De-Identified Medical Datasets and the 2025 Readiness Gap:

Toward Equity, Scale, and Trust in Foundation Model Training

Britney Bennett

Stanford University

Abstract

Foundation models (FMs)—large-scale machine learning models trained on vast, diverse datasets—are reshaping the future of medical AI by powering diagnostic tools, clinical decision systems, and health information summarization. Sometimes referred to as large language models (LLMs) when applied to text, these models are increasingly deployed across clinical contexts. However, the de-identified datasets that form the backbone of FM training are often outdated, demographically limited, and difficult to access. These limitations raise profound concerns about fairness, scientific validity, and the potential for harm—particularly for marginalized populations underrepresented in training data. This paper argues that current de-identified datasets are not adequately representative or accessible for building trustworthy AI in healthcare. It critiques the prevailing assumption that de-identification alone ensures ethical readiness, showing instead how it can obscure structural biases and entrench inequality. Drawing on recent research and emerging technical and policy solutions—including synthetic data generation, automated de-identification, and global benchmarking—this paper explores what it means for datasets to be "2025-ready." It proposes a new standard for responsible dataset design, grounded in demographic transparency, equity-centered governance, and inclusive participation in medical AI development.

Introduction

Artificial intelligence (AI) is increasingly integrated into healthcare through foundation models (FMs)—large-scale neural networks capable of analyzing clinical texts, generating diagnostic summaries, and supporting real-time decision-making across specialties. As these tools begin to inform critical aspects of medical care, from radiology to primary care triage, the data used to train

them becomes a central concern. Most existing FMs in healthcare are built on de-identified medical datasets—patient records stripped of personal identifiers under privacy regulations such as HIPAA's Safe Harbor or Expert Determination frameworks. These datasets are designed to balance patient privacy with the imperative to enable research and innovation. However, de-identification, while necessary, is far from sufficient. Too often, it gives a false sense of neutrality or universality to datasets that are structurally narrow, geographically concentrated, and demographically skewed. Many de-identified datasets were created before the COVID-19 pandemic, rely heavily on records from white, urban patients in high-income countries, and exclude key health domains like obstetrics, psychiatry, or chronic illness management. Their restricted scope undermines the fairness and generalizability of FMs, especially when applied to underrepresented populations such as rural patients, Indigenous communities, or residents of the Global South.

Moreover, de-identification processes are expensive, technically demanding, and governed by uneven standards, limiting who can contribute to or benefit from medical AI development. Institutions in low- and middle-income countries often face steep barriers to accessing training data or developing their own FMs, exacerbating existing disparities in global health innovation. This paper investigates whether current de-identified datasets are "2025-ready"—a term that signals not just technical adequacy but ethical preparedness, global inclusiveness, and demographic transparency. It draws on recent literature and proposed innovations to argue that foundational models trained on today's datasets are likely to amplify structural inequalities unless urgent reforms are enacted. In particular, the paper explores emerging solutions such as automated de-identification, synthetic data generation, demographic auditing, and global governance frameworks as tools for transforming dataset practices from compliance-oriented to equity-centered. By redefining what constitutes responsible data in the age of medical AI, this paper aims to contribute to a future where all patients—not just those in the training data—can benefit from AI-driven care.

Literature Review

The Role of De-Identification in Healthcare AI

De-identification is a foundational practice in healthcare data management, intended to enable the ethical use of patient information while protecting individual privacy. Under the Health Insurance Portability and Accountability Act (HIPAA), two primary methods exist: the Safe Harbor approach, which removes 18 specific identifiers, and the Expert Determination method, which uses statistical techniques to assess the likelihood of re-identification (Rothstein, 2010). These processes are vital for enabling research without requiring patient consent or Institutional Review Board (IRB) approval. However, Sweeney et al. (2017) demonstrated that even datasets

de-identified under Safe Harbor can be vulnerable when cross-referenced with publicly available data. Moreover, de-identification is often resource-intensive, requiring sophisticated tools and substantial financial investment, which can limit smaller institutions' ability to contribute to or benefit from medical AI development (Stubbs et al., 2015).

Despite these limitations, de-identified datasets remain critical for training clinical AI and foundational models (FMs). These models, defined by Bommasani et al. (2021) and further elaborated by Moor et al. (2023), are large-scale neural networks that can be fine-tuned for specific medical tasks such as diagnosis, triage, and treatment recommendation. Their effectiveness, however, hinges on the diversity and quality of training data. AI models built on homogenous or incomplete datasets risk replicating systemic biases, undermining both scientific validity and ethical acceptability (Chen et al., 2021; Mpanya et al., 2021).

Representation and Bias in De-Identified Datasets

Numerous studies have raised alarms about the demographic limitations of commonly used de-identified datasets. Kleinberg et al. (2022) found that dermatological datasets disproportionately include lighter-skinned patients, which leads to reduced predictive accuracy for people with darker skin. Guo et al. (2022) reinforce this point, reporting that fewer than 9% of dermatology datasets disclosed race or ethnicity, and only 4% reported Fitzpatrick skin type, hindering efforts to audit algorithmic fairness. These issues extend well beyond dermatology. Ford et al. (2025) show that psychiatric service records in the UK, especially in free-text fields, are less complete for minoritized groups. Similarly, Marko et al. (2025) conducted a meta-review of 129 studies and concluded that underrepresentation in training datasets consistently produces disparities in AI outcomes and accessibility. They emphasize that the issue is not merely technical but deeply ethical, rooted in the absence of upstream data governance standards that prioritize inclusivity.

Geographic bias compounds these problems. Widely used datasets like MIMIC-III, MIMIC-IV, and eICU originate largely from northeastern U.S. hospitals and include mostly white patients (Johnson et al., 2016). International datasets such as HiRID and the UK Biobank, while expanding geographic scope, still reflect racially homogenous populations and are often drawn from high-income nations. This lack of representation limits model generalizability, particularly for rural communities, Indigenous populations, and patients in the Global South (Desouza, 2023). Without deliberate efforts to ensure inclusion, medical AI risks reinforcing structural inequities under the guise of technological progress.

Technical and Structural Limitations in Data Access and Distribution

Beyond demographic skew, the structure and distribution of data introduce

further limitations. MIMIC-IV, for example, includes a high number of patients across intensive and emergency care, but EHRSHOT—though smaller in patient count—offers more granular longitudinal clinical events (Wornow et al., 2023). This highlights an important asymmetry: no single dataset currently provides the comprehensive coverage needed for robust FM training. Olaker et al. (2025) caution against assuming that large datasets are inherently better; their review of major platforms like TriNetX and Cosmos emphasizes the "big data paradox," where large sample sizes can obscure critical confounders, introduce information bias, and produce misleading results if not paired with rigorous methodology.

Data accessibility also presents a major obstacle. While datasets like MIMIC and eICU are publicly available, many others are restricted by paywalls, institutional firewalls, or complex IRB requirements. De-identification itself can cost hundreds of thousands of dollars, especially when processing large repositories (Stubbs et al., 2015). This high cost restricts the development of equitable AI by concentrating power in a handful of well-funded institutions. Seastedt et al. (2022) argue that restrictive privacy protocols—intended to protect patients—can inadvertently exacerbate global health disparities by limiting low- and middle-income countries' access to training data and AI development tools.

Table 1. Timeline and Readiness of Commonly Used De-Identified Datasets

Dataset Name	Year Published	Data Timeline	2025 Readiness
MIMIC-III	2016	2001–2012	❌ Outdated for post-COVID modeling
MIMIC-IV	2020	2008–2019	⚠ Still widely used, but aging
eICU	2018	2014–2015	⚠ Lacks modern care trends
i2b2	2014	pre-2014	❌ Small, older dataset
HiRID	2020	2008–2016	⚠ High quality, but pre-pandemic
UK Biobank	2012	2006–2010	❌ Demographically skewed & outdated
EHRSHOT	2023	Longitudinal	✅ Current and benchmark-ready
MC-BEC	2023	2020–2022	✅ Timely, COVID-era data
WAVES	2023	2008–2018	✅ Pediatric-focused, newer format

🧪 **Note:** While synthetic datasets like **EchoNet-Synthetic** and **MeMA** offer promising innovations, they are based on limited demographic baselines and should be used with caution to avoid reinforcing existing biases.

Legend:

- ✅ Well-suited to 2025 research
- ⚠️ Still usable, but aging or narrow
- ❌ Largely outdated for today's needs

Emerging Solutions for Ethical and Inclusive AI Training

In response to mounting challenges around bias, privacy, and accessibility in AI training data, researchers have proposed several innovations aimed at improving both the inclusivity and ethical integrity of these datasets. One promising area is automated de-identification using AI models. In recent years, Natural Language Processing (NLP) models—particularly Large Language

Models (LLMs)—have surged in popularity. This rise was marked by the introduction of Generative Pretrained Transformers (GPTs), first released by OpenAI in 2018 (Radford et al., 2018). By leveraging transformer architecture, which enables models to understand the context of language inputs (Merritt, 2022), GPTs expanded the capabilities of earlier NLP tools. Unlike previous rule-based systems, GPTs can perform a wider range of tasks without requiring predefined rules (How BERT and GPT Models Change the Game for NLP, 2020). Their applications span various fields, including medicine, where early results have been promising (Altalla et al., 2025). For example, DeID-GPT, a neural network-based model developed for zero-shot medical text de-identification, has outperformed traditional rule-based systems in removing sensitive information from clinical documents (Liu et al., 2023). This success is attributed to in-context learning, a technique where LLMs learn new tasks from a few examples (Dong et al., 2022; Zhou et al., 2023). In practice, this could involve showing a model examples of de-identified Protected Health Information (PHI) in an Electronic Health Record (EHR). Zero-shot learning, a derivative of in-context learning, allows LLMs to generalize to tasks they haven't been explicitly trained on (In-Context Learning, n.d.). Once the task is defined, prompt engineering (Giray, 2023) is used to provide structured instructions for task completion.

Automated de-identification also reduces the labor burden and costs associated with manual anonymization, enabling more institutions—especially those with limited resources—to participate in data sharing efforts (Dernoncourt et al., 2017). At first glance, GPT-based de-identification may appear to be an ideal solution for generating large, de-identified datasets. However, important limitations remain—particularly regarding re-identification risks. Patsakis and Lykousas demonstrated that GPT-3 could deanonymize 72.6% of text descriptions of celebrities, compared to 26.39% by human participants, based on a prior study by Kleinberg et al. While this high success rate may be due to the extensive amount of publicly available data about celebrities, it underscores a critical risk: GPT models trained on massive, heterogeneous datasets may also contain identifiable information about non-celebrities. In sensitive domains such as healthcare, this could pose a serious threat to patient privacy. Furthermore, the alarming gap between GPT-3 and human re-identification rates challenges current standards. Under HIPAA's Safe Harbor method, de-identification is validated by human experts who deem the re-identification risk "very small" in a given context (U.S. Department of Health and Human Services, n.d.). Yet, if AI models are capable of far greater re-identification, relying solely on human assessment may no longer be sufficient. Ironically, making models better at de-identifying data can sometimes increase their ability to re-identify it—a troubling paradox highlighted by Valacchi (2024). Ford et al. (2025) argue that technical solutions must be coupled with robust governance frameworks to address residual risks. This suggests that current regulations like HIPAA may need

updating to incorporate hybrid approaches, where both human and AI assessments are used in a multi-step de-identification process.

Another emerging solution is synthetic data generation, which uses machine learning to produce datasets that mirror real-world data while aiming to preserve privacy. Techniques such as Generative Adversarial Networks (GANs) have been employed in this space (Altalla et al., 2025). GANs consist of two components: a generator that produces new data and a discriminator that distinguishes between real and synthetic samples. As the two networks compete, the generator improves, ideally producing increasingly realistic data (Goodfellow et al., 2020). However, the application of GANs to medical data presents challenges. While GANs excel at creating continuous data, medical datasets often contain a mix of continuous and discrete attributes (Torfi & Reddy, 2022). Moreover, GANs are vulnerable to bias and membership inference attacks, where adversaries can infer whether specific data points were used during training (Hayes et al., 2017). In one study involving 20,000 skin lesion samples, GANs produced distorted or inaccurate images based on artifacts such as hair in the original data—and in some cases, generated patterns not present in the real dataset at all (Jindal & Singh, 2024). Although researchers attributed this to limited training samples, rare diseases often lack large datasets, compounding the issue.

Projects such as EchoNet-Synthetic (Chen et al., 2021) and the MeMA dataset demonstrate the potential of synthetic data in modeling ECGs and ophthalmologic conditions. These projects use diffusion models instead of GANs. Diffusion models corrupt data with noise and then learn to reconstruct realistic outputs through a denoising process (Yang et al., 2023). While technically distinct from GANs, diffusion models share a common generative architecture. However, like GANs, they often replicate the biases of their training data—such as overrepresentation of Stanford or Shanghai patient populations—raising concerns about scalability and equity. A third pathway involves establishing global data-sharing frameworks and implementing standardized demographic audits. Many disparities in global health data stem from uneven implementation of Electronic Health Record systems, particularly in low- and middle-income countries (Woldermariam & Jimma, 2023). The COVID-19 pandemic exposed the consequences of inadequate local data collection, prompting increased efforts to improve these systems (Zhao et al., 2021; Karthikeyan et al., 2024). To ensure equity, Lahoti et al. (2023) advocate for the mandatory inclusion of demographic metadata—such as race, gender, age, and geography—in datasets. This would enable researchers to assess representational gaps and align foundation model training with ethical standards. In parallel, Desouza (2023) calls for international governance standards that uphold equitable AI development while respecting diverse national legal frameworks. Finally, Olaker et al. (2025) propose a practical 26-point checklist for responsibly using large-scale

de-identified datasets. Their recommendations include bias mitigation strategies, transparency practices, and compliance with reporting standards such as STROBE and FAIR. This framework helps balance the power of big data with the need for methodological rigor and ethical oversight. While the solutions discussed are still developing, they provide a roadmap for addressing persistent shortcomings in AI training datasets. De-identification alone does not guarantee fairness—especially when structural inequalities continue to shape who is represented and who is excluded. The path forward must involve technical innovation, regulatory reform, and a commitment to equity in both data collection and model design.

Discussion

The literature reveals that while de-identified datasets are central to training foundational models (FMs) in healthcare, they are far from comprehensive or equitable in their current form. Three major concerns emerge: the uneven demographic representation across datasets, structural limitations in access and governance, and an overreliance on technical fixes that do not fully address ethical implications. Together, these factors challenge the assumption that de-identified data is a neutral, universally beneficial resource. Instead, the evidence suggests that de-identified datasets risk reproducing and even amplifying health disparities if their design, composition, and usage are not critically examined and reformed.

Representation remains the most urgent concern. As shown in the literature, the majority of large-scale datasets originate from high-income countries and are overwhelmingly composed of white patients (Johnson et al., 2016; Guo et al., 2022). The limited geographic and demographic scope of datasets like MIMIC, eICU, and even UK Biobank renders them insufficient for training models intended for broad or global application. This lack of diversity compromises both the accuracy and fairness of FMs, particularly when applied to underrepresented populations. As Kleinberg et al. (2022) and Marko et al. (2025) argue, these omissions can result in biased predictions, misdiagnoses, or inadequate treatment recommendations for non-white, rural, and low-income patients. Structural barriers compound these limitations. The cost of de-identification—estimated at over $250,000 for large health systems (Stubbs et al., 2015)—as well as restrictive access protocols, mean that data-sharing is dominated by elite institutions. This centralization not only narrows the range of perspectives in FM development but also limits the ability of smaller hospitals, researchers in the Global South, and community health organizations to participate in shaping AI-driven healthcare. As Seastedt et al. (2022) warn, overly cautious privacy regimes may actually worsen inequality by shutting out those who lack resources to meet increasingly stringent compliance demands.

Proposed solutions such as automated de-identification, synthetic data generation, and demographic auditing offer promising avenues for addressing these

challenges—but each comes with tradeoffs. While DeID-GPT and similar tools (Liu et al., 2023) may reduce the costs and burdens of manual de-identification, they still require careful oversight to prevent re-identification and data leakage. Synthetic data, too, risks amplifying existing biases if based on homogenous training sets (Chen et al., 2021). These technical strategies must be embedded in broader governance frameworks that prioritize equity, transparency, and accountability. Ultimately, if healthcare AI is to serve all populations fairly, the foundational step must be rethinking what constitutes a "good" dataset. Size, accessibility, and de-identification status are no longer sufficient metrics. Instead, comprehensiveness must include representation across race, gender, geography, health condition, and social context. Researchers, institutions, and regulators must collaborate to establish shared benchmarks for dataset quality and create mechanisms for inclusive data stewardship. Without deliberate reform, these shortcomings do not remain abstract; they shape diagnostic algorithms, clinical workflows, and patient trust—deepening the very health inequities AI claims it can solve.

Conclusion

De-identification remains a critical enabler of AI development in healthcare, but the literature reviewed here reveals its current implementation is neither sufficient nor equitable. Rather than a neutral safeguard, de-identification often obscures structural inequalities embedded in the data itself. Foundational models trained on such datasets risk perpetuating—and even institutionalizing—biases that have long plagued healthcare systems. As AI becomes more deeply integrated into clinical decision-making, the stakes of relying on partial, demographically skewed, and difficult-to-access datasets grow increasingly high. This moment calls for a fundamental shift: from treating de-identified data as an ethical endpoint to understanding it as a starting point for broader commitments to justice, representation, and inclusion. Tools like automated de-identification and synthetic data generation must be evaluated not only by their technical performance but by their capacity to include diverse populations meaningfully and safely. Similarly, governance must evolve—from compliance-based frameworks to participatory models that empower underrepresented communities to shape the terms of their data's use. The road to ethical, generalizable healthcare AI does not lie in perfecting de-identification alone, but in redefining what responsible data looks like in practice. That means embedding demographic transparency into dataset standards, investing in global data-sharing infrastructures, and aligning technical advances with clear normative commitments to health equity. Without these changes, foundation models will continue to mirror the narrowness of their inputs. With them, we can begin to build models that reflect the full complexity—and dignity—of the patients they aim to serve.

References

Ajayi, D. (2020). How BERT and GPT models change the game for NLP. Watson Blog, Dec, 3.

Altalla', B., Abdalla, S., Altamimi, A., Bitar, L., Al Omari, A., Kardan, R., & Sultan, I. (2025). Evaluating GPT models for clinical note de-identification. *Scientific Reports, 15*(1), 3852.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, *4*, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3), 596–606.

Desouza, K. C. (2023). Toward a global health data commons: Legal and ethical enablers for equitable data sharing. *Health Policy and Technology*, *12*(2), 100673.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., ... & Sui, Z. (2022). A survey on in-context learning. arXiv preprint arXiv:2301.00234.

Ford, E., Pillinger, S., Stewart, R., Jones, K., Roberts, A., Casey, A., ... & Nenadic, G. (2025). What is the patient re-identification risk from using de-identified clinical free text data for health research?. *AI and Ethics*, 1-14.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.

Guo, L. N., Lee, M. S., Kassamali, B., Mita, C., & Nambudiri, V. E. (2022). Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—a scoping review. *Journal of the American Academy of Dermatology*, 87(1), 157–159. https://doi.org/10.1016/j.jaad.2022.01.058

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016).

MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035.

Kleinberg, G., Diaz, M. J., Batchu, S., & Lucke-Wold, B. (2022). Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *Journal of Biomed Research*, 3(1), 42–47. https://probiologists.com/Article/jbr-3-47.pdf

Lahoti, P., Beutel, A., Haghgoo, B., Raghunathan, A., Zhang, K., Lee, H., ... & Chi, E. (2023). Fairness in representation: Quantifying stereotyping as a representational harm. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Liu, Y., Park, Y., Wood, D., & West, R. (2023). DeID-GPT: Zero-shot medical text de-identification using generative transformers. *arXiv preprint arXiv:2301.02322*.

Marko, J. G. O., Neagu, C. D., & Anand, P. B. (2025). Examining inclusivity: the use of AI and diverse populations in health and social care: a systematic review. *BMC Medical Informatics and Decision Making*, *25*(1), 57.

Merritt, R. (2022, March 25). What is a transformer model? NVIDIA Blog. https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model

Moor, M., Banerjee, O., Abid, A., Zhang, M., Jagadeesan, A., Ling, Y., ... & Zittrain, J. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7958), 259–265.

Mpanya, D., Celik, T., Klug, E., & Ntsinjana, H. (2021). Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. *IJC Heart & Vasculature*, 34, 100773. https://doi.org/10.1016/j.ijcha.2021.100773

Olaker, V. R., Fry, S., Terebuh, P., Davis, P. B., Tisch, D. J., Xu, R., ... & Kaelber, D. C. (2025). With big data comes big responsibility: Strategies for utilizing aggregated, standardized, de‐identified electronic health record data for research. *Clinical and Translational Science*, *18*(1), e70093.

Patsakis, C., & Lykousas, N. (2023). Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1), 16026.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rothstein, M. A. (2010). Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics*, 10(9), 3–11.

Seastedt, K. P., Schwab, P., O'Brien, Z., Wakida, E., Herrera, K., Marcelo, P. G. F., ... & Celi, L. A. (2022). Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health*, *1*(10), e0000102.

Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58, S11–S19.

Sweeney, L., Abu, A., & Winn, J. (2017). Identifying participants in the personal genome project by name. *Data Privacy Lab Working Paper*.

Valacchi, M. (2024). Entity De-Identification in Information Extraction Tasks with Large Language Models (Doctoral dissertation, University of Applied Sciences).

Zhou, Y., Li, J., Xiang, Y., Yan, H., Gui, L., & He, Y. (2023). The mystery of in-context learning: A comprehensive survey on interpretation and analysis. arXiv preprint arXiv:2311.00237.