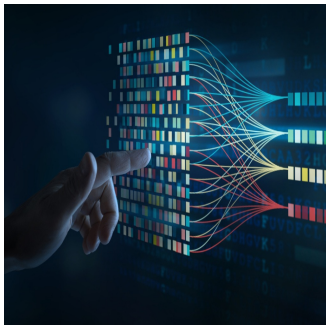
HealthManagement
Promoting Management and Leadership

The Impact of Sample Size on AI Prediction Models



As the integration of artificial intelligence (AI) in healthcare advances, there is increasing emphasis on the need for robust methodological standards in model development and validation. Among these, the role of sample size in AI-based prediction modelling remains critically underappreciated. Unlike traditional clinical research such as randomised trials, AI studies frequently omit justification for sample size, leading to questionable reliability of their outputs. Despite the existence of guidelines such as TRIPOD and regulatory principles like Good Machine Learning Practice, a large proportion of AI models are developed or tested using insufficiently sized datasets. This gap undermines model performance, raises fairness concerns and jeopardises clinical decision-making. By exploring the detrimental consequences of inadequate sample size and potential strategies for improvement, a clearer understanding of the issue's importance emerges.

The Consequences of Small Sample Sizes

The use of small datasets in AI model development presents several challenges that compromise the representativeness and reliability of predictions. Limited sample size often fails to capture the diversity of the target population, which is particularly problematic in ensuring the inclusivity and fairness of AI systems. Under-represented groups are especially vulnerable to poor model performance, as rare predictor patterns and subgroup-specific variations are unlikely to be sufficiently reflected. Even in apparently large datasets, effective sample sizes for particular groups may be alarmingly small, leading to unstable predictions and limiting generalisability.

This lack of representativeness has significant downstream effects. Predictor effects derived from small samples tend to exhibit high variance and instability. Models trained on such data produce inconsistent results with each iteration, rendering explanatory tools ineffective. Furthermore, individual-level predictions become unreliable due to the wide range of outcomes generated by models trained on similar but insufficiently sized samples. Uncertainty intervals around predicted risks become so broad that they fail to support clinical decisions, diminishing the value of such tools in patient communication and care.

Performance and Calibration Challenges

Small training datasets not only affect representativeness but also reduce the discriminatory power of prediction models. With limited data, models struggle to distinguish true signals from background noise, which degrades overall predictive performance. This is evident in metrics such as the c-statistic and explained variance (R^2), both of which improve significantly with increased sample size. In practice, a model with poor discrimination fails to accurately identify patients at risk, limiting its utility in guiding treatment decisions.

Must Read: ***The Next Frontier of AI in Healthcare: Prediction and Proactive Care***

Moreover, insufficient sample size contributes to poor calibration, meaning that predicted risks do not correspond well to observed outcomes. This miscalibration is a critical flaw when deploying models in real-world settings. The variability in calibration observed with small datasets is particularly concerning for complex models such as random forests, where prediction instability directly translates into inconsistent calibration. Poorly calibrated models offer little advantage over simplistic approaches, and in some cases, perform worse than treating all or none of the patients, depending on the clinical context.

The effect is further compounded during model evaluation. Validation of a model on an inadequately sized test dataset yields imprecise estimates of its performance. This imprecision results in wide confidence intervals and may give a misleading impression of model superiority or readiness for implementation. Such scenarios increase the risk of premature adoption, which may lead to suboptimal or even harmful clinical decisions.

Overcoming Barriers and Improving Practice
Addressing the issue of sample size in AI healthcare research requires both cultural and technical shifts. Education and training in data science should stress the necessity of rigorous study design, including sample size estimation for both training and evaluation phases. It is important to dispel misconceptions that modern machine learning methods obviate the need for sample size calculations. Although traditional statistical power analysis may not apply directly, alternative methods exist to determine appropriate dataset sizes for reliable model development and validation.

Several tools and methodologies are available to support sample size estimation, including software modules specifically designed for AI model research. While these approaches often derive from regression theory, they are adaptable to more complex machine learning techniques. When the available dataset is small, researchers are encouraged to seek additional data rather than proceeding with potentially unreliable models. Combining data from multiple studies or using routinely collected data, such as electronic health records, are viable strategies to enhance dataset size and diversity.

Importantly, increasing sample size is not a panacea. Models must still be carefully designed to mitigate biases, use appropriate definitions for predictors and outcomes, and remain generalisable across different populations. Even large datasets can produce flawed models if these aspects are neglected. Moreover, in scenarios where acquiring large datasets is impractical—such as in rare diseases—transparency about prediction uncertainty is essential. Models can still hold clinical value, provided that their limitations are clearly communicated and understood within the decision-making context.

In the development of AI-based prediction models for healthcare, the importance of sample size cannot be overstated. Insufficient datasets compromise representativeness, model stability, calibration and, ultimately, clinical utility. These limitations have real-world consequences, from inaccurate diagnoses to misguided treatment decisions. Therefore, robust sample size considerations must become a foundational element of AI research in healthcare. Emphasising this aspect in training, methodological guidance and regulatory oversight will be essential to ensure that AI models deliver tangible benefits to patients and uphold standards of fairness and quality in clinical care.

Source: **The Lancet Digital Health**

Image Credit: **iStock**

Published on : Mon, 9 Jun 2025