ORIGINAL ARTICLE

Can a Small Sample Size Capture a Big Truth? A Hypothetical Simulation Study on Diagnostic Accuracy of AI vs Radiologist in Tuberculosis Detection through X-rays

Frederick Satiro Vaz
Department of Community Medicine, Goa Medical College, Goa

CORRESPONDING AUTHOR
Dr Frederick Satiro Vaz, Department of Community Medicine, Goa Medical College, Goa 403202
Email: frederickvaz@rediffmail.com

CITATION
Vaz FS. Can a Small Sample Size Capture a Big Truth? A Hypothetical Simulation Study on Diagnostic Accuracy of AI vs Radiologist in Tuberculosis Detection through X-rays. Journal of the Epidemiology Foundation of India. 2026;4(1):84-90. DOI: https://doi.org/10.56450/JEFI.2026.v4i01.010

ARTICLE CYCLE
Received: 02/06/2025; Accepted: 15/02/2026; Published: 31/03/2026

ABSTRACT

Background: Artificial intelligence (AI) tools are increasingly being adopted in diagnostic radiology. However, early-phase evaluations often rely on small sample sizes due to logistical constraints. It remains unclear whether such small studies can reliably detect real differences in diagnostic accuracy between AI models and human experts. **Objectives:** To assess whether a small sample size (n = 30) can reliably detect statistically significant differences in diagnostic accuracy between an AI model and a radiologist for TB detection from chest X-rays under varying performance scenarios. **Methods:** A Monte Carlo simulation was conducted in R (v4.5.0) using 1000 iterations per scenario. Three diagnostic accuracy scenarios tested were: scenario 1. AI (98%) vs Radiologist (70%), scenario 2. AI (90%) vs Radiologist (75%) and scenario 3. AI (92%) vs Radiologist (80%). Predictions were simulated using defined sensitivity and specificity values, and accuracy was computed. A one-sided paired t-test assessed whether AI outperformed the radiologist. Empirical power was calculated as the proportion of simulations yielding p-values < 0.05. **Results:** Scenario 1 with a large accuracy gap (28.2%), achieved 96% empirical power. Scenarios 2 and 3, with moderate and small differences (~14%), achieved only 46.8% and 37.5% power, respectively. **Conclusion:** Small samples can detect large diagnostic performance differences but are unreliable for moderate or small differences.

KEYWORDS

Artificial Intelligence, Radiologist, tuberculosis, Monte Carlo method, Sample size, Reproducibility of results.

INTRODUCTION

The integration of artificial intelligence (AI) into diagnostic radiology represents one of the most promising innovations in modern medicine. In diseases like tuberculosis (TB), where early diagnosis can significantly reduce morbidity and transmission, computer-aided interpretation of chest X-rays offers the potential for timely and scalable decision-making. However, a critical question arises when assessing whether AI performs better than a trained human radiologist: how much data is enough to detect this difference? This question is important and crucial. Many diagnostic studies, especially those in early

phase are limited to small samples due to cost, logistics, or ethics.

Edgar Derricho[1] in his insightful simulation-based essay on treatment efficacy, states that even with a relatively small sample size (n = 30), we can often detect population-level effects if: - effect size is sufficiently large, sample is randomly drawn and variability is not extreme. Extending this concept, this study aims to examine whether small samples can reliably detect a meaningful performance gap between AI and radiologists in diagnosing TB from X-rays when such a gap truly exists at the population level.

Objective: to assess whether small sample studies (n = 30) can reliably detect statistically significant differences in diagnostic accuracy between artificial intelligence (AI) and a human radiologist for tuberculosis (TB) diagnosis from chest X-rays, under three different performance scenarios.

Material & Methods

This simulation study was conducted using R Programming Statistical Software (R version 4.5.0). The approach involved generating a large hypothetical population, simulating diagnostic predictions based on defined sensitivity and specificity for both AI and the radiologist, extracting repeated random samples, and comparing diagnostic accuracy using paired statistical tests (paired t test). Simulation is a computational method where random processes are repeatedly sampled to mimic real-world experiments. Repeated sampling experiments (1000 iterations) were run to calculate the empirical power which is the proportion of simulated studies that successfully detect a statistically significant difference. Monte Carlo simulation methods were to assess how often small samples detect a true difference in diagnostic accuracy between AI and human radiologists.

Conceptual Framework: This study focuses on comparing diagnostic accuracy between two classifiers—an AI model and a human radiologist. The diagnostic accuracy[3] is defined as the proportion of correctly classified individuals (both with and without TB) in a sample. This overall metric, however, is based on two measures[4]: Sensitivity (True Positive Rate): The proportion of actual TB cases correctly identified and Specificity (True Negative Rate): The proportion of non-TB cases correctly classified as negative. The study simulated diagnosis in a controlled environment using known values of sensitivity and specificity for both AI and human radiologist readers.

Steps in the simulation process:

Simulation Setup and Key Parameters: The exercise began with defining three hypothetical scenarios, each representing a different degree of superiority of the AI model over a radiologist in terms of accuracy. The three scenarios were:

1. **Scenario 1**: AI = 98%, Radiologist = 70%
2. **Scenario 2**: AI = 90%, Radiologist = 75%
3. **Scenario 3**: AI = 92%, Radiologist = 80%

These three scenarios reflect large, moderate, and small performance differences, respectively. The goal was to determine whether a sample of 30 paired cases would be sufficient to detect this difference statistically in each case.

We set the following constants:

```
n_population <- 1000   # Size of the simulated population
n_sample <- 30         # Sample size for each simulated trial
n_sim <- 1000          # Number of simulation iterations
```

Simulating Diagnostic Outcomes: The main concept was to simulate how both AI and doctors make predictions in a binary classification task (TB present or absent), using known sensitivity and specificity. For each simulated patient: If the patient has TB (true_label = 1), the model has a probability equal to its sensitivity of correctly predicting positive (1). If the patient does not have TB (true_label = 0), the model has a probability equal to its specificity of correctly predicting negative (0).

This was implemented using the binomial distribution. This function returns a vector of 0s and 1s representing simulated diagnostic predictions:

```
simulate_predictions <- function(true_labels, sensitivity, specificity) {
  sapply(true_labels, function(t) {
```

```
  if (t == 1) rbinom(1, 1, sensitivity) else rbinom(1, 1, 1 - specificity)
 })
}
```

Population-Level Accuracy Assessment: To benchmark each scenario, first diagnosis for an entire population of 1000 patients was simulated. For each method, predictions were compared against the true TB status to compute overall accuracy This step provided the "ground truth" population-level diagnostic performance for AI and the radiologist:

```
true_labels <- sample (c (0, 1), size = n_population, replace = TRUE) # Random TB presence
ai_pred_pop <- simulate_predictions(true_labels, sensitivity_AI, specificity_AI)
doc_pred_pop <- simulate_predictions(true_labels, sensitivity_doc, specificity_doc)
ai_accuracy <- mean (ai_pred_pop == true_labels)
doc_accuracy <- mean (doc_pred_pop == true_labels)
```

Simulating Paired Sample Trials: To evaluate how well a small sample (n = 30) reflects population-level differences, simulation of 1000 paired studies for each scenario was run. Each iteration drew a new random sample of 30 patients. Diagnostic predictions were simulated for both AI and radiologist. Accuracy for each method was computed for each patient (1 = correct, 0 = incorrect). The paired t-test was used to determine whether AI was significantly more accurate than the doctor for that sample:

```
# Storage for results
pvalues <- numeric(n_sim)
diffs <- numeric(n_sim)
for (i in 1:n_sim) {
 true_labels_sample <- sample(c(0, 1), size = n_sample, replace = TRUE)
 ai_pred <- simulate_predictions(true_labels_sample, sensitivity_AI, specificity_AI)
 doc_pred <- simulate_predictions(true_labels_sample, sensitivity_doc, specificity_doc)
 # Accuracy per patient
 ai_acc <- as.integer(ai_pred == true_labels_sample)
 doc_acc <- as.integer(doc_pred == true_labels_sample)

 # Store accuracy difference
 diffs[i] <- mean(ai_acc) - mean(doc_acc)
 # Paired t-test (alternative = "greater" tests if AI > doctor)
 pvalues[i] <- t.test(ai_acc, doc_acc, paired = TRUE, alternative = "greater") $p.value
}
```

Calculating Empirical Power: Empirical power was defined as the proportion of trials where the null hypothesis (no difference) was rejected at a 5% significance level. This value tells us how often a small-sample trial successfully detects AI's superiority under each scenario.

```
emp_power <- mean(pvalues < 0.05)
```

Visualization of Simulation Results: Following visualizations were performed: Histogram of p-values showing how often small studies yield significant vs non-significant results. Histogram and boxplot of accuracy differences showing distribution of observed effect sizes. ECDF (Empirical Cumulative Distribution Function) depicting what fraction of trials yielded p-values below threshold (i.e. 0.05).

Thus, each scenario repeated this full simulation pipeline independently, using its own sensitivity and specificity values for AI and the radiologist.

RESULTS

this hypothetical simulation study was conducted to evaluate whether small sample studies (n = 30) could reliably detect statistically significant differences in diagnostic accuracy between an artificial intelligence (AI) model and a human radiologist for the diagnosis of tuberculosis (TB) on chest X-rays.

Population-Level Accuracy and Simulated Differences: The AI model outperformed the radiologist in all three scenarios (Table 1). In Scenario 1, the AI system had a population-level diagnostic accuracy of 97.6% compared to the radiologist's

69.3%, yielding a large mean accuracy difference of 28.2 percentage points (SD: ±9.0). This large difference was detectable in most of the small samples drawn, with 96.0% empirical power to reject the null hypothesis of no superiority ($p < 0.05$).

In Scenario 2, where the AI and radiologist had accuracies of 89.5% and 75.5% respectively, the mean accuracy difference was 14.0% (SD: ±10.7). Despite the moderate effect size, the empirical power dropped to just 46.8%, indicating that fewer than half of the small-sample simulations successfully detected AI's superiority.

Scenario 3 presented a slightly smaller effect size than Scenario 2, with AI and radiologist accuracies of 93.0% and 79.1%, respectively. The average accuracy gain of 13.9% (SD: ±10.6) resulted in still further lower empirical power of only 37.5%, This depicts the challenge of detecting small-to-moderate differences with small sample sizes, even when the true population difference is meaningful.

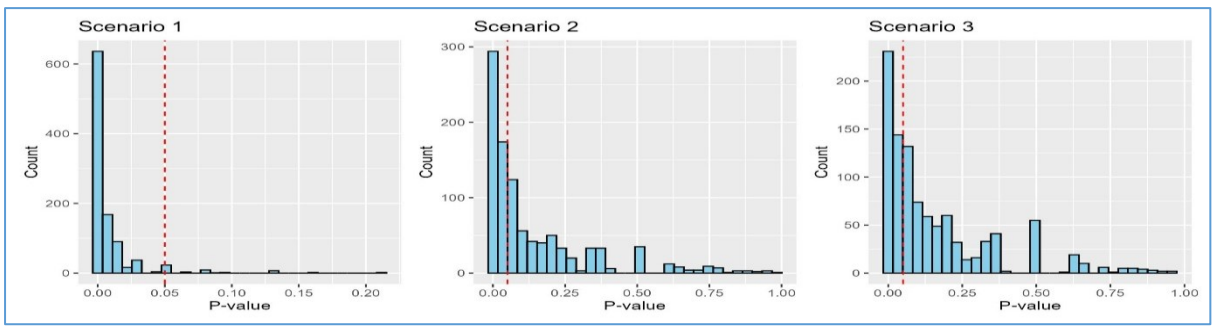
Table 1. Summary of Population-Level Diagnostic Accuracy and Simulation Results Across Three Scenarios (n = 30, 1000 simulations)

Scenario	AI Accuracy	Doctor Accuracy	Mean Accuracy Difference (SD)	Empirical Power (%)	Effect Size	Interpretation
1	0.976	0.693	0.282 (±0.090)	96.0	Large	*Small samples reliably detect the large performance gap between AI and doctor.*
2	0.895	0.755	0.140 (±0.107)	46.8	Moderate	*Detectability is limited.*
3	0.930	0.791	0.139 (±0.106)	37.5	Small	*Effect size too small for consistent detection in small samples.*

Visualization of Simulation Distributions

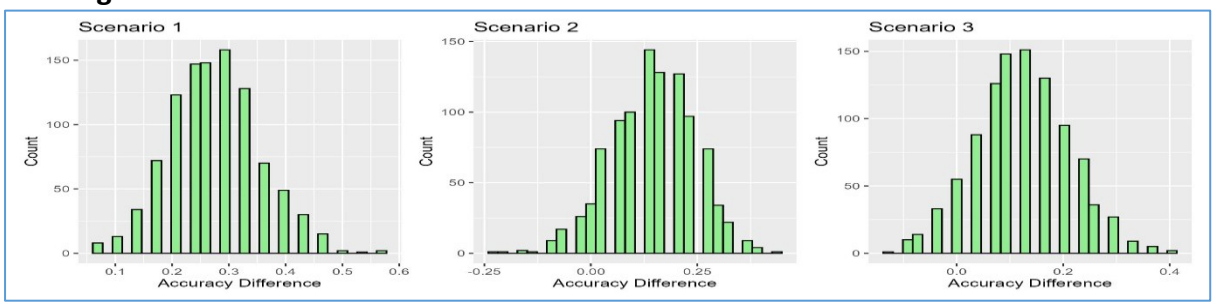
The distribution of p-values for each scenario is presented in Figure 1. In Scenario 1, a substantial proportion of p-values lie below the 0.05 threshold, suggesting strong statistical evidence of AI's superiority. This is visually indicated by the high density of p-values clustering near zero. In contrast, Scenarios 2 and 3 show flatter and more dispersed p-value distributions, with fewer values falling below the significance threshold, illustrating the decline in statistical power.

Figure1: Histograms of p-values from 1,000 paired t-tests comparing AI and radiologist across the three scenarios.



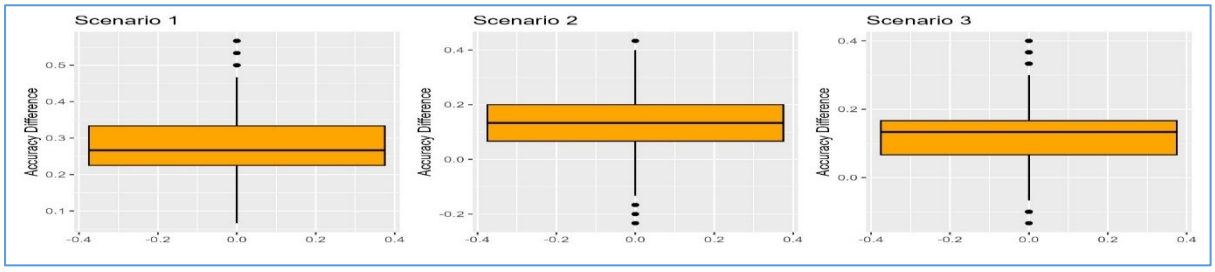
The accuracy difference distributions in Figure 2 show near-normal shapes centered around their respective mean differences. Scenario 1 displays a clear separation, while Scenarios 2 and 3 have greater overlap and spread, suggesting increasing uncertainty as effect size decreases.

Figure 2: Histograms of Accuracy Difference from 1,000 paired t-tests comparing AI and radiologist across the three scenarios.



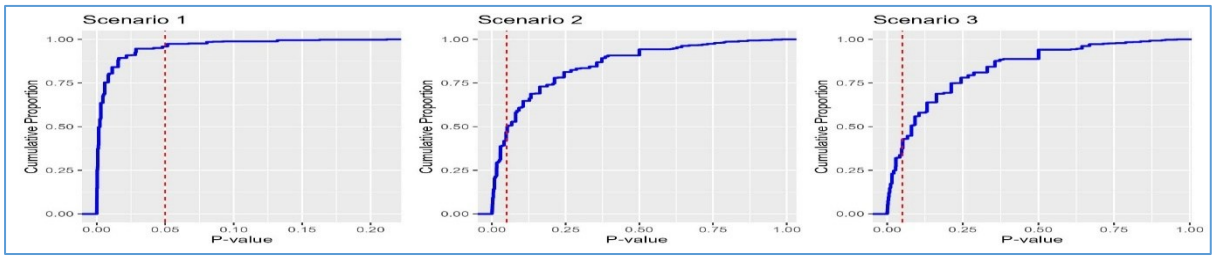
The boxplots of accuracy differences (Figure 3) further highlight the variability across simulations. Scenario 1 exhibits tightly packed interquartile ranges and fewer outliers, consistent with a large and stable effect. In contrast, Scenarios 2 and 3 show wider spreads, occasional negative differences (favouring the radiologist), and greater influence from outliers

Figure 3: Box Plots of Accuracy Difference from 1,000 paired t-tests comparing AI and radiologist across the three scenarios.



Finally, the Empirical Cumulative Distribution Function (ECDF) plots of p-values (Figure 4) provide a visual representation of cumulative statistical significance across simulations. In Scenario 1, over 95% of simulations resulted in p-values below 0.05, whereas this proportion dropped to less than half in Scenario 2 and to around one-third in Scenario 3.

Figure 4: Empirical Cumulative Distribution Function (ECDF) plots of p-values from 1,000 paired t-tests comparing AI and radiologist across the three scenarios.



DISCUSSION

The study set out to answer a fundamental methodological question in diagnostic research: Can small sample sizes (n = 30) reliably detect real and meaningful differences in diagnostic accuracy between an artificial intelligence (AI) model and a human radiologist in detecting

tuberculosis (TB) from chest X-rays? This well-structured simulation has demonstrated that the answer strongly depends on the size of the effects. More specifically, the difference in diagnostic accuracy between the two comparators.

The results show a clear decreasing gradient in empirical power across three diagnostic scenarios. In Scenario 1, the AI model had a substantial superiority over the radiologist (97.6% vs. 69.3%), which was consistently detectable even in small sample simulations (empirical power: 96%). This has also been articulated by Edgar Derricho[1], that small samples can detect population-level truths when the effect size is large, the data is sampled randomly, and the variance is moderate. Scenario 1 serves as a proof-of-concept that when differences are large, even a modest number of observations is sufficient for statistical tests.

Scenario 2 and Scenario 3 offer a more sobering reality. Although AI continued to outperform the radiologist, the ability to detect this difference in small samples fell to 46.8% and 37.5%. Although the accuracy gap would be considered meaningful in clinical practice, particularly for a disease like TB, the statistical evidence from small samples was inconsistent highlighting the need for caution in the interpretation of non-significant results from underpowered studies i.e. absence of evidence is not evidence of absence.

The paired t-test was used considering its suitability for within-subject comparisons. Each simulated sample represented 30 patients whose diagnostic outcomes were independently assessed by both AI and a radiologist. By analysing the paired difference in accuracy per simulated patient, inter-subject variability was reduced. Although, accuracy is a proportion, the t-test remains appropriate due to the near-normal distribution of the differences.

In real-world settings however, sensitivity and specificity often trade off. However, the use of the same value for sensitivity and specificity for each diagnostic agent within each of the scenarios was a deliberate choice to simplify the simulation exercise.

Conclusion

This simulation study indicates that sample size alone does not determine study validity; effect size, variability, and statistical methodology play equally vital roles. When differences in diagnostic accuracy between comparators are large, small studies can confidently detect them. However, this is not true for moderate to small effect sizes. Simulation offers a powerful, flexible tool for navigating these trade-offs, ultimately enabling more reliable, scalable, and interpretable diagnostic research in the era of artificial intelligence.

Recommendation

The findings of this have practical implications for the design, evaluation, and interpretation of diagnostic studies in particular, those evaluating AI technologies.

1. Early-phase evaluations of AI tools often rely on small feasibility studies. Scenario 1 supports this approach for high-performing models where large accuracy gain exists. However, Scenarios 2 and 3 provide abundant caution that smaller effect sizes may be hidden by sampling noise.
2. Negative or null results in underpowered diagnostic studies must be interpreted with caution. In Scenario 3, the true superiority of the AI was statistically undetectable in nearly two-thirds of simulations. This highlights risk of falsely concluding non-superiority when sample size is inadequate.

3. Power analysis and simulation modelling should now be an integral part of study planning. The empirical power values obtained in this study demonstrate the importance of tailoring sample sizes to expected effect sizes. Simulation therefore offers a practical way to perform power calculations.

Limitation of the Study

This simulation assumes ideal conditions: perfectly random sampling, binary ground truth, and no misclassification bias. Real-world diagnostic studies encounter confounders like reader fatigue, variation in image quality, and class imbalance, all of which can affect power and generalizability. We used accuracy as metric of performance; this measure does not capture clinical trade-offs between sensitivity and specificity. In TB screening, missing a true case (false negative) may be more consequential than a false alarm. The simulation fixed sample size at n = 30 across all scenarios (considering the conventional link between sample size of 30 and the Central Limit Theorem), evaluating how increasing the sample size improves empirical power in each case my provide a different viewpoint.

Note: This study does not support the hypothesis that AI performs better than human but is only intended to demonstrate the impact of small sample size in a hypothetical situation where AI is stipulated to perform better than humans with three different effect sizes.

Relevance of the Study

This study adds to the growing literature on artificial intelligence evaluation in healthcare. It demonstrates how sample size and effect size jointly influence statistical detectability in diagnostic research. Since most research on AI performance focus primarily on overall accuracy metrics, this study emphasizes the methodological issue of statistical power in small-sample diagnostic studies.

Financial Support and Sponsorship

Nil

Conflict of Interest

There are no conflicts of interest.

Declaration Of Generative AI And AI Assisted Technologies In The Writing Process

During preparation of this work, the authors used GITHUB copilot integrated with R studio to assist in generation and debugging the R code. Quillbot was used for grammar check. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Derricho E. Can a Small Sample Reflect a Whole Population? A Simulation Study on Treatment Efficacy [Internet]. Medium. 2025 [cited 2025 May 10]. Available from: https://medium.com/@ederricho1/can-a-small-sample-reflect-a-whole-population-a-simulation-study-on-treatment-efficacy-50f0b2a24258
2. R Core Team (2025). R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
3. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009 Jan 20;19(4):203-11.
4. Altman DG, Bland JM. Statistics notes: Diagnostic tests 1: Sensitivity and specificity. BMJ. 1994;308(6943):1552. doi: 10.1136/bmj.308.6943.1552.