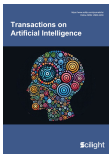
Scilight

*Transactions on Artificial Intelligence*

https://www.sciltp.com/journals/tai





*Article*

# Human-Centric Clinical AI: A Structural Data Validation Approach for Human-Reliable Medical Inference

Marco Roccetti

Computer Science and Engineering Department, University of Bologna, Bologna 40125, Italy; marco.roccetti@unibo.it

**How To Cite:** Roccetti, M. Human-Centric Clinical AI: A Structural Data Validation Approach for Human-Reliable Medical Inference. *Transactions on Artificial Intelligence* **2026**, *2*(1), 204–213. https://doi.org/10.53941/tai.2026.100013

Received: 16 July 2026
Revised: 31 August 2026
Accepted: 1 September 2026
Published: 11 September 2026

**Abstract:** Researchers increasingly rely on Clinical Artificial Intelligence (CAI) to derive predictive insights from massive observational datasets. However, the paradigm of Big Medical Inference often conflates statistical volume with clinical representativeness. This study argues that when methodological rigor is sacrificed for data scale, AI models risk institutionalizing historical biases, thereby compromising patient safety and public health policy integrity. We propose a precautionary, multi-level auditing framework designed to assess the structural architectural integrity of large-scale clinical datasets prior to computational deployment. Rather than contesting specific clinical outcomes, our approach establishes a quantitative prerequisite for CAI: the validation of demographic representativeness and baseline case distribution against official national benchmarks. To demonstrate the validity of our approach, we applied our framework to a prominent, high-profile case study. Our check revealed a tripartite structural divergence: a 32.5% demographic deficit in high-risk elderly (aged ≥ 65 years) strata, an aggregate of approximately 26% cancer incidence suppression, and a 45.1% deflation of expected cases in non-exposed groups. These discrepancies demonstrate that even massive datasets can be fundamentally misaligned with clinical reality. We conclude that structural validation is not an elective procedure but an ethical imperative for Human-Centric Clinical AI. By re-establishing this hierarchy of validation, we ensure that the intelligence of automated AI systems remains subordinate to the structural truth of the data, thereby transitioning from a reliance on Big Data alone toward a more robust, ethically sound, and authentic practice of clinical knowledge.

**Keywords:** clinical artificial intelligence; human-centric AI; medical data integrity; structural validation; dataset representativeness

## 1. Introduction

Clinical Artificial Intelligence (CAI) represents the frontier of modern healthcare, promising to transform vast repositories of medical data into actionable clinical insights. At its core, CAI encompasses a spectrum of computational methodologies, ranging from predictive modeling to automated diagnostic support, designed to augment human decision-making and personalize therapeutic interventions. By synthesizing information across diverse patient populations, these systems aim to optimize diagnostic accuracy and public health strategies, positioning themselves as essential tools in the transition toward truly data-driven, precision medicine [1].

The efficacy of these systems is intrinsically linked to the emergence of Big Medical Inference, a paradigm wherein complex algorithms derive clinical correlations from large-scale, heterogeneous datasets. In this context, the enormous volume of Big Medical Data, encompassing Electronic Health Records (EHRs), administrative databases and longitudinal clinical registries, serves as the substrate for machine learning architectures. The power

of this approach lies in its ability to transcend the limitations of traditional, small-scale clinical trials, offering the statistical depth required to investigate elusive biological signals across millions of patient records.

However, this reliance on Big Medical Inference introduces a fundamental methodological challenge: the clash between statistical significance and clinical representativeness. In the realm of large-scale observational research, in fact, an increase in sample size can paradoxically act as a veil, masking structural instabilities within the cohort architecture. Large datasets frequently suffer from sampling biases, demographic imbalances, or asymmetric under-reporting that, while not invalidating the statistical significance of the model, may severely compromise the clinical validity of the findings. If a CAI system is trained on data that trade clinical plausibility for large volume, the resulting AI models risk perpetuating and scaling these initial architectural instabilities, ultimately producing statistically precise predictions that lack genuine clinical relevance [2].

The implications of this methodological vulnerability extend well beyond technical disappointment; they constitute a problem of profound social impact and humanistic relevance [3]. When CAI systems underpin public health policies, diagnostic guidelines, or treatment recommendations, the integrity of the data becomes a matter of societal trust. If the foundational data is structurally unstable, the derived insights may influence human lives, impacting treatment pathways or policy decisions, based on artifacts of data organization rather than authentic biological reality. Thus, ensuring the integrity of these datasets is not merely a quantitative requirement, but an ethical imperative for protecting the well-being of the populations served by CAI.

To address this systemic risk, this paper proposes a rigorous, precautionary approach for the structural preliminary validation of large-scale clinical datasets. We argue that before any predictive algorithm or deep learning pipeline is deployed, the dataset must undergo a multi-level audit to assess its fundamental architectural integrity. This validation framework is based on three essential pillars, applied as an additional control phase to the original medical dataset (beyond its mere statistical significance): (1) checking demographic representativeness across age strata; (2) ensuring consistency of crude incidence rate (CR) against established benchmarks; and (3) evaluating stratum-specific consistency within non-exposed groups. This preliminary step ensures that the data substrate is robust and representative before it is subjected to an AI/computational analysis.

To demonstrate the efficacy of our validation approach, we have selected a prominent, high-profile case study from the recent literature [4]. Through this exemplar, we illustrate how our proposed method effectively uncovers underlying structural instabilities that would otherwise remain masked by the dataset's large size. Our analysis highlights how such instabilities, if left unexamined, could critically misalign the resulting CAI training process with real-world medical evidence, showing that the system's intelligence is inherently constrained by the architectural quality of its input.

To bridge this conceptual gap, this paper introduces a computational perspective: we frame dataset structural validation not as a traditional epidemiological descriptive task, but as a mandatory pre-training architectural constraint for Clinical AI. By establishing quantitative verification gates prior to model deployment, we aim to prevent data-substrate distortions from propagating into downstream algorithmic optimization and automated decision-making.

Finally, it is essential to emphasize that our selection of this specific case study is purely illustrative, chosen for its exemplary capacity to reveal structural methodological dynamics. We maintain a strictly agnostic position regarding the medical findings reported in the original dataset, particularly concerning the discussed relationships between COVID-19 vaccination and oncological outcomes, a complex topic that remains the subject of ongoing scientific debate. Our study is directed exclusively toward the structural architecture of the data used for the analysis, with the utmost respect for the original authors, the institutions they represent, and the journals that have facilitated this vital area of medical research.

The remainder of the paper proceeds as follows. In Section 2, we review the main concepts at the basis of the CAI paradigm and expose some typical pitfalls that have been historically encountered in this specific field. In Section 3, we describe the methods at the basis of our proposed solution, along with the data from the exemplar case with which we decided to test our approach. Section 4 provides a comprehensive summary of the achieved results, while Section 5 terminates the paper with a discussion.

## 2. Clinical AI: Potential Pitfalls and Solutions

The digital revolution in medical research has ushered in an era where the scale of available data is unprecedented. However, this transition from traditional clinical trials to massive, retrospective observational studies necessitates critical reflection. Data integrity is not merely a mathematical requirement or a technical hurdle; it is a profound social duty. When clinical decisions regarding public health are predicated on structurally unstable datasets, the resulting policies directly impact human lives, often irreversibly. Consequently, protecting

data integrity by ensuring that observed signals are truly representative transcends technical accuracy to become a fundamental social and humanistic concern.

As documented in modern literature, reliance on massive inferential techniques without adequate foundational checks can obscure critical errors, leading to contradictory results that violate the core principles of biostatistics and biomedicine established by pioneers such as Fisher, Pearson, Neyman, and Cox. In fact, in this context of digital health, the pressure to reject the null hypothesis and identify sensationally new associations often overrides the diligence required for basic descriptive analysis and base-rate checks [5–8].

To fully contextualize the gravity of these methodological lapses and demonstrate that unstable dataset construction is not a modern anomaly, it is essential to revisit medical history. Structural data errors have historically misled clinical science, often precisely because researchers neglected to verify the structural validity of their study datasets.

As a first example, for decades, influential observational studies such as the Nurses' Health Study suggested that Hormone Replacement Therapy (HRT) drastically reduced cardiovascular risk, making it a global clinical standard [9]. It was later revealed that these studies suffered from severe healthy user bias, as women choosing HRT were inherently healthier and wealthier. The definitive Women's Health Initiative (WHI) trial in 2002 proved the opposite, revealing that HRT actually increased cardiovascular risk [10,11]. This remains the most prominent example of how structural dataset bias created a protective effect that was entirely non-existent, leading to years of suboptimal clinical practice and demonstrating a cycle of bias that periodically reasserts itself in different clinical sectors.

Secondly, and no less importantly, the most damaging precedent is the 1998 study linking the Measles, Mumps, and Rubella (MMR) vaccine to autism. While data manipulation was the headline, the central methodological failure was the use of a tiny, selectively chosen sample with no valid non-exposed group, an extreme case of selection bias. Subsequent large-scale studies involving millions of children demonstrated exactly a null risk, confirming the lack of association. This case remains a cautionary tale of how incorrect dataset construction can erode public trust in preventive health measures for years [12].

Third, in the 1980s and 1990s, observational studies suggested that frequent use of inhaled beta-agonists increased asthma mortality. In reality, this was a classic example of confounding by indication: patients receiving the most intensive doses were simply those with the most severe, life-threatening disease. The drug appeared to cause the severity, whereas it was merely a marker of pre-existing risk. This structural bias produced spurious findings leading to widespread misinterpretation of clinical risk, demonstrating how procedural errors in establishing the priority of causes and effects can generate misleading signals even in the absence of fraud [13].

These precedents underscore a critical humanistic and ethical concern. When methodological rigor is sacrificed for the sake of Big Data convenience, we risk institutionalizing historical biases within the very heart of Clinical AI. The allure of massive datasets creates a dangerous illusion of objectivity, where the ample volume of records is mistaken for a guarantee of quality. However, when these massive, heterogeneous datasets are fed into sophisticated machine learning pipelines without a preliminary audit of their structural architecture, the resulting models risk learning and amplifying existing methodological instabilities. In this context, the AI does not merely replicate historical errors; it crystallizes them into automated clinical guidelines, transforming unstable selection processes into seemingly precise predictive scores. Consequently, if we treat Big Data as an inherently reliable substrate, we inadvertently delegate our public health decisions to algorithms that may be fundamentally misaligned with the real-world population they intend to serve.

Therefore, our study aims to address this systemic vulnerability by advocating for mandatory structural validation of the dataset as a foundational prerequisite for CAI and Big Medical Inference. We reiterate that the audit of a dataset's architecture, verifying demographic representativeness, baseline risk parity, and strata-specific case distribution, must precede any application of predictive algorithms or deep learning models. By re-establishing this hierarchy of validation, we ensure that the scale of Big Data is leveraged for clinical depth rather than statistical noise. This preliminary validation step is not merely a technical suggestion; it is an ethical and humanistic imperative for secure AI. It ensures that the evidence driving future medical interventions is grounded in structural truth, protecting the integrity of CAI applications and, ultimately, the safety of the patients whose lives depend on these automated insights.

Nonetheless, it should be reiterated that our approach has its position within the computational domain of Clinical AI, since the structural validation of training data is to be considered not as a substitute for predictive modeling, but as a mandatory pre-training architectural constraint. In standard machine learning pipelines, a predictive model minimizes a loss function $\mathcal{L}\,(y, f(x, \Theta))$ over an empirical training distribution $\hat{P}(x, y)$. When the underlying clinical dataset suffers from unaddressed structural distortions, such as a macro-level demographic deficit or a systematic baseline incidence suppression, the empirical distribution $\hat{P}$ diverges sharply from the true

population distribution $P$. Consequently, optimization algorithms (e.g., gradient descent) fit model parameters $\Theta$ to an artifactual data manifold. This results in severe downstream consequences that standard model-level evaluations cannot automatically correct. They can be as follows.

First, suppressed outcomes in high-risk strata skew the penalty weights assigned to minority or vulnerable sub-populations.

Second**,** even if internal validation metrics (such as the Area Under the ROC Curve) appear high, models trained on structurally flawed cohorts suffer from extreme miscalibration when deployed in real-world settings.

Third, neural networks and complex classifiers inherently learn and scale the structural instabilities of their training substrate.

Therefore, it should be clear that our proposed multi-level auditing framework serves as an algorithmic filter, a necessary computational checkpoint executed prior to model training, ensuring that automated inference remains subordinate to the structural truth of the medical data.

## 3. Data and Methods

As anticipated, we developed an approach to test the validity of clinical data before its deployment within a CAI. We also reiterate here that to test on-field the validity of our approach we selected a specific clinical dataset [4] for which a preliminary biostatistical evaluation had revealed a discrepancy: the reported overall crude incidence rate (CR) of cancer (the disease of investigation) within the dataset was lower than the corresponding official national average recorded in the country of interest [14]. Specifically, the dataset's overall CR was 22.3% lower than the established national average rate calculated as an average over the 2020–2022 period [15–17]. When a more contemporary and rigorous baseline was applied, it amounted to a CR of 55.02 per 10,000 individuals, with the suppression of the dataset's overall cancer incidence becoming even more pronounced, confirming a substantial 26% deficit.

Obviously, our present study moves significantly beyond the mere identification of this single statistical divergence, seeking instead to uncover the underlying structural causes that define it. Using that oncological dataset as a prominent case study, we will show how to quantitatively deconstruct the dataset's demographic data architecture against verified, publicly available figures and official national statistics. Following this pathway, the initial signals of increased cancer risk in the vaccinated population recorded in the dataset will emerge as a result of asymmetric selection procedures specifically localized within the high-risk elderly demographic, revealing an instability which is independent of the large size of the dataset from which that data was extracted. Our study shows that, absent a quantitative framework for structural validation like that here proposed, these pervasive selection effects cannot be effectively controlled.

Beginning with relevant data for our study, we maintain that primary data, encompassing age-stratified participant distributions and specific oncological case counts, were meticulously extracted from the original dataset [4]. Moreover, to construct a definitive gold standard for conducting an appropriate comparison, we fetched official statistics from the nation of interest (South Korea), including comprehensive cancer incidence rates and demographic reports sourced from recognized public registries and official governmental databases [15–19]. All this documentation is portrayed in Table 1.

**Table 1.** Structural Comparison: N-size Dataset vs. National Benchmarks.

| Population Stratum | National Weight | National Benchmark CR | Group | Participant | Case | Dataset CR |
|---|---|---|---|---|---|---|
| Total | 100% | 55.02 | Aggreg. | 2,975,035 | 12,133 | 40.78 |
| Non-elderly (<65 years) | 82% | 33.03 | NE | 522,722 | 1373 | 26.27 |
| | | | E | 2,090,888 | 6861 | 32.81 |
| Elderly (≥65 years) | 18% | 155.20 | NE | 72,285 | 616 | 85.22 |
| | | | E | 289,140 | 3283 | 113.54 |

To ensure maximum clarity for the reader, here follows the column-by-column description of the structural data presented in Table 1. In particular, in the first column (Population Stratum), the population is divided into two primary age-based groups: non-elderly (<65 years) and elderly (≥65 years), alongside an aggregate row representing the total population.

In the second column (National Weight), the demographic proportion of each age stratum is represented within the national South Korean population (18% for the elderly and 82% for the non-elderly).

In the third column (National Benchmark CR): the official, verified Crude Incidence Rates (CR) for all cancers, per 10,000 individuals, are reported, which act as the gold standard for comparisons with the corresponding CRs of the dataset.

In the fourth column (Group), the exposure status to the COVID-19 vaccination within the dataset is reported, distinguishing between the Non-Exposed (NE) and Exposed (E) cohorts, as well as the total combined population.

In the fifth column (Participants), the raw count of individuals is listed in each specific subgroup derived from the investigated dataset.

In the sixth column (Cases), the absolute number of oncological cases are reported as observed within each corresponding subgroup of the dataset.

Finally, In the seventh column (Dataset CR): the calculated crude incidence rate per 10,000 individuals for each subgroup is calculated as derived directly from the dataset's figures to facilitate a direct comparison with the national benchmarks provided in the third column.

To conclude, it needs to be emphasized also that while the figures for the Total Population and the Population ≥ 65 were directly sourced from official national registries [18,19], the metric for the Population < 65 was calculated to ensure absolute coherence with the aggregate dataset. Specifically, the Crude Incidence Rate (CR) for the population under 65 years (33.03 per 10,000) was derived by applying the principle of aggregate weighted incidence. Under this methodological framework, the total national cancer incidence is defined as the weighted sum of age-specific incidence rates. To achieve this, we solved the following linear equation using the established demographic weights (18.0% for ≥65 and 82.0% for <65) alongside the official national CR of 155.2 for the ≥65 group.

We come now to the mathematical framework we defined to analyze the data above consisting of three main core formulas. These equations were designed to compute, respectively, the Crude Cancer Incidence (CR), the relative Demographic Deficit or Surplus within the cohort strata, and the percentage Difference between Observed and Expected Cancer Incidence Rates.

It is important to specify that these three key mathematical formulas are considered robust and valid under the fundamental methodological assumption of a fixed and synchronized time window of observation for both the exposed and non-exposed cohorts. This synchronization ensures that the comparison between the oncological trends of the initial dataset and the national gold standards is chronologically and statistically consistent.

They are the following.

$$\text{CR (per 10,000)} = (\text{Number of Cases}/\text{Population at Risk}) \times 10{,}000 \tag{1}$$

$$\text{Relative Deficit/Surplus \%} = (\text{Cohort Percentage} - \text{Population Percentage})/\text{Population Percentage} \times 100 \tag{2}$$

$$\text{Difference (\% incidence)} = (\text{Cohort Incidence Rate} - \text{Reference Incidence Rate})/\text{Reference Incidence Rate} \times 100 \tag{3}$$

We anticipate a succinct but exhaustive explanation for the structural anomaly we previously mentioned as applied to our exemplar case. In fact, to anticipate the scale of this deviation, it would be sufficient to calculate the overall Crude Cancer Incidence Rate (CR) for the entire large-scale cohort analyzed in the original study [4]. By applying our standardized formula to the total reported figures (12,133 cancer cases within a population of 2,975,035 individuals), we obtain an aggregate CR of 40.78 per 10,000. When this cohort metric is contrasted against the established national gold standards, the statistical gap becomes immediately evident. Specifically, comparing the dataset's performance to the national historical average CR for the 2020–2022 period (52.46 per 10,000) reveals a significant downward deviation of approximately 22.3%.

Furthermore, if we utilize the more recent and refined 2022 national benchmark (55.02 per 10,000), the discrepancy widens to almost a 26% deficit. Such a massive under-reporting of expected cancer cases, over a quarter of the anticipated national volume, serves as a definitive indicator of a concerning data selection issue, confirming that the cohort's baseline appears somewhat disconnected from the real-world oncological landscape of the nation under investigation.

*Decision Criteria*

To transition our methodological framework from descriptive comparison to a rigorous, threshold-driven decision tool to check the validity of data prior to training, we established a quantitative three-zone structural validation model (adapting the governance taxonomy proposed in [20]). This model defines explicit mathematical

thresholds of divergence between the study cohort and national benchmark populations, establishing clear pass/fail criteria for dataset suitability in Clinical AI:

Scenario A (The Safe Zone): The cohort incidence and demographic distribution mirror the population benchmark within a minimal margin of error (<10%). The dataset is structurally anchored, selection bias is negligible, and findings exhibit high generalizability. Datasets in this zone pass structural validation.

Scenario B (The Cautionary Zone): Characterized by a moderate divergence (10% to 20%). While not a complete statistical outlier, the cohort acts as a restricted proxy rather than a universal baseline. Datasets falling into this zone require mandatory structural justifications, rigorous sensitivity analyses to prevent the direction-flipping of associations, and explicit labeling as cohort-specific.

Scenario C (The Danger Zone): The deviation exceeds the critical tolerance threshold (20% or higher). Deficits of this magnitude indicate a structural failure in representativeness. Relative risk estimates (such as Hazard Ratios) and low *p*-values obtained in this zone are flagged as potential artifacts of database volume rather than genuine biological signals. Consequently, datasets residing in this zone fail structural validation and are mathematically flagged as incompatible for reliable downstream Clinical AI training.

By formalizing these operational boundaries, our framework provides an objective auditing gate: empirical datasets must clear these quantitative constraints to ensure that subsequent computational modeling and automated inference remain anchored to population-level reality.

Along this line, it should be noticed that, while we acknowledge that minor discrepancies in micro-level censoring, follow-up duration, or precise case-ascertainment definitions can introduce baseline variance, our analytical framework is specifically calibrated to detect structural shifts of a macro-architectural scale. Divergence thresholds exceeding 20% vastly surpass the expected statistical noise introduced by minor methodological differences in ascertainment, thereby confirming that the observed rate differences reflect genuine structural selection drift rather than mere administrative or temporal artifacts.

As to all the temporal issues that are apparently disregarded in our treatment above, it needs to be reiterated instead that both the evaluated cohort and the national reference data are derived from official health and cancer registries centered on year 2022. In this sense, it is confirmed that all baseline eligibility criteria and temporal reference windows were harmonized at the macro-administrative level. This definitely ensures rigorous comparability between the two sources of investigated data.

## 4. Results

To ensure clarity, we begin by exposing in the following Table 2 all the results we have achieved working with the data of Table 1, using the unrounded underlying values, and employing the equations mentioned in Section 3.

We reiterate that we have contrasted the data from the large scale dataset of interest against the established 2022 national standards and that both the dataset and the national benchmark belong to South Korea. Detailed comments on the obtained results will follow Table 2 in the remainder of the present Section.

**Table 2.** Summary of Results.

| Age Group | Group | Weight (Data/Pop) | Abs. Diff. | Rel. Diff. | Dataset CR | Bench. CR | Suppr. (%) * |
|---|---|---|---|---|---|---|---|
| ≥65 years | NE | 12.2% vs. 18.0% | −5.8% | −32.5% | 85.2 | 155.2 | 45.1% |
| | E | | | | 113.5 | 155.2 | 26.9% |
| | Subtotal | | | | 107.9 | 155.2 | 30.5% |
| <65 years | NE | 87.8% vs. 82.0% | +5.8% | +7.0% | 26.3 | 33.0 | 20.4% |
| | E | | | | 32.8 | 33.0 | 0.7% |
| | Subtotal | | | | 31.5 | 33.0 | 4.6% |
| Total | Aggreg. | 100% vs. 100% | | | 40.8 | 55.0 | 25.9% |

* All numbers in this column are preceded by the minus sign.

To better illustrate the results comprised in Table 2, it is worth mentioning the following.

In the first column (Age Group): the age stratum of reference (Elderly-aged ≥ 65 years vs. Non-elderly-aged < 65 years) is specified. Similarly, in the second column (Group): the exposure status (NE, E, and Age-specific Subtotals) is reported.

In the third column (Weight Data/Pop.), the percentage distribution within the dataset versus the expected national population weight are displayed.

In the fourth column (Abs. Diff.), the absolute difference (−5.8%/+5.8%) is quantified between the dataset weight and the national population weight.

In the fifth column (Rel. Diff.): the demographic bias is expressed as a relative percentage of surplus or deficit.

In the sixth and seventh columns (Dataset/Bench. CR), the observed Crude Incidence Rates (CR) are reported as compared to the National Benchmark CR per 10,000 individuals.

In the final column (Suppr. %), the incidence suppression percentage is measured, revealing the systematic under-reporting of oncological events across strata.

Even if the systematic layout of Table 2 would allow for an immediate assessment of how the dataset's internal rates deviate from established national reality, serving as a visual foundation for the subsequent structural validation analysis, to better understand those results, the following considerations can be useful.

Specifically, we emphasize three critical methodological instabilities that, if left unaddressed, would render this dataset unsuitable for robust CAI training. These figures act as a diagnostic report of the dataset's structural adequacy for instructing an AI system.

In particular, the first result to take into account is the 45.1% incidence suppression in the elderly (aged $\geq$ 65 years) non-exposed group. This is the most alarming discrepancy. The fact that the non-exposed elderly (aged $\geq$ 65 years) cohort shows an incidence rate nearly half that of the national benchmark suggests a massive under-reporting or selective exclusion of high-risk cases. In a CAI pipeline, this missing signal would probably force the algorithm to learn an artificial, low-risk profile for elderly (aged $\geq$ 65 years) patients that does not exist in clinical reality.

The second relevant result is the 32.5% demographic deficit in the elderly (aged $\geq$ 65 years) population. The lack of the elderly in the dataset (12.2% vs. 18% expected) creates a critical issue. Since this stratum is statistically the most prone to oncological events, its under-representation could systematically skew a CAI's predictive capacity, effectively blinding it to the most vulnerable clinical demographics.

Finally, of relevance is also the almost 26% aggregate incidence suppression**.** Across the entire cohort, the dataset fails to capture over a quarter of expected oncological events. This aggregate suppression proves that the methodological instability is not limited to a specific subgroup but is systemic. Relying on such a dataset for Big Medical Inference would mean accepting an unstable reality where the CAI learns from a filtered, incomplete version of the disease landscape.

Together, these three indicators demonstrate that our framework has been able to capture the fact that the initial dataset suffered from a non-random structural instability. Moreover, all the parameters would position the data in the Zone C (very dangerous) to be used for training. By prioritizing these foundational checks, we move away from the illusion of Big Data volume, towards a model of Human-Centric CAI that demands structural truth before it allows AI/computational inference.

In the end, these findings suggest that even a massive dataset can yield questionable or, at the very least, unstable results if data extraction is performed inconsistently with the underlying baseline references. Such outcomes persist regardless of the data volume or the sophistication of the extraction models and algorithms employed, such as Cox models and PSM-style algorithms [21–24], emphasizing that structural validity is a fundamental prerequisite for reliable intelligent inference.

*Analytical Impact of Structural Divergence on Supervised Optimization*

To formally demonstrate how dataset architectural divergence propagates into downstream machine learning models, we propose the behavior of a standard supervised classifier trained via Empirical Risk Minimization (ERM).

For example, let a predictive model $f(x;\ \theta)$ be optimized by minimizing the expected loss over the training distribution following the well-known formula:

$$min_\theta \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}(y_i, f(x_i;\ \theta)). \quad (4)$$

When the underlying cohort suffers from a systematic structural distortion, such as the observed 32.5% demographic deficit in the high-risk elderly (aged $\geq$ 65 years) stratum and an aggregate incidence suppression of circa 26%, the empirical distribution $\hat{P}$ is severely biased. In gradient-based optimization (e.g., stochastic gradient descent), the parameter update at step $t$ is driven by the gradient of the loss computed over this distorted sample: $\theta_{t+1} = \theta_t - \eta\, \nabla_\theta\, \mathcal{L}_{\hat{P}}\, (\theta t)$. Because the high-risk elderly population (who carry the highest density of positive clinical outcomes) is under-represented by nearly a third, and total disease prevalence is artificially deflated, the gradient $\nabla_\theta\, \mathcal{L}_{\hat{P}}$ points toward an artifactual data manifold rather than the true population risk landscape $P$. Consequently, we yield all the following. First, the loss function under-penalizes errors on vulnerable sub-populations because their representation in the training manifold is artificially sparse. Second, the optimization trajectory converges to parameter values $\theta^*$ that optimize the filtered, incomplete data distribution, resulting in severe post-deployment miscalibration when the model encounters true, unsuppressed population baselines.

Thus, unverified datasets do not merely provide noisy inputs; they actively bias the gradient descent path, institutionalizing data-substrate distortions into the core weights of the clinical AI system.

## 5. Discussion

The primary strength of our auditing framework lies in its strictly quantitative, agnostic focus on fundamental demographic and oncological incidence metrics. By grounding our investigation in a purely structural analysis, contrasting an initial dataset against established national benchmarks, we demonstrate how selection bias in large-scale datasets can be empirically exposed. This methodological choice is deliberate: by isolating the mathematical architecture of the data, we transcend inconclusive biochemical or clinical debates [21], thereby grounding our conclusions in objective, verifiable reality.

Our findings on a specific but exemplar case have revealed a tripartite mathematical divergence that exposes the fragility of the examined dataset: (1) a 32.5% demographic deficit in the high-risk elderly (aged $\geq$ 65 years) strata; (2) an almost 26% suppression of the overall cancer incidence compared to national standards; and (3) a 45.1% deflation of expected cancer cases within the non-exposed group.

Beyond this specific descriptive result, a crucial critique of our framework concerns the absence of an empirical downstream retraining experiment demonstrating performance shifts before and after auditing. Nonetheless, while individual-level data restrictions prevent the re-training of proprietary deep learning architectures, the computational rationale for our pre-deployment gate rests on foundational machine learning theory. In safety-critical domains such as Clinical AI, post-hoc algorithmic fixes (such as re-weighting or threshold tuning) are mathematically insufficient if the training substrate suffers from macro-level structural distortion, such as the 32.5% demographic deficit and an almost 26% incidence suppression uncovered here. Consequently, our auditing framework is designed to function as an upstream computational acceptance gate. By establishing quantitative verification thresholds prior to model ingestion, the framework prevents corrupted empirical distributions from defining the optimization manifold, addressing the direct relationship between input data anomalies and downstream model behavior.

Beyond this specific result, our analysis highlights a critical intersection between technical rigor and the paradigm of Human-Centric Clinical AI. While scholars often prioritize internal validity through statistical tools like Propensity Score Matching (PSM), our evidence demonstrates that such techniques may be insufficient when external validity is fundamentally compromised [21–25]. If a dataset fails to mirror the demographic and clinical reality of the target population, algorithms, regardless of their complexity, merely scale these initial architectural instabilities [26–28]. In the context of our examined case, this has translated into institutionalizing historical biases, where the AI does not discover new clinical insights but rather automates and obscures pre-existing data deficiencies.

We acknowledge the major limitation of our study: it centers on just an exemplar case study and relies on re-analysis of aggregate data, lacking individual-level granularity. However, we reiterate that this has served as a proof-of-concept where nonetheless the sheer magnitude of the observed discrepancies, particularly the asymmetry of missing cancer cases, was too profound to be dismissed as mere statistical noise. These findings suggest that when methodological rigor is sacrificed for the convenience of Big Data volume, we risk an unstable view of clinical reality.

More specifically, limitations regarding granularity should be acknowledged. Due to reliance on aggregated reporting standards, our current implementation utilizes macro-level strata rather than fine-grained subdivisions of age, sex, regional distribution, comorbidities, or socioeconomic status. While unmeasured residual confounding across these micro-dimensions can introduce local variance in crude rates, our structural auditing model is intentionally calibrated to detect systemic macro-failures. We emphasize that localized confounding cannot mathematically explain population-wide incidence suppressions exceeding 20%. Nonetheless, future extensions of this framework should integrate multi-dimensional demographic and healthcare utilization indices to map structural drift with heightened resolution.

Further, our study serves to advance a broader principle: preliminary validity checks on medical data are not elective, but an ethical imperative. In an era where AI-driven policies impact human lives irreversibly, protecting data integrity becomes a profound social and humanistic duty. We must move beyond the illusion that massive record counts guarantee objectivity; instead, we propose that structural validation (verifying demographic representativeness and baseline case distribution) be established as a mandatory prerequisite for the reporting of any observational medical evidence.

Moreover, without any intention to contest specific studies, our analysis has served as a call for a paradigm shift in how we approach Clinical AI. The intelligence of an automated system remains strictly subordinate to the structural integrity of its constituent sub-groups. By re-establishing this hierarchy of validation, we ensure that the big in Big Medical Inference is leveraged to enhance clinical depth, thereby protecting the integrity of public health strategies and ensuring that future interventions remain grounded in authentic biological truth, rather than in the outcomes of an unstable data organization.

As a last comment regarding possible limitations regarding framework reproducibility and validation breadth, we emphasize that our structural audit model is designed as a foundational, dataset-agnostic screening tool rather

than an exhaustive multi-center epidemiological catalog. The independent mathematical verification of our metrics, as evidenced by the precise audit of our comparative thresholds, directly attests to the high reproducibility and transparency of the proposed methodology. Because the framework relies on macro-level benchmark invariants, its auditing logic remains universally applicable: any structural selection drift or representativeness failure will trigger the same quantitative boundaries regardless of the specific clinical domain. While future large-scale implementations will naturally expand across diverse repositories and benchmark downstream AI performance degradation, the current framework successfully establishes the mandatory auditing standard required prior to clinical model training.

In conclusion, we would also like to reiterate that our proposed framework strictly belongs to the domain of data pre-processing and structural dataset auditing rather than model training. To convince about this, one can look at the typical architectural separation in machine learning pipelines: pre-training gate vs. in-training optimization. Our proposed framework operates strictly upstream of any learning algorithm. In fact, first, it functions as an acceptance/rejection filter applied to the raw data distribution $\hat{P}(x, y)$ before the optimization process begins. It does not modify loss functions, alter model architectures (such as neural network layers), or tune hyper-parameters ($\Theta$). Second, the metrics used (demographic representation weights, crude incidence rates, and benchmark comparisons) evaluate the physical and architectural integrity of the cohort itself (e.g., detecting the 32.5% elderly deficit). This is a purely epidemiological and data-engineering task. Third, just as input sanitization and schema validation ensure that corrupted strings do not enter a database (occurring before business logic execution), our framework ensures that structurally questionable populations do not enter the computational training pipeline. Thus, it is fundamentally a data pre-processing validation gate.

**Funding**

This research received no external funding.

**Institutional Review Board Statement**

No humans, animals and private data are involved in this study. This study is a secondary analysis of already published data in an aggregated form all available within the article and from the cited References. Therefore ethical permission is not required.

**Informed Consent Statement**

This study did not require informed patient consent, as it relies exclusively on pre-existing, anonymized, and fully aggregated data. No individual human subjects were involved, and the data utilized do not contain any personally identifiable information.

**Data Availability Statement**

Data supporting the findings of this study are all available within the article and from the cited References. Supplemental queries can be addressed to the author: marco.roccetti@unibo.it.

**Conflicts of Interest**

The author declares no conflict of interest.

**Use of AI and AI-Assisted Technologies**

No AI tools were utilized for this paper.

**References**

1. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
2. Roccetti, M. Quantifying structural selection bias in observational cohort data: a ponderation analysis of age—specific incidence rates to inform vaccine safety verification. *Front. Pharmacol.* **2026**, *16*, 1754809. https://doi.org/10.3389/fphar.2025.1754809.
3. Roccetti, M.; Delnevo, G.; Casini, L.; et al. Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *J. Big Data* **2019**, *6*, 70. https://doi.org/10.1186/s40537-019-0235-y.

4. Kim, H.J.; Kim, M.H.; Choi, M.G.; et al. 1-year risks of cancers associated with COVID-19 vaccination: A large population-based cohort study in South Korea. *Biomark. Res.* **2025**, *13*, 114. https://doi.org/10.1186/s40364-025-00831-w.
5. Fisher, R. Studies in Crop Variation (I). An Examination of the Yield of Dressed Grain from Broadbalk. *J. Agric. Sci.* **1921**, *11*, 107–135. https://doi.org/10.1017/s0021859600003750.
6. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. https://doi.org/10.1080/14786440009463897.
7. Neyman, J.; Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. A* **1933**, *231*, 289–337. https://doi.org/10.1098/rsta.1933.0009.
8. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. B* **1972**, *34*, 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
9. Wilson, R. *Feminine Forever*; M. Evans and Company: New York, NY, USA, 1966.
10. Anderson, G.; Limacher, M.; Assaf, A.; et al. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: The Women's Health Initiative randomized controlled trial. *JAMA* **2004**, *291*, 1701–1712. https://doi.org/10.1001/jama.291.14.1701.
11. Cagnacci, A.; Venier, M. The Controversial History of Hormone Replacement Therapy. *Medicina* **2019**, *55*, 602. https://doi.org/10.3390/medicina55090602.
12. Deer, B. How the case against the MMR vaccine was fixed. *BMJ* **2011**, *342*, c5347. https://doi.org/10.1136/bmj.c5347.
13. Spitzer, W.; Suissa, S.; Ernst, P.; et al. The use of beta-agonists and the risk of death and near death from asthma. *N. Engl. J. Med.* **1992**, *326*, 501–506. https://doi.org/10.1056/nejm199202203260801.
14. Roccetti, M. Enhancing public health surveillance: A statistical validation of potential sampling bias in large retrospective vaccine cohorts. *AIMS Public Health* **2026**, *13*, 589–597. https://doi.org/10.3934/publichealth.2026031.
15. Kang, M.; Jung, K.; Bang, S.; et al. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2020. *Cancer Res. Treat.* **2023**, *55*, 385–399. https://doi.org/10.4143/crt.2023.447.
16. Park, E.; Jung, K.W.; Park, N.; et al. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2021. *Cancer Res. Treat.* **2024**, *56*, 357–371. https://doi.org/10.4143/crt.2024.253.
17. Park, E.; Jung, K.W.; Park, N.; et al. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2022. *Cancer Res. Treat.* **2025**, *57*, 312–330. https://doi.org/10.4143/crt.2025.264.
18. Statista. Crude Incidence Rate of Cancer in South Korea in 2022, by Age Group (Per 100,000 Population). Available online: https://www.statista.com/statistics/1440818/south-korea-cancer-crude-incidence-rate-by-age/ (accessed on 20 May 2026).
19. World Bank. Population Ages 65 and Above (% of Total Population)—Korea, Rep. Available online: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?locations=KR (accessed on 20 May 2026).
20. Roccetti, M. Unlocking the stochastic parrot: Epistemic obligation and the decline of biological plausibility in clinical reality. *Inform. Med. Unlocked* **2026**, *62*, 101752. https://doi.org/10.1016/j.imu.2026.101752.
21. Chemaitelly, H.; Ayoub, H.H.; Coyle, P.; et al. Assessing healthy vaccinee effect in COVID-19 vaccine effectiveness studies: a national cohort study in Qatar. *eLife* **2025**, *14*, e103690. https://doi.org/10.7554/elife.103690.
22. D'Agostino, R.B. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **1998**, *17*, 2265–2281. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B.
23. Qin, R.; Titler, M.G.; Shever, L.L.; et al. Estimating effects of nursing intervention via propensity score analysis. *Nurs. Res.* **2008**, *57*, 444–452. https://doi.org/10.1097/nnr.0b013e31818c66f6.
24. Austin, P.; Grootendorst, P.; Anderson, G. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Stat. Med.* **2007**, *26*, 734–753. https://doi.org/10.1002/sim.2580.
25. Austin, P.C. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiol. Drug Saf.* **2008**, *17*, 1202–1217. https://doi.org/10.1002/pds.1673.
26. von Elm, E.; Altman, D.; Egger, M.; et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *PLoS Med* **2007**, *4*, e296. https://doi.org/10.1371/journal.pmed.0040296.
27. Meng, X.L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **2018**, *12*, 685–726. https://doi.org/10.1214/18-aoas1161sf.
28. Guo, S.; Wu, D.O. Game Theoretical AI for Precision Medicine. *Trans. Artif. Intell.* **2025**, *1*, 170–196. https://doi.org/10.53941/tai.2025.100011.