



Healthcare AI Needs Contextual Data



www.veritasdataresearch.com

TABLE OF CONTENTS



HEALTHCARE AI NEEDS
CONTEXTUAL DATA

02

HEALTHCARE AI NEEDS
CONTEXTUAL DATA

03

WHY CONTEXT MATTERS
IN HEALTHCARE

04

ADDRESSING THE
CONTEXTUAL DATA GAP

06

REFERENCES

09

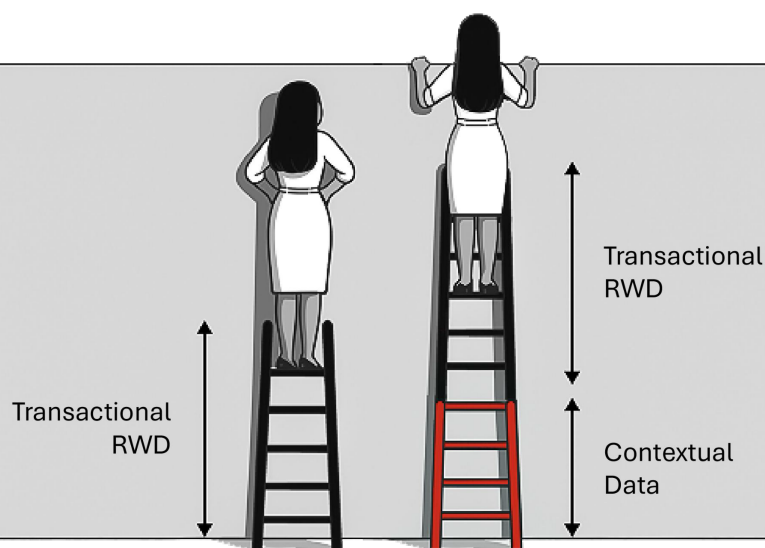
ABOUT

10

HEALTHCARE AI NEEDS CONTEXTUAL DATA

The availability of **real-world data (RWD)** has undoubtedly revolutionized the time, cost, and utility of analytics in healthcare. Spurred by the wide-spread adoption of privacy-protecting technologies, new players like HealthVerity and Datavant have joined industry stalwarts like IQVIA to make “big data” available through their ecosystems (partners) for analysis at unprecedented scale. However, despite this emergence of new business models, approaches, and value propositions, the insight we expect to be inherently available from RWD remains elusive or imprecise.

Why might this be? By “big data” in healthcare, what most people mean is “transactional” data generated from individual healthcare encounters like filling a prescription or visiting a doctor. And while this transactional RWD has become critical in letting us know “what” is happening to patients in near real time, AI and other advanced analytics are going to struggle to answer “why” patients have different experiences and outcomes because we are missing a key ingredient: contextual data.



Why do patients
have **different**
outcomes?

CURRENT HEALTHCARE REAL-WORLD DATA IS TRANSACTIONAL

Most health analytics today are primarily built on **three types of RWD**:



- **Pharmacy claims data:** what drug and dosage was dispensed, to whom, and by which physician.
- **Medical claims data:** request for insurance payment for physician visits, with diagnosis and procedure codes and, sometimes, treatment information.
- **Electronic health records:** a more complete medical record of patient visits, lab tests, and procedures from a health system (often excluding physician notes and other unstructured fields).

Each of these **RWD types are available at scale**, often allowing analysis of millions of patients in a single sample. The data are updated quickly, often with new data becoming available within days of a medical encounter. Researchers have become incredibly adept at working with these data types to answer retrospective questions about medical practice patterns and are now gaining proficiency at building observational study (i.e. no clinical intervention but RWD only) designs that will accurately measure the safety and efficacy of drugs in standard clinical practice. In fact, this observational study paradigm was critical in assisting the FDA in making rapid determinations about questions like “Is hydroxychloroquine an effective treatment against COVID-19?” early in the pandemic¹.

However, prescriptions, claims, and even EHR data are records about specific transactions or encounters. This type of data is great for measuring “what” happened, “where” it happened, and “when” it happened. And while this data unlocks many possible analyses, the market so far remains insight-poor. What transactional RWD is insufficient at representing, and therefore what analytics and AI will struggle to elucidate, is “why” things happen. **For that, we need something more.**

WHY CONTEXT MATTERS IN HEALTHCARE



There are over **1,200,000 practicing physicians** (among over 2,900,000 clinicians) in the United States², spread across more than 6,000 hospitals³ and tens of thousands of urban and rural clinical facilities. Context about a patient's specific clinical environment, then, offers a critical new set of variables that need to be considered in understanding why a patient was diagnosed and treated in a particular manner:

Physician profile data

Each physician was trained through a unique mix of which medical school they attended, where they did their residency, and who they practiced with. And each physician's experience is shaped by what era of medicine they have practiced in, their relationship with pharm. companies, and their approach to innovation. These factors mean that no matter how much standardized continuing medical education they consume, or published best practices they read, every physician's approach to medicine is slightly (or wildly) different.



There are over
1,200,000
Practicing Physicians
in the United States

System profile data

Physicians don't practice in a vacuum. They are constrained to lesser or greater degrees by the capabilities of the facilities they work in, the policies and preferences of the health system or integrated delivery network they belong to, and the referral network available to them. The system norms that surround the physician's clinical pathway thus become deterministic of the treatment pathway they pursue.

Health plan profile data

The rules governing what care a patient's insurance will pay for is a dominant factor in diagnostic and treatment decisions, and including these variables as part of care analyses can illustrate how much patient care is influenced by the physician and how much by the payer.

Additionally, there are all the contextual variables that lie outside of a specific health encounter, which some researchers believe can determine 83% or more of a patient's outcome⁴:

Social Determinants of Health (SDOH)

This broad term encompasses a wide variety of contextual data, and can include demographic variables such as race, ethnicity, religion, education and socioeconomic status, as well as other attributes important in healthcare such as access to food, housing, transportation to (and presence of) health facilities, the presence of a caregiver in the home, and more.

Behavioral data

Patients' behavior, including purchasing activity, is often tracked for consumer marketing uses. This data can be leveraged for health analytics to understand variables like the type of food a patient eats, participation in fitness activities, and use of over the counter (OTC) therapies.

Environmental risk factors

A lifetime of environmental exposures (the "exposome") is also gaining recognition as an important set of health variables, but one that is not reflected in common RWD sources.

Vital status

While survival is the most basic goal for treatment, and a common endpoint in any clinical study, the vital status of a patient (whether they are alive or dead) is a variable that is poorly captured in transactional health data, but can be found outside of health data.

It isn't hard to see that many of these contextual variables could greatly increase our understanding of why some patients are diagnosed correctly or not, why some have access to one therapy versus another, and why some patients avoid complications and others prematurely perish. And yet, contextual data are NOT included in most of today's health analytics.

ADDRESSING THE CONTEXTUAL DATA GAP



The **next revolution in healthcare data** will be bringing contextual data online in ways that are affordable, widely accessible, and delivered in ways that make it easy to incorporate into analytics and AI modeling. Contextual data is widely available (though often in messy and opaque data formats that require expertise to untangle). And while you need to know where to find it, much of this data is very inexpensive to obtain from various local, state, and federal government databases.

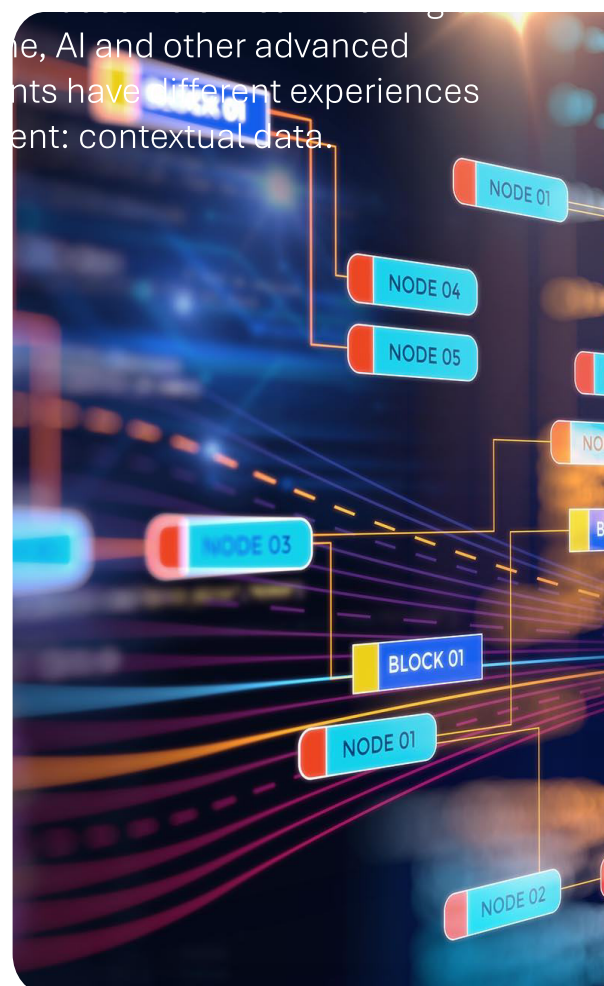
Ironically, however, the public availability of contextual data has greatly reduced the commercial interest in collecting and selling this data, with potential vendors believing that the lack of intellectual property protection will lead to numerous “fast-following” competitors and rapid price erosion.

Thus, while many companies have built these contextual data sets, they have only done so as an internal resource upon which they can append their transactional data. In so doing, the contextual data gets buried within their proprietary data and is never made widely available, meaning the next group who needs it must build their own version of that same data from scratch. Furthermore, the provenance of the contextual data is lost and leads to, sometimes, unnecessary questioning of valid conclusions.

This model of proprietary data creation is highly inefficient and has resulted in most data and analytics groups not bothering to invest in building strong contextual data or, if they do, creating non-uniform data sets that invisibly alter analytic results from one group to the next (i.e. assigning a high-prescribing physician to a different health system in two otherwise similar datasets would create variability in an analysis of those system’s treatment patterns).

To successfully address this gap in contextual data will require **solving the following problems:**

...e, AI and other advanced
nts have different experiences
ent: contextual data.



Expertise

While contextual data is often publicly available, finding those actual sources and accessing them requires substantial subject-matter expertise. Knowledge of the data, where it is generated, how it flows, and where to access it is critical to assembling a complete and accurate contextual database. And knowledge of the compliance barriers to accessing and using the data is required to successfully navigate the regulatory, certification, or other accreditation steps to be approved for access. Finally, once received, the data itself is very “raw” and requires deep expertise to map and organize the disparate formats into a common data model.

Many-to-One” Data Indexing Platform

Contextual data sources are incredibly fragmented and highly varied in format; therefore, the creation of a high-quality data set involves building a scalable and flexible data “factory” that can routinely monitor tens of thousands of sources, ingest and index data in whatever format it is available, and standardize it into a common data scheme. Because much of the data will be redundant (or worse, primarily the same with some variation), this factory also needs to be able to rapidly de-duplicate and rationalize multiple inputs about the same entity and create a single master index record.

Many-to-One” Data Delivery Platform

As a data vendor, it can be tempting to focus primarily on the creation of the data set itself but given the broad span of potential consumers for a contextual data product, it is equally important to deliver that data in a manner easily consumed across many different IT environments and data workflows. Serving this breadth requires building the systems to deliver files via sFTP or cloud hosting, or direct sharing through platforms like Databricks and Snowflake and application program interfaces (APIs). And it means investing in high levels of configurability so that file layouts and formats can match what consumers’ systems can handle.

Radical Data Transparency

Making contextual data available to support a wide range of analytics only improves outcomes if that data is used correctly, and if it can be vetted by consumers to ensure its accuracy and consistency. (It is for these same reasons that FDA has released guidance for transparency in the use of data when generating real-world evidence for regulatory submission). To satisfy this requirement, any contextual data set must be as transparent as possible about the sources from which each individual record is derived, and ideally link the user to those sources directly.

(Historical) Currency

Contextual data is only useful if it is current to the healthcare event under study. To understand why a patient has just been re-admitted, it is important to have near real-time contextual data of what their behavior has been in the days prior to re-admission. Likewise, if studying the path to diagnosis a patient received five years ago, it's important to have contextual data from that period. Notably, most consumer and behavioral data sources do not have historical data readily available, making this gap particularly hard to fill.

Neutrality

To effectively provide contextual data at scale requires serving a wide variety of data vendors and analytics players, which can only happen if those stakeholders do not view the data provider as a competitor. A successful business model for serving this space therefore must be completely neutral and remain strictly within the contextual data niche. Neutrality in this business model can then translate to commonality in contextual data across research projects, leading to more uniform analytic results and increased clarity in research findings that benefit patients.

Veritas Data Research is building best-in-class contextual data sets, which will form the foundation for better understanding of patient experiences. If we are successful, we will help unlock the next stage of progress in big data, AI, and healthcare analytics and allow new insights into why patient care and outcomes differ, and how to improve them.

REFERENCES

- [1] <https://pink.citeline.com/PS142063/RealWorld-Evidence-On-COVID19-US-FDA-Approaching-With-Sense-Of-Urgency>
- [2] Veritas Data Research. (2025) Provider Index and Affiliations
- [3] <https://www.aha.org/statistics/fast-facts-us-hospitals>
- [4] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6827626/>

ACKNOWLEDGEMENTS

Thank you to all of the folks who provided thoughtful commentary in creating this piece, including [Kevin McCurry](#), [Leon Flek](#), [Auren Hoffman](#), [Gillian Cannon](#), [Gaurav Singal](#), [Vera Mucaj](#), [Helen Moran](#), [Andrew Kress](#), [Rob Nagel](#), [Bobby Samuels](#), [Travis May](#), and [Shahir Kassam-Adams](#).

ABOUT



Jason has 25 years of experience in health information and technology. He has a Ph.D. in virology from Harvard, and an A.B. in molecular biology from Princeton.

Jason LaBonte

CEO - Veritas Data Research

ABOUT VERITAS DATA RESEARCH

Founded by **experts in the data analytics industry**, Veritas uses cutting-edge technology and efficient workflow design to collect, curate, and corroborate foundational contextual datasets. Veritas makes critical information accessible to data/analytics teams and companies, their customers, and other stakeholders.

To learn more about us, visit veritasdataresearch.com and follow us on [LinkedIn](#).