AXTRIA
INGENIOUS INSIGHTS

Seeing the Whole Elephant: Integrated Advanced Data Analytics in Support of RWE

November 2022

Seeing the Whole Elephant: Integrated Advanced Data Analytics in Support of RWE

Won Chan Lee, Ph.D., Principal, HEOR/RWE Practice, Axtria Inc.

Introduction

The 21st century has brought about significant technological advancement, allowing the collection of new types of data from the real world on an unprecedented scale. The healthcare industry will benefit immensely from this abundance of patient data from electronic health records (EHRs), patient-reported outcomes (PROs), laboratory, demographic, social media, digital, and even climate data. While conventional statistical methods still play a significant role in supporting the drug lifecycle, machine learning (ML) and artificial intelligence (AI) are assuming a more prominent role in the analysis of this "big data." Moving forward, conventional statistics and ML/AI will work together to support descriptive, diagnostic, and even predictive analytics to further revolutionize drug discovery and development, regulatory approvals, and payer acceptance. In addition, counterfactual prescriptive analytics, such as causal inference analysis using real-world data (RWD) to generate insights that have cause-and-effect conclusions, will gain momentum as a methodology that can stand up against the rigor of regulatory review. Our real-world evidence/health economics and outcomes research (RWE/HEOR) field has evolved in ways that require us to integrate all the methods and data into a single framework that guides a holistic analytic approach and decision-making.

Executive Summary

- The passing of the 21st Century Cures Act in 2016 demonstrated the Food and Drug Administration's (FDA) growing acceptance of using RWD to generate RWE in support of regulatory submissions for pharmaceutical products and medical devices, where

previously, only randomized clinical trials (RCTs) were accepted.

- RWD analytics can help answer the following key research questions and situations:
  1. What happened?
  2. Why did it happen?
  3. What's likely to happen?
  4. What if…?

The conventional statistical approach often addresses questions one and two, forming a contextual background regarding disease epidemiology and clinical, economic, and humanistic burden. Increasingly, ML techniques are being applied to answer the third question.

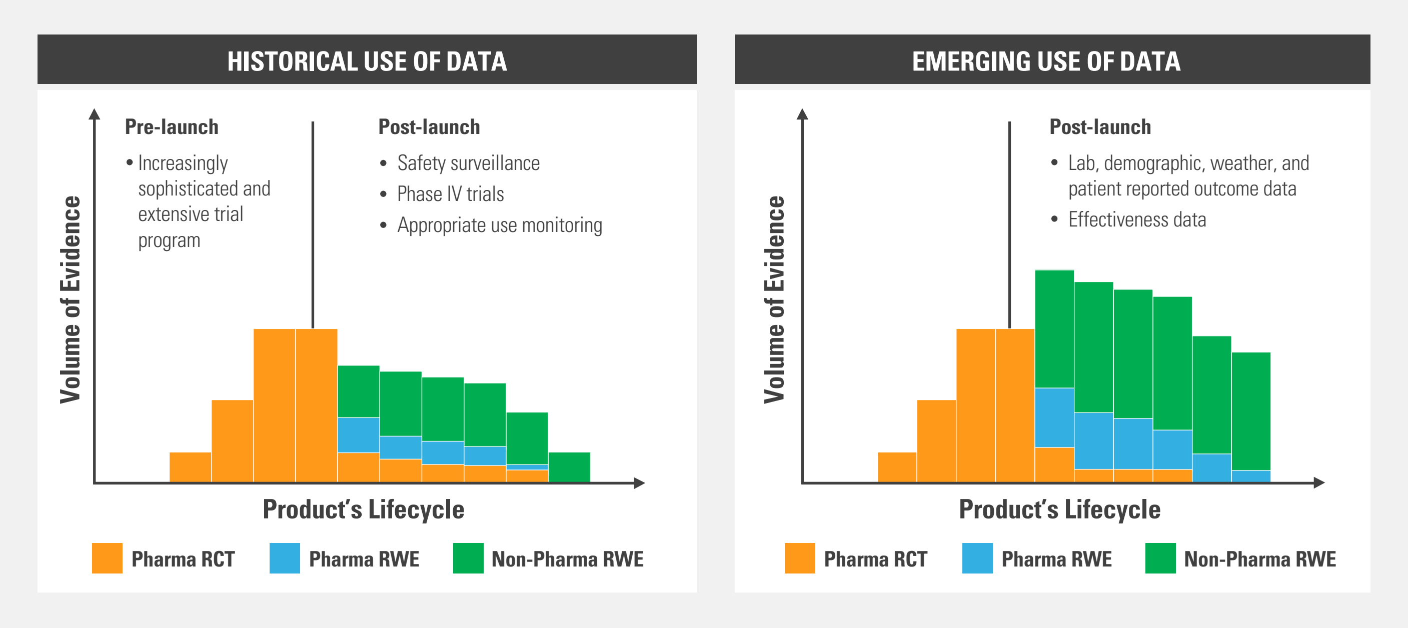
In contrast to the objectives of predictive analytics, causal inference asks questions about the effects of interventions or policies, allowing us to answer the fourth question.

- Counterfactual prescriptive analytics, such as the causal inference model utilizing RWD to generate insights for causal conclusions, will be gaining momentum as a methodology that can stand up against the rigor of regulatory review.

The Healthcare Data That We Touch: A Dilemma in the Fable of "The Blind Men and an Elephant"

RWD available for gaining insights into best clinical practices and improving patient outcomes have become increasingly more abundant, encompassing several forms of data, from closed insurance claims to social media posts by patients. Historically, patient-level data has been limited primarily to randomized clinical trial results owned by the study sponsor. These are generally not available for external researchers to investigate; however, with the current preponderance of RWD, there is ample post-launch data in the form of medical claims, EHRs, biomarkers, labs, etc. The sheer volume of data, coupled with their improved quality, has given rise to an asymmetry of data in the last decade towards RWD sources.[1]

Figure 1: Abundance of RWD and Asymmetry of Data Sources (Illustrative Purposes)



Source: Axtria Inc.

However, RWD sources remain fragmented. As such, it is unlikely that all researchers can access sufficient data to address their study objectives and business needs. Our perspectives are inherently limited due to the incomplete and fragmented data. It is much like the old Indian parable:

"Six blind men approach an elephant in order to learn more about it. The first man touches the side of the elephant and concludes that an elephant is like a wall. The second

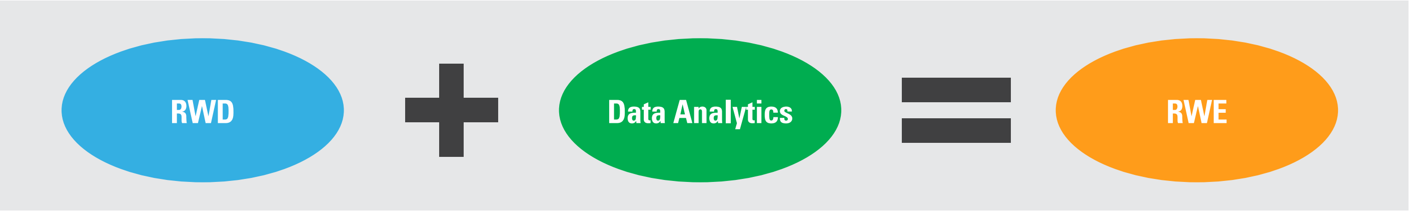
man feels the tusk and deduces that the elephant is like a spear. A third man grabs the squirming trunk and resolves that the elephant is like a snake. The fourth man reaches out and pats the huge leg, thereby determining that the elephant is like a tree. The fifth man touches the ear and thus infers that the elephant is like a fan. Finally, the sixth man seizes the swinging tail and judges the elephant to be like a rope. And, so the story goes, each man vehemently argued for the truth of his perception." [2]

Thanks to the emergence of different types of data and the ability to link them through a variety of techniques, the era of big data has truly dawned on us. Digital personal data, climate data, and social determinants of health are being added to the abundance of data, enabling us to draw a more complete picture of the whole elephant. An unprecedented era has descended, from which life sciences companies can harness as many and varied RWD sources as possible, thereby generating the most comprehensive and deep insights possible.

By passing the 21st Century Cures Act in 2016, the FDA defined fit-for-purpose data for generating RWE from the relevant and reliable framework. [3] A great deal of emphasis has been placed on identifying fit-for-purpose RWD. However, even when the "right" data are identified, the blind men's problem remains when the "right" analytics are not applied. RWE is closely scrutinized as part of the regulatory process, making it necessary to use both the right data and the right methodological approach. When discussing RWE, the power of the analytics used to maximize scientific rigor should be the focal point, overcoming the limitations of the best fit-for-purpose data sources.

RWE is a combination of RWD and analytics. This white paper describes the extent to which analytics will play a role in generating RWE once fit-for-purpose data is available. RWE should be further augmented and strengthened, leveraging the highest level of scientific rigor in analytics in the presence of a "right" and "comprehensive" data set. The following section briefly describes what types of analyses are applied for what purposes in generating RWE. In Section 3, the growing number of scientific studies leveraging ML/AI in the last decade is highlighted, followed by a description of how ML has become more of a complement to the conventional statistical approach and will be more so in the future (Section 4). In Section 5, the emerging "causal inference methods" in generating RWE is discussed as a methodology that can stand up against the rigor of regulatory review. Finally, in the context of the evolving regulatory landscape and analytical maturity we have reached in the HEOR/RWE field, Section 6 highlights what can be called "integrated RWE analytics" that aims to guide analysts and decision-makers.

Figure 2: Analytical Rigor at the Intersection of RWD and RWE



Source: Axtria Inc.

A Spectrum of RWD Analytics

RWD analytics can help answer the following key research questions (Table 1):

1. What happened?
2. Why did it happen?
3. What's likely to happen?
4. What if…?

Conventional analytic approaches often address questions one and two, forming a contextual background in terms of disease epidemiology and clinical, economic, and humanistic burden. Increasingly, ML techniques are being applied to answer the third question. The following section will discuss whether ML should be considered more of a complement to or substitution of traditional statistical approaches. In the past, ML has been used for classification and prediction rather than causal inference. In contrast, causal inference asks questions about the causal effects of interventions or policies.

Table 1: A Spectrum of Analysis Necessary for RWE

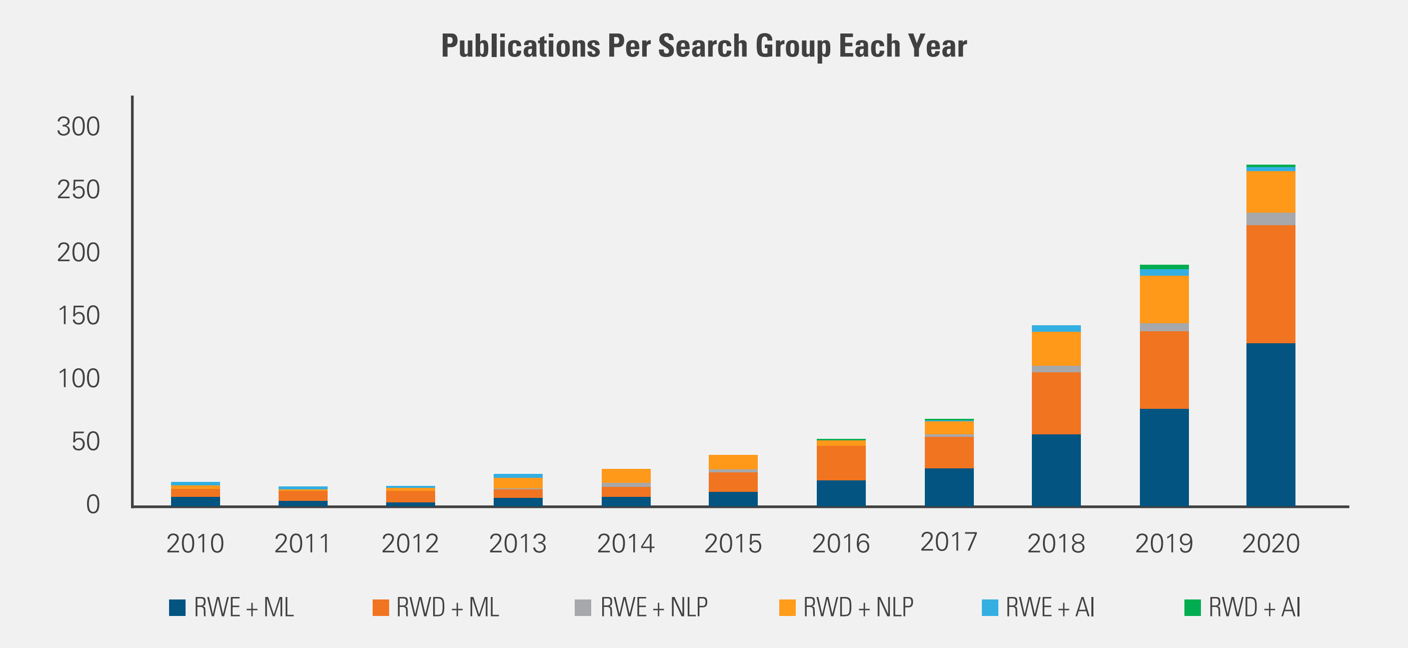
DESCRIPTIVE (What happened?)	**DIAGNOSTIC** (Why did it happen?)	**PREDICTIVE** (What's likely to happen?)	**PRESCRIPTIVE/ COUNTERFACTUAL** (What if…?)
• Simple means, standard deviations, counts, and percentages; t-test/ ANOVA, chi-square test, correlation analysis • Providing high level insights into patient populations, treatment patterns, etc. • Useful for hypothesis generation	• Multivariable regression analysis to identify associations between variables • Survival analysis for time to event • Adjusted for confounding factors to determine independent predictors of outcomes	• **Machine learning methods** to uncover hidden data structures for classification • Implementing algorithms on test/train split data to predict outcomes in external samples • High accuracy, precision, and recall for prediction across models	• **Causal inference analysis** designed to emulate the target trial • Generating a sample of simulated patients assigned to hypothetical treatment arms • Implementing methods such as g-estimation/ g-formula with the causal diagram

Source: Axtria Inc.

An Explosion of ML/AI Analytics

An increasing interest in RWE, driven by increasing data availability and advanced analytics, indicates an explosive capacity for growth in this area. I used PubMed to take a deeper look into the search results for six relevant groups of keyword searches: 'RWE + ML,' 'RWD + ML,' RWE + AI,' 'RWD + AI,' 'RWE + NLP,' and 'RWD + NLP.' The key results from this search are visualized in Figure 3.

Figure 3: The Exponential Growth in the Number of HEOR/RWE Scientific Papers Leveraging ML/AI Analytics



Source: PubMed

Year-over-year (YOY) rates show increasing growth in interest. Some notable observations are:

- Significant spikes in publications from 2012-2013 as interest starts to gain momentum
- Trending in 2017-2018
- Exponential growth in 2019-2020

Traditional Statistics vs. ML: Are They Complements or Substitutes?

Conventional statistics has its merit and is still a valuable tool in the modern-day analyst's repertoire. The comparative statistical analysis between treatment arms still requires the conventional approach of hypothesis testing and reporting of p-values. In classical statistics, we often examine a set of data using measures of center and spread to fit distributions to this data, effectively constraining it to how we believe the data should present itself. When we use ML, however, we let the data speak for itself. In this way, AI and ML can help us uncover great treasures hidden in the data. ML can complement classical statistics by providing additional patient insights that would otherwise be lost.

ML can use large volumes of data to improve the drug discovery process, provide customized treatments for patients based on their unique characteristics, and even predict prognosis or misdiagnosis. Because of ML's ability to handle massive data sets, images, and unstructured data, these insights are possible. Unlike the parametric restrictions or non-parametric assumptions from most classical analyses, deep learning models extract features directly from the data. This enables the identification of associations in the data that may have been previously unknown. [4]

However, just because ML methods operate on big data does not protect against bias. If we rely solely on data patterns, ML can lead us to a point where we make wrong decisions. It is still essential to have a theoretical framework (please refer to Section 5: Prediction vs. Causal Inference). Without it, we are lost and are forced to make guesses based on data patterns only. Conventional statistics is informed by theory, which is used to generate and test hypotheses. ML can be very useful in generating hypotheses. In addition, transparency is essential for regulatory acceptance. How can we make the process as transparent as possible by avoiding "black box" criticism?

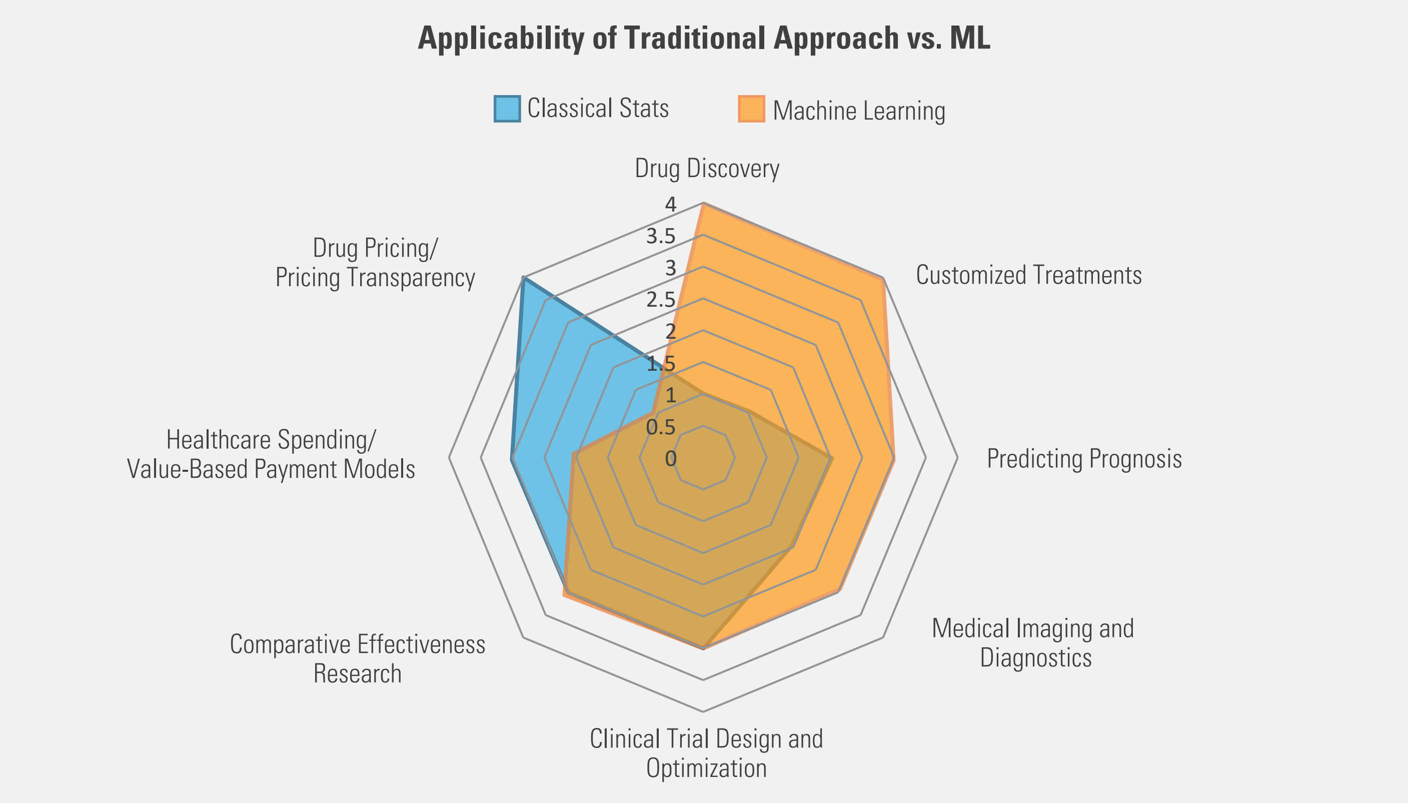
It is important to note that analysis of RWD has been limited to mainly the "easy-to-use data." For instance, about 20% of EHR data is structured, leaving 80% largely untapped, resulting in an incomplete picture. [5][6] We can use the ML methodology known as natural language processing (NLP) to analyze unstructured text data such as physicians' notes within an EHR or from patients' posts on social media platforms. This information can be used in various ways, from gleaning additional symptoms to sentiment analysis for the medical product of interest.

As different types of data, including texts, pictures, voices, etc., have become more available, methods of analysis for this data should be similarly varied, from classical statistical analyses such as basic descriptive analysis all the way to neural networks and NLP. The even more accelerated increase in popularity and application of ML and AI is inevitable. Together with classical statistics, ML/AI will continue to evolve to discern more parts of the pharmaceutical data elephant than would be possible if using these methods alone (Figure 4).

Prediction vs. Causal Inference

Traditionally, ML methods have been used for classification and prediction rather than causal inference. The prediction capabilities of ML are valuable by themselves. However, using ML for causal inference is still evolving.

Figure 4: Applicability of Traditional Approach vs. ML



Source: Axtria Inc.

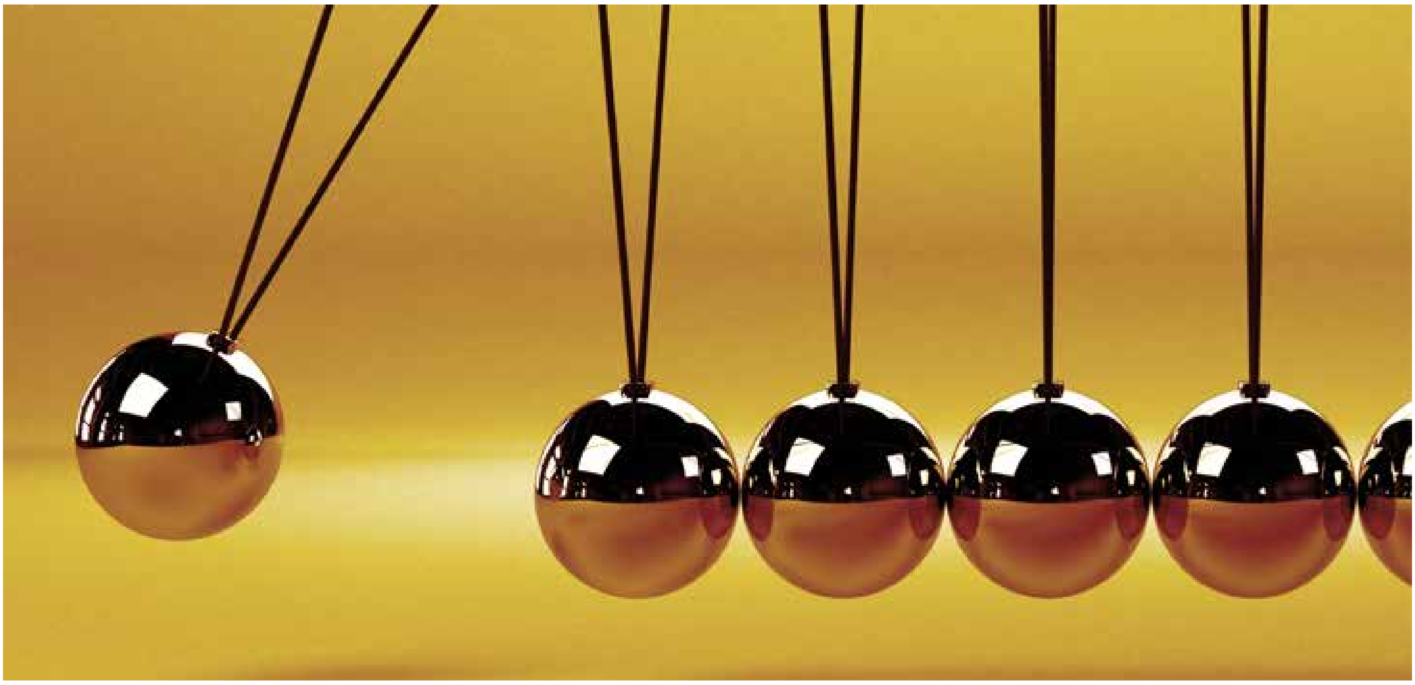
ML can be used for hypothesis generation, followed by the application of traditional causal methods. [7]

Traditional statistics can map inputs to outputs but do not consider how the world would look if circumstances changed. A statistical analysis can be deployed to overcome the challenges associated with the absence of key confounding variables in the available data. For instance, the unobserved variables are outside of our control and can significantly affect our analysis without our knowledge. A variety of statistical techniques, such as propensity score matching, difference in difference, and instrumental variable (IV), have been employed to minimize biases associated with the imbalances between comparison cohorts. However, establishing causality through these models has remained questionable. [8]

Causal inference uses RWD to predict certain features of the world as if the world had been different. One question of key importance to pharmaceutical companies is what the impact on safety and effectiveness endpoints would be if patients in the real world took the company's treatment instead of the treatment they were taking. To answer this question, first, a graphical causal model representing risk factors, treatments, outcomes, and relationships between these variables is designed. Next, a causal model structure based on expert opinion is developed. Then the model is fit to RWD. Finally, patient-level simulations are modeled to account for all post-baseline confounders and outcomes under different interventions. Outcomes are then compared among simulated interventions, from which results can be interpreted as having causal effects.

Evolving Regulatory Acceptance of RWE

RWE is among the most important forces shaping the future of the pharmaceutical industry. More than ever before, we are seeing growing acceptance and favorable regulatory changes for RWE as it relates to patient health on the part of regulatory bodies around the world. In 2021, based on RWE, a new indication for preventing rejection and death in lung transplants was granted to Astellas Pharma for PROGRAF®. In 2022, the FDA's approval of Novartis' VIJOICE® for a rare condition represents an unprecedented case where the approval was based on a retrospective analysis of RWD in 57 patients before a phase II or III clinical trial. Based on a systematic review of FDA approval documents from January 2019-June 2021, there were 116 approvals incorporating RWE in any form, but the RWE only influenced the FDA's final decision in 65 cases (48%). [8] This acceptance rate illustrates the necessity for RWE studies to be designed to stand up against the rigor of regulatory review.



Conclusions

Over the past century, there has been a massive expansion in data availability and in the technology necessary to analyze that data. This presents a unique opportunity for the healthcare industry to expand the range of research questions it is able to address and increase the speed with which those questions can be answered. Conventional statistical methods will continue to play a significant role, but ML will be increasingly important in analyzing big data. Combining these techniques will support not only descriptive and diagnostic analytics but also predictive analytics that will further revolutionize drug discovery, development, regulatory approvals, and payer acceptance. In addition, counterfactual prescriptive analytics, such as causal inference analysis in generating RWE, will gain momentum as a methodology that can stand up against the rigor of regulatory review.

The RWE/HEOR field has evolved such that we need to integrate all the methods and data into a single framework that guides a holistic analytic approach and decision-making process. The RWE/HEOR field may have reached that point. [9]

References

1. Concato J, Corrigan-Curay J. Real-World Evidence – Where Are We Now? N Engl J Med [Internet]. 2022 Apr [cited 2022 Sep 26]. Available from: DOI: 10.1056/NEJMp2200089

2. Snyder CR, Ford C . Coping with Negative Life Events: Clinical and Social Psychological Perspectives. New York: Springer Science; 2013. p. 12.

3. U.S. Department of Health and Human Services Food and Drug Administration. Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products, Guidance for Industry [Internet]. Rockville (MD): Food and Drug Administration; 2021 Dec [cited 2022 Sep 26]. 12 p. Available from: https://www.fda.gov/media/154714/download

4. IBM Cloud Education. What is Machine Learning? [Internet]. Armonk (NY); IBM; 2020 Jul [cited 2022 Sep 26]. Available from: https://www.ibm.com/cloud/learn/machine-learning

5. Li I, Pan J, Goldwasser J, et al. Neural Natural Language Processing for Unstructured Data in Electronic Health Records: A Review. ArXiv [Preprint]. 2021 Jul 7 [cited 2022 Sep 26]. Available from https://doi.org/10.48550/arXiv.2107.02975

6. Chase H, Mitrani L, Lu G, et al. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Medical Informatics and Decision Making (2017) 17:24 DOI 10.1186/s12911-017-0418-4.

7. Crown, William H. "Real-World Evidence, Causal Inference, and Machine Learning." Value in Health, Volume 22, Issue 5, 587 – 592. https://doi.org/10.1016/j.jval.2019.03.001

8. Hernán MA, Hsu J, Healy B. Data science is science's second chance to get causal inference right: A classification of data science tasks. Version: 6. ArXiv [Preprint]. 2018 Apr 28; updated 2019 Apr 7 [cited 2022 Sep 26]. Available from: https://arxiv.org/abs/1804.10846

9. Padula W, Kreif N, Vanness D, et al. Machine Learning Methods in Health Economics and Outcomes Research—The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force. Value Health [Internet]. 2022 Jul [cited 2022 Sep 26];25(7):1063–80. Available from: https://doi.org/10.1016/j.jval.2022.03.022

Won Chan Lee, Ph.D.
Principal, HEOR/RWE Practice
Axtria Inc.
300 Connell Drive,
Berkeley Heights, NJ 07922
E: wonchan.lee@axtria.com

Contact Us

+1-877-9AXTRIA
info@axtria.com

Disclaimer

Axtria® understands the compliance requirements behind personalization and we do not work with any personally identifiable data that can identify an end-customer of a business.

We have the strictest data security guidelines in place as we work with businesses to improve the experience for our customers.

🌐 www.axtria.com
✉ info@axtria.com
🐦 @Axtria
f facebook.com/AxtriaInc/
in Axtria – Ingenious Insights

Founded in 2010, Axtria is a global provider of award-winning cloud software and data analytics to the life sciences industry. Axtria's solutions are used to digitally transform the entire product commercialization process, driving sales growth, and improving healthcare outcomes for patients. Our focus is on delivering solutions that help customers complete the journey from Data-to-Insights-to-Action and get superior returns from their sales and marketing investments. Our cloud-based platforms - Axtria DataMAx™, Axtria SalesIQ™, Axtria InsightsMAx™, Axtria MarketingIQ™, and Axtria CustomerIQ™ - enable customers to efficiently manage data, leverage data science to deliver insights for sales and marketing planning, and manage end-to-end commercial operations.

For more information, visit www.axtria.com

Follow Axtria on LinkedIn, Twitter, and Facebook