The Road to Meaningful AI-Driven Data Curation

Introduction

Artificial intelligence (AI) is fueling innovation throughout the healthcare industry. Whether healthcare facilities are using AI to detect diseases earlier, assist in clinical decision-making, or identify patients suited for clinical trials, AI's impact is becoming increasingly evident.[1,2,3]

However, some often overlooked limitations highlight the importance of proceeding thoughtfully when exploring the use of AI in healthcare. Models that are improperly trained, fed low-quality data, or given tasks that exceed AI's capabilities can lead to incorrect diagnoses and potentially harmful clinical decisions. During the COVID-19 pandemic, for example, errors in training or testing AI tools resulted in models that functioned improperly.[4] In one instance, AI picked up on the fonts hospitals used to label scans, causing the model to falsely correlate fonts with predictors of COVID-19 risk. Diagnostic AI has also revealed biases, such as less accurate skin cancer diagnoses for patients with darker skin due to a lack of diversity in training datasets.[5] Further, an evaluation of a sepsis prediction model found that the model failed to identify two-thirds of patients with sepsis—underscoring the dangers of relying on AI alone.[6]

While AI holds massive potential to transform patient care, the industry is still in the early stages of exploring this technology. AI models cannot work effectively without high-quality clinical data. To develop accurate AI models, we must get the data right first.

This article discusses the value AI-driven data curation offers in hospital settings and clinical research. It explores common pitfalls associated with developing AI models for data curation, along with the factors necessary for successful AI implementation. Lastly, it shares why Q-Centrix is uniquely positioned to explore the use of AI to curate clinical data.

> To develop accurate AI models, we must get the data right first.

The Potential of AI in Healthcare

AI can support the processing of massive amounts of valuable yet unstructured information for a range of different purposes in healthcare, research, and beyond.

Unlock the Value of Clinical Data

AI is poised to help hospitals derive greater insights from the vast amounts of data they have. Patients generate an average of 50 million gigabytes of data every year—and 97 percent of the clinical data hospitals possess go unused.[7] When the vast majority of these data are unstructured, often taking the form of doctors' notes, image scans, and other formats that require interpretation, making sense of this staggering amount of data is an impossible undertaking for any person or team—but an algorithm can be trained to quickly go through massive datasets and extract valuable insights.[8]

Improve Clinical Research

AI also has the potential to make a significant impact in addressing clinical research challenges. Currently, nine in 10 drugs that reach the clinical trial stage fail to receive FDA approval due to challenges in the clinical trial process.[9] Insufficient patient enrollment remains one of the biggest hurdles in clinical trials—and it's the reason why 20 percent of cancer clinical trials fail.[10,11] Finding eligible patients for a study often involves combing through electronic medical records (EMRs) and other information systems not built for clinical research purposes, which is very time-consuming.

9 in 10

drugs that reach the clinical trial stage fail to receive FDA approval.

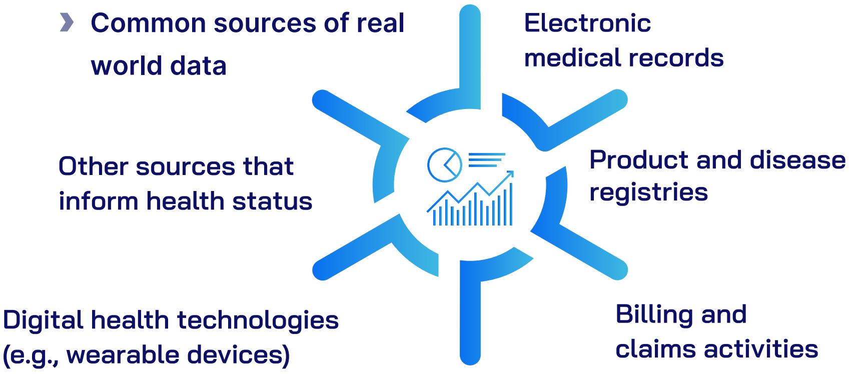
Even research teams that manage to identify and enroll enough patients for a clinical trial may find that their sample is not representative of the general population. Many racial and ethnic groups are underrepresented in clinical research, emphasizing a need to improve diversity among clinical trial patients.[12] With the aid of AI-powered tools to sift through data dispersed across various information systems and find patients that meet trial criteria, research teams may be able to conduct clinical research more efficiently. This can both greatly reduce the time and costs associated with patient recruitment and aid in increasing diversity in clinical trials.[13,14]

Support Observational Studies

In addition to improving clinical research processes, AI can support research that relies on existing patient data, such as observational studies. These studies can be conducted using the real world data hospitals and health systems already have (such as data from electronic medical records, billing and claims data, and other sources of information).

Although unstructured data and inconsistent data preparation practices are common challenges associated with observational studies, using AI-enabled techniques to curate these data allows facilities to overcome these barriers and produce custom, high-quality, research-ready datasets.[15] These datasets can be used for facilities' internal research purposes or for funded opportunities in which healthcare facilities contribute data to retrospective studies for sponsors in the pharmaceutical and life sciences industries. As observational studies are less expensive to conduct than clinical trials—and can be completed much more quickly—AI-driven data curation offers researchers a valuable, cost-effective, and efficient way to gather findings and advance medical research.

Observational studies can be conducted using the real world data hospitals and health systems already have.



❯ **Common sources of real world data**

Electronic medical records

Product and disease registries

Billing and claims activities

Digital health technologies (e.g., wearable devices)

Other sources that inform health status

Considerations in Using AI for Data Curation

While AI excels in repetitive, straightforward tasks, such as scheduling appointments, using AI for data curation is a much more complex undertaking. Clinical terminology is highly nuanced and requires specialized expertise to be interpreted properly, and it can be challenging to provide an AI model with the scale and scope of diverse, accurate data it needs in order to be effective.

Common Pitfalls in AI Model Development

Developing and training an AI model is not an easy feat. An improperly trained model is likely to have deep flaws due to biases or shortcomings in the data used to train it. This, in turn, can have untold effects on the model's performance, accuracy, and utility in a clinical setting.

Some factors that can hinder an AI model's ability to learn and function properly include:

- **Poor data quality.** The standards for clinical data quality are very high, and high-quality data are essential for training and refining AI models to ensure their accuracy and reliability. Many failures in AI tools have been linked to the poor quality of data researchers have used to develop these tools.[16]
- **Outdated data.** New drugs and treatment pathways are developed every year, changing how medicine is practiced. Models trained on data from even a year ago would miss crucial insights from the rapid pace of innovation.
- **Documentation practices.** Documentation practices vary from physician to physician and, on a larger scale, from facility to facility. A model trained on one physician's or one facility's data may not be applicable elsewhere.
- **Distinct EMR setups.** Drastically different data capture practices occur not just across hospitals that use different EMRs, but even among hospitals that use the same EMR. Customized configurations, differing data entry protocols, or unique workflow integrations greatly alter how patient data are entered and stored.

- **Differences across care settings.** The setting of care, whether inpatient, outpatient, academic, or community-based, introduces additional layers of variability. Each care setting may have specific requirements, workflows, and priorities that dictate data capture practices.

Organizations must proceed cautiously when developing AI models for data curation. When AI models lack the rigorous training needed for safe and effective data curation, any number of repercussions can result. For example, training AI models on low-quality data can create data cascades—defined as compounding events causing negative, downstream effects—that can lead to unreliable results or harmful effects.[17]

Elements for Successful AI Implementation

Automation alone is not enough to ensure accurate, efficient data curation. Automation is one component on a continuum of actions that drive efficiency incrementally. To that end, AI-powered clinical data curation hinges on aligning automation with the right combination of actions, processes, and expertise.

Key elements necessary for successfully developing and training an AI model for data curation include:

Access to large volumes of data. Many AI models are trained on 10,000 or fewer examples—but significantly more examples are necessary for more impactful training.[18]

Ability to normalize large amounts of disparate data. As clinical data are largely unstructured and are stored across multiple information systems, these data must first be transformed into a consistent and understandable format before AI systems can process, analyze, and extract insights from them.

Consistently training and refining the AI model for accuracy. With clinical data standards requiring a very high level of accuracy, AI models must be refined enough to meet these requirements—and no technology can yet solve the challenge of accurately processing unstructured data without human assistance. Software and data engineers must work in tandem with clinical experts to ensure the accuracy of the model's output.

Thoughtfully integrating AI technology into the workflow and user dynamics. This requires a substantial investment in software development, particularly in enhancing user interfaces and user experience capabilities to enable healthcare professionals to easily draw insights from their data through dashboards and user-friendly tools.

Consistent quality validation of the model to ensure continued accuracy. Regular testing and quality checks ensure the AI model continues to meet its intended purpose while complying with evolving healthcare standards. Without a robust system in place for consistent quality validation, organizations may struggle to detect and rectify inaccuracies.

Cost-effectiveness. Currently, using an AI model such as Chat GPT4 to abstract clinical data at an acceptably high level of accuracy for a registry such as CathPCI would cost approximately three times more than a traditional data abstraction model that involves a combination of AI and human expertise. Automation technology has not yet advanced enough to replicate the combined capabilities of clinical data experts and technology in a cost-effective way.

Achieving success on all of these fronts is not easy to do—but Q-Centrix is well-positioned to take on this challenge. With the data access, resources, experience, and scale necessary to succeed, Q-Centrix will continue to advance its data automation efforts to further optimize AI-enabled data curation.

Q-Centrix's Unique Role in Leading AI-Driven Data Curation

While the barriers to entry for building a data model may seem low, the barriers to achieving success in this endeavor are substantially high—and the risks of failure can have harmful impacts on patient care.

To pursue AI-powered data curation safely, healthcare facilities should partner with an established organization with a proven track record, longevity, and a trusted reputation.

Q-Centrix is uniquely positioned to explore the use of AI for data curation, for several reasons.

Combining technology with clinical experts. Experts in the technology sector don't often have the medical knowledge necessary to interpret complex clinical datasets—a disconnect that ultimately led to the failure of many AI tools created during the pandemic.[19] Because Q-Centrix relies on a combination of proprietary AI-powered software and clinical experts to curate data and perform quality checks, its approach bridges the gap between medical knowledge and technological expertise. Q-Centrix's more than 1,300 clinical data experts have strong backgrounds in healthcare and abstract millions of cases each year.

Deep understanding of nuanced clinical terms. Clinical concepts can have different definitions depending on the context. For example, some registries define a family history of heart disease as having an immediate family member dying of a heart attack or stroke before age 60 in women and before age 55 in men. Off-the-shelf AI models may react to any mention of a family heart condition—regardless of its severity, the direct relationship, or age. This highlights the need for nuanced clinical context in data curation, which only clinical experts can provide.

Streamlined processes that prioritize data integrity. Q-Centrix is committed to maintaining the highest data integrity standards in the industry. Q-Centrix implements a series of quality checks throughout the data lifecycle, spending over 11,000 hours per month conducting quality-related checks on data.

User-friendly software. Q-Centrix's offerings go beyond data curation to ensure that healthcare facilities have the tools they need to engage meaningfully with their data. Q-Centrix's market-leading clinical data management software provides a comprehensive suite of analytics and reporting tools, empowering clinical and quality leaders to uncover valuable insights that drive clinical decision-making and quality improvements.

> Q-Centrix is uniquely positioned to explore the use of AI for data curation.

Experience. Q-Centrix has over a decade of experience managing clinical data for more than 1,200 hospital partners, making its AI-driven technology extremely well-trained in reviewing data to ensure data integrity.

Conclusion

Hospitals need to trust their data. When clinical data are the cornerstone of patient care, groundbreaking research, quality improvement, and so much more, ensuring the integrity of these data is paramount.

For AI-driven data curation to be meaningful and effective, it must be capable of maintaining the highest data quality standards—which today's technologies can't yet do alone. Due to the complexities of clinical data curation—and the risks inherent in low-quality data—a combination of clinical data experts, software, and optimized processes must be used alongside AI technologies to curate high-quality data effectively.

Q-Centrix is well positioned to lead in AI-driven data curation given our strong commitment to data quality, our experience curating clinical data for more than 1,200 hospital partners, our 1,300+ clinical data experts, and our investments in technology. Through our efforts, hospitals and health systems can improve data integrity and derive deeper meaning from their data. Moreover, life sciences organizations and research institutions can gain valuable assistance in identifying study patients and overcoming common research roadblocks.

As we move forward, we are excited to advance our use of AI technology while recognizing that, like all new technologies, it will require time, investment, and deliberate effort to ensure high data standards and continued progress in the healthcare industry.

References

[1] Ashley Welch. “Artificial intelligence is helping revolutionize healthcare as we know it.” Johnson & Johnson (September 13, 2023). https://www.jnj.com/innovation/artificial-intelligence-in-healthcare.

[2] American Hospital Association. “How AI is improving diagnostics, decision-making and care” (May 9, 2023). https://www.aha.org/aha-center-health-innovation-market-scan/2023-05-09-how-ai-improving-diagnostics-decision-making-and-care.

[3] Brian T. Horowitz. “The current state of AI in healthcare and where it’s going in 2023.” HealthTech (December 16, 2022). https://healthtechmagazine.net/article/2022/12/ai-healthcare-2023-ml-nlp-more-perfcon.

[4] Will Douglas Heaven. “Hundreds of AI tools have been built to catch COVID. None of them helped.” MIT Technology Review (July 30, 2021). https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

[5] J. Bob Alotta. “OPINION: The AI in our healthcare needs a reckoning.” Thomson Reuters Foundation (July 21, 2022). https://news.trust.org/item/20220721103518-9edv5/.

[6] Andrew Wong et al. “External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients.” JAMA Internal Medicine 8, no. 181 (August 1, 2021): 1065-1070. https://pubmed.ncbi.nlm.nih.gov/34152373/.

[7] Q-Centrix. “Achieving reliable, actionable insights through automated clinical data abstraction” (June 16, 2023). https://www.q-centrix.com/achieving-reliable-actionable-insights-through-automated-clinical-data-abstraction/.

[8] Q-Centrix. “Healthcare IT Today: Health care quality, registries, and artificial intelligence at Q-Centrix” (January 14, 2022). https://www.q-centrix.com/news/healthcare-it-today-health-care-quality-registries-and-artificial-intelligence-at-q-centrix/.

[9] Brian Mai, Andrea Roman, and Alondra Suarez. “Forward thinking for the integration of AI into clinical trials.” The Association of Clinical Research Professionals (June 20, 2023). https://www.acrpnet.org/2023/06/forward-thinking-for-the-integration-of-ai-into-clinical-trials/.

[10] Rachana Pradhan. “The business of clinical trials is booming. Private equity has taken notice.” KFF Health News (December 2, 2022). https://kffhealthnews.org/news/article/business-clinical-trials-private-equity/.

[11] Maria Sae-Hau et al. “Overcoming barriers to clinical trial participation: Outcomes of a national clinical trial matching and navigation service for patients with a blood cancer.” JCO Oncology Practice 17, no. 12 (December 1, 2021): e1866-e1878. https://ascopubs.org/doi/full/10.1200/OP.20.01068.

[12] U.S. Food and Drug Administration. “Clinical trial diversity” (November 4, 2022). https://www.fda.gov/consumers/minority-health-and-health-equity-resources/clinical-trial-diversity.

[13] Brian Mai, Andrea Roman, and Alondra Suarez. “Forward thinking for the integration of AI into clinical trials.” The Association of Clinical Research Professionals (June 20, 2023). https://www.acrpnet.org/2023/06/forward-thinking-for-the-integration-of-ai-into-clinical-trials/.

[14] Fanta Cherif. “Harnessing the power of AI in drug development and testing.” Advisory Board (August 9, 2023). https://www.advisory.com/topics/artificial-intelligence/2023/07/harnessing-the-power-of-ai-in-drug-development-and-testing.

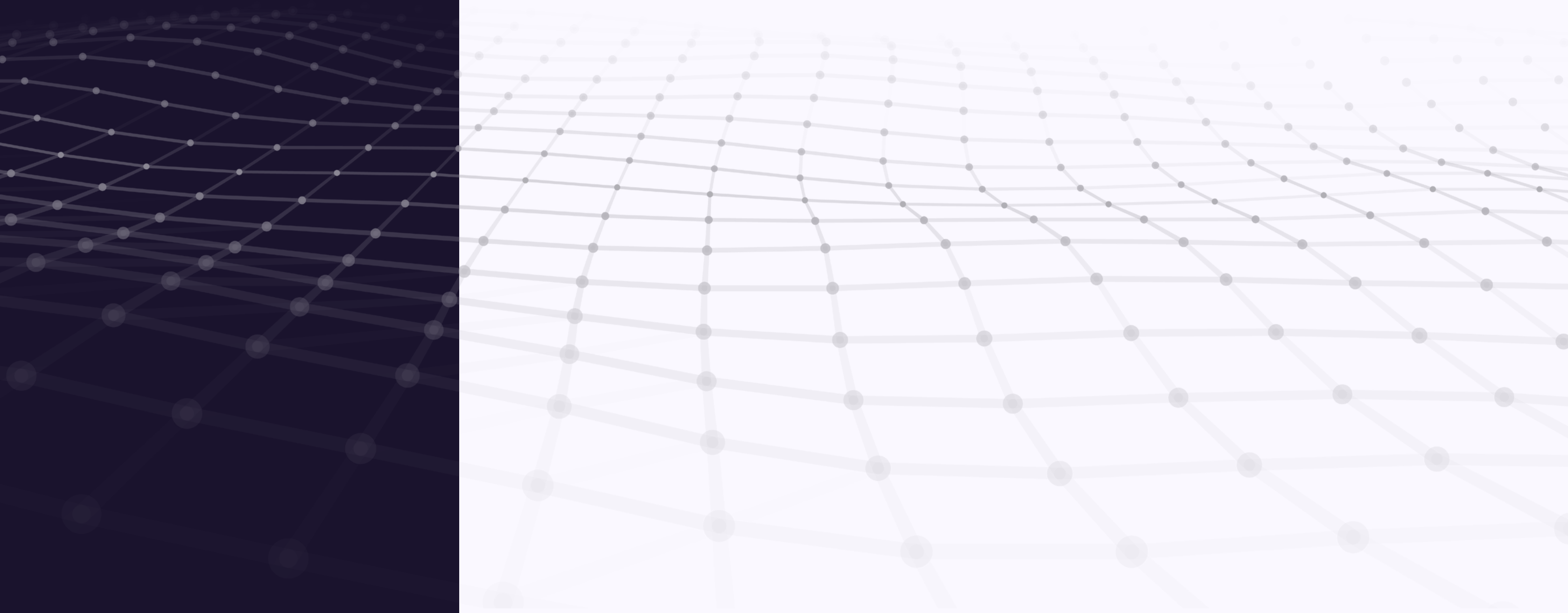
[15] Q-Centrix. “Observational studies: Benefits and challenges in using real world data to advance research.” https://www.q-centrix.com/ insights/detail/observational-studies-benefits-and-challenges-in-using-real-world-data-to-advance-research/.

[16] Will Douglas Heaven. “Hundreds of AI tools have been built to catch COVID. None of them helped.” MIT Technology Review (July 30, 2021). https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

[17] Nithya Sambasivan et al. “‘Everyone wants to do the model work, not the data work’: Data cascades in high-stakes AI.” Google Research (May 8-13, 2021). https://storage.googleapis.com/pub-tools-public-publication-data/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf.

[18] Gil Press. “Andrew Ng launches a campaign for data-centric AI.” Forbes (June 16, 2021) https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/.

[19] Will Douglas Heaven. “Hundreds of AI tools have been built to catch COVID. None of them helped.” MIT Technology Review (July 30, 2021). https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.



About Q-Centrix

Q-Centrix sees clinical data differently—as custom data sets with infinite possibilities.

Providing the industry’s first Enterprise Clinical Data Management (eCDM™) approach, Q-Centrix combines AI-enabled technology, the largest and broadest team of clinical data experts, and insights from its more than 1,200 partners to help improve patient outcomes and drive process and performance improvement, strategic growth, and operational efficiency.

Its solutions address a variety of clinical data needs, including quality measurement and improvement, cardiovascular, oncology, trauma, research, and more.

One North Franklin
Suite 1800
Chicago, IL 60606
q-centrix.com

QCentrix