Key Takeaways

- Today’s healthcare data fails to capture the full diversity of interactions patients have with the healthcare system, leading to incomplete information that can put patients at risk.

- Systematic selection and sampling biases have long plagued healthcare’s largest commercially available data sets, and analytics built on biased data sets can deliver misleading insights.

- If healthcare is going to achieve optimal outcomes for patients, data must be representative and complete.

- Healthcare data that features complete patient journeys will be able to power the highest value applications of AI in healthcare.
In the real world, a patient’s journey through the healthcare system is incredibly complex—health is, after all, affected by a varied combination of chronic disease, comorbidities, genetics, and the dynamics of the social environment. Our understanding of this journey, however, continues to rely on narrow slivers of a patient’s record and a view of healthcare built on availability. And it is this overly generalized view that informs the development of new products, services, care protocols, and policies.

For decades, the best way to bring a patient's information together to understand their journey through the healthcare system has been by manually incorporating primary research, one-off chart reviews, and analyses of secondary data. These costly, time-consuming and manual methods often require reconciling clinical, financial, and transactional data that are scattered digitally over several applications and database servers and physically in the file cabinets of many provider organizations. The data often lacks population stability and loses visibility of patients as they “travel” across multiple services and care settings over time. Reconciling the data's bias and fragmentation also requires intensive manual intervention and interdisciplinary collaboration of subject matter experts.

The alternative—constructing journeys from actual patient-level data at scale—is significantly more troublesome. Robust data across payers and providers is limited; lives are poorly sampled, with no guarantees that the journeys are complete; and the lag between data creation and availability can be significant—often six months or more. Data sets may also be filled with discrepancies. On the rare occasions in which a patient’s healthcare record is complete, specific provider, institution, and payer information is often suppressed—and crucial contextual information along with it.

As a result, it has become impossible for innovators and decision makers to accurately contextualize patients' lives to identify the largest unmet medical needs or to target interventions. Without shifting from a few high-level “personas” to in-depth micro-segments, future behaviors cannot be accurately predicted. Without real-time data availability to track how new protocols and policies are performing in the real world, patients are prevented from reaching their optimal outcomes. Our ability to understand, act on, and predict based on healthcare data needs to significantly evolve.
Healthcare Data Today: Pointing Us in the Wrong Direction

For the healthcare industry to best serve its population, patient-level data sets must represent the actual patients. They must include all of a patient’s interactions with various care providers, facilities, and payers. Geographic, ethnic, gender, and age distributions must be accurately recorded. The data set population must also be stable enough that patient and provider behavior can be tracked over time.

Existing data providers suffer from significant selection and sampling biases, driven by:

**Missing visits**
Occurs when only the data from certain providers are captured for a given patient (e.g., an outpatient visit is captured but a hospital visit is not). Most electronic medical record-based sources and aggregated-claims data sets face this challenge.

**Missing patients**
Occurs when only the data from certain payers are captured for a population (e.g., commercial plans that underrepresent disease burden in the elderly). Most single-payer data sets present this issue.

**Missing longitudinality**
Occurs when a patient drops out of a data set without explanation (e.g., a patient moves to a different physician or payer that is not captured in the data set). Most aggregated-claims data sets have significant gaps in longitudinality.

**Missing linkages**
Occurs when only some patients are attributed to particular providers or facilities (e.g., the specific provider-patient connections are missing or not fully captured). This leads to the underrepresentation of disease burden in areas where connections are unclear. All data sets in the market have missing linkages.
These gaps in the data compound bias, leading to conclusions that may be ineffectual at best and outright dangerous to patients at worst. And because this bias is caused by an inexistence of critical data, it is effectively invisible. This is a “false north” bias. A broken compass makes you believe you are walking north when, in fact, you are being pointed in the wrong direction; worse yet, you are walking with a false sense of security. Similarly, when you look for answers in an incomplete data set, you are unable to ascertain the underlying bias and cannot, therefore, account for it. This invisible bias can lead to a series of missed opportunities for supporting patients and improving clinical outcomes.

False north bias is driven by poor data transparency and poor representation in healthcare data that result from:

**Complex data-licensing arrangements**
The complex arrangements between providers, payers, intermediaries, and data aggregators govern data access to end users and create fragmentation and bias in data sets.

**A lack of advanced technology**
Currently, no systems are in place to provide a real-time snapshot of healthcare, creating systematic lag between event and observation.

**Unavailability of important data**
Though information from national payers (including Medicare) and state-specific Medicaid programs is valuable in understanding the landscape of healthcare, it is excluded or poorly sampled in aggregated data sets.

Unless we address these drivers, false north bias will continue to prevail in healthcare.

### A Modern Prescription for Healthcare Data: Finding True North

As healthcare evolves, the need for deeper insights into new pools of data grows. Trillions of new explicit and implicit data points that paint the picture of patient health are being generated by daily care decisions, patient registries, precision medicine, the Internet of Things (IoT), and the overall consumerization of healthcare. But what is the value offered by these continuous streams of data? To determine that, we must contextualize the information being provided. What patients are represented? What therapies are in use? What provider actions, payer plans, patient behaviors, and therapies contribute to better outcomes?
To leverage these new data streams to power real-world evidence at scale, we must create the backbone that unifies healthcare data and delivers a source of truth: healthcare’s “true north.” This shift to true north requires resolving the aforementioned sources of bias in aggregated data sets, then building a system that maintains robust and complete patient-journey data sets. We prescribe the following steps:

1. **Address data fragmentation with deeper access to payer-complete sources.**

Payer-complete sources include the Centers for Medicare & Medicaid Services (CMS) and the largest commercial payers. Without complete data from the largest risk-bearers in the US, sampling biases from the data derived from intermediaries will continue to cloud the assessment of true cost and outcomes. A de-identified patient master based on these payer-complete sources can serve as the backbone for all additional patient-related data points.

2. **Build an interoperable data platform that preserves a patient’s context over time while protecting their confidentiality.**

This platform needs to ensure that each source, be it traditional (e.g., patient chart) or emerging (e.g., mobile application data), can connect to the patient master while preserving confidentiality. It must also provide true entity resolution, using AI to correctly link a fragment of data to the correct de-identified patient record. Today’s de-identification engines fail to provide entity resolution, exacerbating data biases.

3. **Apply the best AI on top of the best data.**

The highest-value AI applications require context-rich data sets featuring complete patient journeys. The cost and time required to run advanced machine learning models on trillions of healthcare transactions have dropped dramatically over the last few years, fueling the possibilities for new AI applications that can model and predict optimal interventions in real time.

With this prescription, true north in healthcare data sets will usher in a new era of trust—one that is not currently available with the incomplete and hand-curated data dominating patient journey assessments today. True north will create new ways of charting patient journeys that are grounded in reality, predicting the interplay of emerging trends on current behaviors and opening up new possibilities to shape standards of care. It will serve as the backbone for AI applications that, at present, are considered to be impractical or economically impossible to scale. Using this prescription, we can fix the current state of healthcare data and usher in a new era in which healthcare costs are massively reduced while patient outcomes are improved.