Louvain community detection algorithm is used in this exercise, it is most widely used algorithm known for its efficiency and effectiveness in detecting hierarchical community structure.

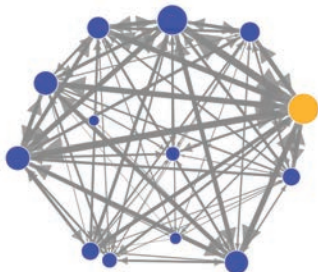


Figure 7: Visual representation of Communities with channel & Topic interactions

The algorithm yielded 12 prominent communities as shown in the *Figure [7]*. Blue nodes indicate communities and yellow node indicates Sales. Each of these communities had distinctive features, for e.g., One of the communities had large number of veeva esales calls with topics on technical drug data on exacerbations, suitable patient profile information etc. Another community was heavy on speaker events and so on.

Clustering Steps

1. From the results of community detection step, for each community TOP Paths (Paths with edge % > 5% of total edges) are chosen
 - a. Example: In community 1, if the total edge weight is 1000, weight of the path $N1 \rightarrow N2$ is 100 (10% of Total edges), it is chosen as one of the Top paths
2. Data Input for the cluster exercise would include the frequency of interactions of an HCP with top paths within each community on their interactions across channel x topics.
3. Upon data reduction, multiple clustering algorithms were tested to derive the latent groups of HCPs based

Gaussian Mixture Model worked better to map the business context upon profiling. *Figure [8]* visualizes cluster data points

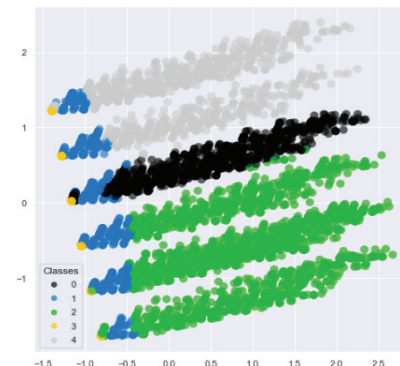


Figure 8: Visual representation of clusters identified Gaussian Mixture Models Algorithm

Cluster Profiling

After mapping the results from cluster exercise back to HCP writing and web engagement, it was observed that the latent groups detected had distinctive characteristics

Group 1: Shortest tenure, low writers, and low web engagement

Group 2: Second shortest tenure, medium writers, and medium market share, lower web engagement

Group 3: Medium tenure, high writer for competitors, highest web engagement

Group 4: Second longest tenure, medium to high writer, high writer for competitors, higher time spent on websites.

Group 5: Longest Tenure writer, Highest market share and high writers, high web visits but lower time spent in websites

Mapping Communities to Clusters:

Upon mapping the identified clusters against the interaction communities of channel x Topics, it was seen that, each groups had high affinity towards specific marketing elements. Some of the smaller communities are merged due to smaller interaction volume. This is visually represented in *Figure [9]*. Based on the mapping, below are the characteristics of the groups.

Group 1: Very Light Interactions, among social media alerts and emails and they are not so specific on topics

Group 2: Strong interactions among digital displays, social media, and veeva emails. Affinity towards Technical product (efficacy, edge over competitors etc.) and patient profile information

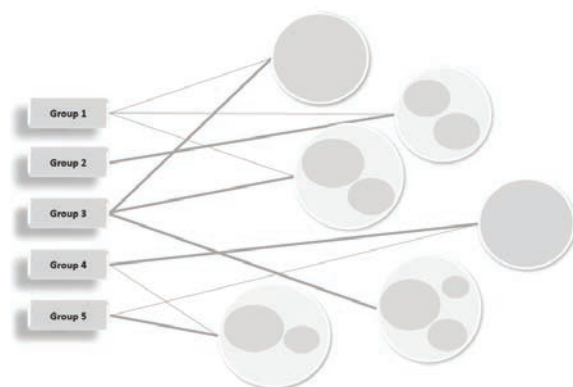


Figure 9: Illustrates Cluster groups and Community mapping, darker edges indicate intense marketing

Group 3: Strong in veeva interactions, attends speaker events, responds to 1:1 emails, Direct mails, and Digital displays. HCPs in this group are interested to learn about, intricacies of drug's mechanism of action, drug administration, coverage, and payment information through eSales aid.

Group 4: Higher response to veeva emails and more time spent on digital detailing, they request samples, outcomes on trials etc.

Group 5: Similar pattern to Group 4 but less intense, advocates the product and speaks in speaker events

Step 2: Group Validation

Identified HCP Groups are tested for statistical significance considering some of the main promotional elements using tests described in previous sections.

	T Statistic	p Value	Group 1	Group 2	Group 3	Group 4	Group 5
Sales	862	0	4.0	7.9	10.7	12.3	22.0
Calls	1,925	0	1.3	4.1	10.5	9.7	10.8
Samples	766	0	0.5	2.4	8.6	8.2	11.5
Emails	446	0	0.3	0.8	1.0	1.1	1.3

Average Values

Figure 10: shows the statistics for Kruskal-Wallis H Test for Case 1 Groups

The test indicates that the groups are well apart and are statistically significant. Average sales and engagement are increasing from Group 1 to Group 5.

In the forecasting exercise, it is observed that Group 1 HCPs have minimal dependence on promotional engagement, Group 2, 3 & 4 HCPs would need continuous engagement without which there would be significant drop in sales, Group 5 HCPs would still need regular check-ins but will carry more lag effect of the previous promotions.

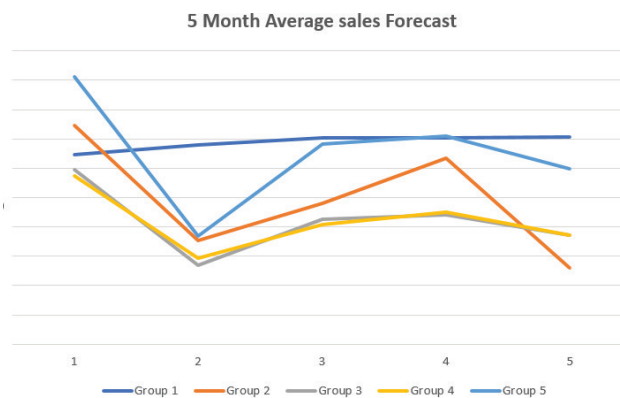


Figure 11: 5 months forecast of average sales

Step 3: eSales aid analysis:

Identified Subgroups	Attribution	Impact (Prescription / Call)
Group 1	0.02%	0.01
Group 2	1.6%	0.33
Group 3	6.8%	0.30
Group 4	5.76%	0.62
Group 5	3.8%	0.39

Result table 1: eSales calls impact

Messages Impact:

Group 1: It is observed from above table that, the HCPs mapped to this group, have negligible impact due to eSales aid calls, however, couple of messages are resonating well among these groups.

Message	Impact	Frequency	Optimal Frequency
Information on Re-Imbursements	0.15	1	1
Coverage on the product	0.1	1	1

Result table 2: message impact for Group 1 HCPs

HCPs in Group 1 have lighter engagement. Messages informing about payment and coverage interests them the most and there is no requirement to change the engagement frequency.

Group 2: Although lower attribution, for the volume of call these group of HCPs receive, they are exhibiting higher impact on sales.

Messages	Impact	Frequency	Optimal Frequency
Information on Re-Imbursements	0.25	4	2
Drug Administration	0.15	3	3
Mechanism of Action	0.02	1	1

Result table 3: message impact for Group 2 HCPs

HCPs in this group are interested to learn about how the drug works and how it can be administered to patients along with payment and coverage information.

Above observations confirms that Group 1 & Group 2 HCPs are at the early stage of product prescription and adoption.

Group 3: Calls for the HCPs in this group have higher attribution towards sales, this explains the intense cross channel marketing engagement communities being mapped.

Messages	Impact	Frequency	Optimal Frequency
Patient Enrollment Forms	0.22	2	2
Disease Progression Information	0.24	3	2
Dosing Information	0.2	1	3
Baseline markers Information	0.15	1	2
When to Hospitalize	0.12	1	1

Result table 4: message impact for Group 3 HCPs

HCPs in this group are more interested to learn technical details of the drug and, they seek information on patient enrollment forms, upon deep dive into NBRx metrics, it is observed that among all groups, HCPs in this group onboard more new patients. Also, it is observed that messages in some combination yield higher Impact. Below are few highly impactful combinations.

Message Combinations	Impact
Symptoms Requiring Hospitalization, how to reduce Disease progression	0.95
Disease Characteristics, how to reduce disease progression	0.8
Progression Reduction, Dosing Reduction	0.9

Result table 5: message combination impact for Group 3 HCPs

It is evident that these HCPs respond and seek technical information, and onboard new patients, naturally growing these customer base becomes very important and hence high intense marketing is observed.

Group 4: Calls for these HCPs have second highest attribution towards sales and highest impact per call. Upon deep dive in the data, it was observed that, these HCPs are targeted specialists, they are influential referrers, do not switch brands very often, also they have higher end of market share for the product.

Messages	Impact	Frequency	Optimal Frequency
Patient Enrollment Forms	0.18	2	2
Patient Types	0.22	3	2
Summary	0.34	5	3

Result table 6: message impact for Group 4 HCPs

These HCPs although need timely technical details on Drug, they prefer to view the summarized information.

Group 5: These HCPs are stronger Referrers; they speak at speaker events and advocate the

product. Their calls attribution although lower but have higher impact

Messages	Impact	Frequency	Optimal Frequency
Videos on Patient testimonials	0.15	2	1
Patient Identification	0.21	4	2
Clinical Trials	0.12	1	3
Disease Characteristics	0.18	2	2

Result table 7: message impact for Group 5 HCPs

Message Combinations	Impact
Disease Characteristics, Summary	0.75

Result table 8: message combination impact for Group 5 HCPs

Key Takeaways & Limitations:

- Data Driven Approach provides latent information that is often underutilized.
- The approach was able to successfully characterize HCP groups based on their progression in product use
- eSales aid messages behave different for HCPs in each journey stage, it is important to identify the HCPs map their progression and target with relevant information.
- Data preparation and interpretation demands huge efforts, this approach cannot be fully automated due to the nature of changing contexts
- Approach will not be well suited for non-writers

Case Study 2: Customer value-based Subgrouping using RFM analysis

Step 1: RFM Analysis

RFM (Recency, Frequency, Monetary) analysis is a marketing analytics technique

used to quantify and evaluate customer value. Originating from direct marketing, this technique has gained widespread application across various business sectors, including retail, e-commerce, and service industries. Recency refers to the time elapsed since a customer's last interaction or purchase, indicating engagement level. Frequency assesses how often a customer purchases or interacts, reflecting loyalty and habitual behavior. Monetary value measures the total amount spent by a customer, serving as an indicator of their economic value to the business. By segmenting customers based on these three dimensions, RFM analysis enables businesses to identify high-value customers, tailor marketing strategies, and optimize resource allocation. This customer-centric approach not only enhances personalized communication, thereby fostering stronger customer relationships and driving growth.

RFM grouping Steps

The analysis steps are modified to suit the multi facets of HCP Promotions along with traditional approach of assessment on sales.

Definitions:

Sales consideration:

Length: Number of days since first writing

Recency: Number of days since last writing

Frequency: Number of times HCP has written a prescription

Monetary: Total Number of Prescription

Promotion Consideration:

Length: Number of days since first engagement

Recency: Number of days since last engagement

Frequency: Number of times HCP has is exposed to specific engagement.

Based on the above definitions, for each HCP LRFM values are computed considering sales, Salesforce digital detailing calls, Samples, Speaker Events, emails. For each of these variables, standard scaler function is applied to standardize the values of LRFM at individual feature level. The final Customer score is computed using below formula,

$$\text{HCP Score} = \sum w * L(x) + \sum w * R(x) + \sum w * F(x) + M(\text{sales})$$

w = weightage. This can be determined using,

- Affinity Scoring
- Attribution from MMX models
- Equal weight assignment

x = Sales, Samples, Speaker Events, Calls, Emails

Once HCP score is computed, HCPs are broadly grouped into four Tiers using ranking and quartile functions. Tier 1: HCPs with score >= 75 Percentile, Tier 2: HCPs with score >= 50 Percentile <= 75 Percentile, Tier 3: HCPs with score >= 25 Percentile and <= 50 Percentile, Tier 4: HCP score <= 25 Percentile

Step 2: Group Validation:

The approach was able to produce distinctive subgroups of HCPs which are different from each other. Organically, Tier 1 HCPs are evidently high valued customers followed by Subsequent Tiers.

	T Statistic	p Value	Tier 1	Tier 2	Tier 3	Tier 4
esales Call	25,994	0	8.90	4.94	2.78	0.26
EMAILS	11,845	0	1.05	0.54	0.26	0.04
Sales	20,961	0	13.28	3.50	0.99	0.21
Speaker event attended	940	0	0.09	0.04	0.02	0.00
Samples	10,000	0	7.81	2.94	1.24	0.13

Average Values

Figure 12: shows the statistics for Kruskal-Wallis H Test for case 2

Step 3: eSales aid analysis

Identified Subgroups	Attribution	Impact (Prescription / Call)
Tier 1	13%	0.65
Tier 2	3%	0.45
Tier 3	-	-
Tier 4	-	-

Result table 9: eSales calls Impact for RFM groups

eSales aid calls for Tier 3 & 4 HCPs did not have statistically significant attribution due to lower sales volume.

Messages Impact:

Tier 1:

Messages	Impact	Frequency	Optimal Frequency
Patient Testimonial Videos	0.25	2	2
Summarized Information	0.48	5	3
Disease Progression	0.6	4	2
Symptoms Reduction	0.35	4	3
Dosing	0.42	3	1
Disease Characteristics	0.2	4	2

Result table 10: message impact for Tier 1 HCPs

Tier 2:

Messages	Impact	Frequency	Optimal Frequency
Enrolment Forms	0.2	2	3
Summarized Information	0.17	2	4
Patient Identification	0.13	4	3
Disease Progression	0.08	3	4
Drug Administration	0.12	2	3
Disease Characteristics	0.07	3	2

Result table 11: message impact for Tier 2 HCPs

Observations:

- Tier 1 customer are advanced writers (like Group 4 & % in Case 1), they would need constant engagement, clearly, they resonate with more detailed information about the product and have higher impact on sales.

- Tier 2 customer base are moderate writers, they would need high frequency engagement, they prefer basic product information as opposed to advanced details.
- The approach can create a subset of HCPs based on their value to the organization, however, the categorization is less objective compared to the data driven approach.
- This subgrouping exercise has quicker turnaround time and is effective, but when larger marketing promotions are considered becomes more of a subjective study losing sophistication, thereby creating similar groups like traditional segmentation.

Case Study 3: Subgrouping based on business questions / context

This is a simplistic approach, where subgroups are formed based on business questions and data logics. Though straightforward, upon proper application this approach can be powerful to study and focus on smaller groups of HCPs when time is of the essence and with minimal efforts / costs.

Step 1: Business Question based Grouping

Problem: Identify customer groups for focused marketing, divide the customer groups into three groups, Group 1 HCPs are those who are critical to business and their churn will cause loss in sales, they are loyal to the product, Group 2 consists of those HCPs who toggle between competitor products, they have affinity towards the product but need an extra push, Group 3 should include the HCPs whom if acquired will create additional value.

Business Logic:

Group 1 (Preserve): These customers are matured product writers, they are influential in referrer networks, talks about the product, requests information, engages in conversations, situated in prime locations, responds to marketing stimuli quite well. Have higher tenure.

Group 2 (Develop):

These HCPs are moderate writers, have requested samples, often enquires about patient enrolment, coverage, and payment information, asks basic drug information, often tries to compare competitor products, dabbles between multiple products, occasionally onboards patients to the product. Medium tenure HCPs.

Group 3 (Acquire):

HCPs in this group are low product writers, they have not written recently. Have larger market share for similar products, of specific specialty, rarely prescribes the product in line of therapy, are influential referrers.

Data must be mined to satisfy above business conditions. There is a possibility that one HCPs being mapped to multiple groups, diligent decisions must be taken to categorize these HCPs.

Step 2: Group Validation

Identified subgroups are statistically significant and well apart from each other.

	T Statistic	p Value	Group 1	Group 2	Group 3
eSales Calls	19,309	0	4.06	3.54	2.14
Emails	979	0	0.25	0.28	0.17
sales	25,069	0	20.05	10.23	2.96
Speaker Events	932	0	0.02	0.01	0.00
Samples	962	0	0.42	0.21	0.10

Average Values

Figure 13: shows the statistics for Kruskal-Wallis H Test for case 3

Step 3: eSales aid analysis

Identified Subgroups	Attribution	Impact (Prescription / Call)
Group 1	5.20%	0.88
Group 2	7.30%	0.46
Group 3	1.20%	0.06

Result table 12: esales calls impact for groups

Messages Impact:

Group 1:

Messages	Impact	Frequency	Optimal Frequency
Summarized Information	0.35	2	2
Safety Information	0.16	2	1
Clinical Trials	0.21	2	2
Study Outcomes	0.21	4	3
Baselines	0.28	1	2

Result table 13: message impact for Group 1 HCPs

Group 2:

Messages	Impact	Frequency	Optimal Frequency
Rate of Progression	0.17	2	3
Disease Education	0.19	4	3
Trial Designs	0.13	2	2
patient journey	0.1	1	1
Dosing and Coverage	0.25	3	2
Side Effects	0.22	3	2

Result table 14: message impact for Group 2 HCPs

Group 3:

Messages	Impact	Frequency	Optimal Frequency
Real World Evidence	0.05	1	1

Result table 15: message impact for Group 1 HCPs

Observations:

- The esales aid calls attribution even though at higher end for Group 2 HCPs, Group 1 HCPs have higher impact can be correlated to higher sales with lesser calls.

- Group 1 customers have preference for high level advanced information, they are expecting reps to keep them updated with latest proceedings regarding trials, studies etc.
- Group 2 HCPs have an affinity towards more intricate details about the product
- Group 3 HCPs are seeking empirical proofs for the product's efficacy via real world evidence

How can the approach and outcomes be integrated with Omnichannel Strategy?

The industry is pivoting towards omnichannel marketing strategies, Omnichannel marketing represents a holistic approach that seeks to empower seamless and integrated customer experience across all channels and touchpoints. This marketing strategy blurs the boundaries of traditional approach, ensuring each channel work in Unison, thereby enhancing

the customer Journey. The approaches and outcomes discussed in this paper can be effectively integrated within 3 main components of omnichannel strategy as highlighted in the *Figure [13]*

Behavior and Attitudes: Approach and outcomes discussed in the paper can be used to understand various behavior associated with the same group HCPs, this narrows the application of specific actions / triggers to the groups rather than directing towards broader segments.

Effectiveness of NBAs, NBX, Sequences, Contents & Topics: In reality there will be plethora of sequences with content/ channel combinations, identifying which are working well or have higher impact among specific groups of HCPs will help to adopt preferential campaigns. Approach discussed in the paper to assess impact and optimal frequency will assist the cause.

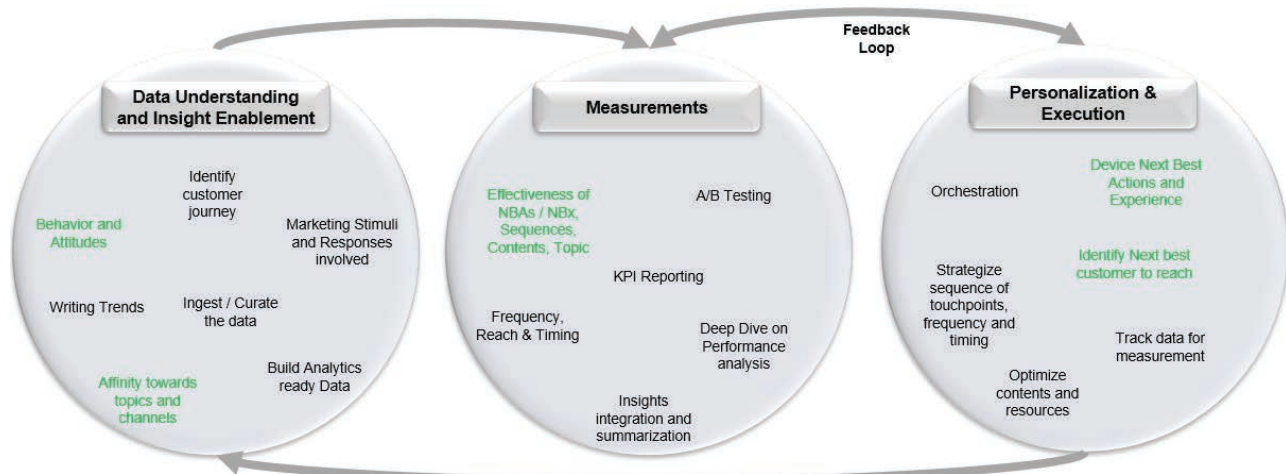


Figure 14: Omnichannel Steps and Components

Strategizing Next Best Actions and identifying next best customers: The recommendations from the analysis can be utilized to plan NBA triggers and optimize resourcing.

Summary and Conclusion

This paper presents an in-depth analysis of Healthcare Professional (HCP) behavior assessment through innovative subgrouping techniques and eSales aid analysis deduced from research and hands-on industry experiences drawn from authors. By employing methodologies such as community detection, clustering, RFM analysis, and business logic-based grouping, the study categorizes HCPs into distinct subgroups, revealing varied characteristics and preferences towards digital detailing and its underlying contents. The findings illustrate that each subgroup demonstrates a unique affinity towards specific topics, underscoring the importance of personalized digital marketing strategies in engaging HCPs effectively. Furthermore, the paper situates its approach within the broader context of omnichannel strategies, highlighting its significance in crafting a cohesive and dynamic marketing framework that resonates with the diverse needs and behaviors of HCPs. The paper's findings advocate for the adoption of sophisticated subgrouping and eSales aid

analysis as essential components in the evolving landscape of pharmaceutical marketing through meticulous analysis and elucidating that through nuanced understanding of data and its components.

Future Work

1. **Longitudinal Studies:** Future research could explore longitudinal studies to track changes in HCP behaviors and subgroup dynamics over time, offering insights into the evolving nature of HCP engagement and the long-term effectiveness of digital detailing strategies.
2. **Cross-Industry Comparisons:** Investigating subgrouping and eSales aid effectiveness across different healthcare sectors could provide a broader perspective on digital engagement strategies, highlighting unique challenges and opportunities in various medical fields.
3. **Integration with Emerging Technologies:** Further studies could examine how emerging technologies, such as Generative AI and machine learning, can enhance subgrouping techniques and eSales aid analysis, potentially offering more dynamic and responsive ways to engage HCPs within the omnichannel framework.

About the Authors

Sachin Ramesh, Manager – Genpact LLC: *Sachin Ramesh works as a Data Science & Analytics consultant at Genpact, building advanced analytics and machine learning solutions for fortune 500 companies. He holds a master's degree in computer science from Syracuse University. He has worked in the field of IT & data science for 8+ years in Life Sciences, Finance and CPG domains. Throughout his career he has successfully planned and executed various end to end AI / ML solutions which have been productionized and generated substantial ROI. He has been leading the delivery of novel analytics solutions across multiple therapy areas. He has received multiple appreciation awards for his excellence in deriving and delivering actionable & timely insights to stakeholders. He has a passion for cooking fine cuisine and enjoys reading in his spare time.*

Karthick Karuppusamy, AVP – Genpact LLC: *Karthick Karuppusamy works as an Assistant Vice President at Genpact, delivering and consulting Data and Analytical solutions for top20 Pharma organizations. He holds a master’s degree in applied mathematics from Anna University, India. He has been delivering commercial analytics solutions for over 16+ years across Pharma and CPG organizations in the areas of sales, marketing, digital, patient and payer analytics using his in-depth domain knowledge and various data science methods to bring most relevant and complete insights. Karthick was awarded **Marketing excellence award** from the customer organization for devising innovative solutions that majorly assisted digital business transformation. Karthick works with his team to continuously explore untapped data and bring innovation solutions that helps business to receive quicker and unseen insights.*

References

1. Jicun Zhang, Jiyu Fei, Xueping Song, Jiawei Feng, “An Improved Louvain Algorithm for Community Detection”, Mathematical Problems in Engineering, vol. 2021, Article ID 1485592, 14 pages, 2021
2. Mahfuza R, Islam N, Toyeb M, Emon MAF, Chowdhury SA, Alam MGR. LRFMV: An efficient customer segmentation model for superstores. PLoS One. 2022 Dec 20;17(12): e0279262. doi: 10.1371/journal.pone.0279262. PMID: 36538513; PMCID: PMC9767363
3. Ozkan, Pinar. (2020). LRFM Analysis as a Customer Segmentation Tool in the Tourism Sector. 10.4018/978-1-7998-3030-6.ch012.
4. Isa Inuwa-Dutse, Mark Liptrott, Ioannis Korkontzelos, A multilevel clustering technique for community detection, Neurocomputing, Volume 441, 2021, Pages 64-78, ISSN 0925-2312
5. Zhou, Xinyi & Zafarani, Reza. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys. 53. 10.1145/3395046.
6. Delvecchio, Devon & Henard, David & Freling, Traci. (2006). The effect of sales promotion on post-promotion brand preference: A meta-analysis. Journal of Retailing. 82. 203-213. 10.1016/j.jretai.2005.10.001.
7. The Long-Term Impact of Promotions on Consumer Stockpiling Behavior, published by the American Marketing Association, Kamel Jedidi, Carl F. Mela, and Douglas Bowman, Vol. 35, no. 2 (May 1998), 250-62.
8. Digital Innovation & Solutioning and Life Sciences business expertise at Genpact LLC

Data X Ray – A Novel Tool that Assesses Data Sources Using ChatGPT

JP Tsang, PhD and MBA (INSEAD), President of Bayser Consulting; Maylis Larroque, Senior Consultant, Bayser Consulting

Abstract: The most challenging issue in data analysis is not the analysis itself (most analyses are rigorous and thorough) but rather the lack of knowledge regarding the shortcomings of the data source. Not only are the insights wrong, they are served up as inconvertible truths as the analysts have been diligent and thoughtful all along to uncover these insights. Needless to say this is a disaster in the making.

This paper addresses this issue head on by delving into 4 fundamental questions regarding data sources: (1) What is this data source about?, (2) Is the data source any good?, (3) Can the data source answer this particular business question, and (4) What are the caveats of the data source? This paper presents Data X Ray, a system we developed leveraging ChatGPT to answer these questions.

Motivation

We have a big problem in our industry and it is that we, as analysts, provide wrong insights and recommendations from the data we analyze. Not from time to time but repeatedly on an ongoing basis. This is hard to accept as our industry invests very heavily in developing and acquiring the best data sources and we employ technically savvy and diligent analysts. What's wrong? In a nutshell, analysts report insights they get straight from the data. They fail to recognize that the insights they cull from the data are off because of the caveats of the data source and these insights need to be interpreted accordingly before dissemination. The sad truth is most of the time the analyses that these analysts perform are beyond reproach. The analysts are simply not aware of the caveats of the data sources they analyze.

Here is a compelling example. The question at hand is to determine how many patients are

eligible for the drug of interest. In this case, the patient must have undergone chemo and radiation therapy (CRT). The analysts went to work and deployed not one but four syndicated open claims data sources which all pointed to roughly the same number of patients that we'll call X. The fact that X was too small did not sit well with Upper Management as it meant that the drug was doomed to failure right off the bat. Also, X flew in the face of Primary Research and other trusted sources which suggested a much larger number of patients. It turns out the analysts got it all wrong and were let go. What was the mistake of the analysts?

Chemo can be administered orally or intravenously. Oral chemo is reported on the NCPDP claims which are very well captured at about 90%. IV chemo is reported on the CMS-1500 claims which are captured at a much lower rate of about 40%. Radiation is performed in the hospital and for patients older than 65, which is the bulk of the population, it is

reported under Medicare Part A which appears on the UB-04 claims which are captured at around 20%. As a result, the capture rate of a CRT patient is 90% times 40% times 20% which is 18%. Out of the gate, the data source is missing 82% of the eligible patients. The fact that four data sources were used does not

mean anything as they are all syndicated open claims and have the same shortcomings. Had the analysts run this simple calculation and realized that the four databases have the same shortcomings, they would have adjusted their findings accordingly and avoided the trauma that ensued.

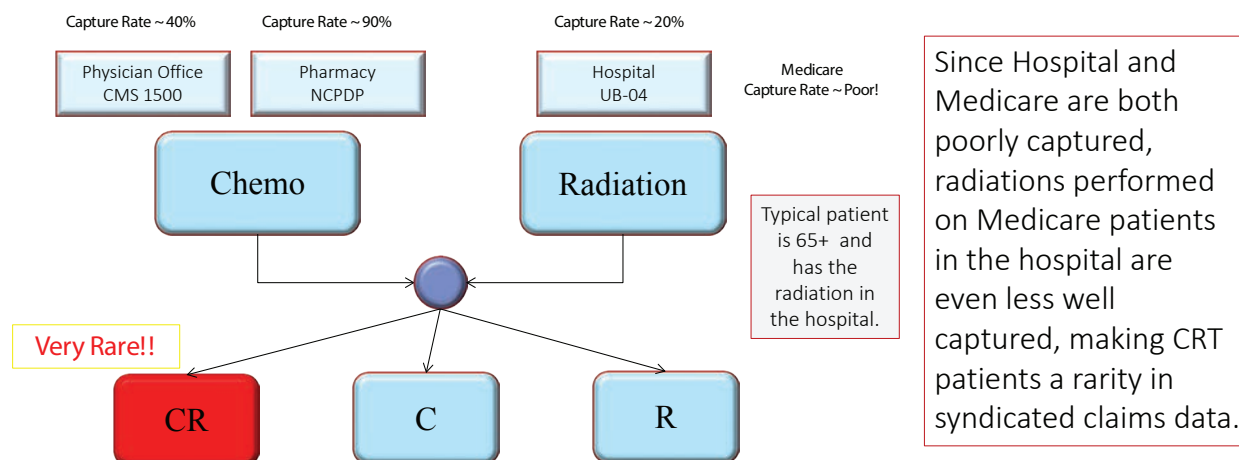


Fig 1 – Capture of CRT Therapy in Open Claims

On a different note, we as an industry need to welcome and encourage data innovation. We need to be able to assess the merits of novel data sources in short order. More routinely, we need to assess and compare competing data sources to make informed go/no go purchase or contract renewal decisions. This can be frustrating as the process of comparing data sources requires expertise we may not have. Also, the process can be arduous and lengthy. What is badly needed is a tool that helps automate the process of evaluating data sources.

ChatGPT is a godsend as it makes programming easier. We do not need to write every single line of code. We can express what we want in

somewhat loose natural language and ChatGPT has no problem understanding what we want and fulfills the request. We can also provide examples of what we are after and ChatGPT gives us what we are after. In sum, the LLM makes it much easier to build such a tool such as Data X Ray.

Do we really need a new tool here? There is a plethora of data assessment tools on the market. A close look at these tools quickly reveals that they focus on data format and syntactical errors. None of them provide an expert assessment of the content of the data source. What we want is a tool that tells us how relevant the data source is for the use case

we have in mind, what's missing in the data source and what the caveats are. That's why we developed the Data X Ray.

Overview of Data X Ray

The Data X Ray tool takes as input a data sample, the data dictionary, and the full data set and produces as output an assessment of the data source in the form of a PDF file.

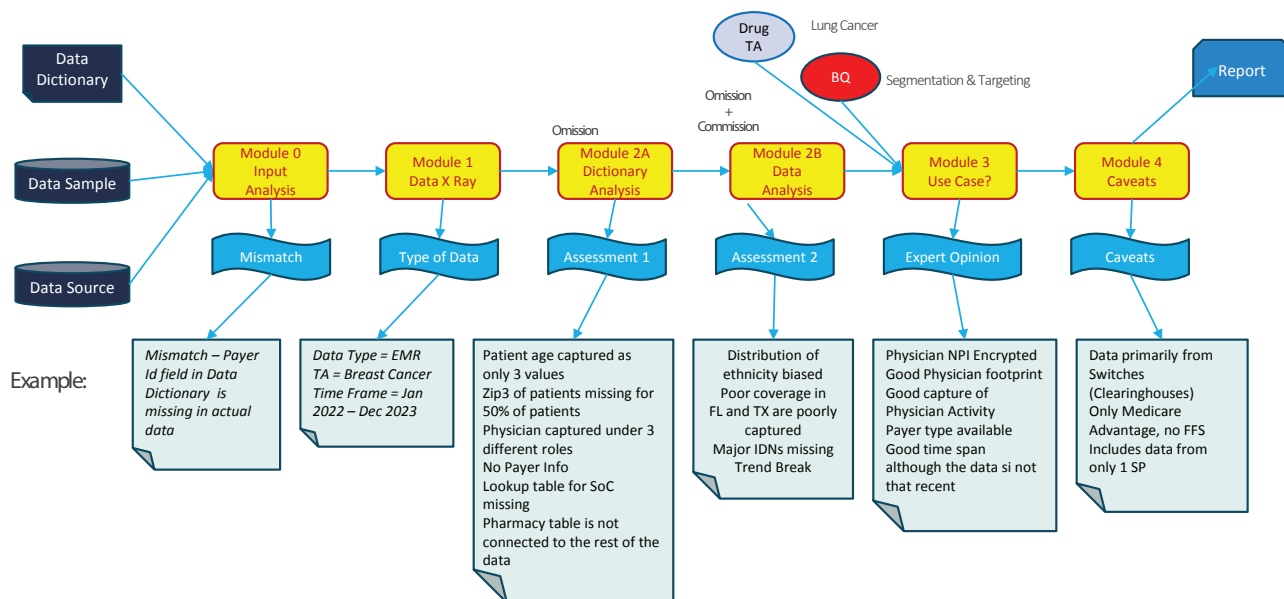


Fig 2 – Input and Output of the Data X Ray Tool

This assessment helps make informed decisions regarding a whole host of data strategy questions. For instance, acquire or drop a data source, keep the current data source, switch to a new data source, or combine two or more data sources together, deploy or not deploy a data source for a specific use case, and continue or halt further exploration of a novel data source.

2.1 Questions

The tool answers four fundamental questions.

First, what is the data source about in a nutshell? This is particularly useful when the data is spread over several tables with field names that are barely recognizable. Oftentimes, we do not have the time or inclination to conduct an exploration of the data source.

Second, is the data source any good? This is about getting assurance that the data source is worth pursuing. We want to know if there are major issues with the data source such as geographic holes (e.g., some states are absent or grossly under-represented), longitudinal

holes (e.g., no hospitalizations are captured), lack of representativeness (e.g., only Medicare Advantage and no Medicare FFS), built-in biases (e.g., no Hispanics or Blacks), lack of granularity (payers are only reported as Medicare, Medicaid, Commercial, or Other), or the data is too old or refreshed too infrequently.

Third, is the data source relevant for the use case of interest? We have a business question to answer and we need to establish in short order if the data source is fit for the job. The danger of using the wrong data source, as we know, is that we'll glean wrong insights and offer misguided recommendations. If the data source is inadequate to answer the business question, we need to know this right away and move on.

Fourth, what are the caveats of the data source? This is the one question that no analysts can afford to ignore and yet they do. Caveats are shortcomings of the data source which suggest that the insights from the data source may not be accurate and need to be interpreted accordingly. Without knowledge of caveats, we may be heading for disaster as we saw in the CRT example earlier. What caveats achieve is confer perspective on the question by forcing us to step back and take in the larger picture.

2.2 Features of Data X Ray

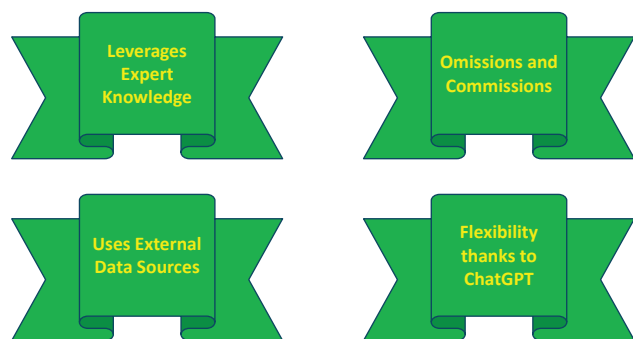


Fig 3 – Prominent Features of the Data X Ray Tool

Expert Knowledge

First off, the tool has expert knowledge on a wide range of subjects. It knows a lot about the content of the common data sources and that's how it identifies a data source as a remit claims data source or an SDOH (Social Determinants of Health) data source. It knows about the most common use cases in commercial analytics and the requirements a data source needs to meet to answer questions pertaining to a use case. It also knows about drugs and how that translates in terms of data requirements. For instance, if the drug is about prostate cancer, it reasons that since the typical onset of prostate cancer is around 67 and Medicare kicks in as of 65 (excluding disability, Lou Gehrig's, and ESRD), Medicare should represent the lion share of the data and as such a 50-50 split between Commercial and Medicare will raise a red flag. Another example: Say we are looking at an IV drug. The tool will look for the corresponding J-code of the drug in the Mx files of the data and will report an issue if it either does not find the J-code or the Mx files.

External Data Sources

To be able to identify holes in a data source or biases (lack of representativeness), one needs to have expectations. This is accomplished by comparing what is in the data with what should be in the data. The latter can be captured through knowledge of certain facts. Knowing that 30 million people in the US have been diagnosed with diabetes type 2 in 2021 (another 8 million undiagnosed), the tool can infer the capture rate of the diabetic population in the data source. In other cases, knowledge lies in a trusted external data source. For instance, using the Open Payments data source (Sunshine Act), the tool can estimate the capture rate of physicians in the data source that received payment from pharmaceutical manufacturers.

External data sources also tell us how trustworthy a data source is by gauging by how much the data source we are looking at is off the mark.

Omission and Commission

Fundamentally, there are two types of errors a data source can make. The first one is omission and this is when the data source does not say anything when it should. The second one is commission and it is when the data source says something that is incorrect.

Here are some examples of omission. The data withholds certain information, for instance, the

NPI of the physicians is encrypted or the name of a Payer is redacted. The data can also present a biased view of the world by not presenting the whole picture. This happens, for instance, when only transactions under medical benefit are reported and not those under pharmacy benefit. A lack of granularity is also a form of omission. For instance, the data may describe the payer as Commercial instead of BCBS of IL. A lack of stability as depicted by a trend break is also a case of omission. A sudden drop in reported transactions that cannot be explained by any market event can only be due to the data vendor losing data suppliers. Likewise, old or infrequently refreshed data fall under omission as more recent data has been left out.

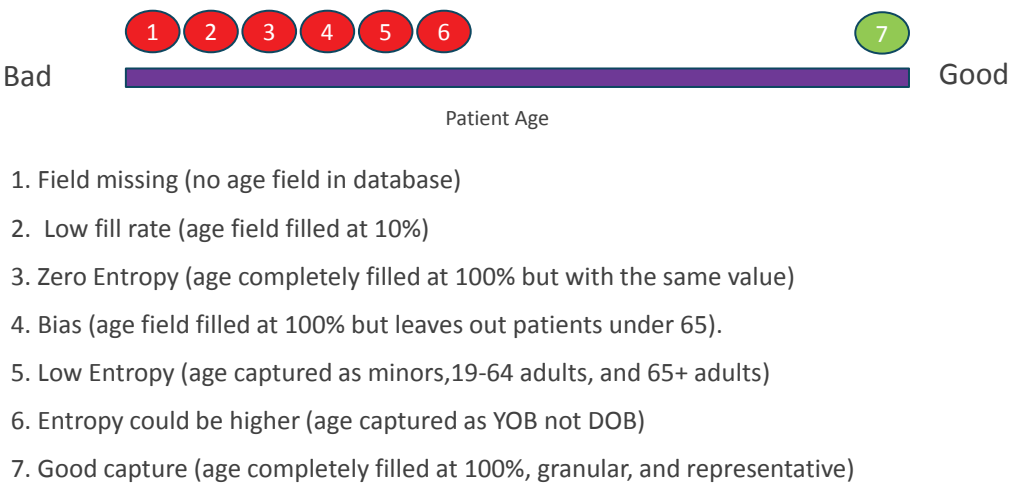


Fig 4 – Types of Omission: Patient Age Example

As for commission, the most compelling example is that of what we refer to as “stated vs. inferred” inconsistencies. This happens when the data source states a value but inference from other variables of the data source points to a different value. The data source may state the price of a drug for a transaction and that

price does not match what we get if we multiply the units by the unit price. More common examples include values that just cannot be: age of a patient, number of patients of a physician, amount of drug a patient takes, multiple patient ids for the same claim, and the list goes on.

Flexibility thanks to ChatGPT

For starters, the tool can take input of different formats and exhibits great flexibility thanks to ChatGPT. That's a mandatory feature otherwise the tool cannot identify the type of data source.

That flexibility does not stop there. The tool needs to interrogate the data source to assess its contents and to that end uses SQL queries. These queries cannot be picked up off the shelf but have to be generated on the fly as each data source uses its own layout to document the data: Field names may be unique to each data source and appear under tables with names that are also unique to the data source.

Data Source in a Nutshell

The question is: "What is the data source about in a nutshell?", and the response is a short description of the data source meant to help the user decide if the data source is of relevance. For example: "This is an EMR about breast cancer and spans 36 months of data from Jan 2021 to Dec 2023".

To explain how the tool accomplishes that, some background on agents, attributes, and id expression is in order. An agent is an

entity to which is assigned an id. Examples of agents include Patient, Physician, Drug, Payer, Diagnosis, Procedure, Claim, etc. Agents have attributes which are captured as fields in tables. Gender and YOB are attributes of a patient. Specialty and address are attributes of a physician. Mode of administration (Oral or IV) and strength are attributes of a drug. As for the id, it can be exposed, encrypted or internal. The NPI of a physician can be exposed in which case we know exactly who the physician is, encrypted in which case we can only tell them apart without knowing who they are, or internal in which case we need a look-up table to get a description of the physician. Without such a look-up table, the internal id is as good as encrypted.

Now each data source is regarded as a collection of agents, attributes, and id expressions. An open claims data source will have patients, physicians, drugs, diagnoses, procedures, and the like. Payers, pharmacies, and sites of care are optional. The physician is identified by an exposed id typically an NPI. An EMR, on the other hand, will not expose the id of the physician although there are a few EMRs that do. More importantly, the EMR carries Family History and Physician Notes.

1. Open Claims
2. Closed Claims
3. Submit
4. Remit
5. SP
6. EMR
7. Patient Narratives
8. Xponent (PTK, LTC, etc.)
9. CDM
10. Lab Data
11. SDOH
12. Payer Formulary
13. DDD (MD, by indication, etc.)
14. NPA
15. NSP
- 16....

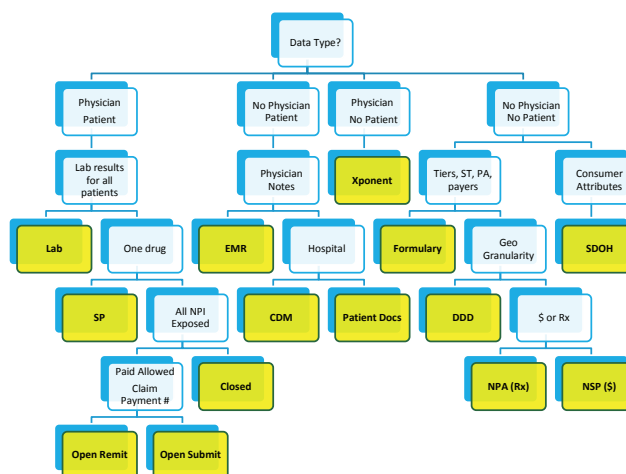


Fig 5 – Decision Tree to Identify Data Type

In essence, the tool uses a decision tree to identify the data source as one of the most popular data sources in commercial analytics. Each node of the decision tree asks about the presence or absence of specific agents, attributes, and expression of ids. If it fails to recognize the data source, it reports the agents that are present in the data source.

Cursory Data Assessment

The question here is: “Is the data source any good?”, and the response is a brief account of the shortcomings of the data source meant to help the user decide if the data source merits further consideration. As such, the emphasis here is more on the omission than commission.

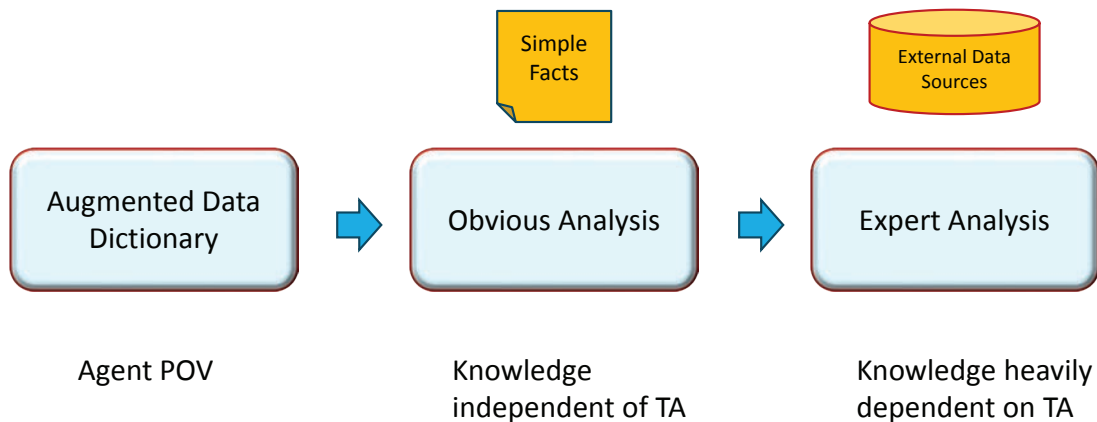


Fig 6 – Establishing Not Good

To that end, the tool looks at each agent in turn and for each agent goes over its attributes to see if they are missing, biased, not granular enough, and the like. Let’s illustrate with a few examples regarding the patient, the physician, and the payer to look at just these three.

For the patient, for instance, a look at the dictionary tells the tool that age and gender are present but not the zip3. By peeking into the values under field names, the tool may notice that what’s reported under age is not DOB or YOB or 5-year age groups. It may also notice that although ethnicity is present, African Americans and Hispanics are grossly under-represented while Asians are nowhere to be

found. This heads-up is very apropos for the analyst who was planning to run an ethnicity analysis as the analyst now knows that the data is biased.

For the physician, for instance, it may report that the id is not the standard NPI but rather an internal physician id for which there is no corresponding lookup profile table, suggesting that the id is no better than an encrypted id. In other instances, it may report that physicians are assigned multiple roles such as rendering, referring, billing, attending, and the like, which may be of particular interest for the analyst who is planning to conduct a referral analysis or establish spheres of influence.

For the payer, for instance, the tool may report that the data does not disclose any plans under the Payer, let alone the Group, Bin, and PCN. Only the method of payment is reported, namely, Medicare, Medicaid, Commercial, and Other. This information tells the user who was planning a deep dive into the payers to drop this data source and look for a more granular data source instead.

The tool also provides some sense of the footprint of the data source. If the data source contains physicians for instance, it will report the number of unique physicians. If the physicians are identified with an NPI, the tool will match them against physicians from an external data source such as Open Payments from the Sunshine Act or NPPES (National Plan and Provider Enumeration System) to give a sense of the footprint.

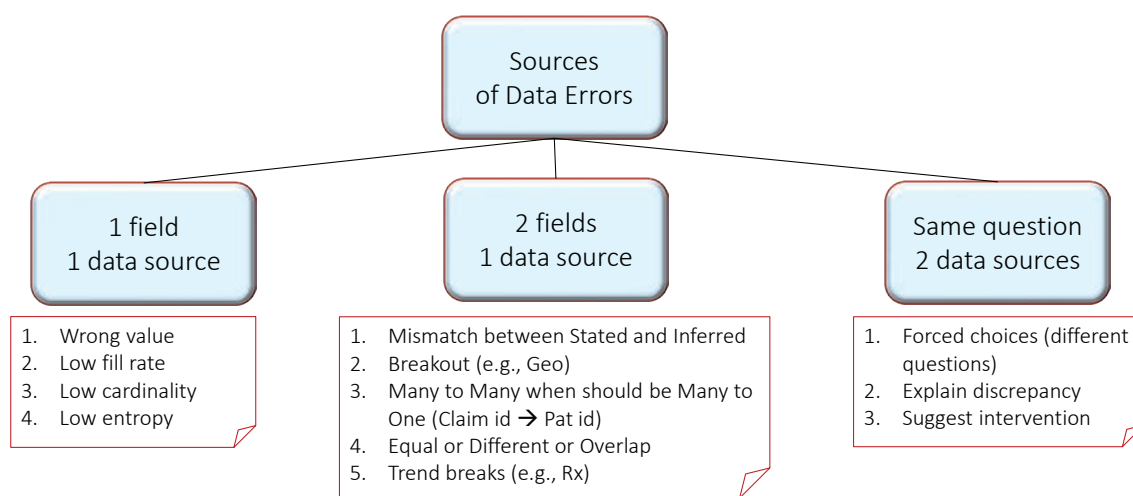


Fig 7 – 3 Ways of Uncovering Errors by Commission

The tool also provides some information on the stability of the data. A good example is that of the trend-break. By tracking the Rx's or new patient starts, for instance, the tool may notice a break in trend. If there is no event in the marketplace to explain the sudden change in trend, we may be looking at a data capture issue where the data vendor lost a data supplier or two.

Adequacy for Data for a Use Case

Now that we know what the data is about and relevant despite the shortcomings, the next question is whether we can leverage that data source for a specific use case. Say we are considering physician segmentation and targeting for a given drug in a therapeutic area.

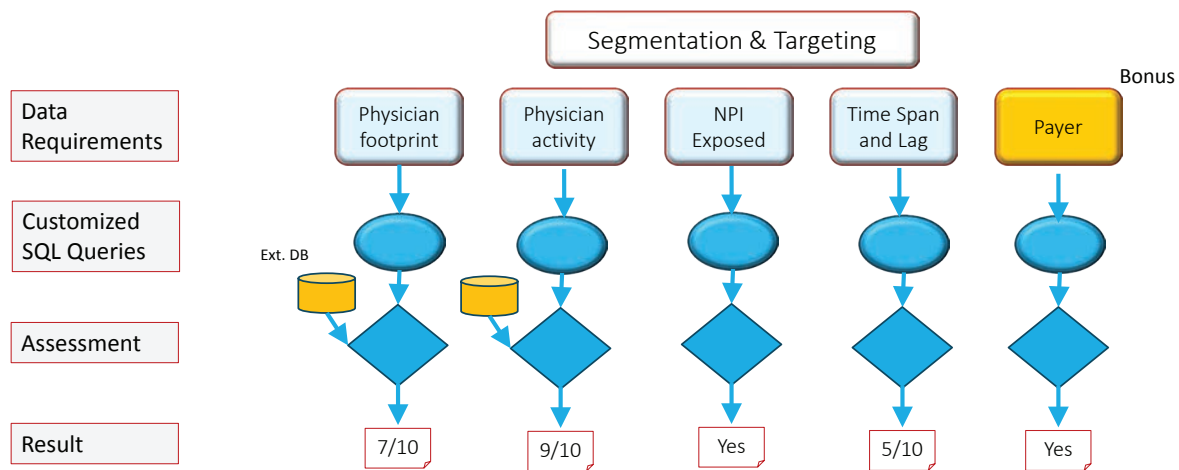


Fig 8 - Assessing Data Requirements

The tool starts off by retrieving of the list of data requirements the data source needs to satisfy to be the relevant for the use case. Below are the most major ones.

Physician Footprint – The data source needs to have a good coverage of relevant physicians.

Physician Activity – The data source must be more than a list of physicians names; it must also capture the activity of physicians so we can tell apart the active from the not so active.

NPI Exposed – The data source should expose the NPI of the physician so we would know the name, address, city, state, zip, and specialty of the physician. If the NPI is encrypted, we will not be able to target the physicians.

Time Span and Lag – The data source needs to span a good time window and be not too old. That's because the activity of physicians changes over time as new entrants come in, physicians retire or die, new physicians join the ranks, new indications are approved, incumbent drugs lose exclusivity, health insurance change coverage policy, and the like.

Payer – This is nice to have since knowing who the payers are at the individual physician level and what kind of OOP their patients are facing can help develop more efficient segmentation and targeting.

The tool then assesses how well the data source satisfies each of the data requirements. To that end, it proceeds in 3 steps. First, it generates and runs SQL code against the data source to establish what is available in the data. Second, it uses external knowledge to shed light on what the ideal should be. Third, it compares what is available in the data source with what the ideal is and provides a score to indicate the level of satisfaction of the data requirement.

If the tool finds out that the NPI is encrypted, for instance, it will report a showstopper to indicate that the data source is not fit for segmentation and targeting. If there is no payer information, the tool reports a setback to warn the analyst that the segmentation may not be as powerful as a key attribute of physician activity is missing.

BUTTERFLY EFFECT

The Butterfly Effect refers to the phenomenon whereby a small change such as the flapping of the wings of a butterfly in Europe may have a huge impact such as a hurricane developing a couple of weeks later in the United States. Indeed, a small change in the business question may lead to very different data requirements. As a result, a data source that was found perfectly adequate for a use case may not be fit for a business question pertaining to that same use case. Let's illustrate that point by considering similar sounding questions that ask about A happening before B where A and B are drugs or events in the patient journey.

Variation 1: Is A before B?

Interpretation: Is A ever before B?

Requirements:

1. A handful of examples where A is before B
2. Database does not need to have a good coverage

Ease of Satisfaction: 1 of 5

Variation 2: Is A always before B?

Interpretation: Knowing that A is most of the time before B, does it ever happen that A is not before B?
Requirements:

1. Lots of examples of A before B (to validate the implied hypothesis that A is before B)
2. Capture A not before B if this were to happen to ensure that not seeing A before B really means A is not before B

Ease of Satisfaction: 2 of 5

Variation 3: How frequently is A before B?

Interpretation: In all cases where A is present, how frequently is A before B?

Requirements:

1. Excellent capture of A
2. Excellent capture of B (otherwise we will not be able to comment on the frequency of A before B)

Ease of Satisfaction: 3 of 5

Variation 4 – When is A before B?

Interpretation: Under what circumstances is A before B?

This can be about the patient profile, the physician profile, the requirements of the payer, events in the patient journey, etc. Also, these circumstances need to be present only when A is before B.

Requirements:

1. Very good capture of peripheral information
2. Good capture of A before B
3. Good capture of A not before B (to establish that the proposed circumstances only hold when A is before B and not otherwise).

Ease of Satisfaction: 4 of 5

Variation 5: What happens when A is before B?

Interpretation: What events take place after B when A is before B? It could be C where C is a drug, a hospitalization, a change in insurance, etc.

Requirements:

1. Good time span after B when A is before B
2. Very good capture of peripheral information in that time window.

Ease of Satisfaction: 3 of 5

Variation 6: Is A immediately before B?

Interpretation: There should be no interloper X between A and B.

Requirements:

1. Agreement as to how much time can elapse between A and B for A to be considered immediately before B.
2. Good capture of X especially between A and B (to establish that A is not immediately before B).

Ease of Satisfaction: 4 of 5

Variation 7 – Why is A before B?

Interpretation: What is the rationale for A before B? Step therapy requirement of the drug formulary? Clinical motivation? Example: A is a coagulant and B is a surgery? Cocktail therapy? A needs to be initiated before B? There should be no interloper X between A and B.

Requirements:

1. The answer most likely lies outside of the database that captures A and B.
2. Need to look elsewhere (e.g., drug formulary of the Payer, Medical journals, etc.)

Ease of Satisfaction: 5 of 5

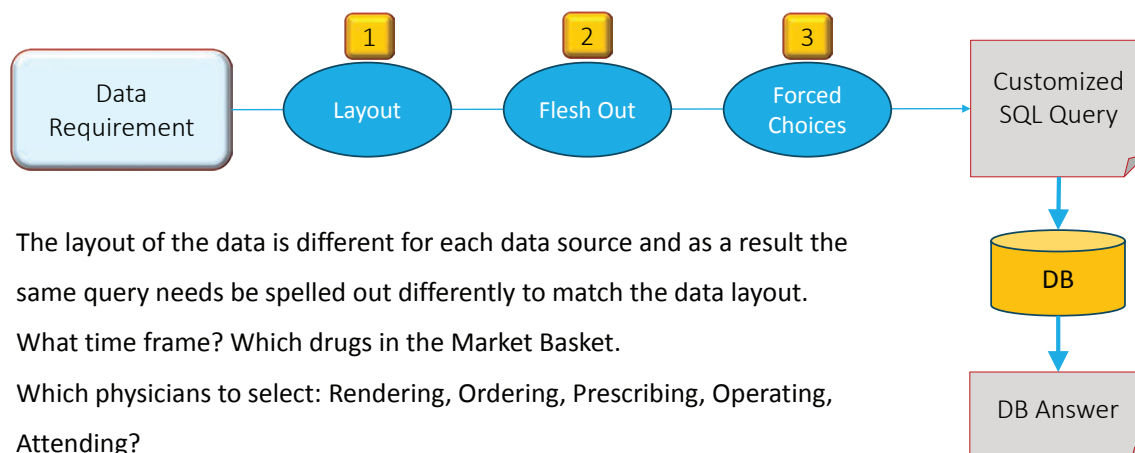


Fig 9 – Query Generation to Interrogate Data Source

Caveats of the Data Source

Caveats are arguably the most treacherous of all the dangers that bedevil the analysts like the CRT example in the introduction. Indeed, ignorance of caveats is far from bliss.

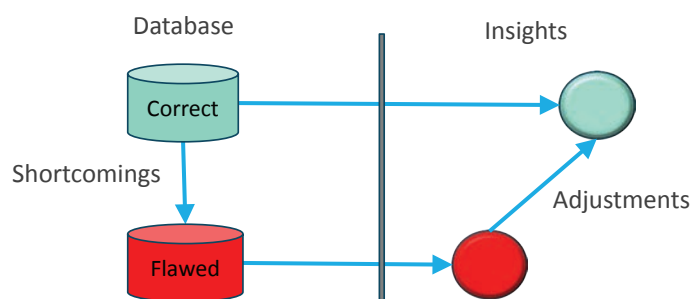
Why are caveats so hard to come by? For one, data vendors do not like to talk about caveats of their data sources as they believe that spelling out caveats underscores the shortcomings of the data source which, as a result, makes the data source come across as less valuable. Have you ever seen a list of caveats accompany a data source upon delivery?

For another, analysts do not for the most part have the luxury of choosing the data source to analyze when answering a business question. The data source has already been chosen before the analyst is asked to answer the question. To make matters worse, the data source may have undergone multiple transformations before consumption. The address, specialty, and affiliation of the physician may have been updated. The profile of the payer may have been curated. Additional data feeds such as SP

data or direct feeds from IDN's may have been brought in to plug holes in the data source. Even the very diligent analyst may have a hard time tracking back the caveats of the resulting data source.

There are two kinds of caveats. Caveats that can be uncovered by zeroing in on the data source. For instance, comparing the number of patients in the data source with the population from the US Census or incidence and prevalence data from CDC may expose the holes in the data. Comparing ethnic distribution of patients with ethnicity data in the public domain may reveal that African Americans and Hispanics are under-represented. As for the second type of caveats, they require explicit knowledge about how the data is captured. Unless told explicitly, we may never know what data feeds were included, let alone in what proportion, or what data sources were left out.

This suggests a two-pronged approach for uncovering caveats of a data source. First, conduct a thorough analysis of the data source using external knowledge to identify holes and biases in the data. Second, gather information



A caveat is a shortcoming of the database that allows the analyst to gain a larger perspective on the insights coming from the database and tweak the insights accordingly.

Comments on Caveats

1. The significance of a caveat is contingent upon the business question at hand. Some questions may be heavily impacted while others not at all.
2. The more accurate the description of the shortcomings of the data source, the better the odds of being successful at adjusting the insights that come out of the database.
3. There are 2 types of shortcomings. The first one can be uncovered by zeroing in on the data and comparing the findings with those from external trusted data sources. The second requires knowledge on how the data is collected to expose the caveat.

Fig 10 – Shortcomings and Adjustments

regarding how the data source is put together. This can be challenging because not only the data vendor has little to gain to reveal the recipe of the data source, but also the data vendor are contractually be bound by several of their data suppliers not to reveal their identity.

One final comment on caveats. Ultimately, what matters is how to tweak the insights drawn from the data source to account for the caveats. Assuming the insight is expressed as a number (e.g., market share of a drug), recognizing the caveat means adjusting the raw number upwards or downwards by a certain amount. In the most favorable case, we would know both the direction and the magnitude of the adjustment. In the worst case, we have no idea if the number is too high or too low let alone by how much. We can only say that the insights need be taken with a pinch of salt.

Conclusion

In this paper, we described a novel approach that produces a domain-oriented assessment of a data source in a very short amount of time. Having such a companion assessment on the

ready will have a transformative impact and far-reaching ramifications. For starters, the assessment carries great educational value as it indicates very clearly what we should be looking for in a data source. Also, this assessment will be accessible to anyone who is interested and that will be a much larger population as no programming skills are required any more. The data IQ of the industry will shoot up. Also, having a standard format to assess data sources means that data sources can be easily compared with each other, thereby giving the upper hand to superior data sources at the detriment of mediocre data sources. Our business questions will have better answers.

Manufacturers will deploy data sources more efficiently to answer business questions. That's because they'd know right away how well a data source meets the requirements of the business question, which will help them decide if they should leverage the data source, replace it with another, or explore merging data sources. Communication with data vendors will be more efficient as manufacturers can clearly spell out the data issues and the kind of fix they are expecting.

Data vendors will be reminded of the shortcomings of their data sources and will feel heightened pressure to address the data issues promptly. The silver lining is data innovation. Nascent opportunities will be easier to recognize, leading to strategic investments,

thoughtful partnerships deals, mergers and acquisitions, which in turn will result in innovative data offering.

In a word, it's a win-win for everyone.

Acknowledgements

Many contributed to this article through engaging conversations, puzzling questions, and thought-provoking reflections. They include Fred Hooey, Ji Xiaohao, Erle Davis, Shunmugam Mohan, and Zhangjin Xu to mention just a few. We apologize ahead of time to all the others who helped us for not calling them out.

References

1. "The AI revolution in Medicine – GPT-4 and Beyond", Peter Lee, Carey Goldberg, Isaac Kohane with Sebastien Bubeck, Pearson Education, 2023, 282 pp.
2. A survey of Data Quality Measurement and Monitoring Tools, Lisa Ehrlinger and Wolfram Woss, Secondary Data Mining and Management, Vol 5, 2022, pages 1-30.
3. "Maladies of Claims Data: Manifestations, Origins, and Cures", JP Tsang, President of Bayser, JPMSA Spring 2020, pages 17-35.

Unleashing the Power of Rep Notes: Extracting Actionable Insights through NLP driven Analysis with LLMs and Generative AI

*Arvind Balaji Gunasekaran, Principal Consultant – Data Science, Trinity Lifesciences;
Nagasubramanya Nabha, Vice President – Data Science, Trinity Lifesciences.*

Abstract: In the pharmaceutical industry, unlocking the wealth of knowledge embedded in a vast array of pharma representative notes is pivotal for enhancing sales strategies, fostering data-driven decision-making, and ultimately optimizing interactions with healthcare professionals (HCPs). Presently, with multiple representative teams operating in isolated silos, the full potential of these valuable notes remains untapped.

This innovative approach aims to achieve the following outcomes:

- **Intuitive Insights Across Therapeutic Areas and Business Units:** By analyzing and processing the representative notes using NLP algorithms, we can uncover valuable insights spanning various therapeutic areas and business units. These insights serve as a compass, guiding strategic decision-making and enhancing overall performance.
- **Identifying Hot Topics, Trends, and Key Client Remarks:** The NLP-driven analysis enables the identification of emerging trends, hot topics, and crucial remarks from HCP interactions. Such real-time insights empower representatives to stay ahead of the curve and tailor their engagements more effectively.
- **HCP Conversation Summarization using Generative AI for Reps via CRM solutions:** Implementing Generative AI within CRM tool empowers representatives with automated summarization of HCP conversations. Augmented by LLMs, this approach streamlines data processing and equips representatives with comprehensive summaries to fuel their engagements.

Trinity's presentation will cover the design, development, and implementation of the entire solution, including real world examples.

Background

In the dynamic landscape of the pharmaceutical industry, the ability to harness the vast reservoirs of data generated from interactions between pharmaceutical representatives and healthcare professionals (HCPs) has emerged as a crucial determinant of success. Representative notes, comprising detailed accounts of these interactions, encapsulate valuable insights

ranging from emerging healthcare trends to nuanced HCP preferences. However, despite their potential, these notes often languish within disparate silos, hindering their effective utilization in informing sales strategies and optimizing engagements.

Traditionally, accessing actionable insights from representative notes has been a labor-intensive and time-consuming endeavor.

Manual review processes, characterized by subjective interpretation and limited scalability, have constrained the ability of pharmaceutical companies to extract meaningful intelligence from these troves of data. Moreover, the exponential growth of unstructured data in recent years has exacerbated these challenges, necessitating the adoption of innovative methodologies capable of unlocking the latent value inherent in representative notes.

The emergence of Generative AI technologies has expanded the horizons of data analysis by enabling the generation of human-like text based on learned patterns and contexts. By leveraging deep learning architectures such as Language Models (LMs), Generative AI can synthesize coherent narratives, summarize complex information, and even engage in conversational exchanges resembling those between pharmaceutical representatives

and HCPs. This capability holds profound implications for augmenting representative productivity, facilitating knowledge dissemination, and enhancing customer interactions.

In light of these advancements, there exists a compelling imperative for pharmaceutical companies to embrace NLP-driven analysis and Generative AI to unlock the untapped potential of representative notes. By harnessing the power of these technologies, organizations can transcend the limitations of manual data processing, unearth actionable insights with unprecedented speed and accuracy, and cultivate a data-driven culture that fosters innovation and excellence. Thus, the convergence of NLP and Generative AI heralds a new era of optimization in pharmaceutical sales strategies, characterized by agility, intelligence, and strategic foresight.

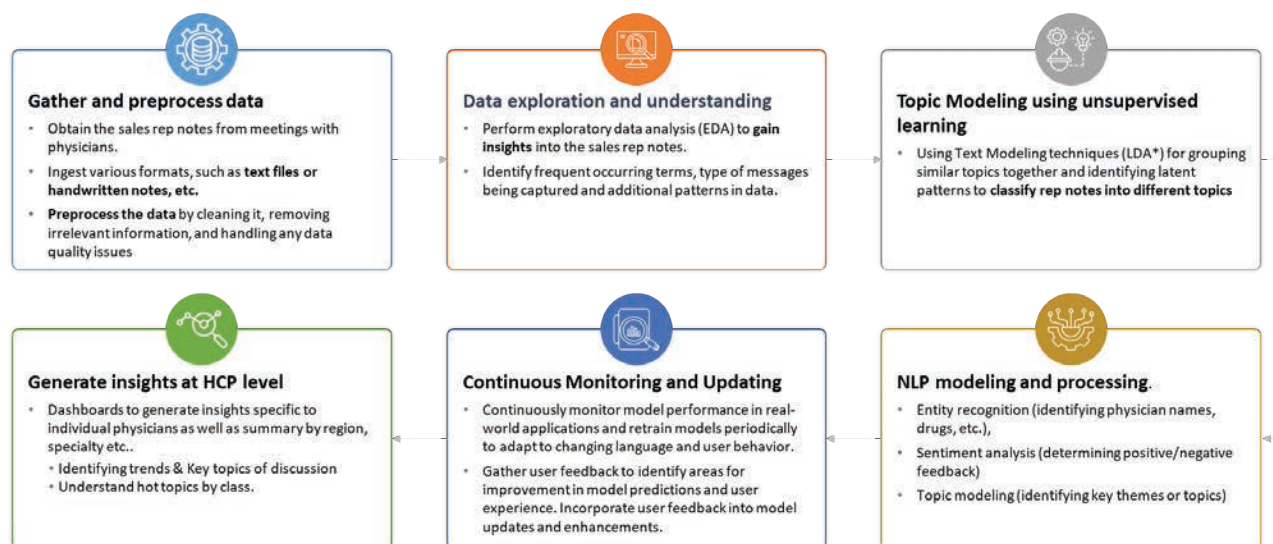


Figure 1: General Approach on Processing Sales Rep data

Traditional Approach (NLP)

In the dynamic landscape of pharmaceutical sales, harnessing the power of Natural Language Processing (NLP) has emerged as a pivotal strategy for extracting actionable insights from sales representative notes. The above image delineates a comprehensive six-step approach aimed at leveraging NLP techniques to optimize sales strategies, enhance customer interactions, and drive business growth within the pharmaceutical industry.

1. Gather and Preprocess Data:

The first step in our approach involves gathering sales representative notes obtained from meetings with physicians. These notes may exist in various formats, including text files, handwritten documents, or digital transcripts. Preprocessing the data is imperative to ensure its quality and relevance. This entails cleaning the data, removing noise, handling missing information, and standardizing formats to facilitate downstream analysis.

2. Data Exploration and Understanding:

Once the data is pre-processed, exploratory data analysis (EDA) is conducted to gain insights into the sales representative notes. Through EDA, we identify frequent terms, patterns, and the nature of messages captured in the notes. This step provides a foundational understanding of the data, guiding subsequent analysis and modelling efforts.

3. Topic Modelling using Unsupervised Learning:

Unsupervised learning techniques, such as Latent Dirichlet Allocation (LDA), are employed for topic modelling. LDA allows us to group similar topics together and uncover latent

patterns within the sales representative notes. By categorizing notes into distinct topics, we facilitate the classification and organization of information, enabling more targeted and efficient analysis.

4. NLP Modelling and Processing:

This step involves the application of various NLP techniques to extract valuable insights from the sales representative notes. Key tasks include entity recognition, sentiment analysis, and topic modelling. Entity recognition involves identifying entities such as physician names, drugs, and medical terms mentioned in the notes. Sentiment analysis enables us to determine the sentiment conveyed in the interactions, distinguishing between positive and negative feedback. Topic modelling further identifies key themes or topics discussed in the notes, providing deeper insights into customer preferences and concerns.

5. Continuous Monitoring and Updating:

The NLP models developed are continuously monitored and updated to ensure their relevance and effectiveness in real-world applications. This involves periodically retraining the models to adapt to evolving language patterns and user behaviour. Additionally, user feedback is gathered to identify areas for improvement in model predictions and user experience. Incorporating this feedback into model updates and enhancements ensures that the NLP algorithms remain robust and aligned with business objectives.

6. Generate Insights at HCP Level:

The final step focuses on generating actionable insights specific to individual healthcare professionals (HCPs) as well as summary insights by region, specialty, etc. Dashboards

are utilized to visualize and communicate these insights effectively. By identifying trends, key topics of discussion, and hot topics by class, pharmaceutical companies can tailor their sales strategies and engagements to meet the unique needs and preferences of HCPs.

The six-step approach outlined in this paper demonstrates the effectiveness of leveraging NLP for pharmaceutical sales optimization. By systematically gathering, preprocessing, analysing, and generating insights from sales representative notes, pharmaceutical companies can enhance their sales strategies, improve customer interactions, and drive business success in an increasingly competitive market.

Generative AI Approach with LLMs

Utilizing Generative Artificial Intelligence (AI) for Healthcare Professional (HCP) conversation summarization is a multifaceted process that involves integrating various components to

distill actionable insights from interactions between pharmaceutical representatives and HCPs. At the core of this approach lies the diverse array of data sources, including representative notes, HCP demographics, call activity records, marketing responses, and HCP target lists. These sources provide a rich repository of information, capturing nuances of discussions, preferences, and historical engagements, essential for generating meaningful summaries.

The engineering aspect of this endeavor is crucial, requiring meticulous prompt engineering to elicit coherent summaries from Generative AI models. Leveraging advanced Language Models (LMs) such as GPT-3.5 Turbo or LLama enhances the sophistication of generated text, ensuring it aligns closely with the context and intent of the original conversations. Additionally, frameworks like Langchain facilitate the synthesis of coherent summaries, streamlining the process of distilling insights from HCP interactions.

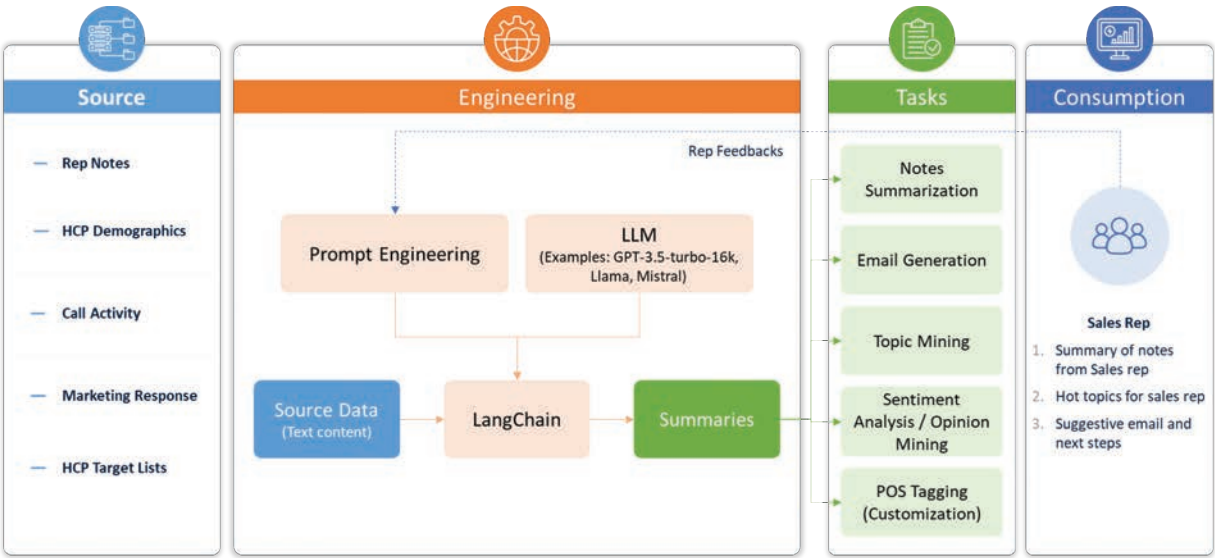


Figure 2: How we use LLMs and designing the output

Generative AI for HCP conversation summarization delivers a spectrum of tasks aimed at extracting actionable insights. These include condensing lengthy representative notes into concise summaries, automatically generating emails based on summarized conversations, identifying prevalent topics and themes, analysing sentiment, and assigning part-of-speech tags to enhance semantic understanding. Each task contributes to the overarching goal of extracting actionable intelligence from HCP interactions, empowering pharmaceutical companies to make informed decisions and tailor their engagement strategies effectively.

The consumption layer plays a pivotal role in facilitating the utilization of generated insights by sales representatives. Providing them with succinct summaries of HCP conversations equips them with actionable intelligence for subsequent engagements. Highlighting trending topics and themes within HCP conversations enables representatives to steer discussions towards relevant subjects, fostering meaningful interactions. Furthermore, offering personalized suggestions for email content and follow-up actions based on summarized conversations streamlines communication and relationship management, ultimately enhancing customer satisfaction and driving business growth.

The integration of Generative AI for HCP conversation summarization offers pharmaceutical companies a powerful tool for extracting actionable insights from representative notes. By leveraging advanced technologies and methodologies, organizations can enhance sales strategies, optimize customer interactions, and drive business success in an increasingly competitive landscape.

Proposed Outcomes

The proposed outcomes represent a comprehensive suite of solutions designed to address key challenges faced by pharmaceutical sales representatives. By harnessing the power of NLP-driven algorithms, these outcomes aim to revolutionize sales strategies, empower representatives, and elevate the quality of interactions with HCPs. From automated meeting preparation to real-time insights on trending topics, each outcome contributes to a holistic approach towards sales optimization and customer engagement.

Gen AI Chatbot to Aid in Meeting Preparation:

One of the primary challenges faced by pharmaceutical representatives is the time-consuming nature of meeting preparation. The Gen AI Chatbot offers a transformative solution by automating and streamlining this process. Leveraging NLP algorithms, the chatbot analyses historical data, representative notes, and HCP profiles to generate personalized meeting briefs. These briefs include pertinent information such as HCP preferences, past interactions, and key discussion points, enabling representatives to enter meetings well-prepared and informed.

Smart Task Curator and Scheduler:

Efficient task management is essential for sales representatives operating in dynamic and demanding environments. The Smart Task Curator and Scheduler leverage NLP-driven algorithms to prioritize and schedule tasks intelligently. By analysing representative notes, customer feedback, and market trends, the system identifies high-priority tasks and allocates resources accordingly. Additionally, the scheduler optimizes time management by

factoring in travel time, meeting durations, and other variables, ensuring representatives maximize their productivity and focus on activities with the greatest impact.

Email Drafts Generation:

Crafting personalized and compelling email communications is a time-intensive endeavour for pharmaceutical representatives. The Email

Drafts Generation feature alleviates this burden by automating the creation of email templates tailored to individual HCPs. Drawing upon NLP-based sentiment analysis and topic modelling, the system generates emails that resonate with recipients, addressing their concerns and preferences. Representatives can further customize these drafts before sending, maintaining a personal touch while saving valuable time and effort.

Analyze discussion topics across various rep – HCP interaction and publish key topics of interest for other teams to leverage

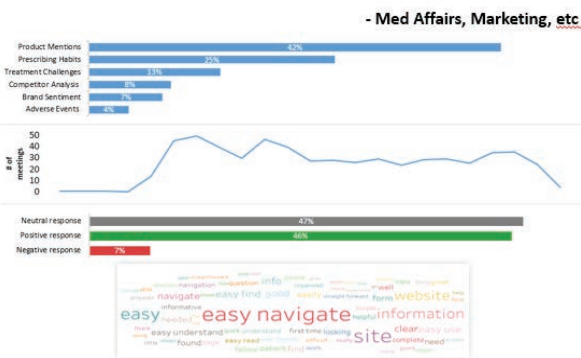


Figure 3: A sample output of the dashboard with top topics and HCP perception

From the rating / other questions on openness to adopt or likelihood to prescribe ☐ Summarize and present results



HCP Perception Analysis:

Understanding HCP perceptions and sentiments towards pharmaceutical products and services is essential for effective relationship management. The HCP Perception Analysis utilizes advanced NLP techniques to analyse representative notes and discern underlying sentiments and attitudes expressed by HCPs. By categorizing feedback as positive, neutral, or negative, the system provides representatives with valuable insights into HCP perceptions, enabling them to tailor their approaches and address concerns proactively. Moreover, longitudinal analysis facilitates the identification of trends and patterns, guiding strategic decision-making and resource allocation.

Real-time Insights on Trending Topics:

In the rapidly evolving landscape of healthcare, staying abreast of emerging trends and topics is crucial for maintaining relevance and competitive advantage. The Real-time Insights on Trending Topics feature leverages NLP-driven algorithms to monitor and analyse industry news, social media conversations, and academic publications in real-time. By identifying key themes, hot topics, and emerging trends, the system empowers representatives with timely and actionable insights. Armed with this knowledge, representatives can engage HCPs in informed discussions, position their products effectively, and capitalize on emerging opportunities.

The targeted outcomes outlined in this section exemplify the transformative potential of integrating NLP-driven solutions into pharmaceutical sales strategies. By automating routine tasks, providing actionable insights, and enhancing communication effectiveness, these outcomes enable representatives to focus their efforts strategically, cultivate meaningful relationships with HCPs, and drive business growth. Moving forward, continued investment in NLP technologies and innovation will further propel the evolution of pharmaceutical sales practices, ushering in a new era of efficiency, effectiveness, and customer-centricity.

Utilizing the Use Case Maturity Continuum for Generative AI in Life Sciences

The integration of Generative Artificial Intelligence (AI) into life sciences presents a transformative opportunity for organizations to enhance decision-making, drive innovation, and improve customer engagement. However, the journey towards realizing consistently high-quality results with Generative AI use cases is often characterized by iterative refinement and feedback loops. In this context, the Use Case Maturity Continuum offers a structured framework for navigating this evolution and achieving optimal performance in Generative AI applications within the life sciences sector.

1. Tech Hooked Up:

At the initial stage of the maturity continuum, Generative AI is employed to interpret queries and provide results based on Language Model (LM) outputs and connected sources or data. However, the quality of answers may vary, and the results may not consistently meet expectations. Despite these challenges, this stage represents a foundational step towards leveraging Generative AI in life sciences, laying the groundwork for subsequent refinement and optimization. Example: If a sales rep asks,

“Show me the recent product preferences of Dr. Smith,” generative AI can analyze the notes, extract relevant information about Dr. Smith’s preferences, and provide the sales rep with a summary of recent product choices made by Dr. Smith

2. Prompting Refinement:

As organizations progress along the maturity continuum, a feedback loop is established with users, enabling iterative refinements in prompt pipelines, prompt engineering, and model selection. This feedback-driven approach allows for adjustments based on user input, leading to improvements in the quality and relevance of generated answers. While results may still be inconsistent, this stage marks a significant advancement towards achieving more reliable outcomes with Generative AI. Example: Sales Rep notes can have 3 primary prompts which can be fine-tuned: Summary, Actions from the notes, Topics of Interest.

3. Prompting Steady State:

In the Prompting Steady State phase, organizations have optimized the performance of Generative AI answers through continuous refinement and fine-tuning of prompt strategies. Confidence increases in the ability to generate acceptable answers consistently, with a higher likelihood of meeting user expectations. This stage signifies a maturation in the use of Generative AI within life sciences, with organizations achieving greater stability and predictability in outcomes. This enhancement in response quality ensures that the information extracted from the notes is more accurate and valuable for decision-making and customer engagement.

4. Data Science Refinement:

At the pinnacle of the maturity continuum, organizations leverage collected user feedback

and ratings to train a supplemental model layer aimed at further improving results quality. This refinement process may involve the development of proprietary algorithms or the implementation of models that learn from client users' feedback. By incorporating user insights into the training process, organizations can refine Generative AI models to better align with user preferences and requirements, ultimately driving higher levels of satisfaction and performance. For instance, when a sales rep interacts with the system and provides feedback on the relevance and accuracy of the responses, this data is used to fine tune the model.

The Use Case Maturity Continuum provides a roadmap for organizations seeking to maximize the value of Generative AI in life sciences. By progressing through stages of refinement, feedback integration, and continuous improvement, organizations can unlock the full potential of Generative AI use cases, differentiate themselves within the industry, and drive innovation and excellence in customer engagement. As the field of Generative AI continues to evolve, leveraging frameworks such as the Use Case Maturity Continuum will be essential for organizations to stay at the forefront of innovation and maintain a competitive edge in the life sciences sector.

Case Study:

Our client, a leading pharmaceutical company, expressed a keen interest in exploring the potential of leveraging free-form notes collected from sales representatives following conversations with Healthcare Professionals (HCPs). The client sought to extract actionable insights from this unstructured data to enhance their sales strategies and optimize interactions with HCPs.

Construct a Vector Database: To facilitate systematic storage and efficient retrieval of

representative notes related to HCP interactions, we developed a robust vector database. This database ensured organized storage of diverse data sources, including text files, handwritten notes, and digital transcripts, enabling seamless access and retrieval for analysis.

Create an Extraction Framework: We formulated an extraction framework designed to identify and extract relevant representative notes associated with targeted HCPs. This framework employed advanced techniques to parse through unstructured data, filter out noise, and extract key insights, laying the groundwork for subsequent analysis.

Leverage LangChain and Prompt Engineering: Utilizing the LangChain framework and prompt engineering techniques, we crafted effective queries and prompts tailored to the pharmaceutical domain. These techniques facilitated the extraction of actionable insights from representative notes by guiding the Generative AI model to focus on relevant topics and themes.

Employ GPT-3.5-Turbo Model: Harnessing the capabilities of the state-of-the-art GPT-3.5-Turbo model by OpenAI, we generated intelligent recommendations based on insights derived from representative notes. This model synthesized coherent narratives, summarized complex information, and provided valuable recommendations to sales representatives based on the analyzed data.

Enhance HCP Engagement Strategy: By integrating these advanced technologies and methodologies into the client's HCP engagement strategy, representatives were equipped to significantly enhance their approach. The data-driven insights derived from representative notes transformed the engagement strategy into a more informed, targeted, and effective approach.

Project Outcomes:

Data-Driven Insights: The project yielded a treasure trove of data-driven insights, empowering sales representatives to make more informed decisions when engaging with HCPs. These insights provided valuable guidance on tailored communication strategies and product positioning.

Enhanced HCP Profiling: Representatives gained a deeper understanding of individual HCP preferences, needs, and behaviours through the analysis of representative notes. This enhanced profiling enabled representatives to tailor their interactions to specific HCP preferences, fostering stronger relationships and driving better outcomes.

Time and Resource Efficiency: With access to actionable insights derived from representative notes, sales representatives were able to optimize their time and allocate resources more effectively. This resulted in higher productivity and improved efficiency in engaging with HCPs, maximizing the impact of each interaction.

Personalized Engagement: Armed with personalized insights gleaned from representative notes, sales representatives could engage with HCPs on a more personal level. Tailored communications and interactions based on HCP preferences and interests strengthened rapport and facilitated meaningful conversations, ultimately driving greater success in sales initiatives.

Challenges and Mitigation

Addressing challenges and implementing effective mitigation strategies are crucial aspects of integrating Generative AI solutions into pharmaceutical sales strategies. One prominent challenge lies in the behaviour of sales representatives, where time constraints during sales visits often hinder the detailed documentation of interactions with healthcare professionals (HCPs). To mitigate this challenge, a comprehensive change management approach is essential. Behavioural trainings and change management processes can be implemented to encourage sales representatives to prioritize note-taking and incorporate it seamlessly into

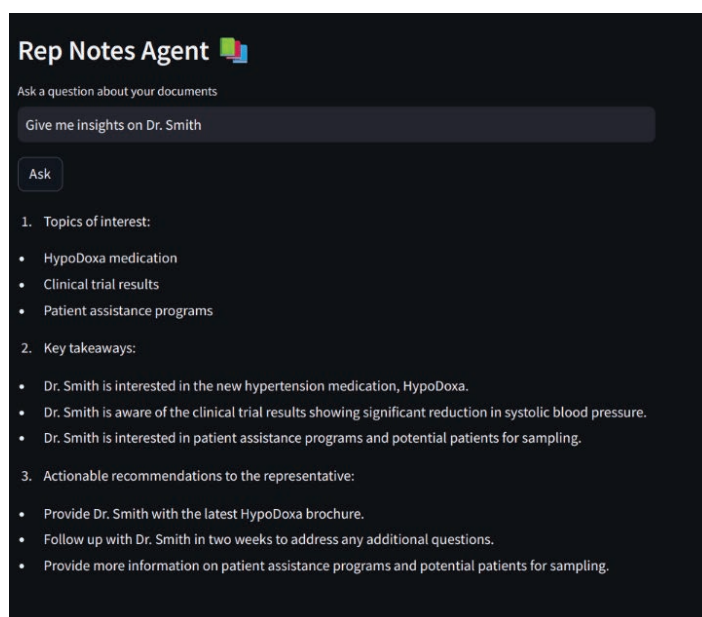


Figure 4: Illustrative output for the chatbot used by the sales representative

their interactions. Additionally, forming smaller focus teams or field input teams dedicated to running this initiative can provide targeted support and foster a culture of data-driven engagement.

Compliance issues pose another significant challenge, with resistance from sales representatives stemming from concerns about entering sensitive data into the system. To address this challenge, a mitigation strategy involves implementing sanitized notes. Leveraging Generative AI, the system can automatically flag any sensitive or non-compliant data within the notes before submission. This proactive approach ensures compliance with regulatory requirements while maintaining the integrity of the data captured during sales interactions.

The accuracy of outputs generated by Generative AI presents a further challenge, as outputs are suggestive and may not be fully vetted results. To mitigate this challenge, integration with human intelligence is crucial. Human cognitive skills are necessary to evaluate and consider the outcomes from Generative AI before utilization. By incorporating human judgment into the decision-making process, organizations can ensure the reliability and relevance of the insights derived from Generative AI solutions, enhancing the overall effectiveness of sales strategies and customer engagement initiatives.

Conclusion

In conclusion, this journal has presented a comprehensive approach to leveraging Natural Language Processing (NLP) and Generative Artificial Intelligence (AI) for pharmaceutical sales optimization. The proposed approach encompasses data gathering, preprocessing, NLP modelling, continuous monitoring, and insights generation, culminating in actionable strategies to enhance sales representative effectiveness and customer engagement. By harnessing advanced NLP techniques and Generative AI models, pharmaceutical companies can unlock valuable insights from representative notes, optimize HCP interactions, and drive business growth. The potential benefits of this approach include data-driven decision-making, enhanced HCP profiling, improved resource efficiency, and personalized engagement. Looking ahead, future directions in leveraging NLP and Generative AI for pharmaceutical sales optimization involve exploring novel applications such as voice-based interactions, sentiment analysis on social media platforms, and predictive analytics for sales forecasting. Advancements in deep learning architectures, model interpretability, and ethical considerations will further shape the landscape of NLP-driven solutions in the pharmaceutical industry, empowering organizations to stay at the forefront of innovation and drive superior business outcomes.

About the Authors:

Arvind Balaji Gunasekaran is a Principal Consultant in Data Science at Trinity Life Sciences. With over a decade of experience in the data science and analytics industry, Arvind specializes in delivering advanced analytical solutions that provide actionable insights and address complex business challenges.

Nabha Subramanya is a Vice President – Data Science at Trinity Life Sciences. 15 years of experience in analytics & data science consultant across Pharma, Med-Devices and CPG. Nabha leads the Trinity's Data Science Practice and has expertise in helping companies apply advanced analytics, including AI and machine learning to solve commercial challenges.

Effectively predict patient discontinuation with AI and opportunities for Rx switching

Ryan Sun, Stuart Kamin

Introduction

Background

Patient adherence is a key metric that commercial and market access teams at life sciences companies monitor and try to appropriately maximize, as it directly affects patients' therapy experience, and ultimately a specialty brand's commercial performance.

Intuitively, patients with certain characteristics or who experience certain events (e.g., with discretionary diagnoses, spotty insurance coverage, adverse effects) are more likely to discontinue therapy prematurely.

By utilizing a broad spectrum of real-time patient data, we aim to deploy AI and machine learning to forecast the risk of patient adherence, thereby facilitating timely intervention by drug manufacturers. Moreover, to validate the potency of such sophisticated analytics, we evaluate the effect of our machine learning insights and their capacity to predict patient behavior on specialty therapy with accuracy.

In addition, the unique market data serves to fill in the gaps in claims and assists in pinpointing prescription switch possibilities. Continuing the journey, exploration is underway on utilizing AI/ML for forecasting these switch opportunities with data. This mirrors the successful prediction of patient discontinuation on the manufacturer's brand.

Goal of this Investigation

The goal of this case study was to predict patient drop-off and measure engagement impact and to be able to capitalize on found opportunities to intervene and keep patients on brand. Leveraging our data set we're able to accurately predict abandonment and discontinuation. Real-time competitive insights market alerts may be an additional opportunity to leverage a new data source to better capture upstream new Rx opportunities.

Goal: Apply machine learning to predict which patients will likely discontinue treatment.

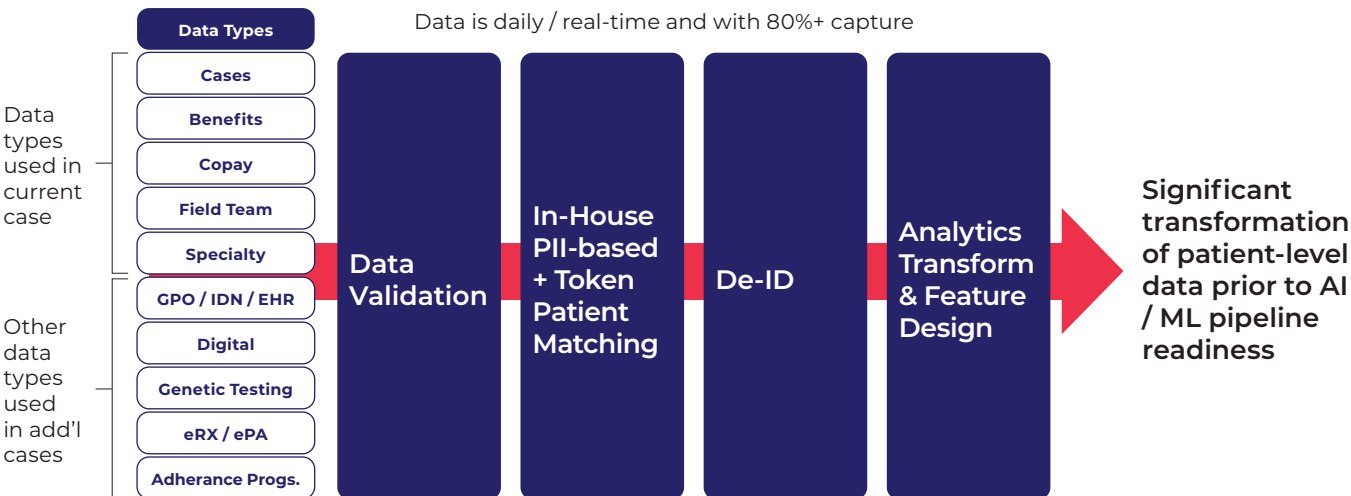
Materials and Methods:

Study Description

To understand how AI/ML can accurately predict patient drop-off enables us to identify at-risk patients and intervene sooner.

To carry out this analysis, we needed to utilize our unique data ecosystem with an impressive capture rate of 80% for brands where our AI/ML has been applied thus far.

Figure 1: Data Ecosystem



Methodology

The AI/ML methods incorporate influential factors from structured data of various dimensions, sources, and tracked longitudinally, including but not limited to:

- Demographics (age, gender)
- Payer (coverage status, payer type)
- Treatment Journey (prior adherence, prior treatment courses, diagnosis)
- Support Service (patient support calls/ email frequency, AE reporting)
- Prescriber (specialty, physician experience with product)

200+ features developed across patient treatment journey as predictive variables.

All distribution models accommodated. The system has the capability to manage an optional hub and the respective patient data coming through the many potential channels to apply predictive analytics appropriately.

No single dominant predictor: many (weak) variables in combination needed to make a stronger model.

Beyond structured data, we have leveraged natural language processing (NLP) methods to extract patient sentiment and other nuances of the patient journey from patient interaction notes. Some examples of how we’ve transformed unstructured data into structured features for

our models include:

- Clinical symptom identification: deriving the presence and timing clinical events in a manner that compliantly does impact further AE reporting requirements.
- Patient sentiment quantification: deriving the feeling of a patient and scoring the relative emotional state, positively or negatively, related to other events happening to the patient in parallel.

Of course, parsing patient notes often has privacy considerations.

Our HIPAA/HITECH-compliant and SOC 2-compliant platform enables us to compliantly manage consented PII, or unconsented PHI and appropriately access and transform unstructured data into features that augment our analytical models.

Takeaway:

Unstructured notes from healthcare providers hold a substantial predictive value upon interpretation. The potential to discover in-depth data and insights from the patterns extracted from these unstructured notes can contribute to the benefit of individual patients and the wider patient population.

Multiple years of patient’s longitudinal data are included in the model development process. Upon completion of the feature engineering to create variables in the aforementioned areas, the dataset is split into training data and test data, where training data consists of 90% of randomly selected patients, and test data consists of the remaining 10% of patients.

We performed various types of machine learning algorithms, including a gradient boosting tree, logistic regression, artificial neural networks, and support vector machines, and obtained the best model form. Hyper-parameter tuning was also performed extensively to fine tune the model. The overall robustness of the model was assessed by running the model on test data. In particular, precision and recall were calculated to support model evaluation and utilization. In the example demonstrated below, the target variable is “patient discontinuation before the first 90 days of therapy”, where 1 means “the patient discontinued therapy during the first 90 days”, and 0 means “the patient stayed on therapy 90 days or longer.”

To truly leverage the model for patient and business impact, model predictions can be pushed to care management teams to help focus the next best action around patient care. Life sciences companies can set a probability threshold based on recall and precision appropriate for the desired level of resourcing for the brand. Commonly a threshold of 50% can

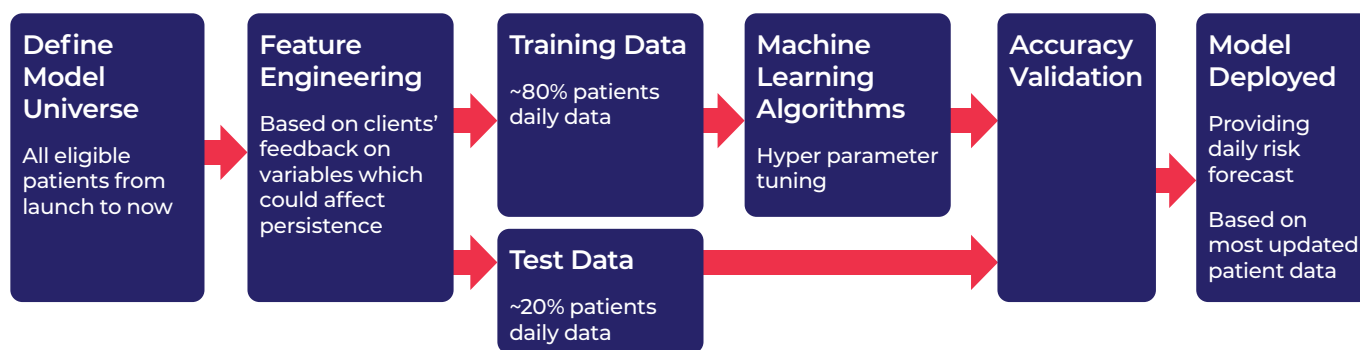
be a baseline to highlight predicted positives for care team resources, meaning any patient with a 50% or greater risk of discontinuation will be evaluated by the manufacturer staff.

It may also be of interest for teams to evaluate a “wide net” versus a “targeted intervention” approach to align model predictions with business objectives.

For example, in a “wide net” approach, manufacturers with extensive patient support staff may decide to set a lower threshold of, for example 35% rather than 50% with the ability to provide targeted interventions to more cases predicted to be at risk (since the threshold for attributing a positive risk of product drop-off is lower). This approach will suggest that most at risk patients be engaged by care management staff, even if the probability of false positives is higher.

For the targeted approach, manufacturers with limited staff may want to set a higher threshold of perhaps 75%, so that there is greater confidence that care management resources are focused on patients who are truly at risk. Below are some initial numbers from the case discussed using the above thresholds.

Figure 2: Model Development Methodology



Training data for our model was from a daily data set of patient events from Dec 20 – June ‘22. The model was then tested against patient daily data from Jul ‘22 – Dec ‘22.

Figure 3: Wide net vs. targeted approach thresholds

We looked at a population set of 2,465 patients between July 2022 and Dec 2022.

We measured precision for predicted true and false readings compared to actuals.

Prob. Threshold	Risk Category	Test Patient Count	Actual DC	Remain Active	True Positive	False Positive	True Negative	False Negative	Precision	Recall
70%	Very High	2,465	690	1,775	359	58	1,716	331	86%	52%
50%	High	2,465	690	1,775	483	170	1,605	207	74%	70%
30%	Medium	2,465	690	1,775	580	580	1,195	110	50%	84%

Evaluation

Leveraging influencing factors of various dimensions, sources, and tracked longitudinally, our machine learning/AI model correctly predicted **70%** discontinuing patients **at least 5 days or earlier** before patient discontinuation, with precision = 74%, recall = 70% @ probability threshold = 50%.

Case Study: Personalized access risk prevention at scale

Background: In 2021 we implemented a discontinuation prediction model for patients on a rare disease therapy, to help the manufacturer identify patients at risk.

In Q2 '22, a patient advocate user of our model found a long-time patient that had moved into an escalated risk of discontinuation. The advocate did not have a relationship with the patient directly, but notified the sales rep that they may want to speak with the patient's doctor (the patient had signed a HIPAA waiver). The sales rep spoke to the doctor who had no idea that the patient stopped their medication.

The doctor learned that while he was on vacation, the patient had been taken off drugs by an

inexperienced HCP due to a minor tolerability issue noted by the patient. The doctor called the patient and restarted treatment right away, resuming a therapy that had provided long term control of the patient's disease.

Takeaway:

From the patient advocate's point of view, this patient would have fallen through the cracks without our predictive model.

Conclusion

The AI predictions and content recommendation engine utilize numerous influencing factors from various data sources. This technology provides accurate predictions of drop-off rates and aids in determining optimal strategies for minimizing patient risk of discontinuing treatment.

Additional Opportunity

Exploring the potential of using AI and machine learning to not only forecast patient adherence, but also to anticipate new-to-brand Rx transitions using more than just standard claim data. This allows for swift action, within days rather than weeks after an occurrence. Comprehensive insights into when and why a patient discontinues treatment enables teams to engage with prescribers about treatment options ahead of the competition.

Our novel, pre-claim data can help teams predict the right time to engage HCPs around adoption opportunities on a daily basis, and with visibility around why patients drop off alternative treatment. See Figure 4.

Supplementing gaps in claims and identifying Rx switching opportunities can also be accomplished with this unique market data, in addition to improving patient adherence through the application of AI/machine learning.

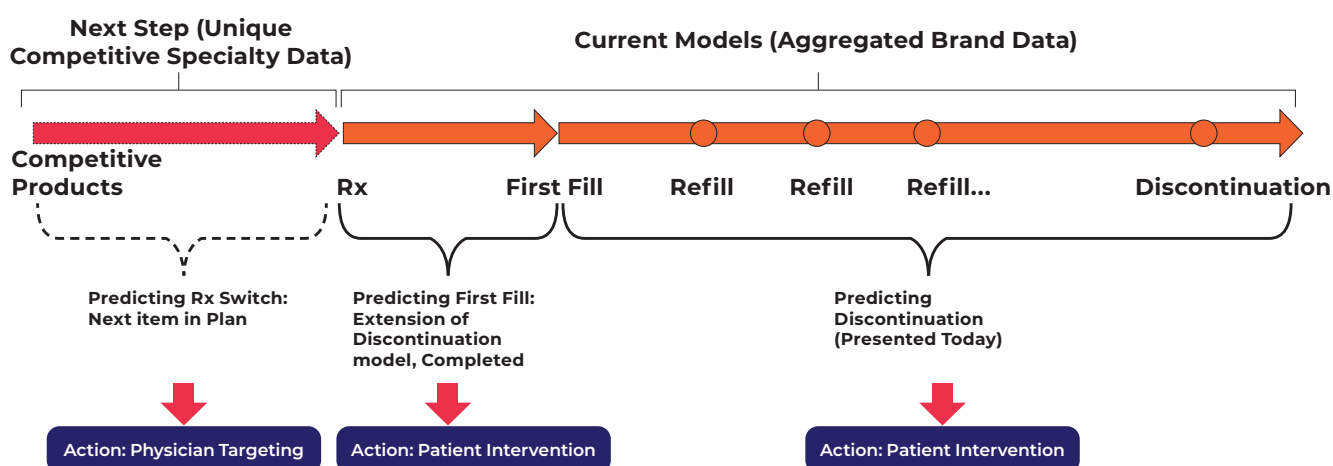
Figure 4: Claritas Rx Market Data

Objective: predict daily when patients will drop off competitor brands using novel specialty market data set

Novel Claritas Rx Data Set	
Daily Data Across Specialty Brands	- Traditional Rx claims have 2-8 weeks of lag from patient event to analytics availability date - Claritas Rx data is on a 1-5 day lag
Visibility From Rx to Discontinuation	- Visibility into pre-dispense milestones - Visibility into Rx abandonment and treatment discontinuation dates and reasons - Payer controls, out-of-pocket, and other access barriers
Can be Integrated with Claims for Max Breadth	Significant (~95%) overlap between Claritas Rx specialty data and best practice Rx claims

Significantly larger patient volumes available for predictive analytics than brand case above

Figure 5: Predicting Rx Switch - Increase the “new patient funnel” by using our daily specialty data



Key Terms used:

Abandonment includes patients who fail to fill a prescription after it has been sent to the pharmacy.

Discontinuation occurs when patients begin the therapy and then stop. Both types of drop-offs have a variety of different factors and causes.



PMSA

PHARMACEUTICAL MANAGEMENT
SCIENCE ASSOCIATION

www.pmsa.org