

WHITE PAPER

Unlocking Real-World Evidence Insights with Data Linking

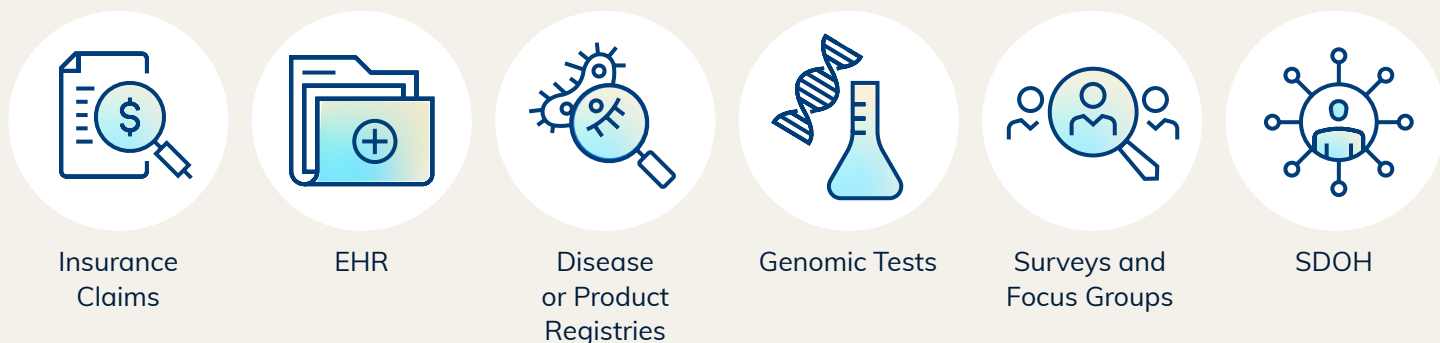
Introduction

Real-world data (RWD) holds incredible potential to accelerate healthcare research and development of innovative medicines to fill unmet medical needs for patients through the generation of RWD-derived clinical evidence of safety and efficacy, or real-world evidence (RWE). While randomized controlled trials (RCT) are considered the gold standard, insights from RWD and RWE have been increasingly utilized by life sciences and healthcare organizations to drive regulatory decisions and guide product life cycle strategies while reducing the high costs and time associated with RCTs.

RWD can enable a broad, longitudinal view of each patient's journey, providing both individual and population-level insights. To maximize this potential, disparate data sources may need to be linked, meaning that various sources of data associated with an individual patient need to be attributed to that specific individual.

Major challenges with data linking stem from the varied nature of RWD sources. Sources can include insurance claims data, electronic health records (EHR), disease or product registries, lab results including genomic tests, hospital charge masters, qualitative data such as surveys and focus groups, and [social drivers of health \(SDOH\)](#).

RWD Sources



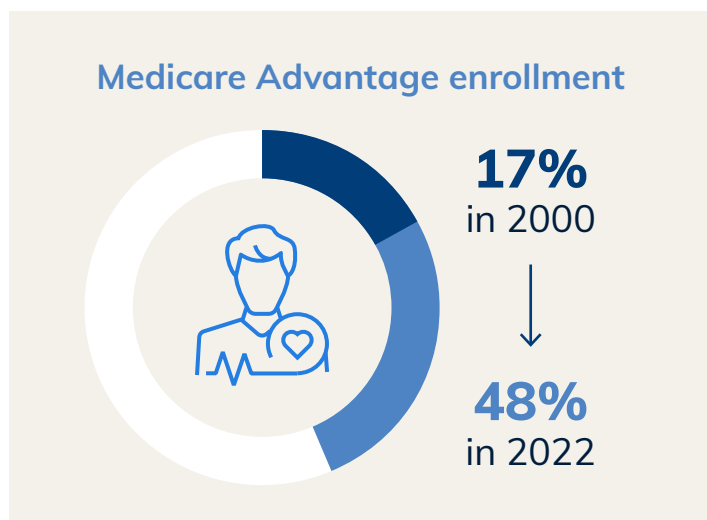
Claims data are a good source of RWD to start with for data linkage and can be complemented by other RWD sources. Claims can reveal what medicines were filled and procedures were performed, when medical events occurred, and how much they cost. Claims can also reveal the clinical and financial impact of events such as changes in insurance coverage.

This white paper discusses the benefits, values, and challenges of data linking, the methods that can be employed to effectively and accurately link RWD across the patient journey, and examples of how linked RWD has been utilized in RWE generation across oncology, autoimmune, and rare disease.

The Value of Data Linking for Payers and Policymakers

Data linking can provide benefits for stakeholders across the healthcare ecosystem, including patients, payers, clinicians, and policymakers. Data linking can enable a longitudinal view of each patient's health history that allows payers to better understand their core customer segments.

A 2023 [study from Inovalon and Harvard](#) utilized two linked-claims datasets to document how Medicare Advantage (MA) customers differ from Medicare Fee-for-Service (FFS) customers, both before and after initial Medicare enrollment at age 65.¹ MA enrollment has grown from 17% of all



Medicare beneficiaries in 2000 to 48% in 2022, but payers and policymakers have historically struggled to understand key differences between these programs such as patient demographics and health outcomes. This struggle has occurred in part due to limited ability to control for baseline differences in customer segments, and limited MA claims availability and quality. Additionally, most Medicare patients will have transitioned from commercial health plans, limiting researchers' ability to track patients younger than age 65 and understand the potential impact of this transition.

Researchers were able to determine a number of clear differences in socioeconomic demographics as well as health history and outcomes. These findings were based on Inovalon's Medical Outcomes Research for Effectiveness and Economics (MORE²) Registry[®] dataset, which links private insurance claims covering 30% of the U.S. privately insured population in a given year, a dataset with all medical and pharmacy claims for Medicare FFS enrollees, enrollment data for Medicare FFS and MA, and a dataset with linked SDOH data.

Findings like these derived from multiple, linked datasets can enable health plans to better tailor their products to their core segments of customers. They can also help policymakers promote health equity by adjusting health policy based on these findings.

The Value of Data Linking for Life Sciences Companies

Data linking can provide value to life sciences companies across [clinical trial](#), [health economics and outcomes research \(HEOR\)](#), and [commercial](#) functions. By linking claims data and SDOH data, companies can identify potential racial/ethnic, socioeconomic, or other disparities related to medication access and adherence, and design clinical trials and observational research studies that reflect the various characteristics of the population with the disease. These integrations can reveal differences in healthcare utilization and practice patterns, enabling companies to focus development and strategies on addressing specific barriers to access that may only affect certain groups and that could have gone unnoticed otherwise.

Linking claims data with laboratory tests provides greater context compared to claims alone. Claims data typically lack certain clinical data that would be measured in observational studies and RCTs as part of enrollment or stratification, such as levels of certain biomarkers. Similarly, genomics data can help define subpopulations with unique prognostic outcomes or responses to therapy. This can be useful in situations where a condition is rare and enrolling enough patients prospectively would be difficult. In some cases, claims data is also limited by ambiguity within the diagnostic code, such as with non-muscle invasive bladder cancer as well as next-generation sequencing (NGS) diagnostic tests. Supplementing claims data with histology data or the results of the NGS test can provide greater context and insight.

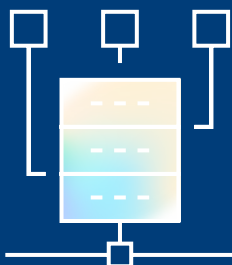
[Integrating RWD with RCT data](#) can potentially make RCTs more efficient and lower associated costs. Inclusion of RWD-derived estimates of efficacy in a network meta analysis led researchers to conclude that the phase III TRANSFORMS, FREEDOMS, and FREEDOMS II studies could have reduced their sample sizes by 40%, potentially saving six months of development and millions of dollars.² These types of integrations have also been used to define endpoints and discover novel markers of progression.

Life sciences organizations can also use claims data or similar RWD to estimate burden of illness, which can be critical to developing market access, pricing, and HEOR strategies. While these functions traditionally used RWD in later phases of product development, early integration of this data can help companies design trials that better address both patient and payer needs. By incorporating data linkage methods into earlier stages of RCT design, sponsors can better understand long-term safety, efficacy, and economic outcomes across and outside RCTs.

Challenges in Data Linking

Combining disparate data sources holds transformative potential for healthcare, but the breadth of data types being used, and need to maintain patient privacy, makes data linking a challenge to implement while maintaining de-identification of the data. The value it generates through improving clinical trial efficiency can be enormous, but especially as more complex data types such as genomics data are increasingly utilized in linked data analyses, organizations need to factor in the need for specialized analysis expertise and licensing to deal with proprietary data formats when weighing the cost of the initial investment. Even for traditional RWD sources like claims or EHR, quality and consistency can be highly variable due to the range of different providers patients will visit throughout their life, potential for operator error when entering information into databases, and life changes that can alter basic details like name or ZIP code.

The breadth of RWD necessitates increasing vigilance for potential privacy breaches associated with transfer and linkage of patient data. While modern algorithms can de-identify many types of RWD, certain types of data such as whole genome and transcriptome data cannot be de-identified. In these cases, special attention must be paid to ensure that consent to use the data was properly documented, and that storage and access to data complies with regulations such as HIPAA.



By investing in data solutions with robust protections for privacy that have been designed to comply with HIPAA's Expert Determination or Safe Harbor methods,³ organizations can ensure that they remain compliant while also reassuring patients who wish to participate in the research process.

Methodologies for Data Linking

Organizations need to assess different methods and data sources available and choose ones that meet standards for quality, privacy, and transparency. Linking solutions must be able to de-identify data in accordance with the HIPAA Privacy Rule while retaining the ability to match patients across datasets with high accuracy and minimal risk of re-identification while providing clear audit trails for regulators.

Methods of data linking largely fall into two main categories:

Deterministic

Deterministic matching attempts to link data through matching of exact identifying information between records, such as name, social security number, or address.

Probabilistic

Probabilistic matching attempts to link data through analysis of multiple different fields to predict which records belong to the same individuals.

While deterministic matching can be highly precise, it also relies on accurate, high-quality data and cannot account for data discrepancies arising from misspellings, name changes, or similar events. Probabilistic methods are typically more robust to these data quality issues, especially when coupled with machine learning (ML) algorithms to improve accuracy and scale. However, it is important to keep in mind that precision and recall ability can vary significantly, so it is critical to consider independent performance validations when choosing any method.

In both methods, modern privacy-preserving record-linkage (PPRL) are used to protect the privacy of patients. These typically rely on two different approaches: hashing and tokenization. In hashing, data is run through an algorithm which scrambles the information and produces a unique identifier string which cannot be traced back to the original input without a key. Tokens are unique identifiers that are created by third parties that enable identification of specific patients across multiple datasets allowing for PHI to be removed from a dataset. This method can be more secure and provides more separation between the life sciences organization and the PHI, but is more costly as it relies on a third-party partner to provide tokens and act as a data matchmaker. Tokenization companies can also create centralized marketplaces with multiple pre-linked datasets for healthcare companies to access and analyze.

When choosing which method to invest in, a best practice is to consider the specific downstream objectives for the linked data as early as possible, to avoid costly mistakes and ensure that the resulting dataset will provide additional value and satisfy the needs of the relevant stakeholders. Companies investing in RWE for regulatory approvals, for example, will need to ensure that the RWE they are using satisfies the published considerations of the FDA⁴ or other regulatory bodies they plan to file with.



Applications in Oncology

In oncology, data linking has been used to drive regulatory approvals. A recent metastudy found that of 254 FDA-accelerated approvals between 1992 and 2020, 119 (47%) were approved for oncology indications using single-arm trials,⁵ demonstrating an increasing need for reliable historical controls to support these trials.

Beyond regulatory approvals, linked claims data have been increasingly utilized to reveal insights into the impact of SDOH on care utilization and health outcomes in oncology. In a recent study, linked data including Medicare FFS claims, commercial multi-payer claims, and SDOH data were used to discover that [Black and Hispanic patients with hepatocellular carcinoma \(HCC\)](#) were visiting the emergency department more frequently relative to White and Asian patients, with further stratification of behavior based on income level.^{6,7}

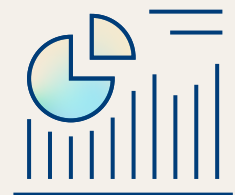
In another recent study exploring the [impact of SDOH on health outcomes](#), patients with HCC who had received surveillance were significantly more likely to receive curative therapy, while Black patients, patients with lower education levels, and patients diagnosed during the COVID-19 pandemic were significantly less likely to receive curative therapy.⁸ By utilizing linked Medicare FFS claims, commercial multi-payer claims, and SDOH data, this study revealed novel insights about how increasing access to HCC surveillance could improve treatment outcomes and reduce total cost of care in underserved patient populations.

Applications in Autoimmune Disease

Similarly, in autoimmune diseases, linked claims data have generated RWE insights including real-world patient behavior, health outcomes, and care utilization. One study found that patients with seropositive (rheumatoid factor [RF]+ and/or anti-cyclic citrullinated peptide+) rheumatoid arthritis (RA) treated with the RA immunomodulatory drug abatacept were more likely to continue their therapy compared to patients on TNF- or JAK-inhibitors, potentially indicating greater efficacy of abatacept in this population, which had also been suggested by exploratory RCT data.⁹ These insights were made possible by linking Medicare FFS Parts A/B/D claims to Prognos laboratory data. A follow-up study utilizing the same types of linked claims and laboratory data further explored treatment persistence patterns in treatment-naïve patients with dual-positive RA, again finding that abatacept users had the highest 12-month persistence rate and median time to discontinuation compared to TNF- or JAK-inhibitor-treated patients.¹⁰

In lupus, linked claims data have been used to explore the efficacy and usage patterns of disease monitoring tests as well as improve understanding of patient demographics. Patients with systemic lupus erythematosus (SLE) often have periods of uncontrolled disease activity that can lead to irreversible organ damage, creating a need for monitoring tests that are more effective at detecting disease activity than standard of care (SoC) tests. A recent study compared a lupus activity monitoring panel with standard tests for measuring blood-based biomarkers of SLE disease activity.¹¹ Data sources included the United Rheumatology-Normal Integrated Community Evidence repository and medical and pharmacy claims from Medicare, managed Medicaid, and commercial plans. Combining data from varied sources allowed the researchers to identify patients in community settings who had either taken at least two lupus activity monitoring panel tests or at least one standard test.

While more research is needed to compare the clinical outcomes of these tests, the study found a link between the size of the practice and the choice of test and provided insights in the variability among patients with SLE.



Applications in Rare Disease

Across healthcare, data linking approaches can broaden the ability to answer complex research questions. For example, rare diseases pose a major challenge for both clinical and commercial development teams. While it is estimated that rare diseases affect 400M people globally, only about 10% of rare diseases have a treatment.¹² In the United States, the FDA has defined rare diseases as a disease or condition affecting less than 200,000 people in the United States,¹³ meaning that recruiting enough patients with the rare disease to carry out a RCT is challenging. Understanding the true prevalence as well as progression is challenging, which makes statistical analysis plans and clinically meaningful outcome measures difficult to design. Additionally, sponsoring organizations may not even get to this stage of development since their understanding of the population prevalence and commercial potential of the product is extremely limited.



In rare disease, linked data have become a powerful tool to help define future trial outcome measures and better understand disease epidemiology. A study using longitudinal data from multiple international sites and database EHR across Huntington's disease gene carriers, early patients, and healthy controls found that serial brain imaging would likely be a more sensitive measure compared to measures related to quality of life, which could be used in future clinical trials.¹⁴ In a U.S.-based study, administrative variables from both public (Medicaid) and private insurer claims and ICD-9-CM codes were used to conduct passive surveillance to identify cases of muscular dystrophy.¹⁵



Future Directions and Opportunities

While data linking approaches have already been used to advance scientific knowledge in a wide range of therapeutic areas including oncology and rare disease, data linking is continually improving, bringing both new opportunities and challenges to the field. For example, incorporating machine learning algorithms into probabilistic PPRL methods can potentially improve linkage accuracy and efficiency. However, these innovations come with new questions about how data is used to train algorithms, and how the resulting outputs will be utilized by life sciences and payers to make product and/or reimbursement decisions that can affect millions of patients.

As data linkage becomes more commonly used, transparency and privacy discussions need to be at the forefront. While third-party oversight is common in RCTs, this has not necessarily been the case in RWD. But the FDA has created a framework¹⁶ to help guide organizations seeking to support regulatory decisions with RWE, and an increasing number of initiatives and task forces focused on key issues concerning the transmission, analysis, and usage of potentially massive, complex, linked datasets have been created.

Life sciences organizations seeking to leverage linked RWD to generate and use RWE in future development decisions should seek to partner with these types of initiatives and task forces sooner rather than later.

Examples of joint initiatives that have proposed ways to promote transparency in RWD usage include the RWE Transparency Initiative, which included representatives from The Professional Society for Health Economics and Outcomes Research (ISPOR), the International Society for Pharmacoepidemiology (ISPE), National Pharmaceutical Council, and Duke-Margolis Center for Health Policy.



Conclusion

Data linking can transform healthcare by improving our understanding of disease epidemiology and progression, enabling more efficient, diverse, representative clinical trials, and streamlining market assessment and payer discussions. Considering the wealth of RWD that is available and being produced, healthcare and life sciences organizations should consider implementing data linking and analysis strategies into their tactical decisions to advance research and healthcare delivery. Establishing partnerships with innovative data partners like Inovalon can help ensure that the data linking strategies invested in today are aligned with organizational goals and healthcare challenges of tomorrow.

Learn how we can help your organization unlock the value of data linking through the industry's largest primary source closed claims dataset, extensive in-house expertise, and a track-record of working with major pharmaceutical companies and other healthcare and life sciences organizations to transform disparate data into RWE.

[Contact Us](#)

Sources

1. "Harvard-Inovalon Medicare Study: Who Enrolls in Medicare Advantage vs. Medicare Fee-for-Service." Bilder S, Brot-Goldberg Z, Jones B, Mohammadi I, Pulangan Z, Su Y, Teigland C, Vabson B. June 2023. <https://www.inovalon.com/wp-content/uploads/2023/06/Harvard-Inovalon-Medicare-Study.pdf>
2. Martina R, Jenkins D, Bujkiewicz S, Dequen P, Abrams K. The inclusion of real world evidence in clinical development planning. *Trials*. 2018. 19: Article 468.
3. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." Created by Office for Civil Rights. Oct. 2022. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharboriguidance>
4. "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." Aug. 2023. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug>
5. Ribeiro TB, Bennett CL, Colunga-Lozano LE, Araujo APV, Hozo I, Djulbegovic B. Increasing FDA-accelerated approval of single-arm trials in oncology (1992 to 2020). *J of Clin Epidem*. 2023. 159: 151-158.
6. Lin D, Tan A, Kilgore KM, Teigland C, Hernandez S, Slay DL, Schroeder AL, Kim S. Impact of Race/Ethnicity and Socioeconomic Status on Emergency Department Visits in Patients With Hepatocellular Carcinoma. *J Clin Oncol*. 2022 40:4_suppl, 385-385.
7. Lin D, Tan R, Teigland C, Hernandez S, Kim S, Kilgore KM. Race/ethnicity and socioeconomic position in emergency department utilization in patients with hepatocellular carcinoma. *Future Oncol*. 2024; Apr 19:1-13.
8. Singal AG, Kilgore KM, Shvets E, Parikh ND, Mehta N, Ozbay AB, Teigland C, Hafez O, Schroeder A, Yang A, Schinkel J. Impact Of Surveillance And Social Determinants Of Health On Treatment Receipt And Healthcare Costs In Patients With HCC. (Podium). *Digestive Disease Week*. Chicago, IL. May 6-9, 2023.
9. Park SH, Schwartz TT, Han X, Robinson S, Kakehi S, Wittstock K, Norris K, Murunga A, Silverstein AR, Sparks JA. Treatment Persistence Among Medicare Beneficiaries with Seropositive Rheumatoid Arthritis Initiating Biologic or Targeted Synthetic DMARDs. *Arthritis Rheumatol*. 2021; 73(suppl 9).
10. Sparks J, Zheng H, Schwartz T, Park SH, Norris K, Patel V, Robinson S, Wittstock K, Khaychuk V, Silverstein A. Treatment Persistence in First-Line Biologic and Targeted Synthetic DMARD Users with Dual-Seropositive RA in the Medicare Population. *Arthritis Rheumatol*. 2022; 74 (suppl 9).n
11. Concoff A, Moore-Schlitz L, Nadipelli VR, O'Malley T, Worley K, Petrilla A, Zack DJ, Robinson SB, Taiyari S, Jones B, Rubin B. The Real-World Utility of a Lupus Activity Monitoring Panel in the United States Community Rheumatologist Practice. *Arthritis Rheumatol*. 2023; 75 (suppl 9).
12. Han Q, Fu H, Chu X, Wen R, Zhang M, You T, Fu P, Qin J, Cui T. Research advances in treatment methods and drug development for rare diseases. *Front Pharmacol*. 2022; 13: 971541.
13. "Rare Diseases at FDA." Dec. 2022. <https://www.fda.gov/patients/rare-diseases-fda>
14. Tabrizi SJ, Reilmann R, Roos RAC, Durr A, Leavitt B, Owen G, Jones R, Johnson H, Craufurd D, Hicks SL, Kennard C, Landwehrmeyer B, Stout JC, Borowsky B, Scahill RI, Frost C, Langbehn DR; TRACK-HD investigators: Campbell M, Milchman C, Coleman A, Dar Santos R, Decolongon J, Sturrock A, Jauffret C, Justo D, Lehericy S, Marelli C, Nigaud K, Valabregue R, Bechtel N, Bohlen S, van den Bogaard SJA, Dumas EM, van der Grond J, 't Hart EP, Callaghan J, Fox NC, Gibbard C, Hobbs N, Lahiri N, Malone I, Patel A, Pepple T, Read J, Say M, Wild E, Cash DM, Berna C, Axelson E, Acharya T, Queller S. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol*. 2012; 11(1):42-53.
15. Smith MG, Royer J, Mann JR, McDermott S. Using administrative data to ascertain true cases of muscular dystrophy: rare disease surveillance. *JMIR Public Health and Surveillance*. 2017; 3(1):e2.
16. "Real-World Evidence." Feb. 2023. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>

About us

Inovalon is a leading provider of cloud-based SaaS solutions empowering data-driven healthcare. The Inovalon ONE® Platform brings together national-scale connectivity, real-time primary source data access, and advanced analytics into a sophisticated cloud-based platform empowering improved outcomes and economics across the healthcare ecosystem. The company's analytics and capabilities are used by nearly 20,000 customers supporting more than 53,000 sites of care. For more information, visit www.inovalon.com.

Inovalon
4321 Collington Road
Bowie, MD 20716
(301) 809-4000

www.inovalon.com