

Reducing Bias in Non-Interventional Research with Greater Visit and Data Density

AUTHORS

Reema Patel, MPH, Epidemiology Solutions Principal, PicnicResearch

Dan R. Drozd, MD, MSc, Chief Medical & Scientific Officer, PicnicResearch

Emily Cibelli, PhD, Quantitative Science Lead, PicnicResearch



Background

Non-interventional research routinely relies on data obtained from standard of care visits to describe the patient journey, to answer research questions on burden of disease, unmet needs, treatment patterns, resource utilization, and clinical outcomes. Traditional methods of medical record retrieval, such as chart abstractions or application programming interface (API) pulls, are dependent on access to specific sites, networks, EHR vendors, or payer coverage networks. Given the fragmented US healthcare system, data pulls using these methods typically generate incomplete datasets.

More advanced data retrieval methods are now used widely but vary across service providers. One such method may involve requesting medical records on behalf of the patient with their consent and authorization and without increased burden. Another method involves multi-site data retrieval across all health encounters for a patient through both electronic data networks (i.e., health information exchanges) and direct-to-facility record requests on an automated and recurring cadence, regardless of EHR vendor or payer coverage network. Advanced data retrieval methods, especially when used in combination, lead to better breadth and density of data and a more complete representation of the patient journey.

Reducing Bias and Confounding With Greater Data Density

There is growing recognition and expectation that real-world data (RWD) used to generate real-world evidence (RWE) should be both relevant and reliable in order to be considered fit-for-use in non-interventional studies.^{1,2} Completeness is fundamental to establishing reliability of RWD, especially when used for regulatory purposes.² More specifically, breadth and density of data are critical and complementary dimensions of completeness.³

Robust, advanced data retrieval methods can support completeness by means of greater visit density, defined herein as the frequency and breadth of healthcare visits recorded for any given patient. Capturing greater visit density enables greater data density and breadth for key outcomes and covariates as more relevant and complete data are captured from the full spectrum of a patient's healthcare interactions. This serves to mitigate the inherent challenges of selection bias, information bias, attrition bias, and confounding that threaten the validity of non-interventional research.

Type of bias or confounding	Examples	Mitigation through advanced data retrieval methods
Selection bias occurring from limited sampling of patients from specific sites of care or specific EHR systems that are not representative of the greater patient population	Medical records are abstracted from dermatology sites, but only severe patients regularly seek care; patients with milder disease are seen by a primary care provider whose medical records are not available in the dermatology site, overestimating risk and severity of disease. Patients from an EHR dataset are screened for eligibility based on available data and excluded due to missing data or not meeting criteria; however, patients received prior or concurrent care from other settings from which medical records could confirm eligibility.	The ability to include patients in a study irrespective of their propensity to seek and access healthcare, geography, coverage, or site affiliation captures a more diverse, representative sample and better supports the generalizability of results. Furthermore, data capture across multiple settings of care reduces bias from overrepresentation of more severe patients and treatment underresponders who are more likely to seek specialized care or patients who receive care in specific healthcare systems only.

Information bias occurring due to poor exposure classification or outcome ascertainment

Patients being treated for an immunodeficiency and are under the care of an immunologist but receive emergency care for acute infections at a hospital results in missing events of interest if records are not shared across sites.

Patients with a complex, chronic disease that is often misdiagnosed may have a medical history of recurring laboratory and pathology results that predated their diagnosis and may not be recorded in the treating specialists medical records.

Coding errors in a specialists' medical records may lead to inaccurate interpretation of patients' condition or progression, but can be validated using medical records obtained from referring physicians and laboratory tests.

Automated and recurring data retrieval can provide continuous and longitudinal data on key exposures and outcomes, identifying signals or anomalies and validating data elements across multiple settings of care.

Data from medical records within one provider or network can be referenced across other providers or sites from which the patient seeks care, thereby classifying an exposure or outcome as being truly absent, present, or simply missing from a set of medical records.

Greater data and visit density reduces measurement errors and increases confidence of accurate classification of exposures and outcomes.

Attrition bias occurring due to patients who choose to seek care elsewhere not being representative of those who remain in the study

Patients whose condition worsens or results in treatment-related adverse events prefer to seek care in emergency settings or with a different specialist; outcomes or disease progression are not documented and underestimate poor outcomes or treatment failures.

Patients with mild severity or improved condition no longer need recurring or specialized care and are lost to follow-up, thereby overestimating poor outcomes and underestimating treatment response based on those who remain in the study.

Automated, longitudinal data retrieval, even after a patient leaves the site or healthcare system under observation, can validate the occurrence or true missingness of outcomes and events of interest rather than erroneously treating them as censored.

Confounding occurring when missing variables cannot be adjusted for when assessing the relationship between an exposure and outcome

Risk factors, such as smoking history and chronic comorbid conditions like asthma, were first documented or diagnosed several years prior to baseline in another care setting and not captured in the current medical records.

The ability to capture a more complete list of comorbidities, precise lab values, or indicators of disease severity over a longitudinal period better supports adjustment to reduce residual confounding, for example, when using propensity score models to balance on baseline characteristics.

Advanced medical record retrieval can be automated, recurring, and applied irrespective of EHR system, geography, site or hospital affiliation, or coverage network. By encompassing inpatient hospitalizations, outpatient visits, emergency department visits, laboratory test results, imaging, and specialist consultations, data are enriched to generate more credible and reliable RWE.

- **Representative patient population:** Patients with varying severity of disease, healthcare utilization patterns, and access to healthcare can be included in non-interventional research because of the ability to retrieve medical records agnostic to these factors. For example, patients from rural and urban settings can be included in a study even if their access to care is characterized by specific types of sites, such as emergency rooms or urgent care clinics, rural health clinics, or a combination of specialists and internists in different healthcare systems.
- **Credible exposure classification:** Patients often receive treatments from various providers – a specialist might prescribe a medication or an independent pathology laboratory may provide detailed disease severity. Capturing specialist notes, pharmacy records, and laboratory results ensures exposures and interventions are fully recorded, reducing misclassification of exposures. Complete medication and treatment histories are critical for non-interventional studies in order to determine eligibility, baseline characterization and matching, and adjustment for risk factors and confounders.
- **Better outcome ascertainment:** Different health outcomes manifest in different settings. For example, serious adverse events may result in hospital admissions, while milder events might only be noted in outpatient visits. Including all care settings in the record retrieval process for complete capture improves outcome ascertainment.

Visualizing Visit Density

We compare two advanced retrieval methods to understand how they may mitigate the discussed biases and confounding in real-world data.

Patients with paroxysmal nocturnal hemoglobinuria (PNH) experience a chronic and complex condition with recurring primary care and emergency room visits early in their journey, followed by often delayed diagnosis, multiple hematology and laboratory visits, and long-term, specialized treatment management. An evaluation of data retrieved by PicnicResearch for a PNH study using PicnicHealth's advanced, best-in-class, patient-mediated and multi-site retrieval method compared to data retrieved by a leading, large health data platform company using an alternative, advanced multi-site retrieval method demonstrates greater visit density with respect to time, frequency, and types of routine clinical encounters. More dense, longitudinally captured data across all of the patients' healthcare interactions improves confidence that any observed outcomes and associations are not artifacts of missing data or hidden biases.

Our data retrieval method yielded a higher density based on both the frequency and types of visits that occurred over the same retrospective time period. For example, we captured an average of 207 total visits per patient compared to 54 using the alternative retrieval method. For capturing critical outcomes in PNH, such as major adverse vascular events and laboratory markers of hemolysis, capturing data from emergency room and laboratory visits is essential. We identified an average of 34 laboratory and 8 emergency room visits per patient compared to 30 and 1, identified using the alternative method. The greater density was due to our multi-site data retrieval method that was not limited to only a prescribing physician (e.g., hematologist/oncologist). We identify the patient's providers who play a critical role in the patient's disease journey to ensure records were requested from across all sites the patients were receiving care.

Figure 1: Comparison of two types of data retrieval: PicnicResearch vs large health data platform company. In addition to greater longitudinal coverage of patients' medical histories, the PicnicResearch method captures a greater variety of visit types.

KEY

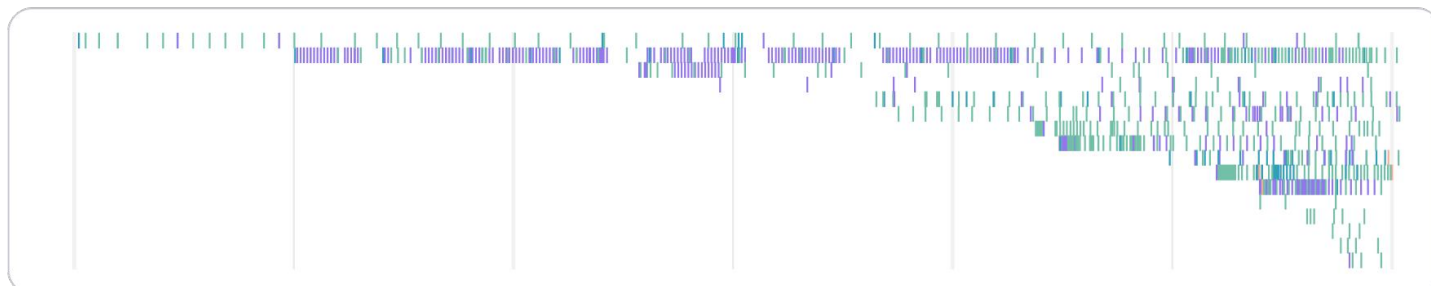
HEM/ONC VISIT

OUTPATIENT VISIT

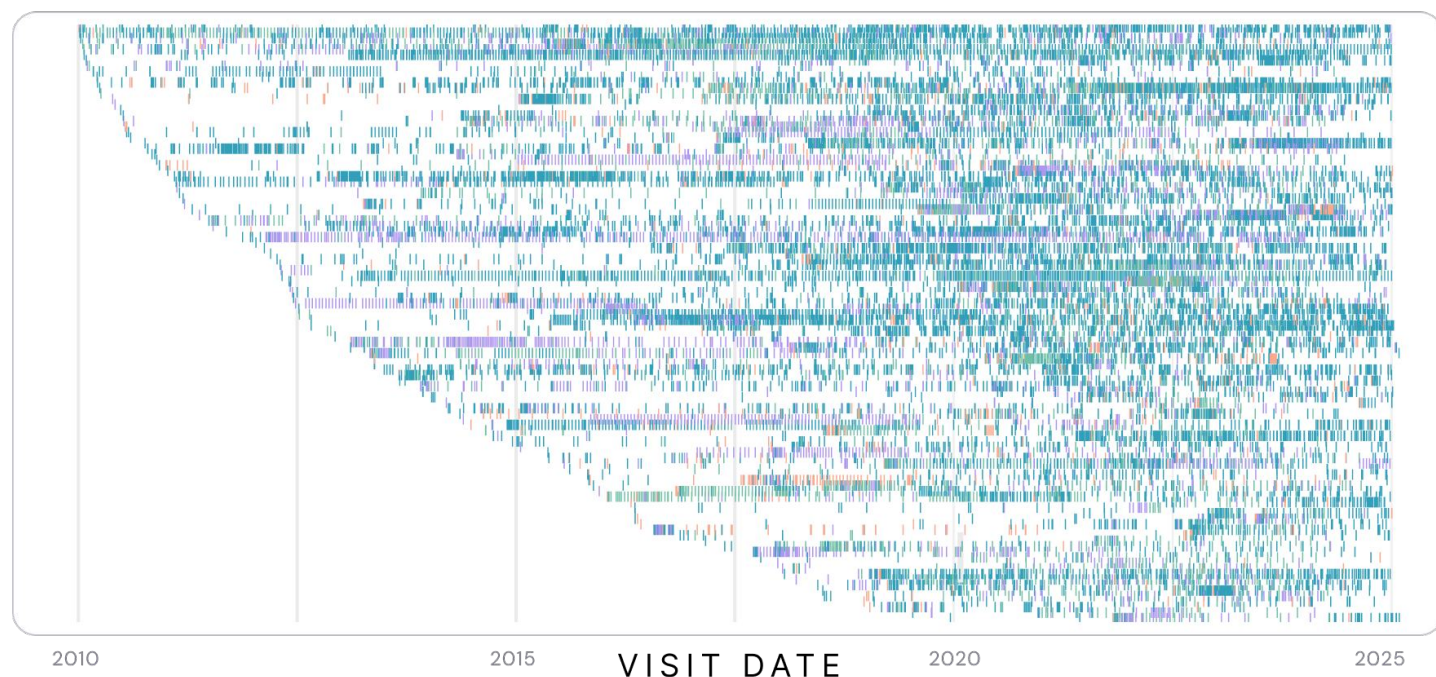
INPATIENT VISIT

LABORATORY VISIT

Visits retrieved by a leading, large health data company



Visits retrieved by PicnicResearch



Conclusion

Greater visit density from advanced, best-in-class medical record retrieval, which includes patient-mediated and comprehensive multi-site retrieval, enables more complete data to support credible baseline adjustments, exposure classification, and outcome ascertainment to mitigate the challenges of bias and missingness that are inherent in non-interventional research. On a spectrum of fit-for-use data to support RWE generation, data captured with greater density and breadth provide more reliable and relevant data for complex research.

Talk to our team about how greater data density can strengthen the evidence in your study.



References

1. U.S. 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. August 2023. Available at: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making>. Accessed May 28, 2025
2. Food and Drug Administration. Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products. Draft guidance. FDA-2023-D-5470. Published 2023. Available at: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-evidence-considerations-regarding-non-interventional-studies-drug-and-biological-products>. Accessed May 28, 2025.
3. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use.



PicnicAI builds intelligence to accelerate human health. The company harnesses the power of AI to address the complexity of healthcare delivery and clinical research through two product lines: PicnicHealth and PicnicResearch. PicnicHealth empowers patients with personalized guidance, actionable insights, and clinical support throughout their healthcare journey. PicnicResearch accelerates the pace of clinical research by making studies faster, more efficient, and more reliable, helping bring new treatments to patients sooner. Together, these efforts reflect PicnicAI's mission to use AI to transform human health and expand access to life-changing therapies. Learn more at picnic.ai.