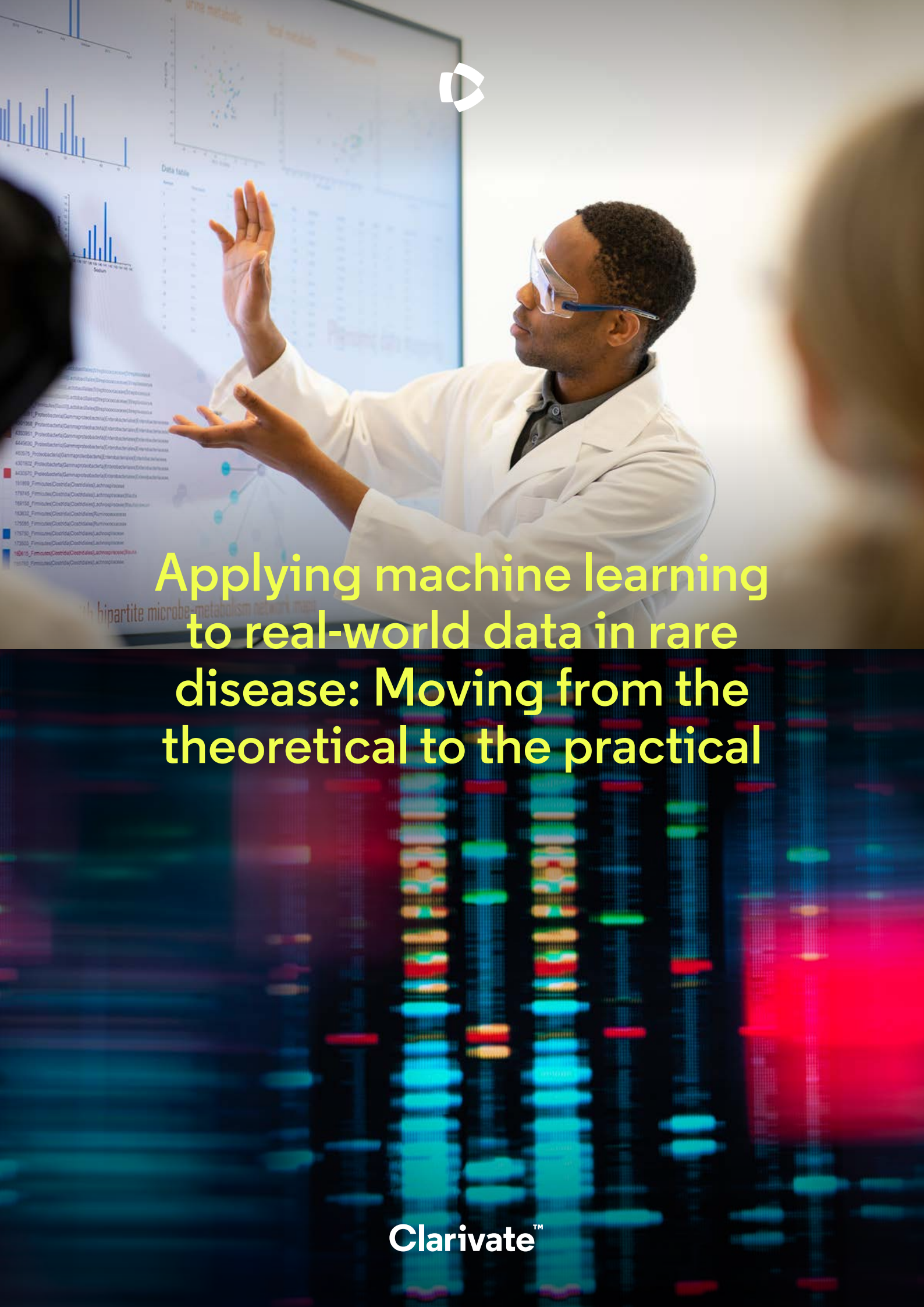


Applying machine learning to real-world data in rare disease: Moving from the theoretical to the practical

Clarivate™

Contributors

Bob Morrison

Vice President,
RWD Strategy and Operations,
Clarivate

Dona Petrozzi

Senior Director,
Product Management,
Clarivate

Cliff Li

Senior Director, Consulting,
Clarivate

Hemanth Nair

Senior Director,
Product Management,
Clarivate

90%+

of rare diseases, classed as any condition which affects fewer than 200,000 people in the U.S., are still without an FDA-approved treatment.

Trainee physicians are often told that when they hear hoofbeats they should think of horses and not zebras, underscoring the importance of considering common and more likely diagnoses before entertaining rare possibilities.

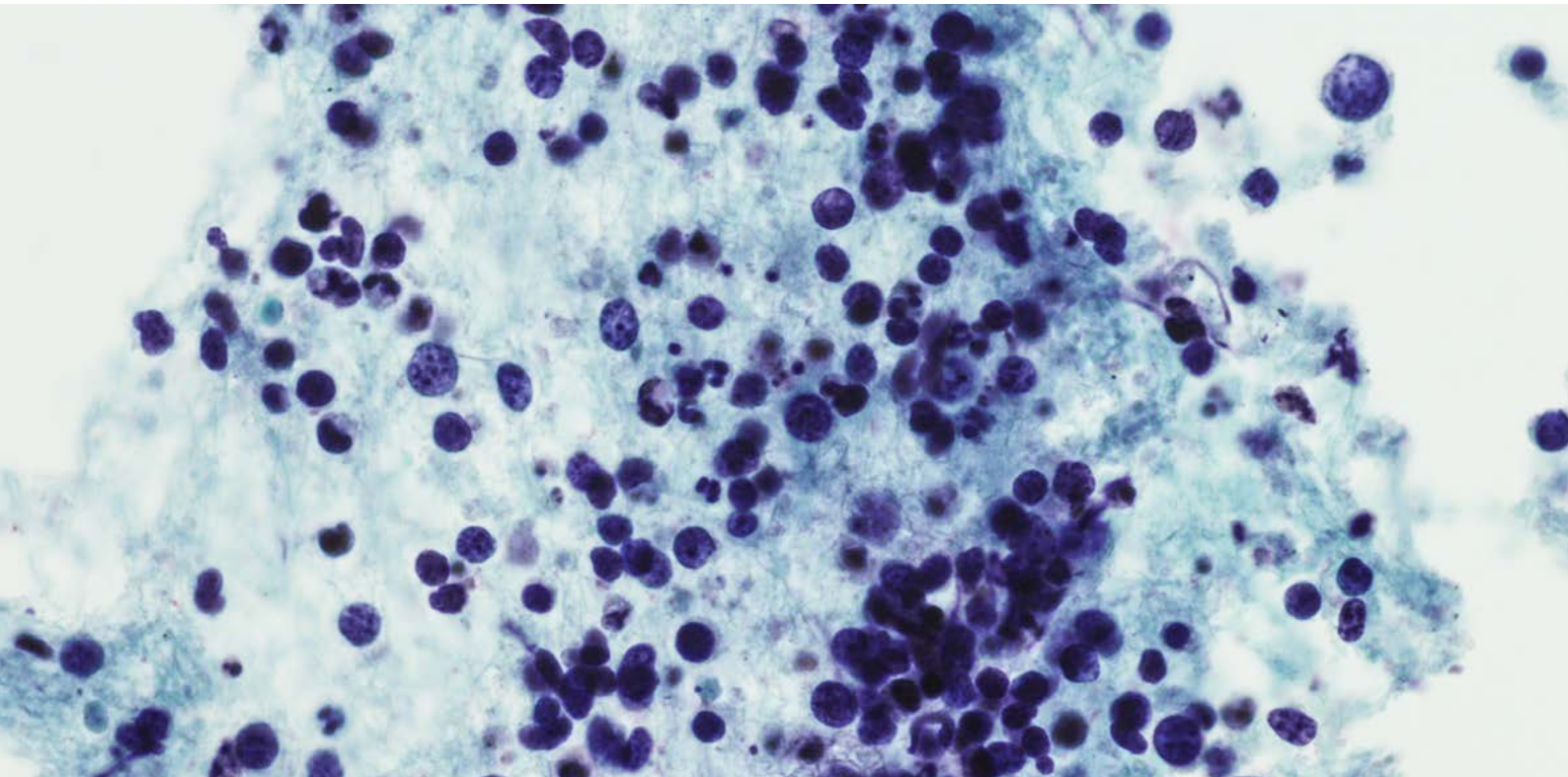
This approach to the diagnosis of rare diseases, coupled with the physiological difficulty in diagnosing rare conditions with heterogeneous symptoms, is evidenced by the figures. More than 90% of rare diseases, classed as any condition which affects fewer than 200,000 people in the U.S., are still without an FDA-approved treatment. Rare diseases affect an estimated 350 million people globally, with 25-30 million Americans living with some form of rare disease.[1]

Diagnosis can take years because of the Occam's Razor Principle applied by medical professionals to encourage simplicity in problem-solving. The difficulty in diagnosing rare diseases begins here, but also stems from limited awareness, a lack of standardized diagnostic criteria, genetic complexity, insufficient research, delayed symptom onset, and high diagnostic costs.

Addressing these issues requires an increase in research funding, improved medical education, and enhanced collaboration among healthcare professionals and researchers to catalyze diagnostic capabilities for rare diseases.

Another option gathering speed in investment and practice is the application of anonymized patient-level real world data (RWD) for diagnosis and the discovery and development of new treatments.

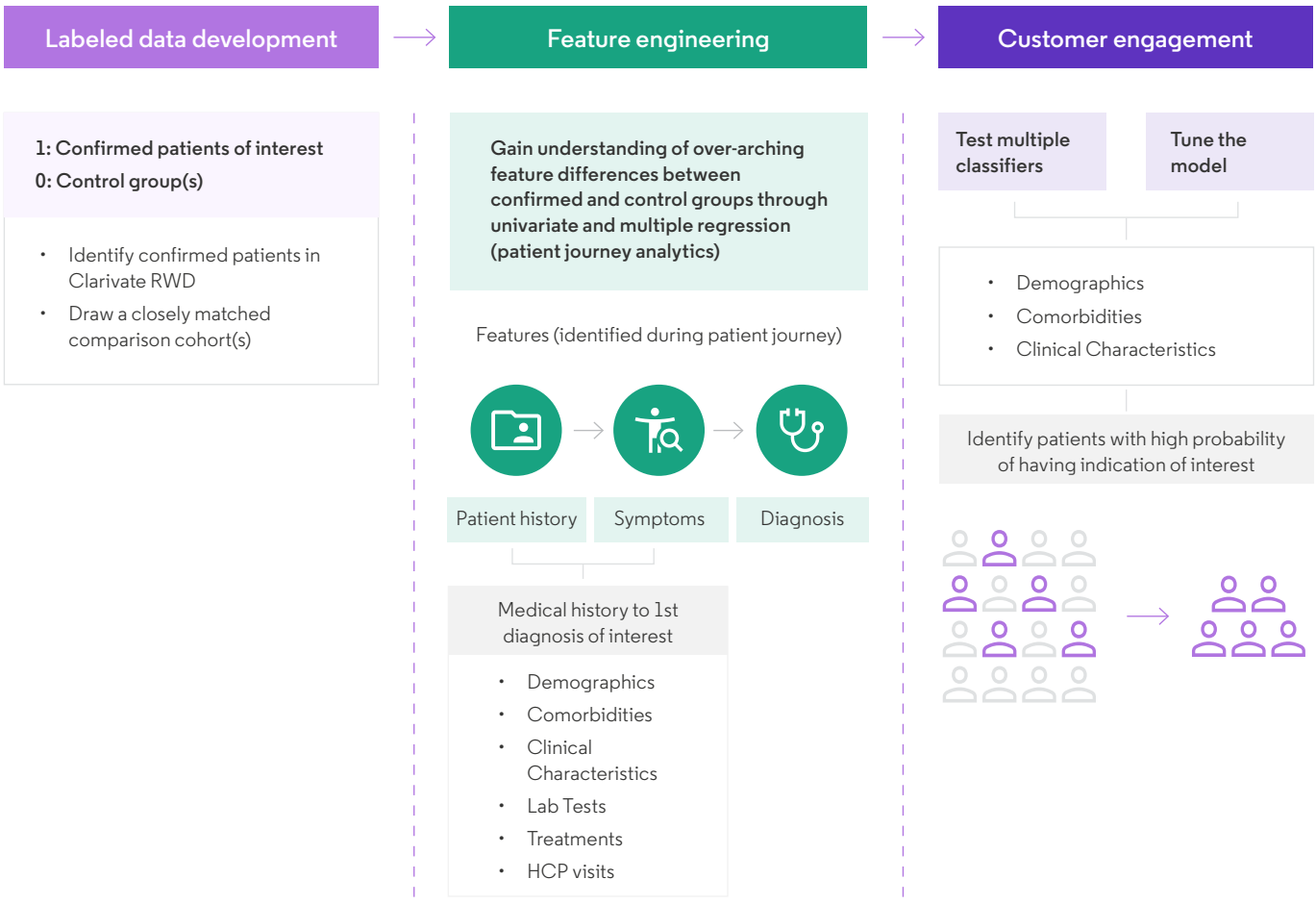
[1] https://rarediseases.org/understanding-rare-disease/rare-disease-facts-and-statistics/



Applying ML to RWD

RWD allows life science companies to take a broader look at the epidemiological aspects of a rare disease, and to interrogate data sources like clinical registries and databases, genomic and molecular data, electronic health records (EHR) and scientific literature. Applying machine learning to these data can uncover hidden patterns and associations, generating usable insights that can be employed for advantage in development and commercialization.

Figure 1: Machine learning approach



Data gaps can pose problems for RWD applications. Despite the promptness of open claims, data stabilization may require 2-3 months. Moreover, external data sources utilized in conjunction with RWD may exhibit several months of lag.

Advanced analytics and machine learning methodologies offer solutions for addressing data gaps. In instances where the application of specific billing codes lacks consistency yet remains crucial for cohort delineation, machine learning algorithms can be employed to identify patients who closely align with the criteria for a particular outcome or product reception, even in the absence of requisite billing codes.

To mitigate data lags, time series forecasting techniques prove invaluable. In such scenarios, employing time series forecasting enables the estimation of counts for the lagging months.

"As machine learning progresses, it will not only advance through refining modeling techniques but also through the ongoing evolution of interoperability with claims data," said Cliff Li, Senior Director, Consulting at Clarivate. "Incorporating third-party data streams will furnish additional features, enhancing the accuracy and robustness of the models."

Algorithms can identify biomarkers related to rare diseases, laying bare the underlying mechanisms of disease and guiding the development of targeted treatments. By mining current data, ML algorithms can examine known pathways and drug interactions to investigate the therapeutic effect of existing drugs on rare diseases.

Everycure, set up by Dr Daid Fajgenbaum, uses AI, including natural language processing, on a multitude of data sources such as PubMed, public data repositories, clinicaltrials.gov, medical record data, pathway, and drug databases to identify the most promising drug repurposing opportunities, before partnering with research organizations to conduct clinical trials. This approach has identified opportunities for therapies for Castleman disease, COVID-19 and angiosarcoma.

Rare disease, by its very nature, makes it difficult for a pharmaceutical company to quantify the commercial opportunity for a disease that, in many cases, has heterogeneous symptoms. RWD comes to the aid of understanding how many patients are suffering from or are at risk of developing a rare disease, what treatment they are receiving, where these patients are located, and if they fit into clinical trial criteria. Put simply, RWD plus machine learning can help to find missing patients.

This ground truth, however, is not universal or complete. By applying machine learning to databases, which may only include 70-80% coverage of the U.S. market, we can discover how many more people could be included in this patient pool. By using this approach, Clarivate™ can ascertain the volume of patients that meet certain criteria at a national level, as well as the number of patients visiting a certain facility in the U.S.

Collaboration between researchers, healthcare providers, and technology experts is essential to leverage these advanced technologies effectively, alongside an understanding of the regulatory challenges associated with incorporating technology-driven systems.

"As machine learning progresses, it will not only advance through refining modeling techniques but also through the ongoing evolution of interoperability with claims data."

Cliff Li,
Senior Director, Consulting at Clarivate.

Regulatory priorities

Top of the list of regulatory priorities when using RWD is data privacy and security. Compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S., the General Data Protection Regulations (GDPR) and the newly formed AI Act in the E.U is essential. The AI Act qualifies AI systems that could pose a risk to health as a high risk, requiring measures such as risk assessment, detailed documentation, and appropriate human oversight. However, when applying ML to RWD, not only is it fundamental to be aware of ethical standards and data quality, but also specific regulatory considerations related to the dynamic nature of ML algorithms and the explicit approach they may have on decision making.

Regulatory agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) emphasize the importance of transparency and explainability of machine learning algorithms.[2,3,4] Developers need to provide clear documentation on how the algorithm operates and how it reaches its conclusions, especially when making predictions or recommendations related to drug development.

Rigorous validation of machine learning algorithms is crucial. Developers must establish the performance metrics used to assess the algorithm's accuracy, sensitivity, specificity, and other relevant measures. In short, demonstrating the algorithm's reliability and reproducibility is essential for regulatory acceptance.

Not only are there regulatory challenges when considering using the machine learning approach to RWD in drug R&D and commercialization, there are practical obstacles that require expert guidance.

In short, demonstrating the algorithm's reliability and reproducibility is essential for regulatory acceptance.

[2] https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development

[3] https://www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf

[4] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Practical challenges

There are foundational actions that must be progressed before using machine learning on RWD, such as ensuring the quality and integrity of the underlying data, which can be messy due to the unique characteristics and sources of RWD, said Bob Morrison, Vice President of RWD Strategy and Operations at Clarivate. "Big data has a lot of challenges that if you're not careful, like with any other machine learning use case, you end up with a 'garbage in garbage out' scenario, biases, and inaccuracies," he explained.

Data must be complete and heterogeneity mitigated as much as possible for machine learning to work to its peak ability. Standardization of RWD collection is complex, especially when it emanates from private and public sources, and results in different forms of structured data.[5] These structures are inherent to the database from which the data is derived, making uniform data collection and merger time consuming.

To counter this, Clarivate has an experienced team that works through base data to ensure when machine learning is applied, the trustworthiness and accuracy of the results are industry-leading.

Ensuring machine learning interoperability and validity is crucial for reliable and ethical deployment of these advanced data models. Best practices and practical safeguards can be employed to address concerns about the quality of data used. There are open standards, such as the Open Neural Network Exchange (ONNX) and Predictive Model Markup Language (PMML), which facilitate synergies by allowing models to be exchanged between different platforms and frameworks.

Developing interoperable APIs also allows different systems to communicate seamlessly, and adhering to widely accepted API standards ensures compatibility between machine learning models and multiple software applications.

Mechanisms should also be deployed to detect and mitigate bias, and to continuously monitor machine learning model performance in real-world settings, allowing for quick detection of data issues and ensuring ongoing validity.

When moving forward with any data integration, it is fundamental to work with a team that has diverse experience, a team that has working knowledge and experience within domain settings, data science, and interdisciplinary collaboration. While more people possess advanced analytics expertise, the challenge lies in making it useful and actionable for specific business needs. To address this, a diverse team is needed, including scientists, clinicians, therapy area experts, AI engineers, and data scientists. Working with a team that can contextualize the problem, understand the AI methodology, and translate it into actionable steps is crucial when beginning a ML and RWD project.

"Big data has a lot of challenges that if you're not careful, like with any other machine learning use case, you end up with a 'garbage in garbage out' scenario, biases, and inaccuracies,"

Bob Morrison,
Vice President of RWD Strategy and Operations at Clarivate

[5] Baser O., Samayoa G., Yapar N., Baser E., Mete F. (2023) Use of Open Claims vs Closed Claims in Health Outcomes Research. J Health Econ Outcomes Res. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10484335/

Case study: Building a ML model to identify undiagnosed patients with rare disease

75%

to 80% of the time, the six ML models accurately IDed rare disease patients.

A North American biotech company approached Clarivate about a data-driven project to identify patients with a specific rare, inherited condition prone to missed or delayed diagnosis.

The company had a non-machine learning algorithm that used EHR data to find potential patients within its disease of interest, but struggled with its lower accuracy, and this approach missed many potential patients.

The Clarivate analytics team leveraged its RWD product by inputting symptoms, diagnoses, procedures, and treatments from its EHR and claims data within a five-year period. Six ML learning models were tested, and a validation study was conducted.

Across the six ML models, accuracy ranged from 75% to 80%, including the validation test. The company can now empower doctors with an ML-based algorithm that effectively identifies potential patients with this rare disease, reducing rates of delayed or missed diagnosis.

If a machine learning algorithm is not fed the right data, it is not learning the correct signs to look for. The Clarivate team knew that laying the groundwork was as important to the project outcomes as the machine learning algorithm itself.

"We studied and characterized the cohort extensively with traditional methods to identify an appropriate control group that could minimize misclassification," explained Hemanth Nair, Senior Director of RWD Product Management at Clarivate.

There are diseases, he continued, that had similar symptoms to the rare disease the project was trying to identify patients for. "If you're not careful, and you don't explicitly put those patients as part of the comparison group and understand their profiles, you easily have a machine learning model that's misclassifying a patient as a genetically based rare disease patient," said Nair.

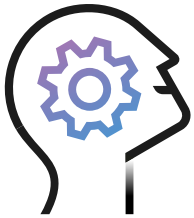
The team then embarked on feature selection, the process of selecting a subset of relevant variables, attributes, or predictors from the original dataset to improve model performance. The goal of feature selection is to reduce the dimensionality of the dataset by retaining only the most informative and discriminative features while discarding irrelevant or redundant ones.

Once feature selection was complete, the model was ready to be trained, tested, tuned, and eventually validated. Because Clarivate has access to timely data, it was able to validate the model on a set of patients that had the rare disease in question in 2023.

The next step for the project, which initially took four months, is to forecast models to find patients before diagnosis, helping the patient, provider and payer to realize improved health outcomes.

Figure 2: Case study

Employing a ML model to identify rare disease patients, exceeding 75% accuracy



Situation	Solution	Impact
• Our client had a proprietary algorithm to identify suspect patients with a specific rare, inherited disease prone to missed or delayed diagnosis, but struggled with its lower accuracy. • They turned to Clarivate to explore data-driven methods, including machine learning, to identify suspect patients with better accuracy.	• Clarivate analytics team developed a machine-learning model for suspect patient identification, leveraging our real world data product. • Symptoms, diagnoses, procedures, and treatments derived from our EHR and claims appearing in the 5-year period prior to index diagnosis were used as inputs. • Six machine learning models were tested, and a validation study was conducted.	• Across the six ML models, accuracy (percentage of patients correctly identified as a true positive or true negative patient) ranged from 75% to 80%, including the validation test. • Our client can now equip physicians with an ML-based algorithm that effectively identifies suspect patients with the rare disease, ultimately reducing rates of delayed diagnoses and misdiagnoses.

Clarivate robust RWD paired with unmatched expertise

The potential for RWD in commercialization

The future trajectory of RWD in therapeutic product planning is poised to shape various aspects of drug development, regulatory approval, market access, and post-market surveillance. In rare disease, it holds greater promise still.

Using data sources including EHRs, patient registries, wearable devices, smartphone apps and connected sensors, the natural history of rare disease can be characterized, including disease progression, comorbidities, and clinical manifestation. By integrating and connecting these various data sources and by adding in diverse data types such as genomic data, imaging studies and patient-reported outcomes, a comprehensive disease profile can be built. These disease profiles can identify underlying disease mechanisms, biomarkers and potential therapeutic targets.

By applying RWD-driven machine learning algorithms, early detection, differential diagnosis, and risk stratification can facilitate and catalyze time to diagnosis, which can take months to decades in some cases, with the average accurate diagnosis taking approximately four to five years.[6]

RWD can complement traditional clinical trial data with real world evidence (RWE) on treatment effectiveness, safety profiles, and long-term outcomes in diverse patient populations. This strengthens regulatory submissions and supports approval decisions. In post-market surveillance, the continuous monitoring of product safety and effectiveness in real world settings using RWD enables the timely detection of adverse events, identification of rare side effects, and compliance with post-market surveillance requirements.

Leveraging RWD throughout the rare disease product lifecycle, from clinical trial design to regulatory submission to post-market commercialization, enhances evidence generation, supports market access, and optimizes treatment strategies, contributing to the successful commercialization and uptake of rare disease products.

[6] Marwaha, S., Knowles, J.W & Ashley, E.A. (2022) A guide for the diagnosis of rare and undiagnosed disease: beyond the genome. Genome Med. Retrieved from: https://genomemedicine.biomedcentral.com

Key takeaways

As beneficial as RWD and machine learning are for the rare disease patient population and the physicians that treat them, there are careful considerations that can maximize the effectiveness and value of data-led initiatives.

Futureproofing for a time when RWD and machine learning is a part of product planning and marketing is necessary today, even if the power of these approaches have not been fully explored within a team or company.

Evaluate the availability, completeness, and quality of the data sources.

If a company is keen to pursue a data-led product plan, it needs to assess databases relevant to the disease of interest, including patient registries, EHR data, claims databases and disease-specific repositories. It should explore the reliability, accuracy, and representability of RWD for the target patient population. Are there data gaps? What are the potential biases? What are the limitations in the RWD databases?

Understand the regulatory requirements and guidelines.

Research the relevant legislation governing the use of RWD in drug development, approval, and post-market surveillance to ensure compliance with regulatory standards and expectations. It is critical to establish robust data governance frameworks, privacy safeguards, and ethical guidelines to protect patient confidentiality, uphold data integrity, and mitigate regulatory risks associated with RWD utilization.

Forge strategic partnerships and collaborations.

By getting to know academic institutions, patient advocacy groups, healthcare providers, and regulatory agencies, rare disease companies can access RWD sources, share expertise, and co-create research initiatives in product planning.

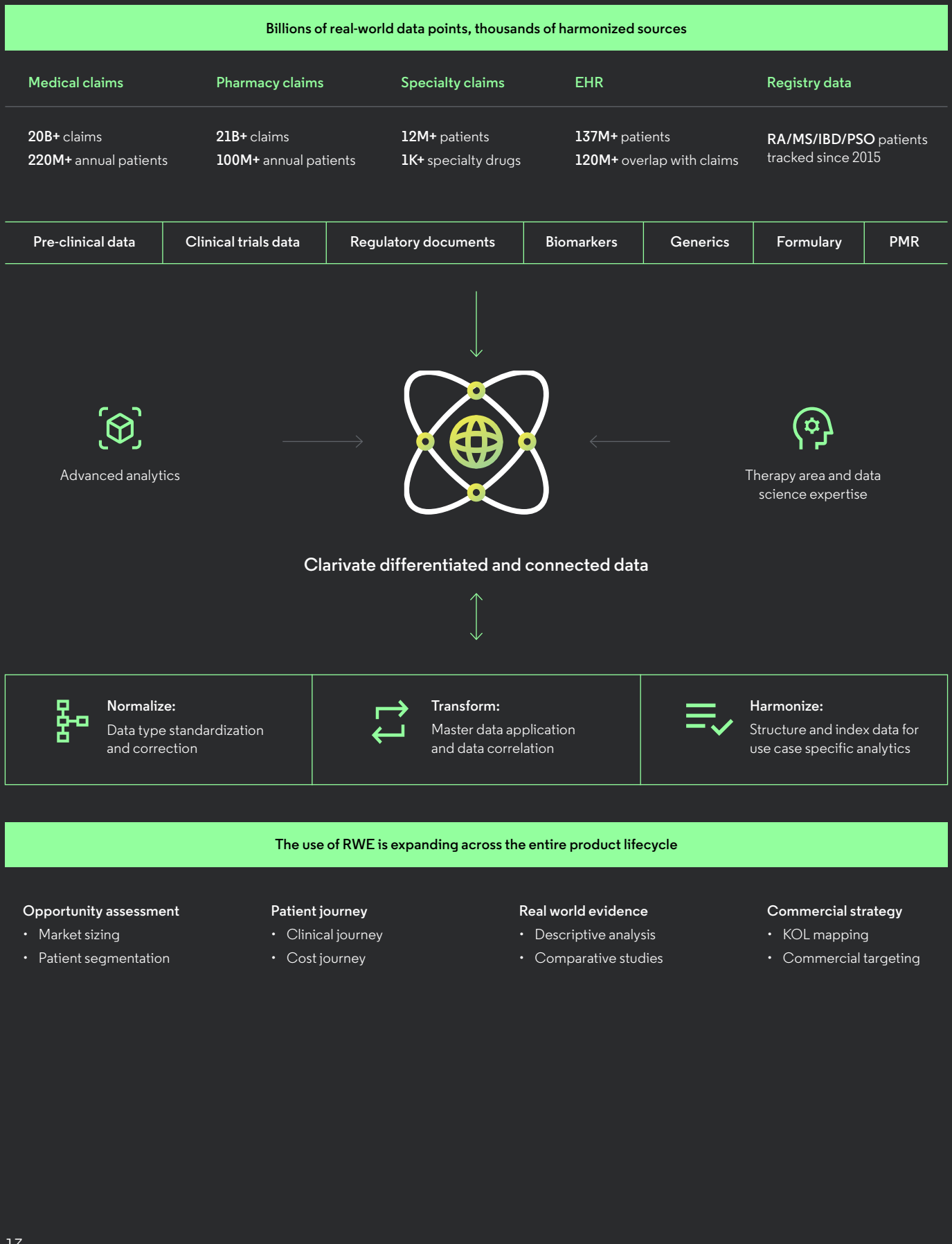
Engage with stakeholders.

It is imperative to embed the company with would-be collaborators, including patients, early in the product development process. This will align research objectives, address unmet needs, and validate the relevance and applicability of RWD-driven insights to clinical practice and patient care, helping to ensure that the company is developing a product rooted in genuine patient need.

Assemble or consult with a multidisciplinary team.

Finding the right partners with expertise in rare disease research, data science, regulatory affairs, and clinical development can foster collaboration and knowledge exchange between different functional areas. By leveraging diverse perspectives and skill sets to maximize the value of RWD to inform strategic decision-making, companies can drive innovation, optimize product development strategies, and improve patient outcomes in the challenging and underserved area of rare diseases.

Figure 3: Billions of real-world data points, thousands of harmonized sources



BXD1307578929 / 04

About Clarivate

Clarivate™ is a leading global provider of transformative intelligence. We offer enriched data, insights & analytics, workflow solutions and expert services in the areas of Academia & Government, Intellectual Property and Life Sciences & Healthcare. For more information, please visit **clarivate.com**.

Learn more about how the Clarivate enriched, connected and actionable data can transform life sciences companies worldwide, visit us:

clarivate.com/blog/unleashing-the-power-of-computational-tools-in-rare-disease/

Contact our experts today:

clarivate.com