JPMSA

JOURNAL OF THE PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

SPRING 2023 | IN THIS ISSUE:

pmsa
PHARMACEUTICAL MANAGEMENT SCIENCE ASSOCIATION

Table of Contents

Official Publication of the Pharmaceutical Management Science Association (PMSA)

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

The key points in achieving this mission are:

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

Please submit correspondence to:
Pharmaceutical Management Science Association
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

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

PMSA Journal: Spotlighting Analytics Research

After a brief intermission due to COVID, PMSA is pleased to announce the return of the *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

Data Curation – Identifying and Fixing Mistakes in the Data

JP Tsang, PhD and MBA (INSEAD), President, Bayser Consulting

Abstract: Data curation has never been more important given the explosive blossoming of data sources our industry is witnessing. It is indeed our only reliable defense against drawing erroneous insights and formulating ill-advised recommendations.

Yet, data curation has received short shrift treatment. For two reasons—first, it requires deep and broad knowledge about the disease, the therapy, data entry and collection, and the like. Second, we tend to assume that the high price these data sources command must mean that they have been fully vetted and curated and ready to use out of the box.

This article attempts to redress the situation. It starts off by explaining why data curation cannot be ignored. It then provides multiple examples of how data curation operates using compelling examples drawn from years of experience on the job. It finally concludes by discussing lessons and takeaways for analysts that want to level up their game.

1. What's Data Curation

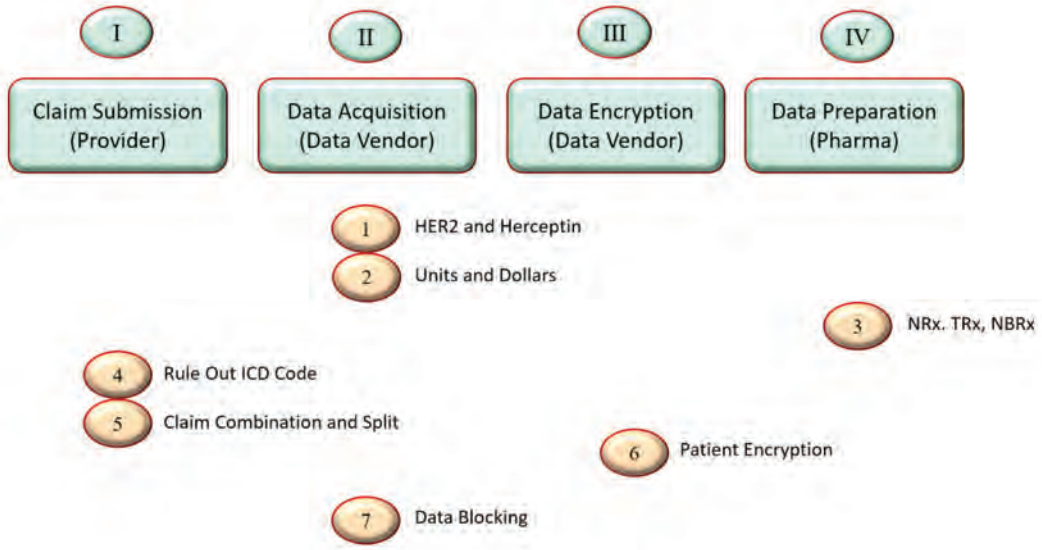
Data curation is about identifying issues with the data and fixing them.

Ideally, all data curation has been carried out by the data vendor and is ready for analysis once in the hands of the analyst. Unfortunately, that's rarely the case and for good reasons. Oftentimes, we can only realize there is a problem with the data when the analysis we ran points to a result that cannot be. This means that the data vendor would have to run a very large number of analyses of all kinds and assess the validity of the results as part of the data curation process and that's simply unrealistic. It is for this reason that data curation ends up being the task of the analyst. Faced with counterintuitive or nonsensical results, the analyst has no choice but to embark on a troubleshooting journey to find out what's going on with the data.

There are 3 common outcomes. First, the analyst understands where the problem comes from and can fix the problem. It is local in nature and only involves a small number of records or a couple of fields. Second, the analyst understands where the problem comes from but does not have a fix. That's for instance when a big chunk of the geography goes missing or longitudinal holes puncture the patient journey following no apparent patterns. The analyst makes a mental note of the problem and accounts for it as a caveat when interpreting the results. Third, the analyst is stumped and cannot locate the problem and the wild goose chase is on. The data curation is a protracted work in progress.

In yet other cases, the analyst does not even realize there is a problem with the data. The finding is plausible and plausible comes across as right even when wrong.

Figure 1: Phases of Data Lifecycle Where Data Issues Emerge



We'll see that curation starts with recognizing that what the data implies is wrong and this is accomplished by leveraging knowledge regarding the disease, the therapy, and data entry and collection to mention just these three. In other words, without this knowledge, there is not much data curation that we can do. Indeed, deep and broad knowledge is essential and this cannot be overstated.

2. Why We Should Care

There are two reasons why we should be obsessed with data curation as analysts.

Avoid drawing flawed conclusions – if the data is flawed to start with, odds are the answers and insights we'll get from the analysis will be off. We'll recommend ill-advised interventions and that's something we cannot tolerate. The messenger deserves to be shot. Indeed, the job of the analyst is not only to run good analyses but also to ensure that the data is good for the job. Can a cook get away with claiming that their cooking is great even though the food tastes awful because the ingredients turned bad?

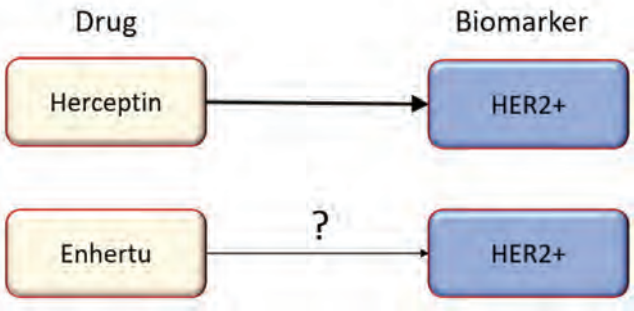
Increase usefulness of the data. By dropping bad records, filling in missing data, resolving inconsistencies using thoughtful business rules, and issuing caveats for the analyst downstream, we make the data more relevant to the organization by enabling analysts to unlock insights that would have otherwise been trapped in the data if it were not curated. See Figure 1.

3. Examples of Data Curation

Curation starts with identifying issues in the data and these issues originate from various steps in the lifecycle of the data. We'll distinguish four of them.

1. Claim Submission – That's when the provider files the claim to the payer.
2. Data Acquisition – That's when the data vendor acquires the data. In the PLD world, that's essentially the pharmacy, the clearinghouse, the payer, and direct feeds from labs, IDNs, GPOs and the like.
3. Data Encryption – That's when the data vendor encrypts the patient in keeping with HIPAA regulation.

Figure 2: Patient is HER2+ if on Herceptin



4. Data Preparation – That's when the data is at the pharma client and is being prepared for analysis to answer business questions.

In what follows, we'll present seven examples of data curation along with the underlying principles. As for the order of the examples, we chose to follow not the steps of the data lifecycle above but rather the intuitiveness of the example starting with the most immediate ones.

3.1 HER2 and Herceptin (Example 1)

Principle – A patient can only get Herceptin if the patient is HER2+.

Example – The data does not report the HER2 status of a patient but indicates that the patient uses Herceptin. We can safely infer that the patient is HER2+.

Explanation – This inference relies on the fact that a patient cannot not be HER2+ and yet have Herceptin. See Figure 2.

Comments

1. This type of inference holds for any drugs that have a mandatory companion diagnostic. If the patient uses the drug, we can be certain that the patient tested positive for the companion test.
2. We may not be able to hold this kind of reasoning for Enhertu, for instance, as Enhertu is indicated for HER2 Low patients. Indeed, the patient could be HER2 Low.
3. In most cases, inferences are not that cut and dried. They work most of the time. Nonetheless, we'll use these inferences knowing all well we may be wrong sometimes.
4. The more we curate the data through such inferences, the further away we may be straying from the truth. So, curation is to be used judiciously.

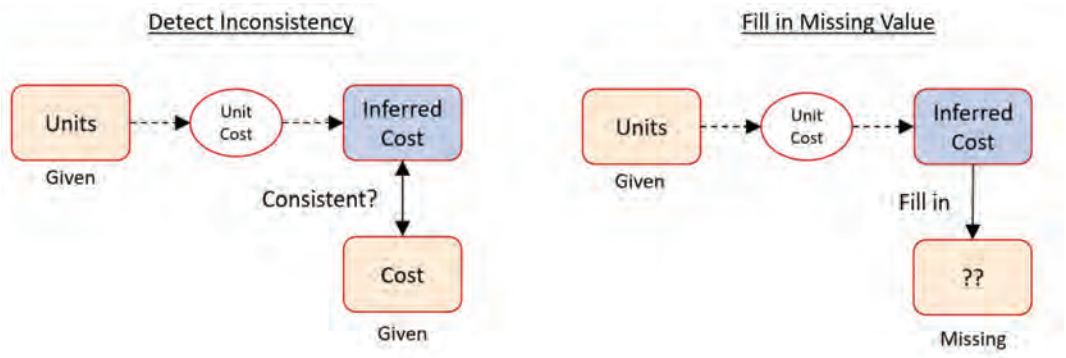
3.2 Units and Cost (Example 2)

Principle – Units and Cost are closely connected to each other in that we can infer Cost from Units and vice versa, given unit cost. Cost (or Units for that matter) can be regarded as having two values: one is stated in the data which we'll refer to as Cost and the other can be inferred from Units, which we'll call f(Units) where f multiplies Units by unit cost. Now, f(Units) and Cost need to be the same, otherwise there is a problem. See Figure 3.

Example 1 – The data indicates Units but not Cost or vice versa. In either case, we can use the unit cost to infer the missing value.

Example 2 – The data indicates Units and Cost but they don't agree with each other given the unit cost of the drug. In that case, we summon a business rule that

Figure 3: Given Cost, we can fill in the value for Units if missing or detect inconsistency if Units are given



accomplishes two things. First, decide which one to believe – Units or Cost – and which one to throw away. Second, how to update the value we threw away. If Cost is missing, we could use Units * unit cost.

Explanation – This reasoning takes advantage of the redundancy in the data. Indeed, when the data gives the Cost, it also indirectly gives the Units and when it gives the Units, it also indirectly gives the Cost.

Comments

1. This reasoning works well so long as we are using the right unit cost, which as we know may fluctuate over time and may depend on the Payer (Medicare, Medicaid, or Commercial).
2. We simplified the example on purpose. Reality is more complex as there may be different versions of Units (paid, billed, etc.) and Cost (submitted, allowed, paid, etc.) and they are connected to each other in a very specific way which allows us to spot inconsistencies and formulate thoughtful business rules to resolve these inconsistencies.

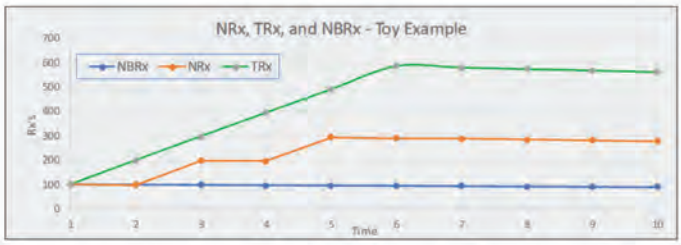
3.3 NRx, TRx, and NBRx (Example 3)

Principle – NRx, and TRx, and NBRx are connected to each other. When NRx drops, TRx also drops as the refill rate tends to be more or less constant. The point here is TRx cannot grow. When NBRx drops, can TRx grow? The answer is yes, because there are new patients getting on the drug even though the trend is downward. So, TRx can grow.

Example – The data says that NRx is dropping but TRx is growing. We know that's impossible. What's going on?

Explanation – What's happening instead is this: NBRx is dropping and TRx is growing. NBRx is the number of new patients on the drug. If NBRx is 0, no new patients are getting on the drug and NRx will not grow. But if NBRx is not 0, NRx may increase as there are new patients getting on the drug and TRx as well, assuming no dramatic drop in the refill rate. In short, a positive NBRx although on the decline is consistent with a growing TRx. And a declining NRx is consistent with a declining TRx, not a growing TRx. See Figure 4.

Figure 4: Toy example to show that TRx may grow while NBRx drops



Time	1	2	3	4	5	6	7	8	9	10
Pat - New	100	99	98	97	96	95	94	93	92	91
1st NRx	100	99	98	97	96	95	94	93	92	91
Refill using 1st NRx		100	99	98	97	96	95	94	93	92
2nd NRx			100	99	98	97	96	95	94	93
Refill using 2nd NRx				100	99	98	97	96	95	94
3rd NRx					100	99	98	97	96	95
Refill using 1st NRx						100	99	98	97	96
NRx	100	99	198	196	294	291	288	285	282	279
TRx	100	199	297	394	490	585	579	573	567	561

Assumptions

1	An Rx is good for one month.
2	The same Rx is used twice before a new Rx is needed.
3	Each patient fills 6 times, 3 are NRx's and 3 refills of the NRx.

Comments

1. Had we not known the relationship between NRx, and TRx, and NBRx we would not have recognized the problem, let alone know where to look to troubleshoot the problem. Now, think of all the relationships that exist between variables that we are not aware of or are only vaguely aware of but not to the point to make actionable decisions. In short, our ability to identify issues with the data may be limited.
2. Not all situations have a happy ending like this one. In many cases, we fail to spot the problem and we keep on using the data unaware that the ensuing answers are not quite right.

3.4 Rule-Out Diagnosis (Example 4)

Principle – The ICD code of the patient may not be the diagnosis the physician is thinking of. Instead, it may be an administrative device meant to satisfy the payer when the payer insists on the presence of an ICD code to pay for the lab test the physician orders.

Example – The patient is treated all along for esophagus issues and the data shows a one-time diagnosis of gastric along with a lab test.

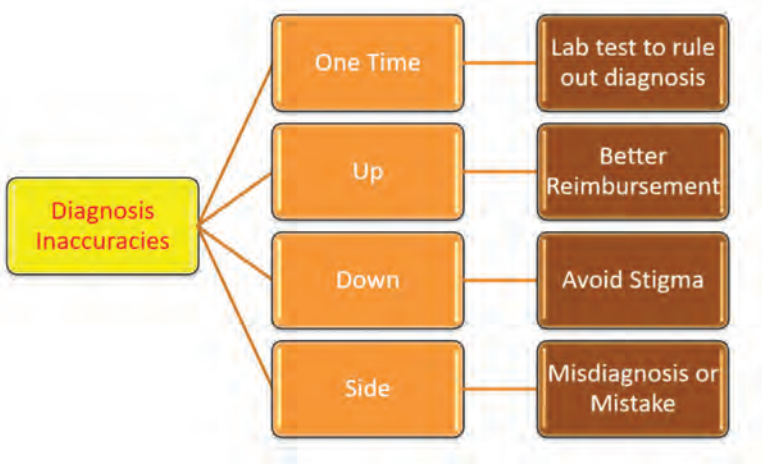
Example 2 – The patient has a one-time diagnosis of hypothyroidism and undergoes a TSH test. The data indicates no treatment of hypothyroidism.

Explanation – The physician here is simply playing along to ensure that the lab test is performed and reimbursed by the payer.

Comments

1. The presence of an ICD code should not be taken to mean diagnosis, especially when it appears only once. It may be meant for the Payer to ensure reimbursement of the corresponding lab tests.
2. That's the reason why a popular business rule used to establish the cohort of relevant patients for a market basket is to require the presence of two ICD codes on different dates. Only one ICD code may be a rule-out diagnosis.

Figure 5: Instances where ICD codes in the data may not reflect true diagnosis



3. More generally, one should be cautious when equating ICD code with diagnosis. Indeed, providers routinely practice up coding (to get a larger reimbursement), down coding (to avoid the stigma of, say, schizophrenia and entering depression instead), and side coding (genuine mistake or wrong diagnosis). See Figure 5.

3.5 Combined Claims (Example 5A)

Principle – There is a limit as to how much of a drug a patient can have. If the data reports the patient is getting more than the threshold, there is something wrong with the data.

Example – The data says that a patient gets an injection in one day in an amount that should kill the patient. Not only does the patient not die, the patient comes back for more the following month. Interestingly, the patient gets the injection only once a month when the patient should be getting 4 injections a week.

Explanation – The physician is grouping the four claims of the patient into one. The physician finds it more convenient to file the claim once a month as opposed four times a month.

Comments

1. No fix is required so long as the analysis does not require a temporal granularity of less than a month. Otherwise, the analyst would have to split the claims and spread them over the month.
2. This example is related to another example we'll discuss later where multiple claims of a patient come from the encryption engine that maps different patients onto the same patient id.

3.6 Split Claims (Example 5B)

Principle – One injection at the physician's office should trigger only one claim, not two.

Example – The data repeatedly reports two claims for the same patient on the same day at the same physician office and not one claim as we would expect.

Explanation – The data capture system assumes that the amount the provider will charge will never exceed a certain amount and the corresponding number of digits to enter the amount is simply insufficient. The workaround is simple. The provider issues to two claims that add up to the requisite amount.

Figure 6: Claims may be combined or split as they are being reported

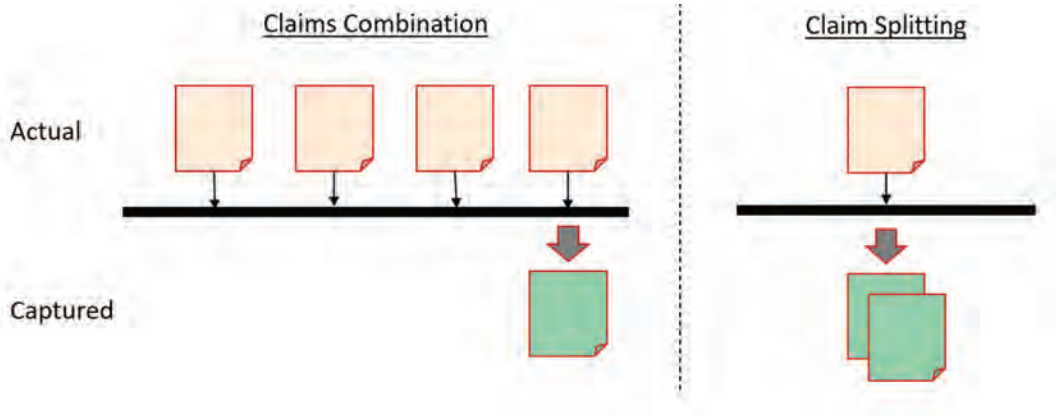
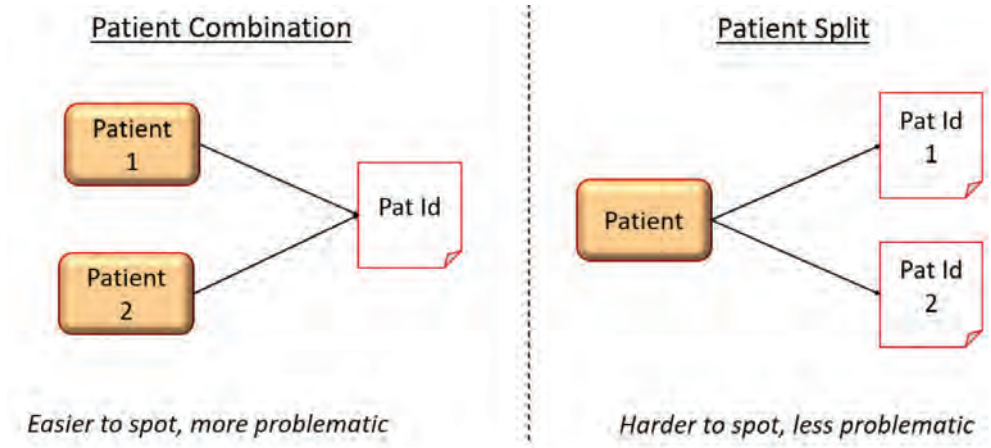


Figure 7: Encryption may map 2 patients onto 1 id and split 1 patient over 2 ids



Comments

1. If we count total number of claims assuming that each claim corresponds to, say, an injection, we'll be overstating the number of injections that were administered.
2. The fix is straightforward: Go through the database and replace the two claims by one claim where the charged amount is the sum of the charged amounts of the two claims. See Figure 6.

3.7 2 Patient Ids for 1 Patient (Example 6A)

Principle – The encryption engine is supposed to map each patient onto a distinct patient id, so one patient id corresponds to one patient.

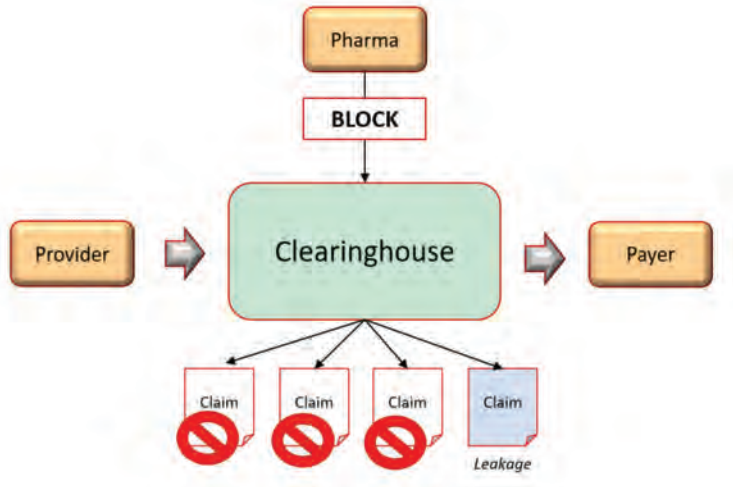
Example – A patient fills too many Rx's over a given period of time, and this cannot be accounted for even if the patient were to fill early.

Explanation – The encryption engine maps two different patients onto the same patient id. That happens when the PII on patients is so scant that the encryption engine confounds them. See Figure 7.

Comments

1. It is theoretically possible to untangle the healthcare interactions and generate two patient ids but the odds of getting it right are slim.

Figure 8: Claims leak through the Clearinghouse although the drug is blocked



2. The best fix is to drop these records.
3. This is related to an example we discussed earlier where the combined claims come from the same patient. In this case, they come from different patients.

3.8 Two Patients for one Patient Id (Example 6B)

Principle – The encryption engine is supposed to map each patient onto a distinct patient id, so one patient id corresponds to one patient.

Example – A patient suddenly disappears and at the same time another appears and the activities of the two patients dovetail with one another. In some cases, the two patients fall under the same physician.

Explanation – There are two well-known cases where that happens. First, when a woman gets married, changes name and relocates. The new PII throws the encryption engine off. Second, when snowbirds take off for a warmer place to get away from the harsh winter. Unlike the woman that disappears for good, the snowbird reappears under the previous id when the temperature warms up.

Comments

1. Even when we are convinced that we found two patient ids that correspond to the same patient based on their activities, it is difficult to prove this is the case. Fluke is always a possibility.
2. The implications are twofold. One, the data suggests that patients stay on the drug shorter than what they do. Second, there are less patients that have dropped therapy than what the data suggests.

3.9 Data Blocking (Example 7)

Principle – In the world of SP drugs, a pharma company has the choice between releasing the data on their drug to other pharma companies or blocking the data.

Example – The data reports very feeble sales for the newly launched drug even after accounting for the fact that the capture rate of the data vendor is far from perfect.

Explanation – The pharma company has issued a block on the data of the newly launched drug. Interestingly, clearinghouses are well known not to fully comply with the request of the pharma company and many claims trickle through. See Figure 8.

Comments

1. This is a case where we can only recognize the problem. There is no fix. The best we can do is to warn the analyst about the caveat.
2. What's insidious about blocking is that it is subject to leakage. It is very tempting to conclude that since we can see a few claims of a drug, we are seeing the bulk of the claims. And that would be wrong.

4. Lessons and Takeaways

1. Realizing that there is a problem with the data is no walk in the park. It's often after we are done with the analysis that we realize there is something wrong with the data. What if the finding is wrong but plausible? Plausible does not mean correct. Counterintuitive does not wrong either.
2. The key to identifying issues with the data is to be knowledgeable regarding the disease, the therapy, data entry and capture, billing and reimbursement, and the like. This not only helps us identify issues but also helps us establish thoughtful business rules to resolve inconsistencies.
3. Curation is best viewed as an enhancement to the data. It increases the usefulness of the data but cannot fix fundamental flaws such as gaping geographic and longitudinal holes.
4. Data Curation is not the panacea. It has its limitations. If the data is deeply flawed, data curation is of little use. Sometimes, the best and only thing we can do is to issue caveats for the analyst to heed while interpreting the results of their analyses.
5. Data curation is to be used judiciously. That's because the business rules are inherently probabilistic in that they are mostly right and sometimes wrong. The more we curate the data, the more likely the curated data drifts away from what it is supposed to capture.
6. Each business question applies "stress" to different parts of the database. Indeed, the data may need to be curated to answer one question but not another. It depends on the question.

5. Conclusion

In sum, it all starts with identifying the problem. Ignorance is not always bliss as we may be unwittingly drawing wrong insights and recommending wrong interventions. What's plausible may be wrong and what's counterintuitive may be right.

There are problems that we can fix and this is the domain of curation. As for those we cannot, the best thing to do is to issue caveats for the analyst downstream.

Analysis and curation share a lot in common. Curation can be viewed as one particular type of analysis where the object of the analysis is the data source itself and the objective is to ensure that the data source is fit for the job. In certain ways, data curation goes further as it recommends interventions such as records to drop, business rules to establish and implement, missing values that can be used to fill in the blanks, and caveats to articulate for the analyst downstream to heed when answering business questions.

When the result of the analysis is a value of a field, we call this curation. However, when it is more general as in: "How many eligible patients?", "What's the market share?",

"How fast do patients go through second line?", we call this analysis.

Finally, powerful data curation can only exist if we have deep and broad knowledge regarding the disease, the therapy, data entry and collection, billing and reimbursement, and the like. Unfortunately, there is no way around this.

6. Acknowledgments

I would like to thank all of you for the insightful conversations we had that one way or another allowed us to connect the dots and spark new ideas. In particular, I would like to thank Erle Davis with Inovalon, Ji Xiao Hao with Kyowa Kirin, Rohit Marzah with Definitive Healthcare and Marc Duey who used to run Prometrics until recently. And all of you who generously shared your data curation ideas and experience with me. I am indebted to all of you.

About the Author

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

References

Data Curation: What You Need to Know to Shine", JP Tsang and Erle Davis, VP of Client Solutions, OncoHealth, PMSA Symposium 2022, Las Vegas, Oct 27-28, 2022.

"Pharma Data As Asset – Moving Roche from an Application Centric to an Information Centric Organization", Martin Romacker, Roche Innovation Center Basel, Elsevier Webinar, July 8, 2020.

"Big Data Curation", Andre Freitas and Edward Curry, New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe, Springer Berlin/Heidelberg, Jan 2015.

"Merging data curation and machine learning to improve nanomedicines", Chen et al., *Advanced Drug Delivery Reviews*, Elsevier, Vol 183, April 2022.

"How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries", Johnston et al., *Journal of Librarianship and Scholarly Communication*, Vol6. 8, General Issue (2018).

The Evolution of Pharma Field Force Deployment and Targeting

Ashvin Bhogendra, Senior Director, Axtria; Abhilash Sain, Senior Director, Axtria; Anjali Attri, Associate Director, Axtria; Monal Tenguria, Manager, Axtria

Abstract: Historically, pharma organizations adjusted their commercial model to accommodate the shift toward specialty portfolios, reduced physical access to HCPs, and the increased complexity of sales roles. The recent COVID-19 pandemic has further accelerated digital promotion and the reduction of personal promotion roles. This article explores how field deployment and targeting approaches are evolving to address these market dynamics and provide superior customer engagement outcomes. We explore two areas where field force deployment is changing: 1) How field deployments are becoming more customer-centric by including hybrid approaches that facilitate better collaboration across roles. 2) How targeting approaches have evolved from static cycle planning to dynamic, multi-channel call planning supported by frequent AI/ML-driven insights that drive high-value actions beyond the call plan. We differentiate these methods across specialty, oncology, and rare disease-focused teams and retail teams. We also examine how dynamic channel scores can ensure effective coordination, channel mix, and messaging over time. Finally, we break the journey to omnichannel transformation into simple steps that pharma organizations can implement easily.

Background

The pharma marketplace has changed rapidly over the past few years, and traditional face-to-face (F2F) meetings between sales reps and physicians no longer fill the needs of modern pharma organizations. The rapid expansion of specialty drug portfolios, reduced physical access to health professionals, and changes in the structure of healthcare organizations have reshaped the landscape of pharma sales – and sales roles are changing, too. Several specialized field roles have evolved to focus on the different types of customers involved in buying pharma products, like specialist physicians, primary care physicians, hospitals, and integrated delivery networks (IDNs). In the past, these roles operated in silos, resulting in potential leakages at each stage. Pharma companies now orchestrate these functions to reduce leakage and provide a better customer experience. Additionally, pharma companies are gradually and carefully bringing sales and marketing teams together to bring an omnichannel experience to customers.

The recent COVID-19 pandemic further accelerated some of these shifts, specifically the following two trends:

1. ***Increased Digital Promotion***: During the pandemic, life sciences organizations' digital promotion spend grew to five times what it was before the pandemic. Marketing mix benchmarking studies show that, compared to personal channels, digital promotion has a better return on investment (ROI) for launch and mature brands but produces a lower impact on overall sales. Despite accelerating digital adoption, the sales force still represents 80% of the non-direct-to-consumer promotional spend across pharma organizations.
2. ***Reduction of Personal Promotion Roles:*** Because healthcare physicians (HCPs) want fewer (F2F) interactions since the pandemic, some organizations are reducing their pure F2F sales

promotional roles and optimizing their coordination with other roles, such as medical science liaisons (MSLs), reimbursement specialists, and nurse educators.

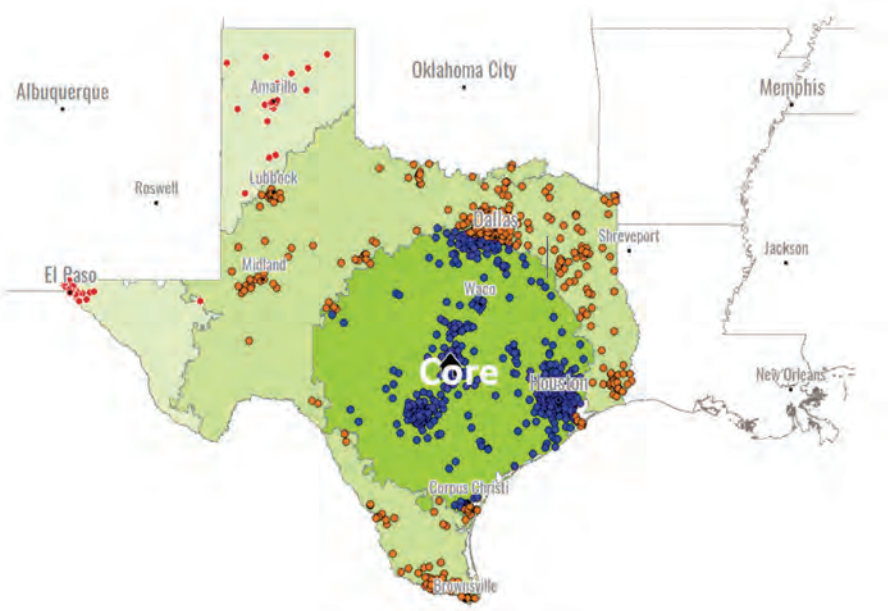
Now, we will explore how organizations are modifying their field deployment and targeting approaches to address these market dynamics and provide superior customer engagement outcomes.

Evolution of Field Force Deployment

Field force deployments and customer alignment models have changed little over the years, with ZIP/brick-to-territory or customer-to-territory mappings employed to define territory alignments. Organizations have explored different commercial models, such as base territory, overlay, mirrored geographies, and differential resourcing, to account for varied portfolios and localized resourcing needs. But they are starting to embrace new approaches to overcome access issues, provider consolidation, and the need for coordinated customer engagement. Some of the most promising are described below:

1. ***Alignment Design That Includes Customer Access and Virtual Engagements:*** With the increase in HCP and account access restrictions on both frequency and total meeting time, along with the increased availability of virtual engagement channels, organizations are incorporating more realistic rep workloads into alignment design. Collecting access segments and HCP channel preferences from industry benchmarks, activity data analytics, and rep feedback helps determine the effective workload of a geography across channels. Using this enhanced workload index to adjust territory boundaries provides better multichannel coverage for customers.
2. ***The Move to Customer-Centric Alignment:*** Traditionally, the alignments for various field roles (rep, MSL, key account manager [KAM], access specialist, etc.) were independently created and managed. These silos led to coordination issues and customer engagement challenges. In the new omnichannel paradigm, we see organizations moving toward customer-centric alignments where the portfolio leaders are accountable for a holistic customer experience. The customer hierarchy and influence networks are clearly identified and defined as ecosystems. The alignments of all field roles are designed to be in sync to ensure clear customer ownership and optimal collaboration. Single product/indication teams within the same business unit are mirrored at the territory or first-line manager level to ease coordination among reps. This technique helps communicate well-coordinated messaging for overlapping targets and leads to superior customer engagement.
3. ***Hybrid Territories:*** With the increase of virtual engagement channels, organizations are exploring hybrid territories that combine F2F and virtual contact. A smaller, defined geography, generally a metropolitan area where a rep will likely be hired, is the "core" where the reps focus on F2F interaction. Cores are surrounded by extended geographies where the reps primarily leverage virtual engagement channels and use F2F follow-ups as necessary. These geographies are generally designed in concentric circles for ease of alignment maintenance and to provide the flexibility to reach out to physicians in-person based on physician preference in the extended geographies. (Figure 1)

Figure 1: Hybrid Territories



4. ***Quarterback Field Role:*** Some organizations are developing "quarterback" roles for the field. These reps become the central point of contact for customers within a more extensive health system, leading and coaching reps as they engage with customers while also helping coordinate across other roles, such as MSLs, KAMs, and access managers. Quarterbacks act as leads who direct reps on how to engage with a customer, but they may or may not be managers.
5. ***Lower Span of Control (SOC):*** The current need for focused planning and coordination among roles is leading to a lower SOC for field managers. The historical range for average SOC in specialty roles was 8 to 10 reps, which has recently dropped to 6 to 8 reps. Today's field managers have extra responsibilities that require them to wear multiple hats rather than being only leaders and coaches for field reps. Some of their new responsibilities include managing relationships within their healthcare ecosystems and ensuring effective cross-functional coordination to meet their localized goals. Organizations are moving toward customer-centric commercial models, making the field manager's role very strategic. They are responsible for providing the best customer experience possible.

Targeting approaches have evolved from static cycle planning to dynamic, multichannel call planning supported by frequent AI/ML-driven insights that drive high-value actions beyond the call plan. Because retail teams and specialty, oncology, and rare therapy-focused teams have specific needs, we approach them differently, as described below.

Most retail organizations have shifted or are in the process of turning from a traditional F2F call plan to a multichannel call plan (MCCP) that ensures planning is aligned with customers' channel preferences. These plans, which also help navigate the post-COVID

reduction in F2F access, are called activity plans to reflect all the actions undertaken by reps rather than only their F2F interactions.

Retail organizations currently create an initial F2F activity plan and allow the field reps to refine these plans across all channels (F2F, remote, phone, email, etc.). Some organizations are exploring new ways to leverage historical call activity data, predictive modeling results, customers' channel preferences, and other pertinent data. This approach provides a channel-level, optimized frequency for targets that field reps can further refine.

However, the specialty, oncology, and rare field is highly complex. Its multiple customer stakeholders and the field roles required to support these customer archetypes make the traditional frequency-based call plan used by retail teams inefficient. Consequently, these teams have historically relied solely on target lists. There has recently been a shift toward HCP target lists in addition to healthcare organization (HCO) lists. Still, these are based on prioritization rules like the contribution of specific HCPs to the brand or market and other business rules.

Some large pharma organizations have invested in rules-based triggers that leverage patient-level data and AI/ML next best action (NBA) capability to provide high-value insights to the field force for both retail and specialty, rare, and oncology teams. Organizations are also taking new approaches like the ones below to improve their targeting strategies:

1. ***Always-On Field Refinement:*** Historically, the field force has had a two-to-three-week window of opportunity to review, refine, and finalize their call plans for the quarter. However, sometimes, a static call plan created before the new cycle begins does not capture changing market dynamics and becomes ineffective as the quarter progresses. One alternative some organizations have adopted allows the field force to provide continuous feedback throughout the planning cycle. This option allows more flexibility when reacting to unexpected market events. Including appropriate guardrails in this process helps prevent large deviations from the overall brand promotion strategy.
2. ***Dynamic Planning:*** Pharma organizations are also upgrading their planning processes to help them react to market dynamics quickly, making them more agile and responsive. Customer engagement plans built on long-term historical data are regularly augmented with recent activity and performance data allowing field teams to respond to current market trends and re-plan for the cycle. Dynamic planning processes differ by the type of teams using them. (Figure 2)

As a first step, strategic objectives and long-term historical data help determine a multichannel customer engagement plan for the entire planning cycle. This planning process can be done quarterly or semi-annually.

The plan is refreshed monthly based on field activity, rep feedback, and customer behavior. This "mini cycle engagement plan" will be closely aligned with the overall cycle plan but may identify specific, high ROI targets that become important for the field to cover before the end of the cycle.

In addition, every week, AI/ML models are run on the latest customer data to gain insights and appropriate actions for orchestrating the best customer omnichannel experience. This exercise produces a subset of very high-value activities for the rep to consider.

Figure 2: Dynamic Planning For Retail Teams

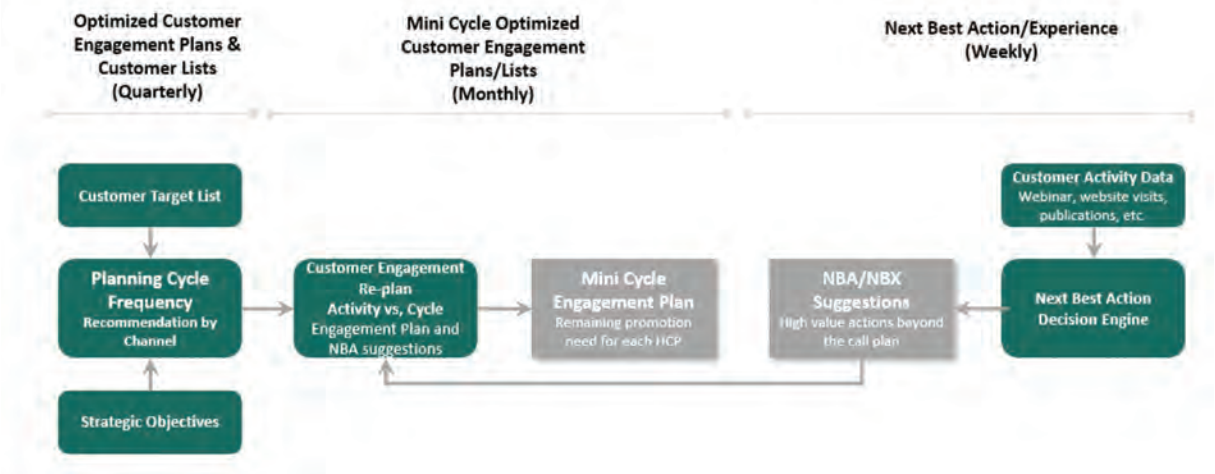
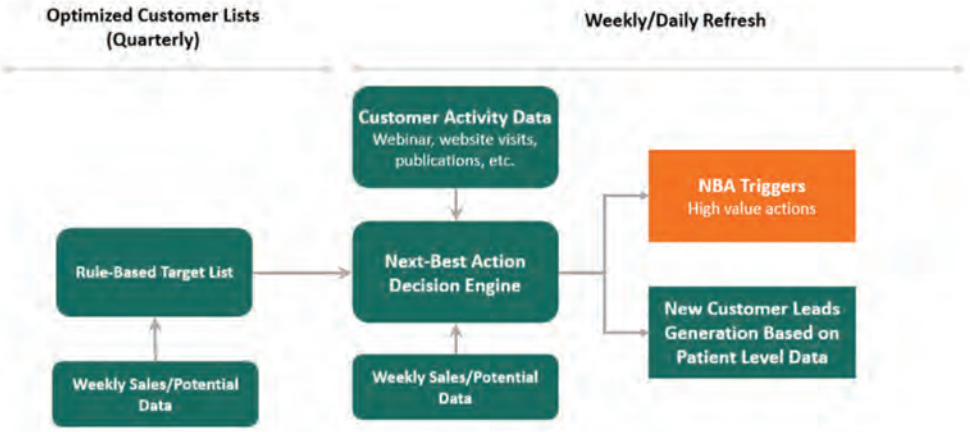


Figure 3: Dynamic Targeting Process For Specialty/Oncology/Rare Disease Teams



ML-based process workflow includes the identification of underperforming HCPs by capturing the gap between predicted sales and actual sales achieved. Such HCPs may indicate decreased writing behavior for the corresponding brand. Business rules and ML techniques, including clustering and regression analysis, are leveraged in the above workflow to predict HCP sales and identify each channel's relative importance. ML models are created based on historical sales and promotions across channels. Rep feedback is collected on these insights and used to improve the AI/ML models.

Implementing this model of a more dynamic and robust plan provides field teams with relevant and timely intelligence that can drive superior customer outcomes. (Figure 3)

In addition to more frequent AI/ML-driven NBA triggers and new customer leads based on patient-level data analysis, specialty, oncology, and rare teams continue to develop quarterly or semi-annual HCO/HCP target lists.

Figure 4: Dynamic Channel Scores

Dynamic Channel Scores: In addition to the HCP's segment and profile, organizations can generate dynamic scores for each interaction channel at the HCP level. These scores are driven primarily by digital behavior, prescribing activity, and the organization's promotional activity, as shown in Figure 4.

In the example above, Dr. Maria's F2F call score is initially high, suggesting that an office visit would be the NBA. Once the sales rep completes the F2F call, Dr. Gonzalez's score for F2F interaction decreases, and the score for email increases, triggering a marketing email. The F2F and email scores now decline, and the digital channel score increases. If Dr. Gonzalez visits the portal, the scores for F2F and email increase, triggering a follow-up call and email. These dynamic scores feed all the customer engagement platforms, enabling effective coordination, channel mix, and messages over time.

Steps Toward Omnichannel Orchestration

While the goal for any pharma company may be a fully omnichannel sales operation, it is possible to break the journey into more easily attainable steps. One option is to explore new and innovative field force deployment models as pilots at the sub-national level, assess feasibility and impact, and then launch nationally. Rather than striving for a "big bang" transformation of the targeting approach, try exploring a more flexible model to ensure that overall strategic objectives are successful and that enough time is built in to integrate what you learn along the way. The table in Figure 5 shows the steps toward full omnichannel orchestration.

Case Studies

Case Study 1: Enabled Dynamic Multichannel Call Planning

Recently, a neurology-based pharma organization re-engineered its deployment models by enabling dynamic multichannel call planning for optimized execution. The client needed to restructure their multichannel promotion strategy to meet evolving market conditions for one of their largest sales forces. The major challenges during the process included dynamic alignment with more than 15% of territories vacant, rapidly changing situations triggered by the COVID-19 pandemic, with physician access and channel preferences impacting reach and frequency. Limited access to high-value physicians for F2F rep interaction in light of COVID-19 restrictions led to a drop

Figure 5: Steps Toward Omnichannel Orchestration

Level 1: Traditional - Slow to Change for Field Teams	• Cycle-based static call plans • Territory target lists
Level 2: Always-on Intelligent	• Enable multichannel planning • Always-on field refinement • Rule-based insights
Level 3: Dynamic Scoring and Planning	• Dynamic planning – allows more frequent adjustments to the plan • Dynamic channel scores • AI/ML-driven insights
Level 4: Omnichannel Orchestrated – Responsive to interaction data from all sources	• Coordination across field roles and digital

of ~40% of targets reached based on field feedback during the call plan refinement cycle. Feedback revealed the need for multiple other channels, like e-detailing, phone, email, and virtual speaker programs.

The multichannel optimization engine assigned optimal calls for each HCP based on rep feedback. The engine also enabled reps to provide continuous feedback on the MCCP through a cloud platform that refreshes the commercial planning and data systems, including the customer relationship manager (CRM), weekly or monthly. These enhancements optimized the customer experience by targeting the HCP's preferred engagement channel, creating an actionable call plan based on actual channel activity.

A channel engagement strategy plan that leveraged customer insights was also created through field feedback and analytics like customer consent and channel preference data, mobility data, channel effectiveness, etc. This simple change helped increase field rep buy-in by enabling users to add or drop planned targets and refine calls across channels throughout the cycle. As a result, sales force engagement increased target reach and frequency by 20% in one quarter. The organization also created a more balanced workload for its sales force by re-distributing calls from traditional F2F channels to less arduous, cost and time-saving virtual and remote channels.

This cloud-based platform also displayed dynamic reports that enabled users to monitor call plan changes, like product, segment, and specialty-wise call execution and target reach. We integrated the call activity dashboard with the CRM to capture reach and frequency trends by channel and geography for timely insights and guidance. This platform also provides actionable insights to field users through weekly reports that track execution vs. guidance and identifies priority targets using a combination of call execution and sales data, including a 360° view of HCP information.

As a result of these enhancements, the company's neurology and central nervous system (CNS) portfolio is now driving sales force efficiency and optimized customer engagement through dynamic multichannel call planning.

Figure 6: Optimized Territory Design Process – Territory Accessibility

Third Party Source	Field Feedback	Field Execution – Calls/Day	Final
Poor Access	Below Average Access	Poor Access	Poor Access
Below Average Access	Poor Access	Below Average Access	Below Average Access
Below Average Access	Average Access	Average Access	Average Access
Average Access	Great Access	Below Average Access	Average Access
Average Access	Poor Access	Below Average Access	Below Average Access

Case Study 2: Optimized Territory Design Process Utilizing Access Information

Because of increased access restrictions and the introduction of virtual channels, a big pharma organization wanted to understand the changing local dynamics of its gastro-focused sales force in the northeastern US. The biggest challenge they faced was the availability of reliable data. Multiple data sources were used to define how easy it was to access a territory. The accessibility data we used included third-party HCP access, third-party contact preference, field execution data, and internal field feedback. Each territory was segmented into low, medium, and high access areas based on each data source. The composite segment for each territory was assigned based on the highest frequency segment across the data sources. For example, if two or three data sources showed a territory had poor access to HCPs, the territory was put into the poor access segment. If there was no common segment, the segment in the middle was assigned as the composite segment. (Figure 6)

Using the segmentation method described above, 68% of geographies had the same final segment as the one calculated using third-party sources, 66% had the same final segment calculated using field feedback, and 55% had the same final segment as that calculated using field execution. This enhanced territory segmentation process helped the organization refine its territories by combining the alignment index and local knowledge of how HCPs react to and prefer contact with reps. For example, poor access is the final segment for one of the territories, so keeping it slightly above the threshold (average index +20%) ensures enough accessible physicians in the territory.

Key Takeaways For Today's Pharma Organizations

Pharma organizations must rethink their field deployment and targeting processes as they foray into the new, digitized commercial deployment era. The following are changes that warrant consideration:

- Establish a regular cadence of field deployment health checks that identify opportunities to focus on the system and make necessary changes. Field deployments and targeting approaches must constantly evolve to respond faster to customer needs and changing market dynamics.
- Evaluate deployment and targeting strategies, capabilities, and systems to create sales team organizations

that can adapt quickly to changing market scenarios and work together in a coordinated manner.

- Plan customer-centric field deployments that utilize field intelligence and local knowledge to better collaborate with field reps and allow them to adapt to changing local market dynamics.
- Deploy agile and integrated systems to enable advanced dynamic planning approaches. Targeting and call planning are shifting from static cycle planning to dynamic and multichannel call planning supported by frequent AI/ML-driven insights that send high-value actions beyond the call plan.

As with any new way of doing things, this dynamic targeting call plan approach may face resistance from some field reps who prefer to use the approach as a series of simple suggestions rather than a methodology that needs to be followed carefully. However, by working with field teams to foster acceptance and providing appropriate field training, these enhanced strategies can help pharma companies deploy well-equipped field forces that can handle rapid changes in the business environment. This approach gives field reps the information they need to reach HCPs and ensure the right treatment regimens quickly reach the patients who need them most.

About the Authors

***Ashvin Bhogendra**, Senior Director, Axtria, has over 18 years of experience in the pharma commercial operations space. In his onshore and offshore roles, he has worked extensively with Top 50 Pharma clients, advising them in commercial excellence for various therapy areas. He has led complex end-to-end sales planning and incentive programs that helped transform the sales operations of large pharma clients. Ashvin is an SME in the pharma commercial model design and operations space and leads a center of excellence focused on driving innovation, capability, and asset development. Ashvin holds a master's degree from the University of Texas at Austin and a bachelor's degree from BITS Pilani India.*

***Abhilash Sain**, Senior Director, Axtria, has over 18 years of experience supporting life sciences and healthcare companies focused on sales planning, operations, and complex analytical engagements. He has led many large and complex projects with top pharma organizations across diversified analytics portfolios like sales force and marketing effectiveness, predictive modeling, and primary market research spaces, generating actionable business outcomes. He holds a Master of Technology in Process Engineering and Design and a Bachelor of Technology in Chemical Engineering from IIT Delhi, India.*

***Anjali Attri**, Associate Director, Axtria, has over 11 years of experience in the healthcare and IT industries. She has led multiple field deployment engagements for various pharma organizations and is a subject matter expert in territory design and people placement. Anjali holds a master's degree in marketing and analytics from the Great Lakes Institute of Management, Chennai, and a bachelor's degree in computer science engineering from Panjab University.*

***Monal Tenguria**, Manager, Axtria, has over a decade of analytics experience in the pharma industry. She has expertise in the commercial effectiveness domain and a focus on call planning across multiple therapeutic areas such as neuroscience, diabetes, cardiovascular, immunology, respiratory, and rare diseases. Monal holds a master's degree in computer science from IIT (BHU) Varanasi.*

Enhancing Patient Classification and Staging in RWD Using Machine Learning

Arrvind Sunder, Principal, ZS Associates; Atharv Sharma, Advanced Data Science Manager, ZS Associates; Priyanka Halder, Associate Principal, ZS Associates

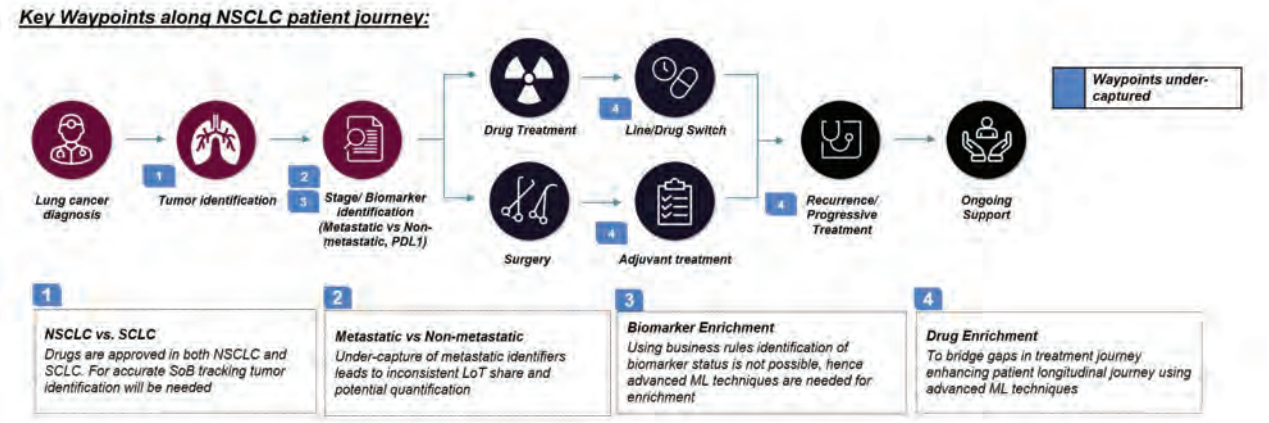
Abstract: Real-world data (RWD) sources like administrative claims house very rich information on patients' real-world interactions with the healthcare ecosystem. This serves as a solid foundation to understanding patients' journeys around diagnostic and treatment sequences, treatment rate, market share, persistence, and compliance. In addition, pharma companies use real-world data for evidence generation, HEOR studies and identifying drivers of disparities in care as well as capitalize on latent demand, etc. In fact, real-world data has started playing a pivotal role in pharma-driven interventions in care delivery, especially driven by drug approvals for niche patient populations – e.g., metastatic Breast cancer with patients receiving prior anti-HER2 therapies or patients with prior surgery in early stage NSCLC – RWD becomes the only way to identify and estimate the relevant population size. However, one of the most common challenges that most organizations face while using RWD is the **incomplete claim capture and biases at sub-national level**, thus restricting the use of real-world data and inhibiting understanding of patient- or customer-level insights. Almost all major pharmaceutical companies struggle to derive robust patient-level insights or do customer valuation, owing to these data limitations or biases. For example, in certain cases, the surgery rate in NSCLC was found to be off by 20-25%[1] when analyzed from real-world open claims data.

Claims data is like a *Swiss cheese* delicacy with holes. While we can derive a few insights from claims data, filling the gaps and painting the *complete* sequence of information requires special art. In this paper, we propose a state-of-the-art, machine-learning-based (ML) approach that integrates knowledge of the therapy area/tumor type through ***gold standard*** **confident cohort identification** and robust feature engineering with supervised/semi-supervised (positive unlabeled) modeling techniques to fill the gaps in the patient journey. Multiple validation techniques – including openly available data sets such as published literature, census data, SEER data, etc. – were used to build confidence and trust in the final result. This approach has been applied across a spectrum of tumor types – one such example of mitigating data gaps and accurately identifying and classifying de-novo metastatic NSCLC patients from claims data has been detailed in the paper.

The proposed approach helps in the systematic mitigation of real-world data challenges as well as moving from generating directional insights to more robust and actionable insights. **Enriched data** can serve as a foundation across different functions, such as the insights/analytics wing of organizations, which can now build better informed strategies and aid robust performance tracking as well as accurate opportunity assessment, while the medical affairs teams can realistically understand and quantify true gaps in care treatment across different patient cohorts, and the Real-World Evidence team can use this data for outcome analysis and evidence generation, etc.

Keywords: Real-world data, incomplete claim capture, biases at sub-national level, gold standard confident cohort identification, enriched data

Figure 1: Under Captured Waypoints Across NSCLC Patient Journey



Background: Challenges in Patient Classification

As the world is moving more toward precision medicine, oncology drugs are getting approved for a niche sub-type of patients, for which accurate classification is seldom available in raw claims data. Due to this, a plethora of domain-driven business rules are needed to identify/ classify and label the patient population, which might introduce bias and at times, it may not even be sufficient to help maximize the power of real-world data sets available with pharma companies. In a few cases, the business rules might not be able to classify the patients across stages/biomarkers at all. All these gaps/limitations may lead to inaccurate insights, hence impacting the strategies multiple functions might end up creating and hindering the purpose of leveraging real-world data.

Key gaps typically observed across the oncology therapy area, for example, might include:

- **Patient subtype/biomarker identification:** Accurate understanding and tracking of a patient's biomarker and subtype is necessary to estimate the accurate market opportunity as well as understand gaps in patient care treatment. Due to the under-capture of these granular details in claims data, we are unable to robustly identify the opportunity and understand market dynamics.
- **Patient stage/tumor identification:** In oncology, identifying the accurate patient stage (e.g. Stage I/II/III/IV) is a crucial piece of the equation. The surgery and metastatic markers (such as presence of secondary sites of metastasis or other progression markers) are under-captured, which leads to heavy reliance on additional business rules to mitigate these gaps. Reliance on business rules introduces biases, which lead to inaccurate insights.
- **Longitudinal drug usage:** Across patients' longitudinal journey, drug usage is difficult to track due to gaps in claims data. This reduces the patient population we can track longitudinally and provides inaccurate LoT (Line of therapy) share estimation, understanding patient persistence and compliance.

Figure 1 illustrates a highly simplified NSCLC patient journey, illustrating major waypoints that matter in the patient journey.

Purchasing more and more data will not solve the problem because each data will come with its own set of nuances and considerations. For example, while EMR data sources in general might capture the longitudinal patient journey better than traditional open claims data sets, they might not be helpful when it comes to customer-level commercial use cases such as targeting, segmentation, sizing, etc. The costs of procuring and maintaining newer and newer data sources also add another layer of complexity for pharmaceutical organizations. In addition, once we start combining multiple data sets, the number of patients for whom capture is complete becomes much smaller, and hence ends up impacting a lot of sub-national analyses.

Hence, there's a need for advanced techniques and a robust analytics framework to mitigate these challenges systematically and enable:

- Better patient classification to identify cohort of interest
- Robust insights from claims data along different waypoints on the patient journey (e.g., adjuvant rate, treatment rate, testing rate, etc.) and mitigate potential factors that lead to these gaps
- Projection techniques to estimate market size
- Understanding true brand performance and comparison of forecast with actuals
- Identification and quantification of care gap, drivers for disparities in care and quantification of at-risk patients as well as identification of geography hotspots
- Opportunity tracking and quantification
- Understanding latent demand
- Real-world evidence generation
- Understanding evolution of market dynamics and so on ...

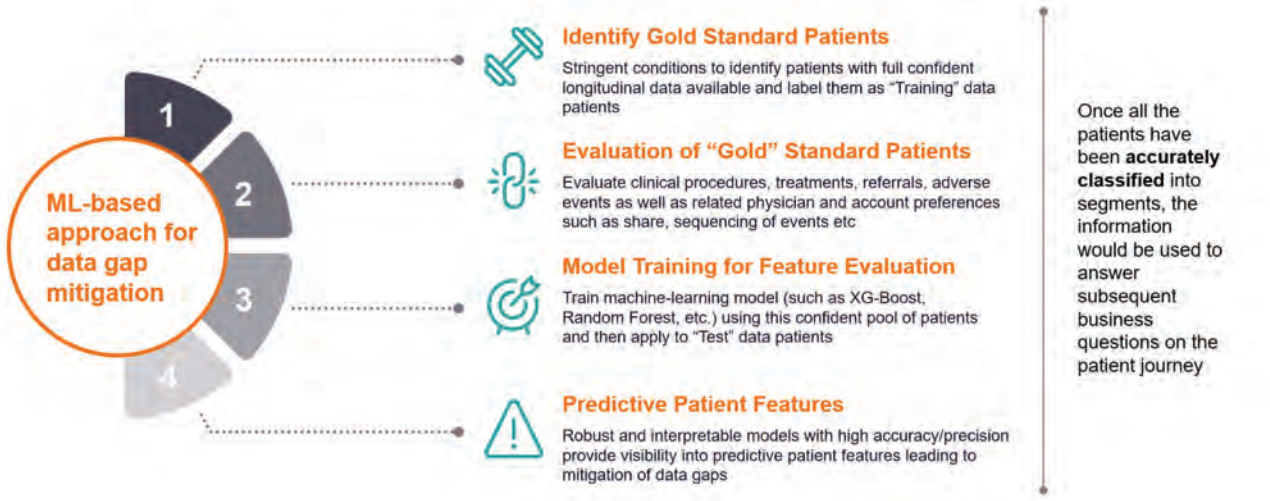
How Can These Data Gaps be Mitigated to Derive Maximum potential from Real-world Data?

Typical business-rules-based approaches to label and classify patients generally fall short to derive accurate insights. Even if we deploy an array of conditions post a thorough study of the tumor type, it ends up reducing the patient population we have confidence in for the analysis, and hence impacts the sub-national insights drastically. This calls for a need for more advanced machine-learning-driven solutions that can decipher trends and patterns from the data and help make it more *complete*, and this enriched data can then be used for all other downstream analytics.

Claims data houses multiple patient-level interactions with the healthcare ecosystem such as diagnosis, treatment, procedures, physician interactions, referral dynamics, OOP cost, etc. While gaps might exist across multiple patient events, as we think about enriching the data, it's important to understand which waypoints are critical for the business, and are highly understated in the data, hence requiring enrichment.

To enrich these key patient waypoints, multiple machine-learning frameworks can be used depending on the data nuances and waypoint selected, e.g., supervised, semi-supervised (Positive Unlabeled), transfer learning, etc. The overall process can be divided into four broad key steps (Figure 2).

Figure 2: Approach to Build ML Models for Waypoint Enrichment



- **Identification of gold standard patients**: This is the first step in any machine-learning model development – selecting patients from the data set whose journey and pattern of events can be analyzed as a reference for the model. Business-rules-based approach, medical and domain expertise, learnings from syndicated reports and research papers can generally be used to define initial labels for these patients. While these patients might have certain other events missing in the data set, they do have enough relevant information (through continuous longitudinal capture) to confidently establish them as gold standard patients for that particular event. Please note the gold standard patients defined here show presence of event of interest but the pattern of missingness of other events is similar to other patients not labeled as gold standard patients, hence allowing for pattern mining from gold standard patients and scoring on the remaining patients. Before we go ahead with model training, additional validations are performed to ensure the patients selected for model training do have enough signals captured in the data set. For example, they should have continuous capture in the data set and should not have breaks in capture in data over long periods of time. They should also possess good capture of events relevant to the particular therapy area, etc. This of course ends up impacting the number of patients that can be used for model building but appropriate trade-offs need to be considered to ensure we keep patients with decent event captures so that the model can robustly learn patterns from those patients, in addition to maintaining n-size of labels required by the model and ensuring decent imbalance for the model. Once finalized, this patient population is used to train the model and identify *look-a-like* patients for the event of interest from the remaining pool of patients.
- **Evaluation of gold standard patients and feature engineering:** Permutation of various patients' transactional claims[2], which include clinical events (treatment/procedure/diagnosis), physician visits and referrals,

ER visits, physician attributes (specialty, location), patient demographics (race/ethnicity) and other patient attributes such as CCI (Charlson Comorbidity Index) across different time periods are evaluated through different aggregators and converted to features that will aid with model training. In addition, physician and hospital treatment preferences/pathways can also be coded to create features that can help with the waypoint enrichment. Apart from coding these generic feature journey events, certain openly available publications can be explored to help define what sequence of events may help with patient classification.

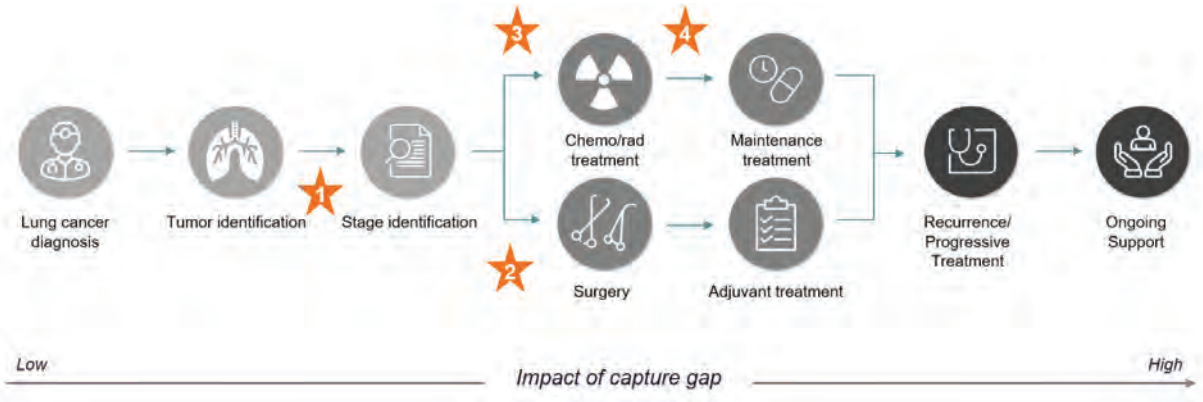
Depending on the modeling technique to be used, the features can be maintained at transaction claim level or data can be converted into a tabular format by deploying aggregators on top of the events. For example, at a patient-time period level, the frequency of hospital/ER visits in last 30 days, etc. can be quantified as a feature. Once we have completed the feature engineering, we can proceed with the model build.

- **Model Setup, Training and Prediction:** A key part of this exercise is the problem formulation and depending on the patient journey waypoint to be enriched and the problem formulation, modeling approach, model parameters, prediction window, training window, etc. are determined.
    - **Case 1:** In situations with high confidence in all classes being passed into the model, a supervised learning model can be trained. For example, in open claims, only a subset of patients can be classified under HR+ or HR- and the rest of the patients are ambiguous, which means they don't demonstrate any obvious markers for analysts to classify them under HR+ or HR-. Once these gold standard HR+ and HR- patients have been identified and these labels have been vetted, a supervised model (traditional ML or deep learning models) can be trained that can later be used to predict ambiguous patients and classify them with reasonable confidence under HR+ or HR-.
    - **Case 2:** In situations with confidence in only one class (henceforth called positive class), and not on the negative class, a semi-supervised or Positive Unlabeled (PU) learning-based modeling approach [3] can be used. For example, CPT codes in administrative claims can help define if and when a patient has undergone surgery. But the absence of these codes doesn't necessarily mean the patient has not undergone surgery, since it can very well be an artifact of a missing claim in the data set. In such scenarios, a PU learning-based approach can help identify additional lookalike surgery patients that exhibit a pattern similar to the known surgery patients.
    - **Case 3:** In situations with no obvious markers available in claims data to define the labels, transfer learning-based approaches can help with patient classification. For example, there are very limited markers to define PDL1 staining in administrative claims. Electronic Medical Record/Electronic Health Records-based data sources directly provide this information in lab tables or can be mined from physician notes. This information can then be used to train a model,

which can later be expanded to predict on administrative claims data. This of course requires additional sophistication to account for the differences in event distribution across data sources, and requires a more robust validation process to ensure accuracy of the patient classification.

- **Validation of Results:** This is the most important part of this exercise to ensure robustness in ongoing insights generation. To help drive trust and confidence in the results, as well as to provide additional quality assurance, a human in the loop approach combining statistical and business validations is performed.
    - **Model parameters and KPIs**
        - Typical model performance parameters such as Precision, Train vs. Spy recall, F1, etc. to check the performance of the model on train data and test/holdout data
            - A certain set of *gold standard patients* are kept aside to validate model performance. These patients do have event of interest and the model is used to score these patients. Depending on the type of model that has been built, relevant model parameters can be analyzed. For example, if we use a semi-supervised (PU learning- based) approach, where we only have information about one class of patients, recall becomes the most important parameter to analyze. Since we won't be able to assess precision/accuracy in such a scenario, more weight is given to other validations.
            - However, if we go ahead with supervised models, precision/accuracy and F1 can be estimated as well.
            - As we do this model assessment, enough consideration should be given to n-size of patients that are present in test/holdout data. For example, if test/holdout data have a small number of patients, some of these model parameters might be highly volatile and this validation alone might not be sufficient to drive confidence in the model. This approach works well in cases where we have enough patient n-size (e.g., stage of the patient) but in cases of highly nuanced biomarkers where n-size is quite low, additional validation techniques need to be deployed.
        - Statistical tests such as T-test to check the feature distribution and similarity between labeled patients and additionally identified/classified patients. We can also analyze deviation across top events with the remaining or true negative class patients and drive home the confidence in patient classification.
        - Top sequences/events/feature summary that aid in patient classification: These can be vetted with medical and domain experts, published literature to ensure the model has latched on the right signals.
    - **Trend analysis, expert and knowledge-driven business validations (typically understood from published literature or deep domain expertise)**

Figure 3: Key NSCLC Waypoints Enriched as Part of Case Study



- A combination of macro- and micro-level validations are performed here to ensure accuracy of the model.
- A lot of macro information, such as incidence/prevalence rate of events, can be sourced from open literature (such as SEER/registry data, etc.), while micro validations at the patient cohort level require a much deeper understanding of the therapy area. This is the place where validation with medical experts/ KOLs can help. For example, they can look at a sequence of events or top events that have aided in patient classification or change in prescribing pattern over time to ascertain model robustness.
    - Thorough secondary desk research can also be done based on published research papers to ascertain the validity of the identified top features/sequences as well as key trends.
- Additional business validations such as trends over time can be assessed to ascertain the robustness of the model predictions. Trends can include (but should not be limited to) patient progression rate when defining patient stage, split of patients by stage/biomarker across time period, etc., timing of involvement of different specialties across different points in the patient journey (e.g., surgeon involvement when surgery is being enriched, etc.) and corroborating with published literature.
- Medical team can help review and rationalize top model features.
- Benchmarking against known market/drug/class shares when enriching systemic drugs across the longitudinal patient journey.

Case Study for Enriching NSCLC Patient Waypoints

Recent approvals in NSCLC indication are nuanced, which make accurate KPI tracking difficult. To mitigate this and make claims data the source of truth, key NSCLC waypoints were identified and enriched. We will talk through the high-level approach followed below and how the results panned out. See Figure 3.

Deep dive into approach for classifying stage of the patient:

Objective: Identify de-novo metastatic NSCLC patient from open claims data.

Identify Gold Standard Patients: The majority of the rules to classify patients as de-novo metastatic are reliant on a *'secondary neoplasm'* diagnosis code, which indicates a distant metastasis of the tumor or certain targeted therapies that the patient should ideally only receive once they are metastatic. Hence confident de-novo metastatic NSCLC patients are classified through a combination of events that these patients demonstrate, for example, a patient with their first secondary metastatic diagnosis within 30 days from the first primary NSCLC diagnosis or patients starting on regimen approved only in metastatic setting.

Due to under-capture of these diagnosis codes/drugs in administrative claims data, the percentage of de-novo metastatic claims data is generally under-reported. As per various syndicated reports and research papers, de-novo metastatic for NSCLC is ~45-50%[4] of all incident NSCLC patients, whereas from claims data we can only classify half of these patients with confidence.

Approach: A Positive Unlabeled framework was leveraged to identify patients with similar events and sequences, to confidently classify additional potential de-novo metastatic patients who had missing secondary neoplasm claims in the data set.

Evaluation of Gold Standard Patients and Feature Engineering: Features involving transactional claims data, capturing procedures, diagnosis, drug treatment, referral pattern, specialty involvement around patient diagnosis, patient comorbidity burden, physician and hospital treatment pattern, etc. were evaluated in two years of patient history (prior to patients' first primary NSCLC diagnosis). Permutations across different time periods, events and aggregators were created, and the newly created cross-sectional/tabular data was used for model training.

Model Training and Prediction: In this scenario, we are confident of patients that are de-novo metastatic, given they exhibit clinical events/sequences that are typically observed in de-novo metastatic patients. However we cannot say with confidence that the patients that don't exhibit these events in administrative claims are truly non-metastatic or early stage, since data might simply be missing capture of these key events. Therefore we leveraged PU learning[3] (One class identification) based approach to identify additional potential de-novo metastatic patients from the pool of incident NSCLC patients, who demonstrate sequences of events more prominently observed in metastatic patients. While multiple other classifiers were tried, optimally tuned XGBoost, a state-of-the-art ML model was finally used for training the model. Use of XGBoost also helped with ensuring model explainability.

Validation of Results: Model results were pressure tested through multiple angles. Robust validations included both statistical as well as business validations to drive confidence in the classification of additional patients. Few validations included:

- Post ML, ~15% additional metastatic patients were identified, taking metastatic vs. early stage split closer to published literature (~40% vs. expected ~45-50% from literature[4])
- Robust model performance (>75% train vs. spy recall)

Figure 4: Rule + ML Model-based Patient Identification

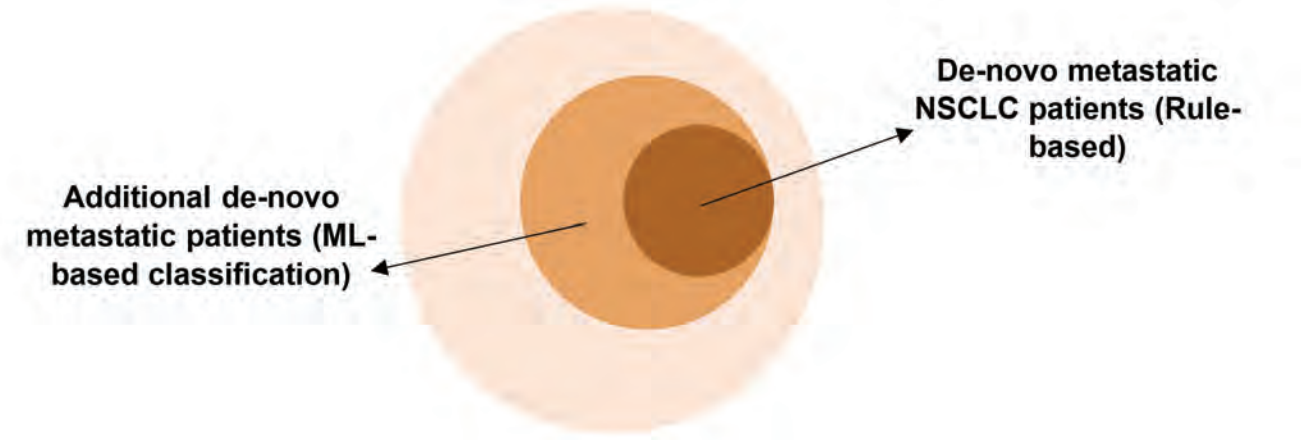


Table 1

#	**Key waypoints enriched along patient journey**	**Impact realized using ML** (compared to current data-driven understanding)	**Few key top events aiding prediction**
1	Identification of patients receiving surgery in early-stage NSCLC patients	**~2x** surgery patients identified bringing additional robustness for early-stage surgery rate	Recency of biopsy, radiation procedure frequency[5], high imaging test frequency[5]
2	Treatment rate: Identification of patients receiving chemo and radiation in early stage	**>1.4x** more patients receiving treatment identified	Recent infusion procedure, frequency of prescriptions for drug of interest

- YoY trend analysis to ascertain trend of de-novo metastatic patients
- Analysis of regimen/drug level summary
- Patient journey analysis and distribution of key events, such as imaging tests, frequency of visit to hospital/infusion centers, etc.
- Model identified top predictive events/sequence of events validated with medical experts, and published literature

Similarly to stage identification, additional waypoints can be enriched to enhance the claims data set and drive higher accuracy in all the analysis across commercial/medical functions. Table 1 summarizes the high-level impact observed for an administrative open claims data source across additional waypoints.

Considerations

While ML-driven analytical techniques can help set real-world data as a robust source for the majority of downstream analytics for commercial and medical teams, there are few considerations to be kept in mind as one starts to use it on an ongoing basis:

- Caveats with model predictions and/or analysis: Owing to multiple nuances, especially in complex therapy areas such as Oncology, some waypoints can be enriched better

than others. This can be a factor of multiple reasons such as data capture, confidence in initial patient labels, how quickly the event happens in the patient journey, number of claims, data source, etc. Hence a robust model assessment and understanding of related nuances/caveats is essential when leveraging the real-world data for downstream analytics. If we don't have a good understanding, we might end up drawing the wrong conclusions.

- Timing of data availability: Some data sources have a lag in data availability. Hence, having a clear understanding of when the output would be available from real-world data is essential while developing models so that relevant adjustments can be made accordingly.
- Triangulation across data sources: It's important to ensure a robustness in benchmark assumptions, and hence it becomes important – especially in a launch scenario – to pressure test outputs from real-world data through multiple angles. One of the validations is to triangulate across sources to ensure no major bias/caveats exist when setting up real-world data as a source of truth for assumptions.

Conclusion

For deriving robust insights and enabling use-cases such as targeting, forecasting, segmentation –especially in nuanced markets like oncology – requires a strong analytical foundation to get it right. This enrichment of real-world data sets can enable multiple other functions such as R&D, RWE, HEOR and commercial. Real-world data such as claims data can help us understand the real-time evolution of trends and market dynamics. However, claims data full of pitfalls requires the analyst to have a deep understanding of the caveats of the data, and deploy smart analytical/ML-based solutions to derive robust insights from the data. Please remember: No data is perfect and buying new data is not always the solution. Rather, the strong analytics and use of ML techniques to discover the undiscovered will help save the day. This approach can not only help with quantifying various waypoints in the funnel with more confidence and robustness, but also with validation, understanding evolving trends and calibrating forecast with actuals when required.

About the Authors

Arrvind Sunder Arrvind *is a Principal in ZS' Evanston office and has been with ZS for 16 years. Arrvind is a leader in ZS' Real-World Data & Insights venture where he focuses his time on enabling clients realize competitive differentiation through data, digital and AI.*

Atharv Sharma Atharv *is an Advanced Data Science Manager with ZS Associates, based in Philadelphia. Atharv has over eight years of experience in applying machine-learning solutions in healthcare, and primarily helping the commercial pharma industry realize the impact of real-world data leveraging advanced analytics and ML-based techniques.*

Priyanka Halder Priyanka *is Associate Principal from ZS Associates, based in Pune, India. She has 13 years' experience working across life science clients. Priyanka leads the advanced analytics for ZS' Real-World Data and Insights venture with a focus on AI and technology-enabled solution and product development.*

References

1 Blom EF, Ten Haaf K, Arenberg DA, de Koning HJ. Uptake of minimally invasive surgery and stereotactic body radiation therapy for early stage non-small cell lung cancer in the USA: an ecological study of secular trends using the National Cancer Database. BMJ Open Respir Res. 2020 May;7(1):e000603. doi: 10.1136/bmjresp-2020-000603. PMID: 32404305; PMCID: PMC7228566.

2 Chilukuri S., and Madgi S., Key Drivers for Successful Patient Event Prediction: Empirical Findings on What Matters and to What Extent, 2020, Journal of the Pharmaceutical Management Science Association.

3 Bekker, J., Davis, J. Learning from positive and unlabeled data: a survey. Mach Learn 109, 719–760 (2020). https://doi.org/10.1007/s10994-020-05877-5

4 Niu, FY., Zhou, Q., Yang, JJ. et al. Distribution and prognosis of uncommon metastases from non-small cell lung cancer. BMC Cancer 16, 149 (2016). https://doi.org/10.1186/s12885-016-2169-5

5 Chen D, Wang H, Song X, Yue J, Yu J. Preoperative radiation may improve the outcomes of resectable IIIA/N2 non-small-cell lung cancer patients: A propensity score matching-based analysis from surveillance, epidemiology, and end results database. Cancer Med. 2018 Sep;7(9):4354-4360. doi: 10.1002/cam4.1701. Epub 2018 Jul 29. PMID: 30058192; PMCID: PMC6143945.

Individualized Customer Journeys Using Bayesian Statistics: Mapping Optimal Sequences of Interaction

Teis Kristensen, Project Lead, Axtria; Kritika Singhal, Data Science Manager, Johnson & Johnson; Ramesh Krishnan, Principal, Axtria

Abstract: Patients and healthcare providers are embedded in a media ecosystem that is characterized by intense competition for attention. Pharmaceutical marketers must emphasize customers' experience over siloed marketing channel exposure if they want to deliver impactful messages. A Bayesian network methodology can predict the most impactful tactic for customers at each step in their journey. We discuss a Bayesian framework that can map journeys with a high probability of an outcome to happen, predict the next best action in a specific customer journey, and allow the attribution of tactics to customer outcomes. This paper first describes the theoretical underpinning of Bayesian network statistics and methodological considerations when examining customer journeys, and then provides results that illustrate the type of quantitative insights that can be generated. Our framework informs the design of customer journeys and omnichannel orchestrations while facilitating the exploration of customer journeys among decision-makers. Practical implications for pharmaceutical operation teams are highlighted throughout the paper.

Keywords: Omnichannel, NBA, Customer Journey, Bayesian Statistics, Individualized Experiences, Commercial Operations

1. Individualized Customer Journeys Using Bayesian Statistics

The pharmaceutical ecosystem has continuously been moving towards customization and digitalization in the engagements between manufacturers and healthcare providers. The COVID-19 pandemic has illuminated the link between public health and people's daily lives, and it has brought the importance of medicines to the forefront of people's minds. Pharmaceutical marketers can utilize the reinforced focus on health to drive better healthcare outcomes for their patients. However, when doing so, they are faced with strong competition for attention by other commercial actors, consolidation restricting access to HCPs, and a continuous move towards digital interactions across stakeholder groups (4). Pharmaceutical marketers must stand out in a crowded and noisy media ecosystem in order to inform HCPs and patients of life-improving medicines.

To assess and navigate the complexities of the pharma environment, operation specialists have turned to data-driven methodologies that help measure and optimize the impact of promotional initiatives. By using observed data, marketers are able to make empirical and proactive decisions. Several data-driven tools exist to inform strategic, tactical, and operational decision-making. For example, marketing mix analyses are often used as a strategic tool to inform budget allocation across promotional channels, ROI analyses provide a tactical overview of the most profitable approaches to reach forecast objectives, while at the operational level, machine learning models are deployed to predict the next best action between HCPs and pharma companies. As analytic methods

and capabilities mature, more and more data-driven insights are becoming available.

Continuously with the method development, a move towards personalizing promotional activities and an increased focus on omnichannel orchestration have gained traction. These initiatives aim to take a customer-centric perspective that emphasizes the holistic experience of health administrators, providers, and patients. A customer journey is the sequence of interactions between a healthcare provider and pharma company over time. The concept has gained traction because it seeks to understand the customer from their first interaction and throughout their complete engagement with a company.

This white paper describes how a Bayesian Network methodology (5) can be used as the engine for a data-driven examination of customer journeys. The method combines the focus on interactions seen in digital attribution with an emphasis on customer journeys. In fact, the basic unit of observation used in the analysis (6) is the unfolding of events, over time. A Bayesian network calculates the probability of an outcome based on a sequence of observed events (7). The insights generated include 1) the most frequently observed customer journeys, 2) identification of customer journeys with high impact, 3) attribution of outcomes to interaction points in the customer journey, and the next best interaction point following an observed sequence of promotions.

The following sections aim to provide an overview of a data-driven customer journey analysis based on a Bayesian Network methodology. Section 2 describes how a customer journey framework can inform promotion orchestration both within and across channels, reviewing methods used for data-driven examination of customer journeys, and highlights the considerations used in applying a Bayesian Network methodology to generate insights. Section 3 provides the theoretical and mathematical underpinnings of the work. Subsequently, Section 4 presents the Bayesian Network methodology results, specifically the structure of the simulation data, central elements of model development, and the built network's queried results. The last section discusses the benefits and limitations of a Bayesian network method to generate customer journey insights and the broader implications for pharmaceutical marketers.

2. Data-Driven Customer Journey Design

The concept of a customer journey has been used in a multitude of contexts. The popularity of the concept has resulted in ambiguity, which requires defining what a customer journey is from a pharmaceutical operations perspective. At the core of the concept is the understanding that optimization of promotional initiatives must be done with an emphasis on a positive customer experience. The application of a customer journey perspective is done to enable marketers to map pain points and opportunities for improvement. From a Bayesian Network perspective, the points of interaction that unfold over time between customers and companies are the elements that make up a customer journey (8). Figure 1 illustrates how a healthcare provider has multiple interactions points before a contract is signed.

When applying a customer journey framework to understand promotional activity, it is essential to clearly identify who is considered a customer and what channels to include in the analysis. Even though these questions may seem trivial, the pharmaceutical industry has multiple

Figure 1: Journey of a Healthcare Provider

stakeholders that could be considered customers, such as patients, providers, and administrators. Therefore, specifying the customer beforehand functions as a funnel that reduces the promotional channels that must be considered. A customer journey can consist of interaction points both within and across channels. For example, a within-channel examination may focus on the sequence of content delivered within a channel, while a cross-channel examination hone in on interaction points such as executed details, email contacts, etc. Thus, the focus of a Bayesian network analysis moves from estimating intensity levels of channels towards an emphasis on the customers' experience across channels over time.

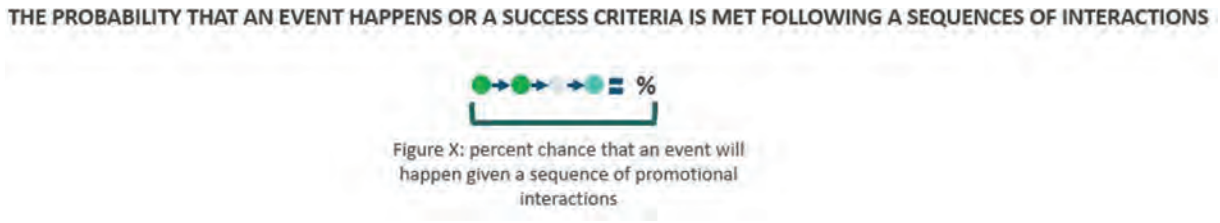
In this paper, the customer experience refers to the interaction between healthcare providers and pharmaceutical manufacturers (business-to-business). Five channels are used as an example throughout the article. The channels included are Face-to-Face Detailing (C5), Virtual Detailing (C4), Mobile Notifications (C3), Direct Mail (C2), and Email (C1). Bayesian network modeling should be limited to channels that can be initiated and pushed by the manufacturer, as this will allow recommendations to be operationalized. The content of the tactics are excluded from the reviewed examples as to reduce complexity, but tactic and content combinations can be treated as individual variables in the modeling process.

Bayesian statistics are suited for customer journey analytics because they encompass time as a central element in the modeling process (5-7). Bayesian methods, such as Markov and Bayesian network models, calculate the probability of an outcome based on a prior sequence of events (5). Specifically, Markov and Bayesian network models ingest a data format that mirrors the points of interaction between companies and customers (6). This approach fits with a customer journey framework as a sequence of events are the points of interaction between a Life Science company and its customers. (Figure 2)

Both Markov chain and Bayesian network methods allow for the examination and active design of promotional sequences (7). The difference between Markov and Bayesian network methods lies in the assumption made about the relational directionality between model variables. Markov network models assume that only undirected relationships exist, while Bayesian networks are directed acyclic graphs, assuming directionality in the probabilistic impact of events (6). Since there is an assumption of directionality in Bayesian networks, it is possible to a priori specify the direction of relationships that can/cannot exist when these are learned between promotional sequences and outcomes. Only Bayesian networks are expanded upon in this paper due to the ability of the modeling framework to benefit from domain expertise by specifying directional relationships (5).

Figure 2: From Customer Journey to Data Components

Figure 3: Bayesian Network Insight



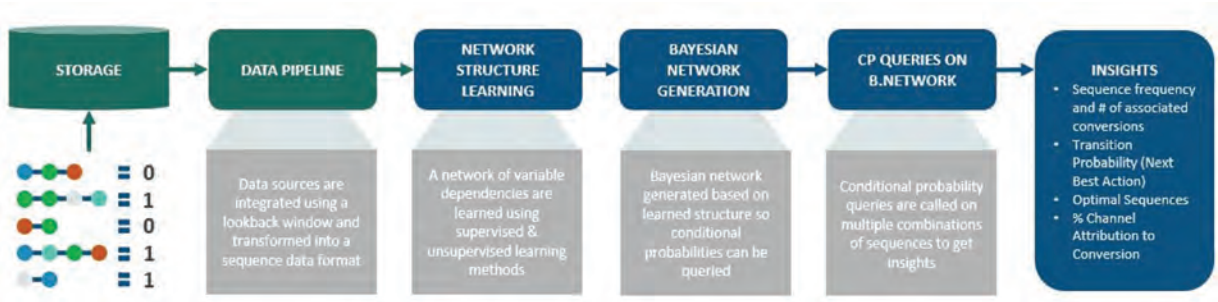
3. Theory & Method: The Foundation of Bayesian Networks

A Bayesian network is a mathematical representation of the probabilistic relationships between random variables. The objective of a Bayesian network is to model the posterior conditional probability distribution of an outcome variable given a series of observed evidence (6). Figure 3 illustrates the insights generated from a Bayesian network. Bayesian networks have gained traction in multiple contexts, such as health outcomes research and medical decision analyses, as the method supports the examination of uncertainty and causality surrounding sequences of events.

A three-stage process is used to generate Bayesian network insights (6-7) that relies on both supervised and unsupervised learning. Before initiating supervised learning, the modeler can specify which relationships that are allowed to exist within the network, such as no relationships going back in time or only allow outgoing ties for attribute variables. This is done by specifying whether a directed acyclic relationship can be formed in the model. These stipulations on learning are applied when supervised learning algorithms are run to identify relational structures among observed sequences of interaction that directly are tied with an outcome. After identifying observed sequences of events, unsupervised learning is used to estimate the outcome probability of unobserved sequences. The combination of supervised and unsupervised learning allows probability queries on all potential combinations of interactions in a customer journey. The relational structures that are learned are used to generate a complete Bayesian network. The Bayesian network contains the structure and strength of relational dependence between variables and can be queried to the probability of an outcome when prior interactions are provided. Queries can be made on observed and unobserved sequences of interactions, as Bayesian networks draw on observed data to estimate the impact of unobserved sequences. (Figure 4)

A Bayesian network has three central components; nodes representing variables from the dataset, edges between nodes

Figure 4: Components of the Analytic Framework



indicating causal relationships, and the conditional probability distributions associated with each node (7). If a causal relationship exists between two variables, the corresponding nodes in the network have a directed edge between them. The conditional probability distributions of the interaction points (nodes) are determined by Bayes' Theorem: for events ***A*** and ***B***.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Given the conditional probabilities of prior events, it is possible to approximate the posterior distributions of the nodes. In other words, the Bayesian network model calculates the dependencies between variables and creates a network of causal relationships that can be queried (5). Each variable is treated as conditionally independent of all its non-descendants so that probabilistic relationships are assumed only along the directed edges of the network, giving:

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i|parents(X_i))$$

Here, *X_1, …, X_n* is a set of nodes of the network. Queries are run on the network by moving through the directed acyclic edges.

This approach captures the conditional dependencies between variables without the need to prune all possible relationships existing in the Bayesian network. The probability of an outcome is calculated by plugging in the parameters from the conditional probability tables. Var1 and VarX are different variables or nodes of the network:

$$P(Outcome = 1|X_1 = 1, \cdots, X_n = 1) = \frac{P(Outcome = 1, X_1 = 1, \cdots, X_n = 1)}{P(X_1 = 1, \cdots, X_n = 1)}$$

The following sections present the insights generated from a Bayesian network. The presented Bayesian network framework was run on both a simulated and e-commerce dataset. Only simulated results are reported. The use of a simulated dataset was done to ensure the validity and reliability of the modeling framework, as a simulated dataset only contains specified relationship patterns that can be compared to model results.

4. Results: Bayesian Network Insights

It is possible to understand the customer journeys that have the highest probability of a desired outcome to happen. The probability score should be interpreted as the probability that an HCP will Rx a specific brand whenever writing a market Rx. The Bayesian Network calculates the probability of an Rx to happen depending on a single or a set of interaction points. The difference in

Table 1: Optimal Sequence of Promotion by Customer Journey Length

Journey Length of One		Journey Length of Two		Journey Length of Three	
Journey	Outcome Probability	Journey	Outcome Probability	Journey	Outcome Probability
F2F Detail (C5)	21.85%	C5 > C4	34.33%	C5 > C1 > C4	83.33%
Virtual Detail (C4)	20.26%	C4 > C5	29.41%	C4 > C5 > C1	80.00%
Direct Mail (C2)	19.93%	C2 > C4	28.57%	C4 > C3 > C1	66.67%
Mobile Noti. (C3)	18.97%	C1 > C1	26.23%	C1 > C2 > C5	66.67%
Email (C1)	18.74%	C5 > C3	25.64%	C3 > C1 > C2	60.00%

Note: Only the five most impactful customer combinations of promotion are included.

probability across possible sequences are compared to identify and rank which journey, or next steps in a current journey, that have highest probability of desired Rx writing. The limitation of the method is that it does not provide the optimal resource allocation of channels at a national level, but the most impactful follow-up sequences based on a HCPs current customer journey. The insights generated from a Bayesian network can be used to understand the impact of individual channels depending upon their placement in a customer journey, and thus help design the optimal sequence of promotions across channels.

Table 1 shows the five most impactful customer journeys from a single interaction point to three interaction points. The results indicate that the most impactful point of interaction is channel five with a 21.85% probability for a positive outcome. However, the results also illustrate that repeating the same channel over time does not hold the same effect. The benefit of using a Bayesian network to understand optimal promotional sequences is that interaction effects and channel position in a customer journey are considered when outcome probabilities are calculated. The insights allow for a tactical evaluation of the optimal sequence of promotional channels. In this case, a sequence of F2F detailing (C5), email (C1), and virtual detailing (C4) has the highest outcome probability of a Rx to happen at 83.33%. Even in cases where the optimal sequence of customer interaction is not executable, the Bayesian Network can be queried to identify the next best alternative sequence of events. Table 1 shows alternative sequences such as initiating with virtual detailing (C4 > C5) or sending emails and direct mail before moving toward a F2F detail (C1 > C2 > C5). The knowledge of which customer journeys that are impactful enable management teams to support sales force and marketing initiatives as they develop, adapt, and maintain uniquely designed customer experiences.

Another central insight generated from a Bayesian network model is the calculation of outcome probabilities associated with the next interaction point in the customer journey - the next best action. This type of insight can be invaluable for pharma operations personnel, as recommendations point to the interaction sequences with the highest probability of a positive outcome. The insights can, for example, be delivered

Table 2: Transition Probabilities from a Bayesian Network with Five Channels

Observed Customer Journey	Outcome Probability
Email (C1) > C2 (Direct Mail) > C3 (Mobile)	Current Journey
Next Best Action for Customer Journey	-
C1 > C2 > C3 > C1	16%
C1 > C2 > C3 > C2	21%
C1 > C2 > C3 > C3	18%
C1 > C2 > C3 > C4	10%
C1 > C2 > C3 > C5	24%

Note: Outcome Probability is the increase in the probability of an outcome to happen based on adding a specific step in the customer journey. C is used as an abbreviation of a channel interaction points (F2F Detail (C5), Virtual Detail (C4), Direct Mail (C2), Mobile Push Notifications (C3), and Email (C1)).

to the sales rep before engaging with an HCP or used by the home office to give guidance on tactic usage.

Table 1 shows a customer journey consisting of three interaction points. Queries on the Bayesian network were run to calculate the probability associated with an additional interaction point in the customer journey, while using the current outcome probability as a reference point. The results in Table 2 show the currently observed customer journey would benefit the most from having Channel 5 as the fourth interaction point, as it would increase the probability of the desired outcome by 24%. In this way, a Bayesian network can support operational decision-making by establishing the next best action across channels for a given HCPs based on the most successful customer journeys that has historically been observed.

Lastly, a Bayesian network can be used to attribute the impact of a channel to an outcome variable. This is done by using the difference in probability between configurations of promotion sequences. The method of calculation follows the principles applied to game theory methods for coalition attribution (8). Axioms to derive a Shapley value were used to estimate channel attribution.

Table 3 reports the attribution results. The impact of a channel was defined as the loss of probability if the channel was excluded from the customer journey. The results show the percentage of outcomes attributed to each channel. The results show that channel five continues to have a strong impact on the probability of an outcome. Channel five can be attributed to 21.52% of the outcomes, compared to the 18.57% attributed to channel one. These insights can form the basis for further analyses. For example, the financial impact of a change in the sequence of promotion for a set of HCPs can be calculated. The insights facilitate the strategic planning of promotional activities by providing a quantifiable understanding of unique customer journeys and thus support omnichannel orchestration dynamics.

Table 3: Bayesian Network-Based Channel Attribution

Channel	Attributed Outcomes	Attribution Percent
C1	4116	18.6%
C2	4261	19.2%
C3	4453	20.1%
C4	4562	20.6%
C5	4768	21.5%

Note: Game theory axioms are used to calculate attribution that follows the assumptions of cooperative games

In addition to the use of simulated and observed data, several machine learning models were used to test and validate the Bayesian network approach. Logistic regression, random forest, gradient boosted regression, and naïve bayes models were tested on the simulated data. The machine learning models' predictive power was hindered by skewness in the data towards non-conversion entries. The best machine learning models, tested on multiple transformed datasets, had a classification error rate of approximately 50%, which suggests a Bayesian network approach is more appropriate for solving the problem of understanding the impact of sequences on outcomes. Traditional machine learning models may be more applicable with data structures less focused on promotional sequences. The strength of a Bayesian network approach is the ability to better understand the dynamics between promotional channels by highlighting the impact of specific promotion sequences from a probabilistic lens.

5. Conclusion: Practical Implications of Using Bayesian Network

The use of a conceptual framework based on probability acknowledges the uncertainty with which an outcome happens. By taking a probabilistic approach to understanding customer journeys, time and sequence of events are included as fundamental components in the statistical modeling process. The probabilistic queries run on Bayesian networks can be used to provide operational, tactical, and strategic insights that help the orchestration of omnichannel efforts.

Examining a customer journey using Bayesian networks provides insight into how different combinations of promotional interactions can impact the probability of an outcome such as a sale or increased patient adherence. The outcome sought can be any success criteria associated with a customer over time, such as clicking an email link, downloading an information brochure, and placing Rx orders. The generated insights about customer interactions enables an active design of customer journeys that ensure HCPs have the information and inventory they need to serve their communities.

Bayesian networks should be seen as a way to provide insights that are underpinned by a historical understanding of customer interactions while allowing for the exploration of unobserved promotion sequences and how these sequences can increase the probability of a positive outcome. The combination of a customer-

centric approach with Bayesian network modeling can help optimize intermediate campaign successes, Rx, and customer satisfaction by quantifying the impact of promotional interactions. Bayesian network modeling informs the next best action in a current customer journey, allowing the mapping of the impactful sequences of promotions, and the attribution of promotional channels to outcomes. The goal is to create a data-driven approach to customer journeys that can help facilitate and guide discussions among commercial operations teams.

Bayesian networks provide insights based on the lived experience of customers that facilitate the design and orchestration of promotional activities across channels.

About the Authors

Teis Kristensen *is a Project Lead at Axtria Inc. He has a PhD from Rutgers University. His experience includes marketing mix, salesforce structure & sizing, territory alignment, and people placement projects for Fortune 500 companies. He has developed omnichannel frameworks, built next best action models, customer segmentations, demand forecasting, a/b testing, market size and competitor share estimations, and market access analytics. He acts as an experience advisor in executive decision processes, has a proven record of leading technical teams, and ensures operational excellence.*

Kritika Singhal *is a Data Science Manager at Johnson & Johnson MedTech. Previously, she was a Senior Associate at Axtria Inc. She has a PhD in Mathematics from Ohio State University. Her experience includes using descriptive and predictive analysis methods for budget optimization, price optimization, segmentation and targeting, RFP mapping and omnichannel orchestration in both pharmaceutical and medical devices industry. She believes in continuous learning and likes engaging with thought leaders and SMEs on developments in healthcare and technology. She is open for discussions on this article.*

Ramesh Krishnan *has a PhD from Lehigh University. He has more than 22 years of life-science experience and an extensive track record of leading mission-critical and highly visible projects for both Fortune 500 clients and start-ups. He has a deep predictive analytics background and is reputed as a thought leader in the analytics and business intelligence community. He blends a strategic focus with on-the-ground realities to deliver enduring solutions.*

References

1 Cohen O, Fox B, Mills N, & Wright P. COVID-19 and commercial pharma: Navigating an uneven recovery. Available from: https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/covid-19-and-commercial-pharma-navigating-an-uneven-recovery [Accessed July 15th 2022]

2 Deloitte. The Great Consolidation: The potential for rapid consolidation of health systems. Available from: https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/great-consolidation-health-systems.html [Accessed December 2nd 2022]

3 Murphy K. A Brief Introduction to Graphical Models and Bayesian Networks. Weblog. Available from: http://www2.denizyuret.com/ref/murphy/intro_gm.pdf [Accessed September 20th 2022]

4 Scutari M, Strimmer K. Introduction to graphical modelling. arXiv. [Pre-print] 2010. Available from: https://arxiv.org/pdf/1005.1036.

5 Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge: MIT Press; 2009.

6 Anderl E, Becker I, Von Wangenheim F, Schumann JH. Mapping the customer journey: Lessons learned from graph-based online attribution modeling. International Journal of Research in Marketing. 2016;33(3): 457-74.

7 Digital Attribution. Google Analytics. Available from: https://support.google.com/analytics/answer/9397590?hl=en#zippy=%2Cin-this-article [Accessed September 23rd 2022]

8 Shapley LS. Stochastic games. Proceedings of the National Academy of Sciences. 1953;39(10): 1095-1100.

Boosting Commercial Performance Through Creation of No-Code AI Pipelines

Gili Keshet, MBA, Head of Content, Verix and Shahar Cohen, PhD, CTO, Verix

Abstract: As the pharma industry is shifting towards specialized medicine, commercializing a brand is becoming increasingly intricate. Companies search for ways to optimize their commercial operations to appropriately deal with this complexity. Artificial intelligence has already proven to allow commercial acceleration, though building long-lasting processes, based on AI is still challenging. This paper presents the concept of deploying a vertical AI platform for pharma commercial operations. Such a platform generates capabilities on top of key pharma commercialization elements that are common to many commercial use cases. These capabilities enable the generation of robust artificial intelligence pipelines through a simple, no-code user interface.

Keywords: AI platform; AI Pipelines; commercial performance; machine learning

Introduction

The pharmaceutical industry is rapidly shifting towards specialized medicine, with new drugs more targeted and effective than ever. However, while focusing on more specific indications, the size of the addressable market for any specific drug is shrinking. When pharma companies are competing on shrinking targets, they must adopt new, increasingly accurate methods for designing their commercial operations.

Traditionally, pharma companies used simple heuristics to design their commercial processes. These simple heuristics, such as prioritizing physicians according to their volume-based metrics (TRx, patients' volumes, etc.), were very effective for "blockbuster" drugs, but often don't scale well to dense, specialized and highly competitive markets, since in such markets the behaviors are more complicated and driven by a multitude of varied parameters.

Artificial Intelligence (AI) and specifically, Machine Learning (ML) allow for training of complex models based on abundance of data. These models can be embedded in commercial processes to significantly optimize them. Although the promise of AI & ML in building optimal decisions based on data is already well proven, establishing continuous AI-based processes is still a considerable challenge. Handcrafted models do not scale well and often become a one-and-done. These models can be built once to prove a concept but fail to evolve and maintain long-lasting business value.

One of the key differences between a one-time handcrafted model and a continuously optimized, model based, business process, is the reliance on robust pipelines. A pipeline is a technology-based implementation of a sound process that streamlines all requirements to ensure continuous model health:

I. Data preparation and data QC.
II. Model training (and re-training), including hyper-parameter optimization.
III. The ability to use the model for serving production.

IV. Monitoring the performance of the model.
V. Delivering end-to-end business applications to different business users.

The capabilities for building such pipelines are typically available on top of cloud services platforms that embed the necessary data foundation and AI facilities. However, when attempting to build AI pipelines on top of horizontal, general purpose cloud services platforms, they turn out complex to consume, manage and integrate, and end up implying extremely high total cost of ownership.

In this paper we present the concept of developing a vertical AI / ML based platform, dedicated specifically for pharma commercial acceleration. Such a platform leverages its vertical centricity by focusing on key process elements that are common to many commercial use cases in pharma companies and building automations around these elements. With these automated key elements, the platform enables building robust AI / ML pipelines without the need to write a single line of code. We describe the key elements that the platform automates and demonstrate the building of commercial applications, based on these key elements.

The rest of this paper is organized as follows: Section 2 describes the engineering aspect of this type of platform; Section 3 introduces the concept of key pharma commercialization elements that facilitates the quick and easy definition of pipelines without writing code; Section 4 demonstrates how these elements can be integrated to build an application; Section 5 describes a real-life example of an implementation of this type of platform in a pharma organization; and finally, Section 6 summarizes and concludes the paper.

1. Platform Engineering

The platform we present is based on three key pillars that are elaborated on in the following sub-sections: Data Foundation; AI Engine; and Workflow Builder.

1.1 Data Foundation

The basis of the platform is a data foundation, a set of capabilities that support data collection, integration, and management. First and foremost, the data foundation includes a set of databases and database schemas that describe key entities that are typically involved in commercial operations of pharma companies. Specifically, these schemas include tables to describe customers (see Section 2.3 below). The data foundation also includes system tables and other technical tables that support the smooth operation of the entire platform: Tables to manage AI / ML models, tables to store models' results and tables to support reports and dashboards. In addition to data tables, the data foundation includes mechanisms that support data ingestion, integration, and preparation. Some aspects of these mechanisms are further described in Section 2.1, below.

1.2 AI engine

The platform's AI Engine is comprised from a set of technologies that support the execution and maintenance of the AI / ML models. The AI models operate on top of standard containers that run Python images. Each container has an API that is used to invoke it and run predefined procedures that use the content of the container. These procedures perform functions such as training ML models, applying a model on new data, calculating model health quantities, and so on.

1.3 Workflow builder

The workflow builder is an application generator that facilitates the design and implementation of software to support business processes. The core of the workflow builder is a studio environment that allows the user to design screens, reports, and dashboards, and define a flow between them. The flow may include user navigation between screens, application of various calculations, such as the training of models, sending results to database tables, and using timers to schedule different tasks. The studio generates user interfaces to the business users and includes all the needed maintenance processes in a seamless manner.

2. Key Pharma Commercialization Elements

The platform was designed specifically to optimize commercial operations in pharma companies. It leverages its vertical centricity by focusing on key Pharma Commercialization Elements (PCEs) that are common to many commercial use cases in pharma companies. By addressing these key PCEs, the platform produces a pipe, or a pipeline component that is created and continuously maintained in the backend. Using the workflow builder (see Section 1.3, above), pharma companies can assemble these pipeline components by using a simple, no-code interface. In this section we give examples to some of these key PCEs.

2.1. Data ingestion and integration

Commercial operations in pharma companies attempt to leverage data, from a wide variety of sources, to support business decisions in an educated manner. The different data sources can be classified into three key types:

I. Internal data sources: Data sources that are collected by the company's information systems, such as the organization's CRM system.
II. Acquired data sources: Data sources that are purchased from 3rd party life sciences data aggregators.
III. Service data sources: Data sources that are collected by ad hoc activities, like primary research data.

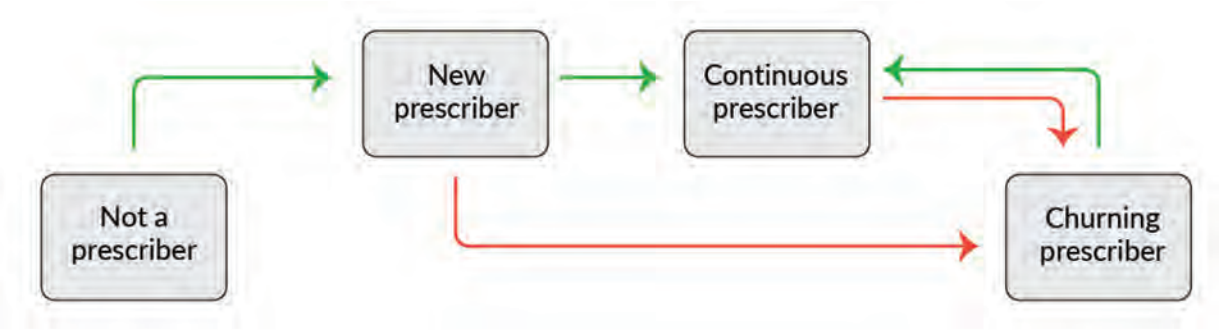
Although there are many different combinations of data sources and data files, there are specific structures of data that are very common and cover a significant part of the industry. For example, many pharma companies use the same popular CRM system that produces specific data structures. Moreover, many times it is the case that competing data sources produce compatible data conventions, as in the case of Anonymous longitudinal Patient-Level Data (APLD). The APLD data structures of competing aggregators tend to have a similar form.

Based on these commonalities, the platform can automatically ingest data files from any significant data source in the industry. For each such data source, the platform automatically recognizes the data's granularity as well as all data variables and can apply pre-designed procedures for quality control and integration of the data. When a non-recognized ad hoc data source has to be incorporated, the platform supports its integration with the other sources through standard SQL or Python based scripts.

2.2 Customer lifecycle modeling

By nature, commercial operations are highly focused on customers. The platform includes a simple, general model that portrays customer behavior. For different brands, the identity of the customer might differ. In general, there are several customer entities, where the most common entities in

Figure 1: CLM: Stages and the Possible Migration Paths Between Them for the Physician Customer Entity



the pharma industry are patient, physician, and account. All types of customers evolve through a series of lifecycle stages. For example, the physician customer entity typically evolves through the following stages (Figure 1):

I. "Not a prescriber" - a physician who is relevant to the brand in question, yet still never wrote it to any patient.
II. "New prescriber"- a physician who just recently prescribed the brand for the first time.
III. "Continuous prescriber" - a physician who writes the brand on a continuous basis.
IV. "Churning prescriber"- a physician who used to write the brand but is not doing so anymore.

Physicians, and customers in general, can migrate between some of the stages, though not between others. The platform supports a flexible definition of lifecycle stages for multiple customer entities.

2.3 Customer 360 tables

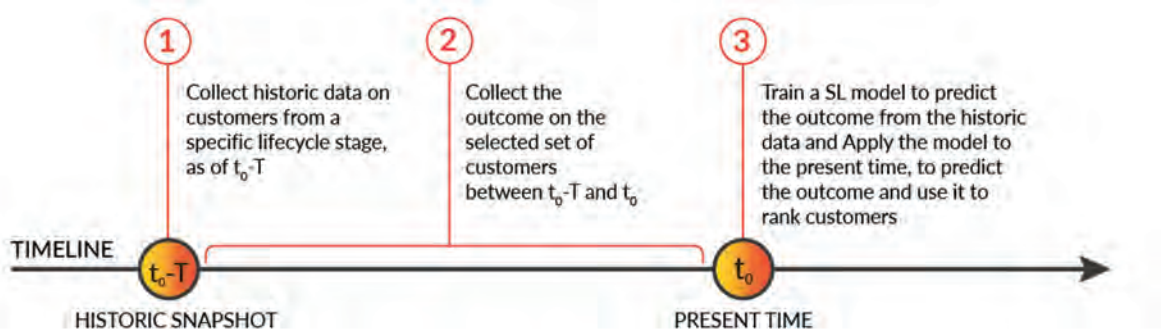
Given the centricity of customers in commercial operations, the data foundation (Section 1.1) includes dedicated customer panoramic tables, called customer 360 tables. Several 360 tables can be built automatically from the ingested data (Section 2.1) to represent different customer entities in different lifecycle stages (Section 2.2). Each row in a 360 table represents a specific customer at a specific point in time, and the columns hold all customers' characteristics.

2.4 Supervised learning with auto-ML

The relatively standard ML task of Supervised Learning (SL) is well suited for many commercial processes, such as the prioritization of marketing activities directed at physicians. The platform allows SL models to be trained on data that describes customers from a specific lifecycle stage and measure quantity or binary outcome that is either advantageous or dis-advantageous to the customers in that stage. For example, when considering the physician customer entity, a SL model can be trained on historic data of physicians, where the explaining variables are any variables available for non-prescriber physicians (their specialty, volumes of patients, competition market shares, etc.), and the outcome might be the event of migrating to the "New prescriber" lifecycle stage.

Upon the ingestion of raw data, its integration, and the building of 360 tables, the platform automatically detects the

Figure 2: Training a SL Model on Customers in a Specific Lifecycle Stage



variables that might contain outcomes, and allows the user to automatically train SL models accordingly. On the course of optimizing such a model, the platform automatically runs a heuristic to select the best SL algorithm and the optimal set of algorithms hyper parameters. (Figure 2)

2.5 Cluster analysis and segmentation

Cluster analysis techniques have been well-proven for customer segmentation in commercial applications in other vertical industries. The platform includes several cluster analysis algorithms that can be applied on the customer 360 tables, to create customer segments. Prior to the cluster analysis itself, different data normalization techniques are applied to support different types of customer variables. When applying the algorithms, the platform uses a heuristic that recommends the optimal number of clusters.

Cluster analysis and its corresponding segmentation can be defined on multiple sets of customer variables, called segmentation dimensions. Each customer can be assigned to a different segment in every dimension (i.e., if two customers are assigned to the same segment in one dimension they will not necessarily be assigned to the same segment in another dimension). The platform may receive importance-ranking between dimensions, and automatically create a set of customer micro-segments that reflects similarity in multiple dimensions.

2.6 Time series

The platform implements several time-series algorithms (e.g., ARIMA, Holt-Winters, Prophet and others). The algorithms can be applied to any time series. Using these time series algorithms has been shown to produce a powerful bottom-up forecast that supports the process of planning and incentive compensation goals. The platform also automatically evaluates different sorts of calculation error, including root mean square error, mean-absolute error and mean-absolute-percentile error. Similar to the SL auto-ML mechanisms, the platform includes a built-in heuristic to select the optimal time series algorithm.

2.7 Markov model and LTV

LifeTime Value (LTV) of customers, is an important metric that characterizes customers. It is the Net Present Value (NPV) of the (discounted) future flow of profits from each customer. The company seeks to maximize that NPV, by taking optimal marketing and sales decisions in its interactions with the customer. The platform includes a Markovian model

to project customers' LTV. It observes the lifecycle journey of customers over a relatively short series of time periods (typically months or quarters), looks at the immediate reward and the transitions between states, and extracts a LTV figure for each customer.

3. Building Pipelines Using Key PCEs

In the previous section, we described the concept of key PCEs, and showed several important examples for such PCEs. Each key PCE is equipped with an automation mechanism that produces a pipeline element. Using the workflow builder, the arsenal of key PCEs can be used to build end-to-end applications, based on a set of robust backend pipelines.

For example, to build a targeting application, a pipeline can be built using the following PCEs:

I. **Data ingestion and integration:** supply the desired data sources, let the platform automatically map the variables in the different files, integrate and QC it. If needed, manually integrate additional ad hoc data files.
II. **Customer lifecycle modeling:** define customer entities and lifecycle stages for each entity.
III. **Customer 360 tables:** let the platform automatically produce 360 tables for the defined customer entities. If needed, manually add more customer descriptive variables.
IV. **Supervised learning:** select the different predictive variables that you want to model. Invoke the auto-ML capability to run a search heuristic and conclude which would be the best-performing model for this pipeline.
V. **Build a tiering mechanism** that uses the model outputs.
VI. **Execute, measure, and iterate** on an ongoing basis

Notice that although it is a highly modular procedure, such a targeting application may take many different forms, due to the definition of different customer entities, different lifecycle stages, SL models to different outcomes, and the application of different tiering logic. Variations are created in minutes and easily evaluated on historic data. Once a variation is verified, it is immediately implemented.

Adding LTV to such a targeting application enhances the precision of the decision engine by adding a segmentation perspective to the application or by extending it to digital channel optimization. Moreover, one can easily see how other key PCEs can be integrated into the platform to create other applications, such as forecasting.

4. Case Study

In this section we describe a real implementation of the platform at a Top 5 US pharma company for dynamic targeting of physicians, through several channels, including both personal and nonpersonal promotions. The described case study is focused on an oncology, targeted-therapy brand that has two main competitors. The following sections describe the key PCEs deployed in the case study, and the pipelines that were built in the back-end by providing the specifics of these PCEs.

4.1 Case study data ingestion and integration

Several dozens of different datasets were ingested by the platform, including:

I. Anonymous Patient Longitudinal Data (APLD), from multiple sources
II. Special Pharmacy (SP) data

III. Special Distributer (SD) data
IV. Call activity data
V. Drug samples data
VI. Non-personal channel data
VII. Anonymous patient diagnosis data
VIII. Participation of physicians in speaker events
IX. Territory alignment data
X. Data from primary research

Automatic data QC procedures were applied on the input datasets, followed by the application of the automatic integration methods. Very little manual intervention was needed in the process. Once validated, these processes produced a data ingestion and integration pipe.

4.2 Case study customer lifecycle modeling

Two different customer entities were defined: physicians and patients. The default lifecycle stages were used, consisting of four different stages for each entity.

- Physician lifecycle stages are: Non-prescriber, New prescriber, continuous prescriber and churning prescriber
- Patient lifecycle stages are: Non-user, New user, Continuing user and Former user

The definition of the lifecycle stages created a pipe that sorts the different customer entities into their appropriate lifecycle stages.

4. 3 Case study customer 360 tables

Two different sets of Customer 360 tables were automatically created, corresponding to the two different customer entities. Some of the variables in the 360 tables only apply to specific lifecycle stages. For example, customer variables that are based on SP data only apply to physicians that already prescribed the brand. The 360 tables include hundreds of different variables. For example, the physician dataset includes variables such as different volume indicators (TRx, NRx, existing patients, new patients, volumes of competitors, etc.), physician specialty, preferred treatment approach, distribution of patients' ages and lines of therapy, information on discontinuous patients, call activity, non-personal activity and more. Patient's dataset included variables like gender, age, duration of treatment, line of therapy, payer information and more. More than 90% of the variables in the 360 tables were created automatically. The variable creation pipe was joined to the previous pipes in the process.

4.4 Case study SL models

A dozen of SL models were trained to predict the probabilities of different events to occur. Each such model was automatically optimized. As part of the optimization process, the models automatically went through a cross-validation procedure that examines several important health parameters, such as area under the ROC curve, accuracy, recall, precision, and different uplifts. The SL models were assembled as another pipe.

4.5 Case study segmentation

Several dimensions of segmentations were defined over the physician entity. The dimensions were designed according to a set of variables that were recognized as important to the specific drug and its competition, as well as variables that capture general volume-based characteristics. The segmentation part was attached to the previous pipeline, as a parallel line that enriches the Customer 360 tables.

4.6 Case study physician tiering

The different SL models were translated into ranking rules and specific lists for different personal and non-personal campaigns. Physician segmentation was used to match the most appropriate messaging and type of campaign to each physician, and the results were assigned to field reps for execution. The tiering pipe was attached to the models and the Customer 360 tables and provided a source to the user interface.

4.7 Implementation time

The described case study was designed and implemented over a period of 7 weeks. After several weeks of POC, the solution went through several limited adaptations, and required 4 more weeks before moving into production.

4.8 Offline evaluation of results

To evaluate the effectiveness of the SL model-based dynamic targeting produced by the platform, we compared the resulting lists of prioritized population with those produced by the brand's existing traditional targeting logic. Both methods were applied on historic data, implemented at the same historic timepoint. The results showed a discrepancy of 15% between the lists. The performance of this 15% incompatible population of physicians was analyzed, based on the different campaigns they were part of, and evaluated according to the desirable results of each campaign. The comparison showed a clear advantage to the SL model-based target lists, with an average uplift of 32% in the KPIs defined to evaluate the campaigns' success, and subsequently, the field reps' success in engaging with these physicians.

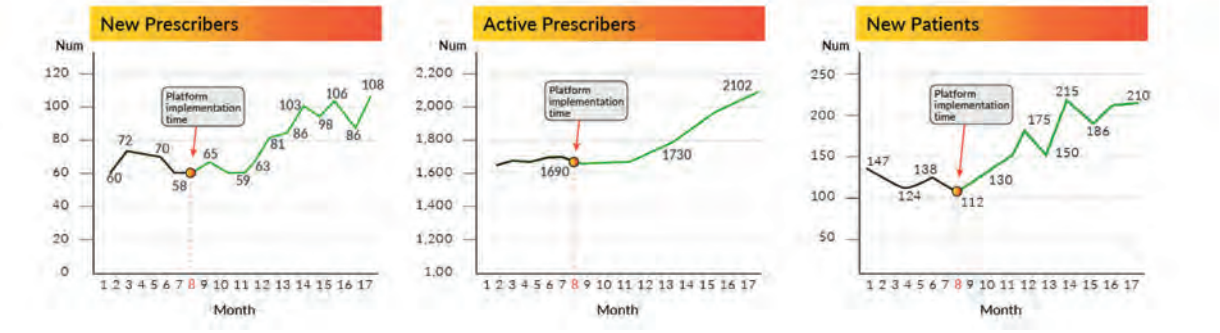
4.9 Long-term results – After 18 months

18 months after the model has been implemented at production level, we designed a statistic test, to check the impact of the new targeting mechanism. The test was based on a synthetic-controls methodology. We compared the performance of a test group of field reps that adopted the SL based recommendations, with the performance of a control group of field reps that adhered to the old methods. The control group was synthetically assembled from multiple reps, to assure similarity to the test group. Performance KPIs on the test group were compared to a synthetic weighted combination of the different control groups. This examination showed that the test group of field reps that adopted the SL model recommendations achieved an uplift of 20% -30% in leading KPIs, clearly outperforming the control group of low adopters. The synthetic control study also showed a significant return on the company's investment in the new technology: 32x ROI in 6 months. The statistical results also align with an actual improvement in the main business KPIs: more new prescribers, more continuous prescribers, and more new patients to the brand. (Figure 3)

Summary and Conclusions

In this paper we present an AI / ML platform that was specifically designed to support the optimization of commercial processes in pharma organizations. The platform is built around key capabilities that are common to many use cases in pharma commercial operations. With that approach, it exposes a simple, no-code user interface, which allows business users to

Figure 3: Synthetic Randomized Controls for Different Groups of HCPs



build end-to-end applications and their corresponding backend pipelines, quickly and easily. As seen in the real live case study from a Top 5 US Oncology franchise, this platform managed to lower the burden of actively developing AI/ML based automated commercial processes, and significantly boosted the commercial performance of brands, on a long lasting, continuous basis.

Gili Keshet, MBA *is Head of content at Verix, provider of commercial optimization solutions for the Pharma industry, and is a marketing veteran with over 20 years experience in hi-tech marketing, strategic planning, and business development. Gili is the founder of Collage Marketing, a strategic marketing consulting firm that helped dozens of hi-tech companies shape their strategy, develop new markets, and acquire key accounts. Gili has an MBA from Santa Clara University, CA and B.Sc. in Mathematics and Computer Science from Bar Ilan University, Israel.*

Shahar Cohen, PhD *is CTO at Verix, provider of commercial optimization solutions for the Pharma industry. A data scientist researcher, practitioner, and serial entrepreneur, Shahar has always focused on improving and automating core business processes using data and algorithms. He is the founder of three AI-focused companies, and has a PhD (summa cum laude) from Tel Aviv University, Industrial Engineering.*

A Causal Modeling Approach for Estimating the Impact of Predictive Customer Recommendations

Sri Krishna Rao Achyutuni, Senior Data Scientist, ZS Associates and Srinivas Chilukuri, Principal Data Scientist, ZS Associates

Abstract: In recent years, predictive models based on machine learning have become increasingly popular for recommending target customers for field execution. While these predictions can be impactful, quantifying and establishing the statistical significance of these predictions has not been straightforward. Typically, the test-control method using double-difference is used, but this has several limitations - volume of activity is not considered, subjectivity is involved in finding controls, interactions between different interventions are not considered, to name a few. On the other hand, incorporating these factors into statistical attribution models also has limitations - they are not meant for inferring causal dynamics, they cannot handle data sparsity well, and when the field execution deviates from the recommendation, which inevitably is the case, they won't be able to control for the observed and unobserved confounders. As a result of these limitations, it has not been possible to generate counter-factual scenarios i.e. alternative execution scenarios to estimate the impact and revise the planning accordingly. Causal models address the above limitations and can provide a solid foundation for not only measuring the impact but also for enabling planning for future execution. In this paper, we lay out the approach for such a causal model and demonstrate how this works with some examples. We hope this paves the way for data scientists and analysts to use causal models for robust impact measurement and field planning.

1. Background

1.1 Introduction and Motivation

Pharmaceutical companies strive to enhance their engagement with healthcare providers (HCPs), to enable them to make the right decisions for their patients and thereby drive sales and revenue. Increasingly, predictive models based on machine learning algorithms are used to identify the right customers and to recommend best ways to engage them. This is driven by a few factors like growing competition, adoption of digital channels in addition to the in-person sales channels and also the evolution of treatments being targeted towards a niche group of patients e.g. rare disease / orphan drugs. In these cases, it is critical for the pharmaceutical brands to reach the right customer at the right time through the right mix of channels. A typical commercial field execution plan involves a list of target HCPs recommended by predictive models based out of machine learning algorithms. These models are developed over a diverse set of behavioral attributes such as historical engagements, channel preferences, market influence, patient accessibility, demographics and so on. These models are usually optimized for one or more business key performance indicators (KPIs) such as new prescriptions (NRx), total prescriptions (TRx), new-to-brand prescriptions (NBRx, if it is a newly launched drug) to name a few. The expectations are set for the various channels to reach the recommended customers for the recommended frequency. However, the execution does not always adhere to planning. HCPs who are recommended are sometimes skipped and instead replaced with non-recommended HCPs which leads

to some challenges in reliably quantifying the impact of the predictions underlying the recommendations.

Business leaders want to understand the impact of these predictive models on the brand performance and decide whether to continue to invest in scaling and refreshing them and/or if they need to course correct. Typically, this is done using test-control for one-time during a pilot. This requires careful selection of test and control territories/regions and poses challenges in understanding the impact at a granular level.

For more granular understanding, typical statistical models like regression are used. However, these approaches rely on association/correlation and therefore could be overestimating the actual impact. Besides, when the signal is quite sparse in cases like rare disease treatments, the sparsity in the signal further conflates the underlying assumptions of regression-based approaches and leads to misleading conclusions.

To summarize, there are two questions to address that are of business relevance– (a) What is the true impact of recommendations in driving business outcomes (*factual analysis*)? And (b) What would have been the impact, if the execution occurred as per the recommended plan (*counter-factual or "what-if" analysis*)? This counterfactual helps in proving the importance of adherence by field teams and as well validate the utility of machine learning models in driving HCP engagement and outcomes. This paper aims to discuss different approaches towards estimating the post-execution impact of recommendations and provide empirical guidance for business leaders and data scientists who would be undertaking such endeavors.

1.2 Methods for impact measurement

A typical field execution plan is designed on a Target list of customers (HCPs in our case). To evaluate the effectiveness of the Target list, a one-time randomized Test-Control is performed. Ideally, the customers in Target and Non-Target list are each randomly stratified into groups of Test and Control. The Target and Non-Target customers of Test group receive the treatment (field activity in our case), whereas the Control group receive none. However, due to factors such as marketing budget constraints and limited availability of field time, it becomes nearly impossible to exactly execute such as an ideal experiment. Brands end up running a limited version of this experiment pertaining to the Target customers only. Here, the Target customers are stratified into groups of Test and Control, and those belonging to Test are exposed to the field, whereas those in Control are withheld, even though they are eligible for exposure. A simple approach to evaluate the effectiveness of this experiment can be "*difference-in-difference*" (also known as "*double-difference*"). It is an analytical approach which facilitates causal inference. Here, the first difference is performed over time for each Test and Control groups, individually capturing their time-varying effects. Then, a subsequent difference is performed between the time-captured effects of Test and Control groups, to estimate the final impact. The estimation and reliability of these impact estimates relies on multiple assumptions such as (a) no time-varying differences exist between Test and Control groups (*equal trends assumption*), (b) no presence of any selection bias induced during execution of the experiment, (c) availability of data in both pre vs/ post experimentation windows,

Table 1: Summary of Recommended v/s Reached in Observational data (exact numbers perturbed a bit to maintain anonymity)

Model Recommended	Field Execution	# HCPs	% Prescribed
Yes	Reached	1000	3% [A]
Yes	Not Reached	5000	0.3% [B]
No	Reached	7000	0.7% [C]
No	Not Reached	10,000	0% [D]

(d) setting aside a deliberate control group can arguably leave opportunity on the table just for the sake of impact measurement, which can sometimes be difficult to operationalize.

However, the use-case we are considering for this paper has scenarios which contrast some of the above assumptions - (a) the recommender model used in generation of target list is available right from the beginning of the promotional campaign launch (for e.g. in the case of a new launched drug), and / or (b) a rare-disease drug execution where the target customers are in a long tail and don't have any ongoing engagement unlike typical mass market / specialty market brands, which means there is no historical activity or prescribing behavior to look at and / or (c) though the promotions were active prior to the launch of the recommender model but the capture rate was too sparse for any machine learning (ML) use-case (i.e. existing drug). In these scenarios, all we have is the promotional activity and the prescribing behavior post the roll-out of the ML based recommendations. Table 1 summarizes the observational data considered in this paper.

Following are different methods considered for impact measurement -

- **Test-Control (Post-Hoc Comparison)** is a macro level association-based multiple comparisons method, using statistical significance tests like T-Test, Chi-Square test and so on. Using the above Table 1, comparing [A] with [B] helps us estimate the missed opportunity of field execution not adhering to the recommended target list. Similarly, comparing [A] with [C] elucidates the value of the recommender model in identifying target list for effective returns (higher % prescribed). Though this approach is simple, intuitive, and business friendly, the inference from this analysis has limitations – (a) they don't always conform to population or randomization. Even a simple coincidence can sometimes show statistical differences [30] (b) they don't account for selection bias, especially in presence of any mediation. In our use case, field execution mediates the recommended target list with the outcome. Without field execution, there is limited benefit one can achieve using the recommender model.
- **Regression** is a micro level association based statistical attribution model, trained at an HCP level. It is often used for marketing mix planning. In the context of analyzing the impact of recommendations (*Prediction*) using the field execution (*Activity*), we have to incorporate the following equation –

$$Y = \alpha + \beta_1 * Prediction + \beta_2 * Activity + \beta_3 * Prediction * Activity + \epsilon$$

The intercept α defines expected prescription writers in absence of both *Prediction* and *Activity*, whereas the error ϵ indicates unmeasured uncertainty. In this we are not only looking for the individual impact of *Prediction* and *Activity* alone but are also evaluating the "*interaction effect*" where there are synergies between the two. Practically, *Prediction* alone won't lead to prescribing behavior because it needs field *Activity* (mediation) to influence customer prescribing behavior. So, we can expect β_1 to be statistically insignificant. However, if β_3 is positive and significant, it shows that prediction enhances customer engagement and leads to better prescribing behavior. In case of binary outcomes, logistic regression can be an option, where the exponent of the coefficients explain the odds of the outcome (% drug adoption). Although these parametric models are simple to implement, easy to interpret and robust to large sample sizes, they have their own set of limitations – (a) assumes a constant effect across the population, which is usually a rare observation in reality, (b) returns inconclusive results in cases of data sparsity especially in rare-disease scenarios, (c) not robust to multi-collinearity especially in the presence confounders such as HCP specialties, and (d) predominantly learns association which can either overestimate or underestimate the true impacts.

- **Causal Inference** is a micro level impact estimation method which is designed to uncover the presence of causal relationships between features of importance (Prediction) and the outcome (Drug adoption). By definition, these models estimate the causal impact, by systematically eliminating any spurious correlations and bring out reliable signals. In the next section, we deep dive into the specific formulation of the causal model that we propose for our use case.

1.3 Primer on Causal Models

Causal models are mathematical models used in evaluating and establishing the presence of causal relationships. Structural equation models (widely known as SEMs) help characterize causal systems with a set of variables and equations. These equations determine the causal linkage between each variable with its immediate predecessors. Also known as structural causal models (SCM), SCMs are widely accepted for its adeptness to combine structural equations with directed acyclic graphs (DAG) to better estimate the causal effects. A structural causal DAG is a representation of directed nodes with no cyclicity. Representing the knowledge as a DAG not only helps in visually validating the domain understanding but can also be used to generate artificial or synthetic data using data-generation process (DGP) [9]. Figure 1 illustrates a simple DAG of a Treatment *T*, Outcome *Y* and Confounder *W*. Here, we assume that *W* may affect *T* and both *W* and *T* may affect *Y*. Additionally, we assume that all the three (*W, T, Y)* may all share some random unmeasured common causes (*U*) [9]. See Figure 1.

SCMs can be represented using the following equations. The directed edges allow for the existence of causal relationships between the variables, with very minimal assumptions on the distribution of unmeasured factors (*U*) and functional form on causal relationships (*f*).

Figure 1: Illustration of a Simple Causal Graph, with a Treatment (exposure) T, Outcome Y and Confounders (Co-variates) W [9]

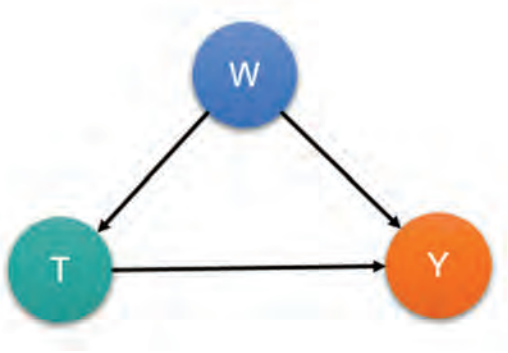
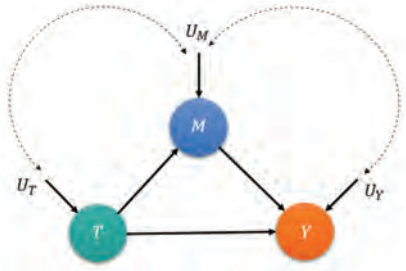


Figure 2: Illustration of a Simple Confounded Mediation Graph, with a Treatment (Exposure) T, Outcome Y and Mediator M, with Dependence Between the Unmeasured Factors of U_M with U_T and U_Y



$$W=f_w(U_w)$$
$$T=f_T(W, U_T)$$
$$Y=f_Y(W, T, U_Y)$$

where U_w, U_T, U_Y represent unrestricted random distributions and f_w, f_T, f_Y represent arbitrary functions

However, these causal graphs fail to unravel effects in the presence of any underlying factors mediating the association of Treatment with the Outcome. As described under Context (section 1.2), the field execution team mediates the causal linkage between the variable of interest (i.e. Customer Target list - Treatment) and its response (i.e. Drug adoption - Outcome). In other words, until the field execution team intervenes with their promotional outreach, the establishment of causal linkage between customer Target list and Drug adoption is futile. Here comes structural causal mediation models.

[16] In the simplest case, a structural mediation model shall have a Treatment T, Mediator M and Outcome Y, of any discrete or continuous random distribution, as shown in Figure 2. In this directed acyclic graph (DAG), we assume that *T* may affect *M* and both *T* and *M* may affect *Y*. Following illustrates its functional forms:

$$T=f_T(U_T)$$
$$M=f_M(T, U_M)$$
$$Y=f_Y(T, M, U_Y)$$

where U_T, U_M, U_Y represent unrestricted random distributions and f_T, f_M, f_Y represent arbitrary functions

Representing the causal linkage as a DAG has multiple benefits – (a) helps in visually validating the domain understanding and (b) can be used to generate artificial or synthetic data using a data-generation process with defined functional forms.

In CMA [6][8][2][11][12], causal effects are usually measured relative to a specific contrast of interest in the treatment *T*. In our case-study, the Treatment is an ordinal attribute of 4 levels – Tier 1 (T1), Tier 2 (T2), Tier 3 (T3) and No Tier (NT), with T1 and NT representing the best and worst recommendations for field outreach. The treatments of interest are T1, T2, T3 in contrast to NT. For the sake of below explanations, let's consider the transition from *T=NT* to *T=T1*. Broadly, there are 4 types of causal effects –

(a) Natural Direct Effect (NDE) – It is a measure of expected change in outcome *Y*, for the entire population, as the treatment transitions from *T=NT* to *T=T1*, while setting the mediator *M* for each individual prior to the transition, i.e. at *T=NT*.

$$NDE=E[Y_{T1,M_{NT}} - Y_{NT,M_{NT}}]$$

(b) Natural Indirect Effect (NIE) – It is a measure of expected change in outcome *Y*, for the entire population, if the treatment is held constant, either at *T=NT* or *T=T1*, and the mediator *M* (for each individual) changes from the value what it would have been at *T=NT* to whatever value it would have attained at *T=T1*.

$$NIE=E[Y_{T_t,M_{T1}} - Y_{T_t,M_{NT}}] \text{ where } t \in \{NT, T\}$$

(c) Total Effect (TE) – It is a measure of expected change in outcome *Y*, for the entire population, when the treatment transitions from *T=NT* to *T=T1* , while the mediator *M* is allowed to track the changes in treatment as per its functional form f_M.

$$TE=E[Y \mid do(T=T1)] - E[Y|do(T=NT)]$$

(d) Controlled Direct Effect (CDE) - It is a measure of expected change in outcome *Y*, for the entire population, when the treatment transitions from *T=NT* to *T=T1*, while the mediator *M* is set to a pre-defined level (*m*) of interest uniformly across each individual.

$$CDE_m=E[Y \mid do(T=T1,M=m)] - E[Y \mid do(T=NT, M=m)]$$

Theoretically, NDE measures the portion of TE which is explained only by *T* without the interference of *M* influencing *Y*, while NIE measures the portion of TE which is explained only through the mediation *M*, without any direct influence of *T* on *Y*. Hence, in terms of structural equations –

$$Y= \alpha_1+ \beta T+ \varepsilon_1 \quad (1)$$
$$Y= \alpha_2+ \beta' T+\gamma M + \varepsilon_2 \quad (2)$$
$$M= \alpha_3+ \delta T+\varepsilon_3 \quad (3)$$

Where α_i denotes intercept and ε_i denotes uncorrelated residual where $i \in \{1,2,3\}$

Now, substituting Eq. (3) with Eq. (2),

$$Y= \alpha_4+(\beta'+\delta\gamma)T + \varepsilon_4 \quad (4)$$

where $\alpha_4= \alpha_2+ \gamma\alpha_3$ and $\varepsilon_4=\varepsilon_2+ \gamma\varepsilon_3$

As per Eq. (4), the parameters β' explain NDE and $\delta\gamma$ explain NIE of *T* on *Y*, mediating on *M*.

2. Experiment Setup

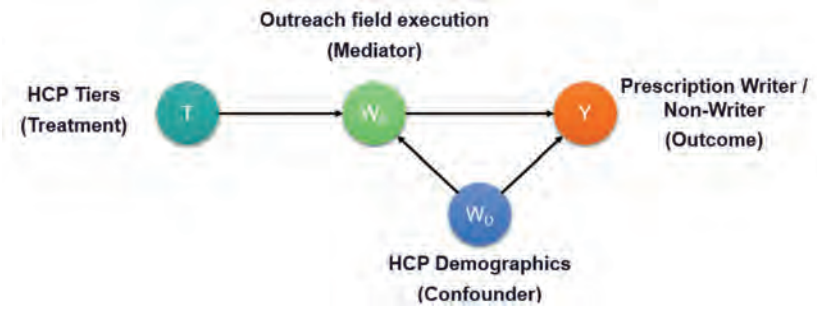
2.1 Data

The experimental results in this paper are based on two data variants – (a) an artificial data using Data Generation Process (DGP)

Table 2: Summarizing Data Types of Different Attributes of Causal DAG, Used in this Paper

Variable	Data Type	DGP Data	Observational Data
Treatment	Ordinal	Variable of 3 levels • Tier 1 (T1) • Tier 2 (T2) • No-Tier (NT)	HCP Target Tiers • Tier 1 (T1) – best cohort • Tier 2 (T2) • Tier 3 (T3) • No-Tier (T4)
Mediation	Continuous Integers	4 Variables	# Outreach via 4 channels
Outcome	Binary	1 variable w/ 2 levels (0,1)	Drug adoption (0,1)
Confounders	Binary	4 variables each w/ 2 levels (0,1)	Specialty (One-hot encoded)

Figure 3: DAG Used for the Causal Analysis of Both Artificial Data (using DGP) and Actual Observational Data



and (b) an actual observational data from a pharmaceutical field planning team, illustrated in section 1.2. Each dataset DAG consists of four kinds of nodes, as summarized in Table 2.

2.2 Inference Objective

Our inference objective is two-fold – (a) What is the quantifiable impact of customer recommendations in driving drug adoption (*factual analysis*)? and (b) What would have been the expected uplift if the field reached out to T1,T2,T3, instead of reaching to NT (*counter-factual or "what-if" analysis*)?

To address the objectives, we considered the following DAG, as shown in Figure 3. HCP Tiers was considered Treatment *T* (ordinal), the field outreach as Mediator *M* (continuous), HCP Specialty as Confounders *C* (binary) and Drug adoption as Outcome *Y* (binary). The Treatment directs the Mediator, and the Mediator directs the Outcome. On other hand, the confounders direct both the Mediator and the Outcome.

2.3 Impact Estimation Models

To estimate the Average Treatment Effect (ATE), we evaluated two different models – (a) Single stage Logistic Regression

(LOGREG) and Natural effect model of causal mediation analysis (NECMA). We leveraged logistic regression of *scikit-learn* [33] python package to build model (a) and CMAverse [32][21][14] R package to build model (b).

As the outcome is binary, we computed NDE, NIE and TE at the odds ratio (OR) scale [2]. Following equations illustrated computation of OR assuming the Treatment transitions from No-Tier (NT) to Tier 1 (T1).

As the outcome is binary, we computed NDE, NIE and TE at the odds ratio (OR) scale. Following equations illustrated computation of OR assuming the Treatment transitions from No-Tier (NT) to Tier 1 (T1) [2] [8].

Odds Ratio of Total Effect (RTE)

$$OR^{TE}_{T1,NT|C} = \frac{P(Y_{T1} = 1|C) \,/\, \{1 - P(Y_{T1} = 1|C)\}}{P(Y_{NT} = 1|C) \,/\, \{1 - P(Y_{NT} = 1|C\}}$$

Odds Ratio of Natural Direct Effect (RNDE)

$$OR^{NDE_{NT}}_{T1,NT|C} = \frac{P\left(Y_{T1,M_{NT}} = 1\middle|C\right) \,/\, \{1 - P(Y_{T1,M_{NT}} = 1|C\}}{P\left(Y_{NT,M_{NT}} = 1\middle|C\right) \,/\, \{1 - P(Y_{NT,M_{NT}} = 1|C\}}$$

Odds Ratio of Natural Indirect Effect (RNIE)

$$OR^{NIE_T}_{T1,NT|C} = \frac{P\left(Y_{T1,M_T} = 1\middle|C\right) \,/\, \{1 - P(Y_{T1,M_T} = 1|C\}}{P\left(Y_{T1,M_{NT}} = 1\middle|C\right) \,/\, \{1 - P(Y_{T1,M_{NT}} = 1|C\}}$$

The causal model presented in this paper has been implemented using a combination of both R and Python programming languages. In R, we have utilized the CMAVerse package, while in Python, we have employed several packages including dowhy, econml, sklearn (specifically, LogisticRegressionCV), numpy, pandas, string, math, and networkx. By leveraging these diverse libraries and tools, we have effectively constructed and analyzed our causal model to provide valuable insights for businesses.

2.4 Approach Evaluation and Selection

On DGP data, we evaluated the performance by comparing the true TE with the model estimated TE. On the observational data, we performed sensitivity analysis using the measure called "E-value" [1]. E-value is defined as minimum strength of association, on the risk ratio scale, required for an unmeasured confounder to fully counter any specific association between the treatment T and outcome Y, conditional on the measure covariates C [1]. A larger E-value, which is preferred, indicates a considerable requirement of unmeasured confounder to counter the causal impact. Whereas a smaller E-value indicate little unmeasured confounder to counter the causal impact. However, the threshold is subjective and varies across use-cases [1].

3. Results

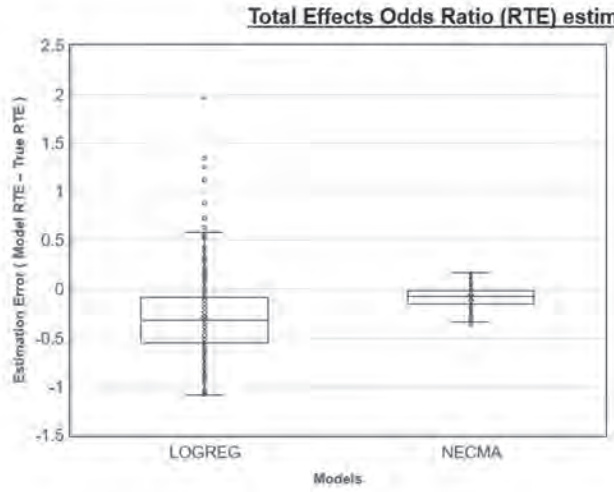
3.1 Results of Artificial data (DGP)

Using DGP and the DAG, as illustrated in Figure 2, 300 artificial datasets of 5000 records each were generated. Each dataset is composed of 1 ordinal Treatment (3 levels with 1 control included), 4 continuous (integer) Mediators, 4 one-hot encoded binary Confounders and 1 binary outcome. Figure 4 illustrates the distribution of Total Effects Odd Ratio (RTE) estimation errors (predicted RTE – true ATE) across LOGREG and NECMA models.

3.2 Results of Actual Observational Data

The actual observational data is composed of (a) an ordinal Treatment of 4 levels of HCP Tier (*Tier-I, Tier-II, Tier-III, No-Tier*), (b) 4 continuous Mediators (*exposure via different outreach channels*), (c) 4 one-hot encoded Confounders (*HCP specialty*) and (d) a binary Outcome (*new prescription writers or non-writers*). Figure 5 illustrates the

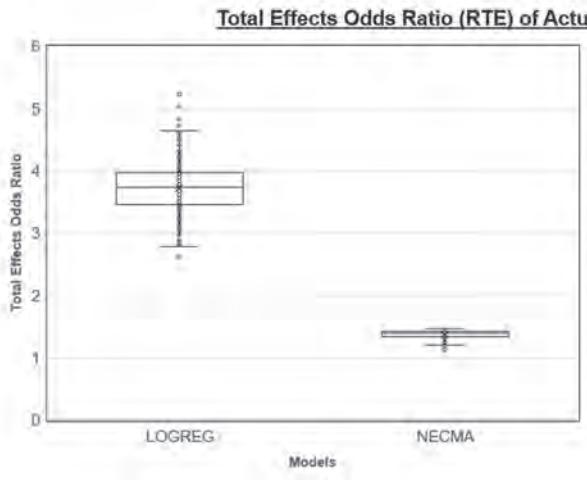
Figure 4: DGP's RTE estimation errors



	LOGREG	NECMA
Average	(0.2627)	(0.0814)
Standard Deviation	0.4288	0.0980
Minimum	(1.0870)	(0.3663)
25th percentile	(0.5469)	(0.1468)
50th percentile (Median)	(0.3184)	(0.0759)
75th percentile	(0.0831)	(0.0170)
Maximum	1.9707	(0.1862)

LOGREG : Logistic Regression with Interaction variables
NECMA : Natural Effects model of causal mediation analysis

Figure 5: RTE on Actual Observational Data, for Treatment Transition of NT to T1 (Best)



	LOGREG	NECMA
Average	3.7246	1.3675
Standard Deviation	0.4143	0.0729
Minimum	2.6287	1.1275
25th percentile	3.4618	1.3290
50th percentile (Median)	3.7313	1.3955
75th percentile	3.9568	1.4176
Maximum	5.2328	1.4624

LOGREG : Logistic Regression with Interaction variables
NECMA : Natural Effects model of causal mediation analysis

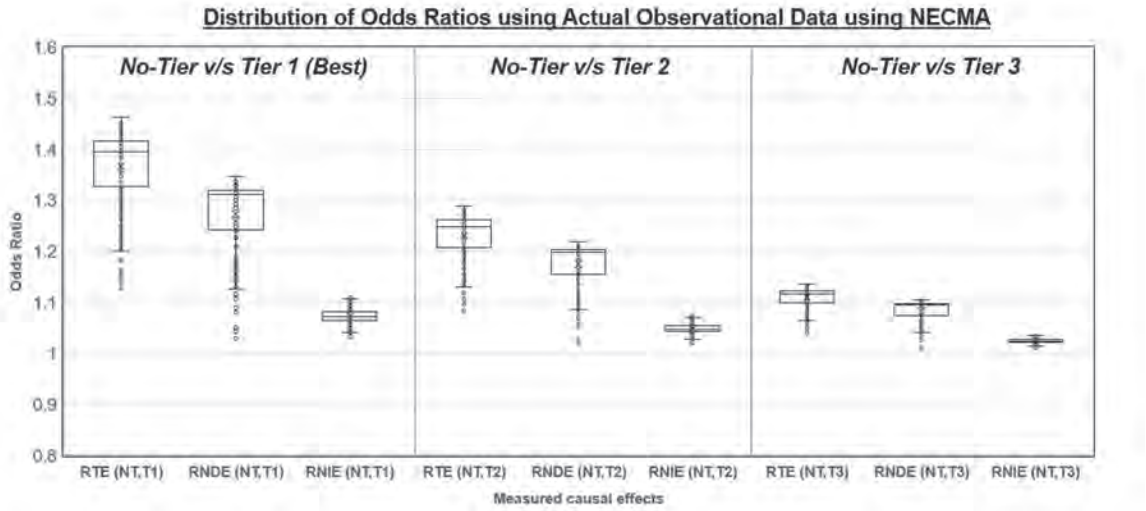
estimates of Total Effects (RTE) using both the LOGREG and NECMA models, bootstrapped over 300 randomly sampled datasets.

Additionally, we generated RNDE and RNIE estimates using NECMA approach. Figure 6 illustrates the distribution of RNDE, RNIE and RTE for 3 different scenarios of Treatment transitions – (a) NT to T1, (b) NT to T2 and (c) NT to T3.

4. Discussion

In this paper, we compared two different approaches for estimating the causal impact between recommended predictions and drug adoptions through the mediation of field outreach programs. The approaches are (a) simple single stage logistic regression model (LOGREG) and (b) natural effects causal mediation analysis (NECMA) from *CMAverse* [32][14]. The models were initially evaluated on a synthetically generated data (DGP). The total effects odds ratio (RTE) of each model was compared with the true-RTE from DGP and the difference in the estimation errors are plotted in Figure 4. NECMA

Figure 6: Measure Causal Effects on Actual Observational Data



The median impacts are observed to be highest for T1 HCPs (RTE: 1.3955, RNDE: 1.3116, RNIE: 1.0719), followed by T2 HCPs (RTE: 1.2489, RNDE: 1.1983, RNIE: 1.0474) and T3 HCPs (RTE: 1.1176, RNDE: 1.0947, RNIE: 1.0235).

outperforms LOGREG, both in terms of bias and variance. NECMA's central tendency measures (average, median, percentiles) are closer to 0 compared to LOGREG's. Additionally, the standard deviation of NECMA (0.0980) is lower than LOGREG (0.4288), elucidating a tighter distribution. We can safely infer that NECMA offers directionally reliable and robust estimates compared to LOGREG.

We then ran both the models on actual observational data. Though we observe higher RTE using LOGREG, we prefer inference based on the results from NECMA because of its performance on DGP. NECMA's RTE indicates a positive causality. The field outreach to Tier 1 HCPs increases the odds of drug adopters by 39.55% compared to the odds of drug adopters among non-Tier HCPs. The E value is 2.138, which means that the observed RTE of 1.3955 could be explained away by an unmeasured confounder that was associated with both the treatment (Prediction) and the outcome (drug adoption) by a risk ratio of 2.138-fold each, above and beyond the measured confounders[1]. Similarly, field outreach to Tier 2 increases the odds by 24.89% and Tier 3 by 11.76%. We believe these causal estimates attribute 18.6% of drug adopters to the top 3 Tiers recommended by the predictive model. We used attribution analysis as described in Appendix A, that can enable return-on-investment (ROI) calculation and provide guidance on how much uplift one can expect by segregating customers across Tiers for field outreach.

One of the key questions from business is on "*what-if*" - what would have been the expected performance if the field reached out to priority Tiers (1,2,3) instead of reaching to non-Tier? To answer, we performed counterfactual simulation analysis using propensity score [35] regression approach [36], as detailed in Appendix B. By simulating a similar outreach to non-reached priority HCPs

based on their reached peers (Tier and specialty), an additional of 37 (median) adopters could have been expected, with a lower and upper bound of 25 and 50 respectively.

5. Conclusion and Future Work

In this paper, we have demonstrated the effectiveness of causal modeling in estimating the influence of recommended predictions on driving customer adoption through the mediation of field outreach and promotional campaigns. This approach has several significant implications for businesses:

1. The adoption of causal modeling provides a more accurate assessment of the impact of predictions on business performance. Greater accuracy leads to better-informed investment decisions and resource allocation.
2. Implementing this method is much more streamlined and efficient. Instead of designing an experiment and meticulously rolling out test and control groups, causal modeling can be incorporated seamlessly while the business operations continue unhindered.
3. Traditional test-control measurements are often conducted as one-time analyses or at a low frequency due to the challenges in implementation. In contrast, causal models can be executed as frequently as needed, allowing for more timely and relevant insights.
4. Beyond mere measurement, the causal drivers identified through this approach can also enable "what-if" simulations. These simulations allow businesses to explore various scenarios and test potential outcomes, providing valuable insights for strategic planning and decision-making.

In terms of future work, our current analysis focuses on cross-sectional field activity. However, we recognize that longitudinal execution of reach may also play a significant role in driving customer engagement and, consequently, behaviors such as adoption. As a next step, we plan to incorporate longitudinal information, including the duration between reach, channel combinations, and channel sequences, while considering the timing between them. This addition will enhance the sophistication of omni-channel orchestration in next best action models.

Moreover, our present analysis considers only HCP specialty as a confounder. To further improve the robustness of causal estimates, we aim to incorporate external factors such as payer details, patient information, regional and territorial attributes, along with attitudinal factors and social opinions as additional confounders. By taking these factors into account, we expect to refine our understanding of the causal drivers and better inform strategic decision-making.

Appendix

A. Attribution Analysis

We used propensity score linear regression approach [36] [34] [35] to perform the attribution analysis. Though this approach can be scaled for logistic regression, we leveraged linear regression primarily for ease of interpretability. The following equations illustrate additive functional form of outcome *Y*, which is a binary flag of drug adoption. *Prediction* represents the recommender model tiers; *Activity* represents field execution. Following are the steps performed for attributional analysis-

i. Stratify the HCPs into 4 groups –

a. T-R : Tier 1,2,3 HCPs with 1+ field reach

$$Y_{TR} = \alpha_{TR} + \beta_{1TR} * \text{Prediction} + \beta_{2TR} * \text{Activity} + \beta_{3TR} * \text{Prediction} * \text{Activity} + \beta_{4TR} * \text{Specialty} + \epsilon \quad (1)$$

b.T-NR : Tier 1,2,3 HCPs with no field reach

$$Y_{TNR} = \alpha_{TNR} + \beta_{1TNR} * \text{Prediction} + \beta_{4TNR} * \text{Specialty} + \epsilon \quad (2)$$

c. NT-R : non-Tier HCPs with 1+ field reach

$$Y_{NTR} = \alpha_{NTR} + \beta_{2NTR} * \text{Activity} + \beta_{4NTR} * \text{Specialty} + \epsilon \quad (3)$$

d. NT-NR : non-Tier HCPs with no field reach

$$Y_{NTNR} = \alpha_{NTNR} + \beta_{4NTNR} * \text{Specialty} + \epsilon \quad (4)$$

ii. Using backfitting-like procedure where a higher-order model is built on the residual outcome of the lower-order model, train the 1st regression model on (4) and use the model to generate predictions on Eq. (2) and (3).

$$Y'_{TNR} = Y_{TNR} - \widehat{Y_{NTNR}} = \beta_{1TNR} * Prediction + \epsilon \quad (5)$$

$$Y'_{NTR} = Y_{NTR} - \widehat{Y_{NTNR}} = \beta_{2NTR} * Activity + \epsilon \quad (6)$$

iii. Train 2nd and 3rd regression models on Eq. (5) and (6).

iv. Generate predictions on Eq. (1) using models trained on Eq. (4), (5), (6).

$$Y'_{TR} = Y_{TR} - \widehat{Y_{NTNR}} - \widehat{Y'_{TNR}} - \widehat{Y'_{NTR}} = \beta_{3TR} * Prediction * Activity + \epsilon \quad (7)$$

v. Train 4th regression model on Eq. (7)

vi. Finally, generate predictions using the 4 trained models on the full data. Following are its interpretations
   a. $\hat{Y}'_{TR}$ drug adopters attribute to both *Prediction* and *Activity*
   b. $\hat{Y}'_{TNR}$ drug adopters attribute to *Prediction* only
   c. $\hat{Y}'_{NTR}$ drug adopters attribute to *Activity* only
   d. $\hat{Y}'_{NTNR}$ drug adopters attribute to *None*

vii. Run 1000 bootstraps, with 80% of randomly stratified samples for training the 4 regression models.

B. Counterfactual Simulation Analysis

We leveraged the architecture illustrated in Appendix A to perform the counterfactual simulation analysis. Here are the steps-

i. Stratify the HCPs into 4 groups based on Predicted Tier : Tier1, Tier 2, Tier 3 and non-Tier
ii. Within each Tier, further stratify HCPs by Reach : R1 : Reached w/ 1+ channels and R0 : Reached with No channel
iii. For each Tier, using the distribution of reach among R1 HCPs, simulate on R0 HCPs across all channels. Make sure that the total simulated reach across all the R0 HCPs in 3 Tiers is less than or equal to the actual reach among non-Tier HCPs
iv. For each Tier 1, Tier 2 and Tier 3,
   a. Generate predictions on R0 HCPs using the R0 and R1 trained linear regressions models
   b. The predicted estimates of R1 model conform to the expected drug adopters if these R0 HCPs were reached at the same distribution of their peers (i.e. counterfactual)
   c. The predicted estimates of R0 model conform to the base value with no reach (i.e. predicted factual)
   d. The delta of (b) with (c) represents the additional expected drug adopters.
   e. Run 1000 bootstraps with 80% of HCPs in each Tier used for training as well as simulations.

About the Authors

Sri Krishna Rao Achyutuni *is a Sr. data scientist with ZS Associates at Artificial Intelligence Center of Excellence in New York. He has 8+ years of experience in applying machine learning solutions across various industries, including healthcare.*

Srinivas Chilukuri *is a Principal data scientist with ZS Associates in Evanston. He leads the Artificial Intelligence Centers of Excellence in New York and Bengaluru. He has close to 20 years of experience in building machine learning solutions across various industries.*

References

1 VanderWeele, T.J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167:268-274.

2 Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010 Dec 15;172(12):1339-48. doi: 10.1093/aje/kwq332. Epub 2010 Oct 29. PMID: 21036955; PMCID: PMC2998205.

3 Nguyen TQ, Webb-Vargas Y, Koning IM, Stuart EA. Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. Struct Equ Modeling. 2016;23(3):368-383. doi: 10.1080/10705511.2015.1062730. PMID: 27158217; PMCID: PMC4855301.

4 Samoilenko M, Lefebvre G. Parametric-Regression-Based Causal Mediation Analysis of Binary Outcomes and Binary Mediators: Moving Beyond the Rareness or Commonness of the Outcome. Am J Epidemiol. 2021 Sep 1;190(9):1846-1858. doi: 10.1093/aje/kwab055. Erratum in: Am J Epidemiol. 2022 Aug 22;191(9):1670. PMID: 33693467; PMCID: PMC8536873.

5 Jung SJ. Introduction to Mediation Analysis and Examples of Its Application to Real-world Data. J Prev Med Public Health. 2021 May;54(3):166-172. doi: 10.3961/jpmph.21.069. Epub 2021 May 7. PMID: 34092062; PMCID: PMC8190553.

6 Zhang Z, Zheng C, Kim C, Van Poucke S, Lin S, Lan P. Causal mediation analysis in the context of clinical research. Ann Transl Med 2016;4(21):425. doi: 10.21037/atm.2016.11.11

7 Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010 Dec 15;172(12):1339-48. doi: 10.1093/aje/kwq332. Epub 2010 Oct 29. PMID: 21036955; PMCID: PMC2998205.

8 Valeri, L. and VanderWeele, T.J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods, 18:137-150.

9 Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014 May;25(3):418-26. doi: 10.1097/EDE.0000000000000078. PMID: 24713881; PMCID: PMC4077670.

10 Nguyen TQ, Webb-Vargas Y, Koning IM, Stuart EA. Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. Struct Equ Modeling. 2016;23(3):368-383. doi: 10.1080/10705511.2015.1062730. PMID: 27158217; PMCID: PMC4855301.

11 Li, Y., Yoshida, K., Kaufman, J. S., & Mathur, M. B. (2022, March 28). A Brief Primer on Conducting Regression-Based Causal Mediation Analysis. https://doi.org/10.31219/osf.io/jath7

12 Vansteelandt, Stijn, Bekaert, Maarten and Lange, Theis. "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects" Epidemiologic Methods, vol. 1, no. 1, 2012, pp. 131-158. https://doi.org/10.1515/2161-962X.1014

13 Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. Am J Epidemiol. 2012 Aug 1;176(3):190-5. doi: 10.1093/aje/kwr525. Epub 2012 Jul 10. PMID: 22781427.

14 CMAverse: a suite of functions for causal mediation analysis - https://github.com/BS1125/CMAverse/

15 Shi, Baoyi; Choirat, Christine; Coull, Brent A.; VanderWeele, Tyler J.; Valeri, Linda. CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses. Epidemiology: September 2021 - Volume 32 - Issue 5 - p e20-e22 doi: 10.1097/EDE.0000000000001378

16 Pearl, J. (2014). Interpretation and identification of causal mediation. Psychological Methods, 19(4), 459–481. https://doi.org/10.1037/a0036434

17 VanderWeele TJ, Vansteelandt S (2014). Mediation analysis with multiple mediators. Epidemiologic Methods. 2(1): 95 - 115.

18 Steen J, Loeys T, Moerkerke B, Vansteelandt S (2017). Medflex: an R package for flexible mediation analysis using natural effect models. Journal of Statistical Software. 76(11).

19 Imai K, Keele L, Tingley D (2010). A general approach to causal mediation analysis. Psychological Methods. 15(4): 309 - 334.

20 Ryosuke Fujii, Shuntaro Sato, Yoshiki Tsuboi, Andres Cardenas & Koji Suzuki (2022) DNA methylation as a mediator of associations between the environment and chronic diseases: A scoping review on application of mediation analysis, Epigenetics, 17:7, 759-785, DOI: 10.1080/15592294.2021.1959736

21 Conor James MacDonald, Pauline Frenoy, Marie Christine Boutron Ruault, Response to Boonpor et al: Types of diet, obesity, and incident type 2 diabetes: Findings from the UK Biobank prospective cohort study, Diabetes, Obesity and Metabolism, 10.1111/dom.14813, 24, 11, (2277-2279), (2022).

22 Ohler, A.M., Braddock, A. Infections and antibiotic use in early life, and obesity in early childhood: a mediation analysis. Int J Obes 46, 1608–1614 (2022). https://doi.org/10.1038/s41366-022-01155-7

23 Tamayo Martinez N, Xerxa Y, Law J, et al. Double advantage of parental education for child educational achievement: the role of parenting and child intelligence. European Journal of Public Health. 2022 Oct;32(5):690-695. DOI: 10.1093/eurpub/ckac044. PMID: 35554528; PMCID: PMC9527951.

24 Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014 May;25(3):418-26. doi: 10.1097/EDE.0000000000000078. PMID: 24713881; PMCID: PMC4077670.

25 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37–48.

26 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–560.

27 Pearl J. The Causal Foundations of Structural Equation Modeling. In: Hoyle RH, editor. Handbook of Structural Equation Modeling. New York: Guilford Press; 2012. pp. 68–91

28 Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82:669–710.

29 van der Laan M, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. Berlin, Heidelberg, New York: Springer; 2011.Hitchcock, Christopher, “Causal Models”, The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/causal-models/>.

30 Curran-Everett D, Milgrom H. Post-hoc data analysis: benefits and limitations. Curr Opin Allergy Clin Immunol. 2013 Jun;13(3):223-4. doi: 10.1097/ACI.0b013e3283609831. PMID: 23571411.

31 Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. doi: 10.1080/00273171.2011.568786. Epub 2011 Jun 8. PMID: 21818162; PMCID: PMC3144483.

32 Shi, Baoyi; Choirat, Christine; Coull, Brent A.; VanderWeele, Tyler J.; Valeri, Linda. CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses. Epidemiology 32(5):p e20-e22, September 2021. | DOI: 10.1097/EDE.0000000000001378

33 Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

34 Imbens G.W. Nonparametric estimation of average treatment effects under exogeneity: A review. The Review of Economics and Statistics. 2004;86:4–29.

35 Rosenbaum P.R., Rubin D.B. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983a;70:41–55.

36 Imbens, Guido W. “Matching Methods in Practice: Three Examples.” The Journal of Human Resources 50, no. 2 (2015): 373–419. http://www.jstor.org/stable/24735990.

Impact of Applying SDOH on Prescription Fill Rate Analysis

Russell D. Robbins, MD, MBA[i], Chief Medical Information Officer, PurpleLab, and Douglas Londono, PhD, VP of Advanced Analytics, PurpleLab

Abstract: Patients requiring prescription medications, particularly for specialty pharmacy medications, may be subject to plan design and other factors in determining whether they are eligible for the medication, and if so, what the out-of-pocket cost will be to them. The patient journey has been segmented into three components, the dispensed or fill rate, the abandonment or reversal rate, and the rejection rate. Understanding the potential underlying causes of prescription rejection rates is important to mitigate them. Traditional analysis focuses only on age, gender, and/or marital status. HealthNexus™, an analytics platform, integrates medical and pharmacy claims records with Social Determinants of Health (SDOH), shedding new insights into healthcare delivery and disparities. Their impact can be seen at each portion of the journey.

This analysis looked at approximately 3.5 million Americans taking an anti-coagulant drug. The number of people having their prescriptions rejected, abandoned, or filled were evaluated. Demographic information, such as age and gender, as well as SDOH information such as race, ethnicity, marital status, and income level were evaluated independently and together to identify disparities in care delivery. Further analysis was done on the raw residuals, which were standardized to produce a common scale, an Adjusted Standardized Pearson Residual (ASPR).

In 2017, CMS found that 3.5% of Part D prescriptions were rejected[1]. For the prescription in this study, using a traditional view, 1.2% of prescriptions are denied, or approximately 43,000 people. With deeper analytics looking at SDOH, results show that Hispanic patients are 1.25 times more likely and African Americans are 1.49 times more likely to have their prescriptions rejected compared to patients from other ethnic or racial groups. When income is factored in, low Income (<$20,000/year) Hispanic patients are 1.18 times more likely, and African American patients 1.65 times more likely to have their prescription rejected. Traditional views of rejection rates miss the discrepancies of race and ethnicity. Our findings show that African Americans are more likely to have their prescriptions rejected no matter what their income. However, this is more pronounced in the lower income African American population. Hispanics exhibit a higher rate of rejection, but income does not appear to play as big a role, suggesting that other factors may need to be considered.

Introduction

Administrative claims data are used to evaluate medical and pharmacy trends. Unlike medical claims, non-specialty pharmacy claims contain only limited amounts of patient information such as name, address, date of birth, and date written and dispensed. While this information is useful, it is limited with regards to understanding the individuals receiving the medication. With the initiation of privacy laws, such as HIPAA, even this type of information is very limited in how it can be utilized for clinical or any other type of investigation. Currently, there are new data sources available which allow for de-identified patient information to be accessed and used for analysis.

Recently, a great deal of attention has focused on other factors that can influence an individual and their ability to access

[i] Corresponding Author: rrobbins@purplelab.com, 600 Lee Rd, Suite 100, Wayne, PA 19087

medical and pharmacy care. According to the CDC, social determinants of health (SDOH) are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a whole range of health, functioning, and quality of life outcomes and risks. SDOH can be grouped into five domains: economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and finally, social and community context.[ii]

The patient journey with regards prescription medications identifies the different touchpoints encountered. A prescription is presented to the pharmacy, and it is either filled and dispensed, filled, and abandoned (not picked up), or rejected. Understanding which groups of patients fall into each of these categories enables health care policy makers, pharmaceutical companies, and others to see what issues need to be addressed to close the gaps when a valid prescription is presented, but the medication is either abandoned or the prescription is rejected. As more medical information is being evaluated, further insights into the prescriptions need to be undertaken. Through the deidentification of patient information, one can gain further insights into the prescriptions and the patient journey.

Methodology

An analysis of a national open claims medical and pharmacy database examined pharmacy claims data for an anticoagulant from 12/01/2018 and 11/06/2021. In addition, Social Determinants of Health (SDOH) data, and mastered information regarding Healthcare Provider and Organization contact addresses, financials, clinical trial experience, and more. Open Claims data is sourced and aggregated from switch, clearing house and RCM providers. A total of 3,540,758 individuals met the criteria to be included in the study. Patient ages ranged from 18 to 85 years old. These people are all deidentified, and only the patient attributes are linked to the prescription through a series of tokenization steps using the Datavant software for linking and disaggregating patient information. Further statistical analysis was then conducted after the reports were generated.

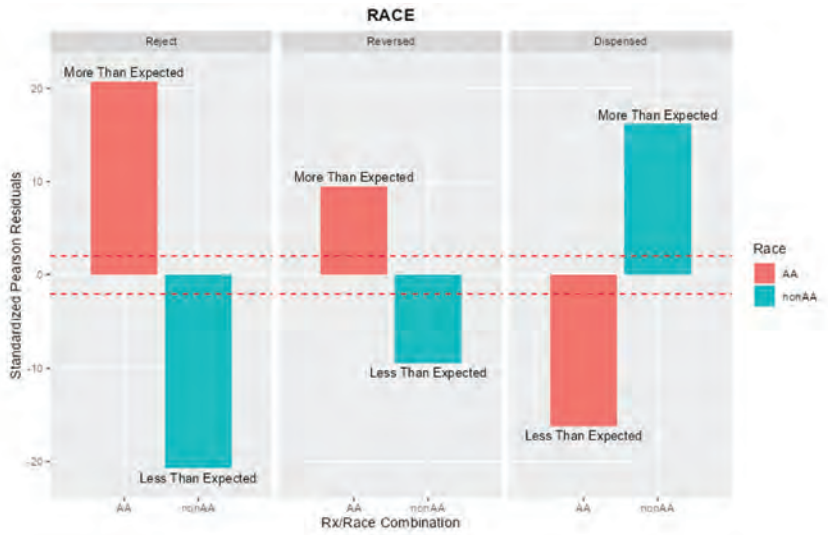
Initial evaluation of the data was conducted looking at three main components of the prescription journey, Dispensed, Abandoned, and Rejected. All claims information for each of these three phases is provided in the data sets obtained by the clearing houses based on determinations by the plans receiving the claims. In addition, further analysis within each component looked at age, gender, race, ethnicity, and income. focuses on the following characteristics for race: White, African American, Asian American, and other. For ethnicity, Hispanic and Non-Hispanic are the two variables considered. For any table cell, a raw residual is estimated as a function of the difference between observed and expected values. Of particular interest are those cells with the largest residuals. However, raw residuals are not immediately comparable. They first need to be standardized to produce a common scale. A type of such residual is termed an Adjusted Standardized Pearson Residual (ASPR). The ASPR that in absolute value exceeds 2 indicates a significant discrepancy between observed and expected values that cannot be explained by randomness alone.[2] This methodology was applied to each cell being analyzed.

[ii] https://health.gov/healthypeople/priority-areas/social-determinants-health

Figure 1: Outcome by Gender

	Reject	Abandoned	Dispensed	Total
Female	1.23%	8.51%	90.26%	100%
Male	1.34%	8.84%	89.82%	100%
Total	**1.29%**	**8.68%**	**90.04%**	

Figure 2: Impact of Race on Prescription Fill Rates



Each variable was compared to the outcomes for dispensed and rejected. Any incidents of statistical significance were noted. In each of the graphs in this article, the dashed lines represent the barriers for statistical significance. Any bar extending beyond this line is considered statistically significant.

This paper focuses on several main areas of observation period, the first being the impact of age, gender, race, ethnicity, and income values. The last variables were combined so that race and income and ethnicity and income were also evaluated to determine if there was any change in statistical significance for individuals earning less than $20,000 per year.

Results

Overall, the rates for prescription being filled, abandoned, or rejected were 90.04%, 8.68%, and 1.29%, respectively. We observe that males exhibit a significantly higher rate of having their prescriptions abandoned (Chi-Square Goodness of Fit p-value < 2.2e-16) or rejected (Chi-Square Goodness of Fit p-value < 2.2e-16) when compared to female patients. (Figure 1.)

With the introduction of the SDOH variables to the data, new trends begin to emerge. For example, when comparing Dispensed vs Rejected outcomes by Race, an exact test yields a p-value < 2.2E-16 with an Odds Ratio of 1.49 with a confidence interval (CI) of (1.42, 1.54). This means that the odds of having a prescription rejected is 1.49 times more likely in African American

Figure 3: Impact of Ethnicity on Prescription Fill Rates

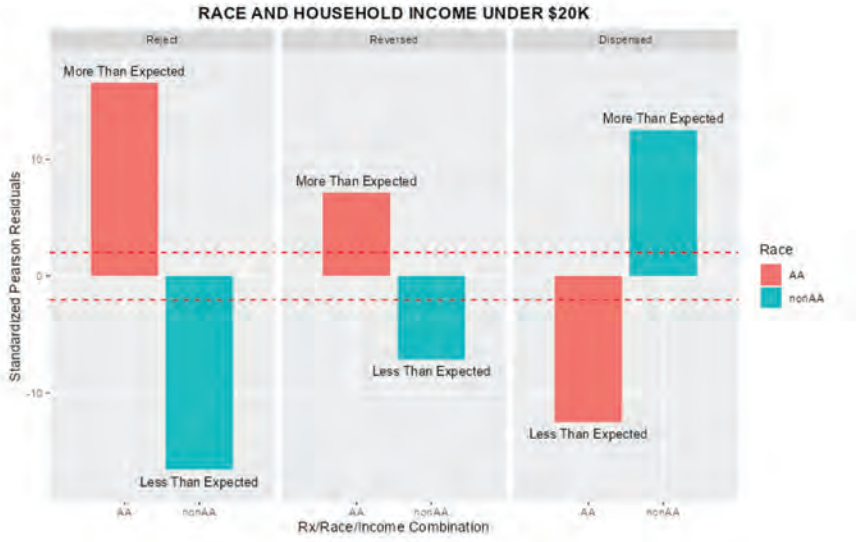
Figure 4: Impact of Ethnicity & Income on Fill Rates

patients than in non-African American patients (Figure 2.)

When looking at Ethnicity, for Hispanic patients, comparing Dispensed vs Reject outcomes, an exact test yields a p-value < 2.2E-16 with an OR of 1.25 CI (1.20, 1.31). This means that the odds of having a prescription rejected is 1.25 times more likely in Hispanic patients compared to patients from other ethnic groups (Figure 3.)

When looking at low-income groups, defined as those earning under $20,000 per year, we observed no significant difference in terms of rejection of prescriptions. We did observe a significant difference when comparing Dispensed vs Reversed outcomes. An exact test yielded a p-value = 1.94E-08 with an OR of 1.03 CI (1.02, 1.04). This means that patients who earn less than $20,000 per year are 1.03 more likely to have their prescriptions abandoned

Figure 5: Impact of Race & Income on Fill Rates



compared to patients making more than $20,000 per year.

When (low) income and ethnicity are considered together, the results shift. For Household Incomes of less than $20,000 a year, comparing Rejected vs Dispensed outcomes among Hispanics vs. patients from other ethnicities, an exact test yields a p-value of 1.67E-05 with an OR of 1.18 CI (1.09, 1.28). This means that Hispanic patients who earn less than $20,000 per year are 1.18 times more likely to have a prescription rejected than non-Hispanic patients who also make less than $20,000 per year (Figure 4.)

African American patients exhibit a larger change when income is also factored in. For Household Incomes of less than $20,000 per year, comparing Dispensed vs Rejected outcomes, an exact test yields a p-value < 2.2E-16 with an OR of 1.65 CI (1.55, 1.75). This means that African American patients who earn less than $20,000 per year are 1.65 times more likely to have a prescription rejected than non-African American patients who also make less than $20,000 per year (Figure 5.)

Discussion & Conclusions

The traditional approaches to evaluating the patient journey show that 1.2% of all prescriptions are rejected. While this percentage may appear to be small, it represents almost 43,000 individuals. There are many reasons why prescriptions may be rejected. Plan design, prior authorization, step therapy, generic substitution, drug/drug interactions, as well as other factors may indicate reasons why the prescription presented would not be accepted for fulfillment. The reasons for each of these rejections were provided as a part of the data file obtained from the clearinghouse. Other factors outside of the plan design must also be considered. For example, patient education about the medication, information in other languages, rebates, and other variables need to be considered. When using medical and pharmacy claims alone, many of these factors are not known. Rather knowing that these barriers exist is often enough to start the remediation process.

In recent years, great advances with regards to claims attribution and linking of individuals to their SDOH determinants enable deeper analysis into the patient journey. Traditional views of the patient journey miss the discrepancies related to race and ethnicity. In this study, we observed that African Americans have a statistically significant rate of rejection, regardless of their income level. Also, in Hispanic population we observed a statistically significant rejection rate compared to other ethnicities. Hispanic patients are 1.25 times more likely, and African American patients are 1.49 times more likely to have their prescriptions rejected compared to other ethnic or racial groups. When low income is factored in (<$20,000/year), Hispanic patients are 1.18 times more likely to have their prescription rejected. Lower income African American patients have an even more marked increase to 1.67 times more likely for having the prescription rejected than non-African American patients. The reasons for the poorer segment of the African American population to have greater rejection rates needs to be evaluated further.

By understanding inequities and disparities in prescription fill rates utilizing SDOH determinants, the healthcare industry can implement changes for the patients and the physicians treating them with programs such as better education, rebates, or understanding why the barriers exist. The use of an analytic platform highlights the importance for moving beyond the traditional view of age and gender alone. Further work will need to be done to understand practice patterns, prior authorization policies and other factors to address and rectify the inequities in dispensing this medication.

References

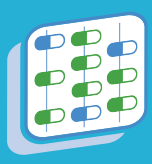
1 Murrin, S. (2019) Some Medicare Part D Beneficiaries Face Avoidable Extra Steps That Can Delay or Prevent Access to Prescribed Drugs. https://oig.hhs.gov/oei/reports/oei-09-16-00411.pdf

2 Agresti, A. (2013) Categorical Data Analysis. 3rd Edition, John Wiley & Sons Inc., Hoboken.

About the Authors

Russell D. Robbins, MD, MBA *is the Chief Medical Information Officer at PurpleLab. He is responsible for the HealthNexus platform's clinical rules and development of new modules, including Episodes of Care (EoC). In addition, he leads the team enabling PurpleLab to become a CMS Qualified Entity. Dr. Robbins is a nationally recognized thought leader on EoC having spent over twenty years working on and developing a variety of healthcare groupers on physician efficiency and quality, gaps in care, and predictive models. Prior to joining Purple Lab, he was the Medical Director and AVP Clinical Informatics at Blue Health Intelligence. He has served as the Chief Medical Officer at FairHealth. He has also done a great deal of work in health and benefits consulting, as the Chief Medical Officer at Cambridge Advisory Group as well as a Principal & Senior Clinical Consultant at Mercer. Dr. Robbins has also served as the clinical consultant to a Fortune 50 company for over a decade designing, implementing, and monitoring their health and wellness strategies. He was in private practice as a Urologist in Schenectady, NY. He received his MBA in Health Sciences from Union College, his MD from NYU School of Medicine, and his BA in Biology, from Swarthmore College. Dr. Robbins was an intern, resident, and Chief Resident, at NYU Medical Center, Bellevue Hospital, Manhattan VA Medical Center, and also trained at Memorial Sloan Kettering Cancer Institute, Cabrini Medical Center, and Long Island Jewish Medical Center.*

Douglas Londono *is the VP of Advanced Analytics at PurpleLab. Douglas obtained his PhD in Biostatistics in 2007 from Case Western Reserve University. He then joined Rockefeller University where he did his postdoctoral work on statistical modeling of genomic data. In 2009, Douglas joined the Department of Genetics faculty at Rutgers University. He has published extensive work on statistical modeling for the analysis of cross-sectional as well as longitudinal medical data. His work has been successfully applied for the identification of genes that confer increased disease risk in Cystic Fibrosis, Bronchopulmonary Dysplasia, Adolescent Idiopathic Scoliosis, and Alcohol/Opiate Addiction, among others. In 2017, Douglas joined PurpleLab, where he is responsible for the design and statistical modeling of research studies involving electronic medical records (EMRs).*



PMSA

PHARMACEUTICAL MANAGEMENT
SCIENCE ASSOCIATION

www.pmsa.org