VIEW POINT

LEVERAGING BIG DATA ANALYTICS IN HEALTHCARE TO DRIVE VALUE-BASED CARE

Abstract

The value-based care model in healthcare systems ties provider payment directly to patient outcomes and cost efficiency, unlike the fee-for-service reimbursement model. Big data analytics makes that transition operationally viable. Without it, the promise of value-based care stays largely theoretical. Under Value-Based Care (VBC) contracts, providers are rewarded for outcomes, which implies predicting what a patient will need before something goes wrong. That demands a fundamentally different relationship with data. This analysis examines how big data analytics in healthcare enables that shift, from population-level predictive risk stratification to AI-driven HCC risk adjustment coding, and what tends to get in the way.

Infosys®
Navigate your next

The North American market for healthcare big data analytics is projected to grow at a 19.3% CAGR between 2024 and 2030, which reflects the scale of investment going into solving this problem. But investment alone does not close the gap.

Healthcare generates an enormous volume of data from EHRs, claims files, lab results, imaging reports, pharmacy records, and wearable device outputs. The volume is not the problem. The problem is that most of this data sits in inconsistently coded siloed systems that do not communicate and are structured for billing rather than clinical insight.

This matters because value-based care is, at its core, an outcomes-and-accountability model. When a payer or employer contracts with a health system under a capitated arrangement, they are asking that system to manage a defined population's health for a fixed payment. The system performs well financially when it keeps people healthy. It loses money when high-cost events, like hospitalisations, emergency admissions, or preventable readmissions, occur at higher rates than projected

What value-based care actually requires from data

To function effectively under VBC, a health system needs three things from its data infrastructure: comprehensive longitudinal patient records that consolidate clinical, claims, and social determinants of health (SDoH) data; the analytical capacity to identify high-risk individuals before a crisis; and performance metrics that enable ongoing measurement against contract benchmarks. Most organisations have elements of all three.

Note that 80% of healthcare organisations cite provider resistance to workflow changes as a barrier to VBC growth. Resistance is often a symptom of tools that add administrative burden rather than reduce it.

From volume to value: the role of big data

Big data analytics in healthcare refers to the collection, processing, and interpretation of datasets that are too large or complex for conventional tools. In VBC, what distinguishes it from standard reporting is the ability to work across data types, such as structured EHR fields, unstructured clinical notes, sensor data, social care records, and derive predictive, rather than purely descriptive, insight.

A 2022 study published in the Journal of Big Data found that medical facilities are increasingly drawing on both structured and unstructured data across clinical, administrative, and business functions, with sources ranging from transaction records and device sensors to email communications and documents.

How risk stratification models work

Predictive risk stratification allows the healthcare framework to move from reactive to proactive care. These models use historical patient data, diagnostic codes, medication histories, prior utilisation, lab values, SDoH indicators, and run them through machine learning algorithms to generate a probability score. The output is a ranked list of patients by their likelihood of a high-cost health event within the next 30, 90, or 180 days.

A well-calibrated risk stratification model recognises a patient with underlying health conditions before the acute event. The care team can then intervene with a care management call, a scheduled follow-up, or a medication review.

This is the mechanism by which big data analytics in healthcare converts population health management from a theoretical aspiration into a clinical workflow.

What the evidence shows

The outcomes from health systems that have deployed these models are substantial. A 2025 study published in the American Journal of Managed Care found that predictive algorithms reduced hospital readmission rates from 27.9% to 23.9% in a safety-net hospital using targeted follow-up interventions. Embedding AI-driven predictive analytics into care pathways for patients with heart failure, COPD, and pneumonia also proved to reduce readmissions.

Risk stratification and the SDoH layer

One of the more significant recent developments in predictive risk stratification models is the integration of SDoH variables like housing instability, food insecurity, and transport access, alongside clinical data. A patient discharged following a hip replacement who lives alone with no reliable transport to a follow-up appointment is at materially higher readmission risk than one with immediate family. Traditional clinical models could not see that. Models that incorporate SDoH data can.

Researchers and clinicians actively disagree about which SDoH variables carry the most predictive weight, how frequently those data points need updating, and whether population-trained models adequately account for individual variation. These are not settled questions, and health systems should approach SDoH-enriched models with appropriate scepticism alongside appropriate ambition.

AI-driven HCC risk adjustment coding

Hierarchical Condition Category (HCC) coding is the mechanism by which patient risk is quantified for capitated reimbursement, primarily in Medicare Advantage and similar risk-based contracts. Each patient receives a Risk Adjustment Factor (RAF) score derived from their documented chronic conditions. That score determines how much a health plan or provider receives to manage that patient's care over the year.

The problem is that HCC coding depends entirely on clinical documentation. This documentation is frequently incomplete. A physician managing 15 to 25 patients daily, each carrying 10 to 15 chronic conditions, will not catch every nuance of every ICD-10 clinical documentation opportunity. Consequently, the RAF score is underestimated. The reimbursement falls short of what the true complexity of the patient population warrants.

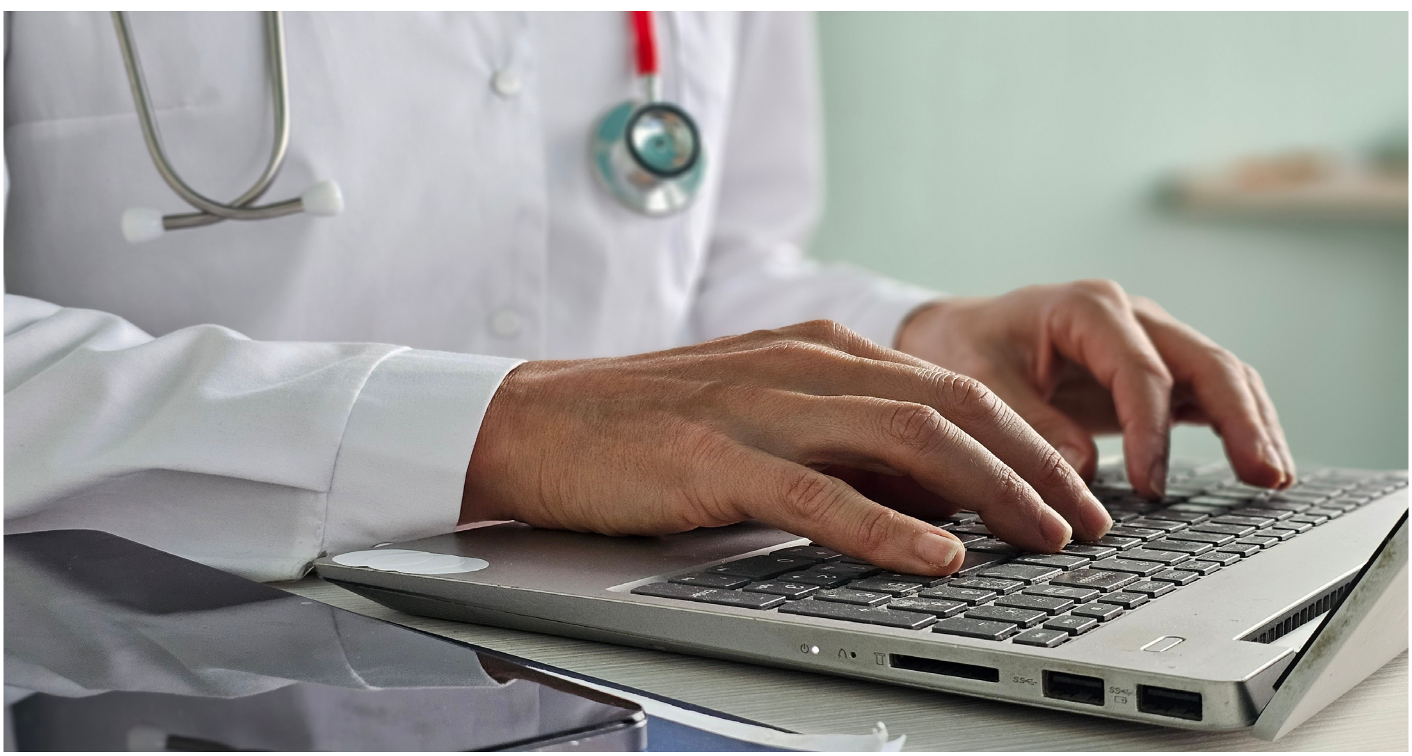
What AI-driven HCC risk adjustment coding actually does

AI-driven HCC risk adjustment coding uses Natural Language Processing (NLP) to scan clinical notes, medication lists, lab results, and prior claims data for conditions that are clinically present but not formally coded.

The commercial results are significant. Digital Scientists, whose AI-powered RAF and HCC platform has been deployed across more than 20,000 patients in seven US states, report improving the annual recapture rate of known chronic conditions from 60–65% to 85–90%. Closing a documentation gap of that scale, across a population of any meaningful size, translates directly into revenue integrity. But the less-discussed benefit is clinical: a more complete diagnosis record drives more accurate care planning.

The compliance dimension

AI-driven HCC risk adjustment coding is often framed as a revenue tool. The compliance angle is equally important. The Centers for Medicare and Medicaid Services (CMS) audits risk adjustment accuracy, and over-coded patients create legal exposure just as under-coded ones create financial loss. The value of AI in this context is not only that it identifies more conditions, but that every recommendation links to specific source documentation, creating an audit-defensible record rather than a black-box output. That distinction has grown considerably more important since CMS increased its scrutiny of Medicare Advantage risk adjustment practices from 2023 onwards.



The implementation barriers

A 2025 systematic review published in the Journal of Big Data draws on 35 peer-reviewed studies across ten major health databases, identifying three consistent barriers to big data analytics adoption in healthcare:

- Data privacy and governance challenges
- Technical complexity in integrating disparate systems
- A shortage of qualified staff

The interoperability problem

Data interoperability is the precondition for nearly everything described so far. Predictive risk stratification models need comprehensive longitudinal records. AI-driven HCC coding tools need simultaneous access to clinical notes, claims histories, and lab data. Neither works well when data is scattered across incompatible EHR systems, legacy claims platforms, and siloed social care databases.

The governance layer

Data quality is downstream of governance. If the rules for how data is entered, validated, and maintained are not defined and enforced at the point of capture, no amount of downstream analytics compensates. Inconsistent documentation, a condition coded as "diabetes" in one encounter and "type 2 diabetes with peripheral neuropathy" in another, produces noise that degrades model performance over time. The organisations that get the most from big data analytics in healthcare tend to have invested in data governance before the analytics tooling.

End note

The shift to value-based care is critical. CMS continues to expand Medicare Advantage, commercial payers are extending capitated models further down the market, and providers are increasingly held to outcomes they cannot manage without better data. [Big data analytics in healthcare](https://) is the infrastructure that makes those outcomes manageable. It does not replace clinical judgement, but makes risk more visible.

Predictive risk stratification models and AI-driven HCC risk adjustment coding are two of the most mature and commercially consequential applications of that infrastructure. Both have documented outcomes. Both improve as data quality improves.

The organisations that will perform best under VBC contracts over the next five years are probably already doing this work.