Informatics in Medicine Unlocked 62 (2026) 101752

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

# Unlocking the stochastic parrot: Epistemic obligation and the decline of biological plausibility in clinical reality

Marco Roccetti

*Department of Computer Science and Engineering, University of Bologna, Via Mura A. Zamboni 7, 40127, Bologna, Italy*

ARTICLE INFO

*Keywords:*
Big medical data
Epistemic injustice
Biological plausibility
Medical AI
Algorithmic auditing

ABSTRACT

*Purpose:* The current reliance on medical Big Data is producing a *mostro ibrido* (monstrous hybrid) of statistical significance that lacks biological anchoring. This investigation builds upon the framework of epistemic injustice and epistemic obligation from (Herzog and Branford, 2025) to analyze the systemic marginalization of human expertise in computational medicine.

*Approach:* This work performs a critical audit of three high-profile studies where algorithmic outputs generated from large-scale databases stand in direct contradiction to national gold standards and established clinical benchmarks.

*Results:* The analysis reveals profound discrepancies, including the disappearance of approximately 50,000 cancer cases per year in a nationwide database and the creation of statistically significant but biologically impossible correlations. These findings demonstrate that: a) the epistemic obligation toward algorithms compels practitioners to ignore clinical common sense in favor of bureaucratized data anomalies, b) medical AI often operates as a stochastic parrot, capable of recombining vast amounts of data into seemingly coherent patterns that possess no actual contact with underlying pathological truths.

*Conclusion:* This investigation denounces the definitive farewell to the proof of causality in favor of a mechanized version of medical truth. It highlights the urgent necessity to reclaim epistemic agency, arguing that biological reality cannot be fully replaced by a database, regardless of its scale. We must prevent the statistical ghost of a database from replacing the living patient in the clinical decision-making process.

## 1. Introduction: beyond relational ethics in computational medicine

In their recent contribution, Herzog and Branford [1] provide a compelling relational ethics approach to the deployment of Artificial Intelligence in medicine, identifying a critical epistemic obligation that compels practitioners to defer to algorithmic outputs. They argue that *"AI-based decision* support *risks excluding patients and medical personnel from relevant epistemic processes vital to good medical practice"*. While we fully resonate with their ethical concerns, this article contends that the crisis is not merely relational but deeply rooted in the technical and statistical foundations of current Big medical Data. Our intervention is intended to move beyond philosophical insights and investigate the hard, and often contradictory, reality of computational medicine.

We contend that the epistemic exclusion identified by Herzog and Branford is rooted in a epistemological deeper crisis: the transformation of statistical inference into a black box, a mechanized guillotine that severs the connection between data and clinical meaning. Rather than proposing a purely technical fix, this study advocates for a return to auditing through clinical plausibility, where a statistical coefficient is no longer allowed to override the pulse of the patient and the specific clinical references that constitute the historical and empirical record of human pathology. By examining the divergence between algorithmic outputs and clinical best practices, this investigation seeks to expose how the erosion of biological plausibility leads to an algorithmic reduction of pathological truth.

In this context, the primary objective of this paper is to validate the clinical evidence hidden behind statistical anomalies. To achieve this, we employ a methodology of rigorous auditing, applying it to three significant case studies where algorithmic outputs from large-scale national databases are contrasted with established clinical gold standards. Through this comparative analysis, we aim to question if the algorithmic degradation of clinical evidence occurred, where the administrative representation of the patient effectively has replaced the biological reality (See Fig. 1).

In pursuing this critique, we wish to state clearly that the examples

*E-mail address:* marco.roccetti@unibo.it.

https://doi.org/10.1016/j.imu.2026.101752
Received 6 January 2026; Received in revised form 26 February 2026; Accepted 4 April 2026
Available online 4 April 2026

selected for this investigation are not intended as an attack on the specific authors, their scientific commitment, or the journals that hosted their work. We maintain the utmost professional deference toward these colleagues and institutions. Our objective is strictly methodological: we use these high-profile cases to demonstrate how even vast amounts of data, if not correctly organized and anchored to clinical reality, can become detrimental to medical practice and decision-making. Our aim is to unlock the clinical truth hidden behind statistical anomalies, offering a path toward a more robust integration of informatics and human expertise.

The remainder of this paper is organized as follows: Section 2 explores the historical and statistical roots of the friction between p-value culture and biological significance. Section 3 presents the detailed audit of our three case studies, highlighting the divergence between digital cohorts and national benchmarks or clinical plausibility, as well. Section 4 discusses the necessity of reclaiming epistemic agency in the age of AI. Finally, Section 5 offers concluding remarks on the future of medical truth in the digital era.

## 2. Background: the mostro ibrido and the fury of the new

To understand this structural injustice, we must recognize that modern science has birthed a mostro ibrido (better said monstrous hybrid). In 1925, when Ronald Fisher [2] published *Statistical Methods for Research Workers*, he did not intend the p-value to be an on/off proof of truth. For Fisher, the p-value was an *index of surprise.* If one obtained $p < 0.05$, the signal was: *"This result is unusual enough to suggest that something interesting is happening here"*. It was never a verdict; it was an invitation to replicate the experiment, to reproduce the results elsewhere or otherwise, a signpost suggesting that the matter deserved further investigation. Statistical significance was a gateway to reproducibility, not a final destination. If a result is not reproducible, that initial $p < 0.05$ is merely a stroke of luck, just a Type I error.

This investigation contends that the modern chaos stems from the forced fusion of two rival schools of thought. While Fisher viewed the p-value as a continuous measure of evidence, Neyman and Pearson introduced a rigid, binary decision-threshold (alpha) designed for mechanical accept/reject choices [3]. Today, researchers (and scientific journal editors, as well) live within this hybrid monster: they perform Fisher's calculations but apply Neyman-Pearson's mechanical guillotine. This has generated a *p-value culture* where the difference between publishing in *The Lancet* or discarding a year of work lies in the trivial gap between a p-value of 0.049 and 0.051.

This cultural shift has inverted the Popperian ideal of science [4]. For Karl Popper, the scientist's duty was to attempt to tear down their own theories, to falsify them. Today's researcher, armed with Big Data, no longer asks, *"Can I prove this idea wrong?"* but rather "*How much data must I add until I am shown to be right?"* We have moved from falsification to data-dredging, where the goal is to find that tiny fragment of signal that allows one to shout: *"εὕρηκα (Heúrēka): I found it!"*

In the era of Big Data, rejecting the null hypothesis has become trivial. When the sample size (N) is enormous, even an irrelevant difference or a systematic error can produce a statistically significant p-value [5]. We have confused statistical significance (a mathematical calculation) with scientific and biological significance, i.e., the reality of the facts. This study argues that this surrender to the *furia del nuovo* (better said, fury of the new) and scientific neophily is driven by a system where academic journals and media outlets prioritize the discovery of new, because that is where the business lies. A paper stating, *"We checked and found no difference between the groups (the null hypothesis holds)"* is often rejected as uninformative. Conversely, a paper claiming, *"We found a link between X and Y"*, even if biologically impossible or based on inconsistent data or not easily reproducible, is rushed to the cover of prestigious journals.

This represents the ultimate epistemic injustice analyzed in this work: rewarding the mirage of the new, while punishing the rigor of clinical reality. This phenomenon is fueled by a systemic obsession with novel associations, and farewell to the proof of causality, that attract high-impact publications. As the sample size grows to millions, the index of surprise vanishes; every minor noise becomes a significant signal. The mechanized guillotine cuts through the complexity of the patient's life-world, replacing it with a mathematical significance that often masks a biological absurdity. We have traded the Popperian ideal of falsification [4] for a culture of data-dredging, where a black box filled of data is interrogated until it confesses a narrative that satisfies the industry's neophilia. This Big Data mirage creates the illusion of certainty through sheer volume, masking the technical depths of its internal contradictions. When, for example, the database ignores national disease registries or clinical benchmarks, it creates a reality that exists only inside the server (or the sample of data collected for that specific retrospective analysis). This leads to a kind of solitude of the clinician. Practitioners are forced into an epistemic obligation to trust a machine that claims an

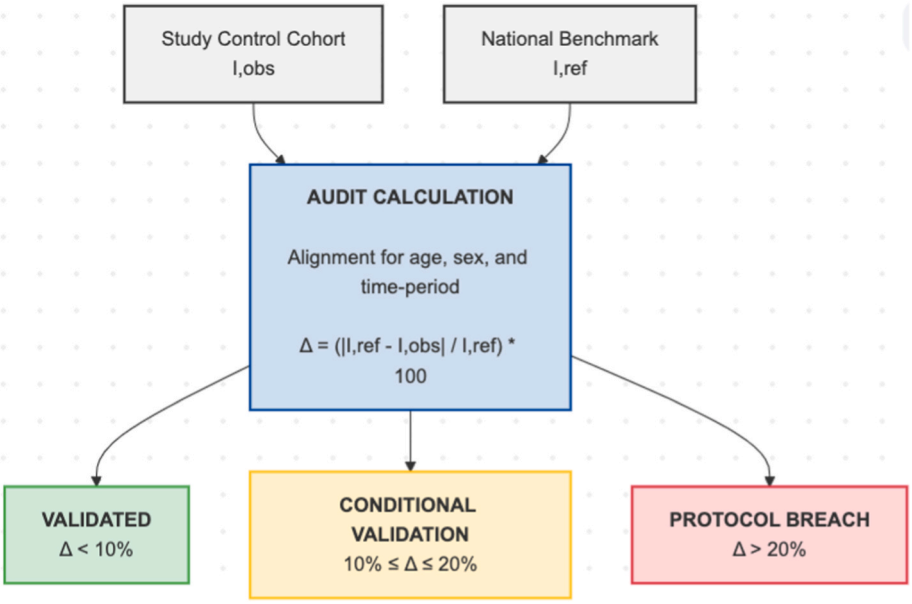


**Fig. 1.** Practical workflow.

over 70% reduction in a disease, even when their own eyes and the historical record say otherwise. This obligation is not merely psychological but fundamentally ethical: as suggested by the relational ethics framework of Herzog and Branford, the clinician feels a duty of deference toward the perceived superior objectivity of the large-scale model. To challenge a statistically significant p-value ($p < 0.05$) derived from millions of records feels like an act of epistemic injustice against the truth of the machine, even when that truth is biologically implausible. Truth is no longer verified against the human body, but against the internal consistency of a (unstable) database. In this epidemiological representation, the formal staging of algorithmic rigor creates a symbolic authority that effectively overrides clinical judgment.

To provide concrete weight to these claims, this present article utilizes a method based on the critical investigation of three recent cases where the algorithmic machine, relying solely on medical Big Data, has extrapolated clinical conclusions that stand in direct contradiction to reality. By reality, we mean here the confrontation with gold standards, best practices, and clinical common sense, the neglect of which has caused distortions in media narratives and hesitation among decision-makers.

In this context, it is crucial to specify that while gold standards and good practices may evolve and take different forms over time, the diagnostic anchor of our investigation remains strictly grounded in the comparison of incidence and prevalence rates of investigated diseases. In epidemiology, in fact, these parameters are not mere administrative labels; they represent the most reliable quantitative account of the actual presence and burden of disease within a living population. They are the very vital signs of public health.

Therefore, any audit of algorithmic findings, which are often epistemically unjust when derived from unrepresentative digital cohorts, must begin by measuring the *fever* of the population. By contrasting the statistical inferences of a specific cohort with the established national epidemiological benchmarks, we can determine if the digital representation maintains a pulse of biological reality or if it has drifted into an unanchored data anomaly. This quantitative confrontation provides the empirical foundation for the auditing method proposed in the following Section.

## 3. Methods: the population-anchored auditing framework

The transition from data-rich to knowledge-rich research requires thus a shift in how we define methodological innovation. We argue that the next frontier of medical innovation is not a more complex algorithm, but a more robust auditing culture. Innovation without accountability is merely a digital staging of a kind of epidemiological representation. To restore the promise of Big Data, we must introduce tools that force a dialogue between the digital artifact and the biological reality.

The proposal that will be presented here is not to be merely seen a statistical check; it is a conceptual innovation designed to protect the integrity of the scientific narrative from the seductive traps of unvalidated large-scale data.

Hence, with the practical intention to build upon the necessity to measure the *fever* of a population, we start with a kind of framework that transitions from qualitative reporting to empirical verification by mandating a formal dialogue between a given Big data-based study's internal results and the external epidemiological ground truth.

This method could be structured as a two-stage verification process that acts as a gatekeeper for clinical truth. We should start from the realistic consideration that the core of the problem we are facing lies in the relationship between the observed incidence within a given study's control group, *I(obs)*, and the established national epidemiological incidence benchmarks, *I(ref)*, which represents the ground truth for the general population.

If the statistic of a given control group (which will constitute the baseline of all the consecutive associations and inferences) is deflated (that is $I(obs) \ll I(ref)$) due to diagnostic under-reporting, or inflated (that is $I(obs) \gg I(ref)$) due to selection bias, the resulting ratio will mathematically predetermined to show an effect, regardless of any actual biological mechanism.

Hence, the core of a possible proposal lies in quantifying the distance between the digital representation of a patient (trough the big data of a given cohort) and the biological reality of the citizen (the national population). This could be formalized through the calculation of a kind of Relative Deviation Index $\delta$ which serves as a clinical thermometer for data integrity.

More specifically, given a study's control or treated cohort, the observed incidence (or prevalence) of a clinical outcome *I(obs)* should be harmonized with the authoritative national benchmark for the same period *I(ref)*, which represents the ground truth. The deviation is calculated as follows: $\delta = |I(ref) - I(obs)|/I(ref)$ x 100.

This index $\delta$ would measures the magnitude of the epidemiological drift. In a healthy and representative study, $\delta$ should remain close to zero, accounting only for minor stochastic noise. When $\delta$ grows, it indicates that the study is no longer observing the population, but rather a "statistical ghost" manufactured by selection bias or data-cleaning artifacts.

Obviously we should not be very strict with the numerical instantiation of the value of $\delta$, in the sense that in turn we could imagine $\delta = O(\varepsilon)$, where $\varepsilon$ represents a small, acceptable threshold of tolerance (typically 0.05 or 5%). Using the *O(.)* notation only emphasizes that we are just constraining the order of magnitude of divergence, not just a single numeric instance When $\delta$ significantly exceeds this order of magnitude, the cohort would exhibit a systematic divergence that invalidates population-level inferences.

Nonetheless, to operationalize this audit for reviewers and clinicians, we could establish a formal hierarchy of evidence based on the value of $\delta$. Rather than a binary valid/invalid choice, a kind of traffic light system could be adopted that weights the generalizability of the findings based on the following scenarios:

- Scenario A: $\delta \leq \varepsilon$ (The Safe Zone). The cohort incidence mirrors the population benchmark within a minimal margin of error (e.g., <10%). The cohort is reasonably representative. Findings are highly generalizable and the risk of selection bias affecting the final estimates is negligible. In simple words, the data is anchored; findings are highly generalizable.
- Scenario B: $\varepsilon < \delta < 0.20$ (The Cautionary Zone). There is a moderate divergence (e.g., 10% to 20%). The cohort is not a perfect proxy, but it is not a statistical outlier. Hence, researchers just need to provide a transparent justification for the divergence (e.g., specific age-group recruitment). Sensitivity analyses can be useful to ensure that this moderate drift does not flip the direction of the reported associations. In essence, he findings must be labeled as cohort-specific. A deeper audit of exclusion criteria and specific demographic recruitment would be mandatory.
- Scenario C: $\delta \gg \varepsilon$ (The Danger Zone). The deviation is an order of magnitude higher than the tolerance threshold. We are here in presence of a critical inferential overreach. Deficits of the order of 20/30/40% or more indicate a structural failure in representativeness. The cohort is systematically different from the target population, rendering relative risk estimates (like Hazard Ratios) likely artifacts of sampling rather than biological signals. Not only that, but any statistically significant p-value obtained in this zone is likely an inferential overreach, a signal generated by the volume of the database rather than the truth of the pathology.

It should be clearly intended the indicative, non exhaustive, and non-normative nature of this proposal. Rather than a rigid bureaucratic rule, it represents a possible operational model to transform an epistemic alarm into a practical clinical audit. We explicitly state that while other, more technically sophisticated solutions for data validation are certainly possible, they remain out of the scope of our current investigation. Our

goal is to provide a robust, common-sense gateway to detect structural failures (Scenario C) where the digital artifact has definitively lost its biological pulse.

To further summarize what is proposed, here, along the line of giving a practical sense to our discourse, it could be useful the following pictorial representation.

## 4. Results: inside the black box vs. outside in the world

To provide concrete weight to our claims, this article presents a critical investigation of three recent cases where the algorithmic machine, relying solely on medical Big Data, has extrapolated clinical conclusions that stand in direct contradiction to reality. By reality, we mean here the confrontation with gold standards, best practices, and clinical common sense, the neglect of which has caused distortions in media narratives and hesitation among decision-makers.

The following analysis demonstrates that what is true inside the black box is not necessarily true outside of it, and requires rigorous empirical demonstration. In doing so, we intend no disrespect toward the authors of these works, whose scientific commitment we fully respect; rather, we aim to expose the structural discrepancies that emerge when statistical power is decoupled from biological plausibility.

In this sense, these specific cases must be considered highly significant, as they represent high-profile examples currently at the center of international scientific debate. While this selection does not aim to be exhaustive, an impossible task given the scale of medical Big Data, these studies serve as emblematic archetypes of the epistemic risks inherent in unanchored digital cohorts.

### 4.1. Case A: The epistemic exclusion or the unrecorded population

This study identifies a profound instance of epistemic exclusion, a concept most evident when a black box produces findings that are biologically impossible, yet are accepted due to their sheer statistical power. This is precisely the case in a recent analysis of a South Korean database regarding cancer incidence following COVID-19 vaccination [6].

Using nearly 3 million records, the study suggests a rising cancer risk among the vaccinated. However, a fundamental check against the national gold standards reveals a significant discrepancy. One simply needs to compare the raw cancer incidence rate (CR) of 40.78 per 10,000 individuals reported in the study [6] with the official CRs of the South Korean population (averaging 52.46 per 10,000 for the 2020–2022 period) [7–9].

Applying our $\delta$ audit framework, the study's overall cohort deviates downwards by 22% ($|40.78 - 52.46|/52.46$). Even more critical is the analysis of the unvaccinated control group, which recorded an incidence of only 33.43 per 10,000 [6], resulting in a $\delta$ value of 36% ($|33.43 - 52.46|/52.46$). According to our analysis, a deviation between 22% and 36% places the study firmly beyond the cautionary zone [10].

The clinical implications of this statistical anomaly are profound. The data would paradoxically imply that vaccination protected people by making approximately 50,000 cancer cases per year simply remain unrecorded in the dataset, given a population of circa 52 million. This biological impossibility has been recently scrutinized and confirmed also in Ref. [11], which emphasizes how such unanchored datasets lead to clinical irrelevance.

In simpler words, the study's population was administratively filtered to the point where it no longer represented the real nation. This represents a form of hermeneutical injustice: using the authority of millions to mask a sample that defies the biological reality of an entire country. Inside the black box, 50,000 patients disappeared; outside in the world, they continue to require care. To mistake this administrative accounting for a pathological discovery is to succumb to a mechanized version of truth that excludes the living patient from the epistemic process.

### 4.2. Case B: The epistemic obligation or erasing the peaks of reality

This second case explores the epistemic obligation that compels practitioners to prioritize algorithmic outputs over their own clinical expertise. This obligation becomes a trap when the signal generated by the machine is a byproduct of structural bias rather than biological truth.

A prime example is the reported link between COVID-19 vaccination and Vitiligo [12], which claims an augmented risk among vaccinated individuals compared to the non-vaccinated (Hazard Ratio = 2.714), utilizing millions of health records. However, a rigorous auditing of the data reveals a systematic overestimation driven by annualization errors and age-bracket discrepancies. The original study reported results on a quarterly basis (0.67 for unvaccinated and 2.2 for vaccinated, per 10,000 individuals), which should be annualized to allow for a direct comparison with national epidemiological benchmarks, yielding respectively 2,68 and 8.88 per 10,000. Furthermore, the contribution of the population bracket under 20 years of age was omitted in the original cohort, even if essential for establishing the true national gold standard for vitiligo. Readjusting the national incidence rate without that age bracket yields 1.96 per 10,000 [13].

When we apply our based $\delta$ audit framework to these adjusted figures, the statistical implausibility becomes undeniable. In fact, of 8.88 per 10,000 in the vaccinated group can be contrasted with the weighted national gold standard which is only 1.96 per 10,000. By comparing these two values, we obtain a relative deviation of 3.53 (calculated as $|8.88 - 1.96|/1.96$). This means a $\delta$ value of 353% which places the starting data in a state of extreme danger as to extendibility of the relative findings.

Such an extreme deviation indicates that the vaccinated cohort captured an implausible cluster of cases, likely due to selection artifacts or errors in data annualization. This is further compounded by a demographic bias: the vaccinated cohort was 11 years older than the non-vaccinated group. Since vitiligo incidence peaks after age 50, this age gap artificially concentrated cases in the vaccinated group, while the unvaccinated group's incidence was recorded at a mere 2.68 per 10,000 (annualized), still disconnected from the national baseline.

Detailed auditing of this discrepancy is now formally confirmed by the work [14]. In closing, under the epistemic obligation to trust the p-value, the clinician would be forced to accept a signal that is not a biological reality, but a demographic anomaly created by excluding an excessive number of cases from the baseline and failing to anchor the digital cohort to the weighted reality of the national population.

### 4.3. Case C: The erosion of epistemic agency or the implausible miracle

This third case exemplifies what Herzog and Branford [1] describe as the erosion of epistemic agency. When an algorithmic machine produces a result that contradicts the foundational principles of human pathology, the clinician's ability to intervene as a rational agent is stifled by the overwhelming symbolic authority of the database. It is the paradox of an algorithmic truth that becomes unassailable simply because it is supported by millions of records, even when it is biologically absurd.

The case in point involves a high-profile retrospective study on a vast neuropsychiatric cohort [15], which reports an incredible over 70% reduction in schizophrenia risk (HR = 0.231) following COVID-19 vaccination. This statistical miracle defies the stable epidemiological dynamics of chronic psychiatric disorders. A deeper analysis reveals the core of this artifact: the vaccinated group was, on average, 10.5 years older than the unvaccinated control group (which served as the reference). In any real-world population, an older cohort would naturally be expected to show stable or slightly higher incidence rates for chronic psychiatric conditions, yet here the signal suggests a massive protective effect.

To audit this result, we apply our auditing protocol based on $\delta$. The original study reports a quarterly incidence rate for the vaccinated group

of 0.51 per 10,000, while the historical national benchmark from the National Health Insurance Service (NHIS) stands at 2.07 per 10,000 [16]. By comparing these two values, we obtain a relative deviation of 0.75 (calculated as $|0.51 - 2.07|/2.07$). A $\delta$ value of 75% projects this study directly into a "danger zone", signaling a plausible violation of biological reality.

This discrepancy, now formally documented and confirmed by recent analysis [17], demonstrates that the algorithm perceived protection simply because the database, through administrative exclusion criteria or misclassification errors, effectively stopped recording real-world patients. Accepting this statistical miracle requires surrendering one's epistemic agency to a bureaucratization of truth that has lost all contact with the clinical reality of mental health.

It is fundamental to reiterate that our analysis takes no stance for or against vaccination; the deliberate selection of examples showing both perceived harms and incredible benefits serves to demonstrate our scientific neutrality. Our thesis is strictly methodological: the digitalization of clinical practice, when inflated by uncurated databases, replaces pathological truth with administrative accounting, where truth is verified against the internal consistency of a server rather than the biological reality of the human body.

This phenomenon of biological implausibility can be formally understood through a novel extension of the *stochastic parrot* metaphor, originally coined in Ref. [18] to describe how large language models (LLMs) mirror linguistic patterns without grasping underlying meaning. We argue that our Cases have illustrated a similar mechanism of statistical over-inference occurs in large-scale predictive medical models.

The boundary between linguistic regurgitation and epidemiological over-inference lies in the nature of the tokens processed: while an LLM recombines words based on probability, an uncurated medical database recombines administrative tokens, such as medical codes and billing data, without anchoring them to the biological constraints of human pathology. Just as a stochastic parrot generates a grammatically perfect but factually impossible sentence, an uncurated epidemiological model can produce a statistically 'significant' result (e.g., the 75% reduction in schizophrenia) that is clinically non-existent. In both cases, the failure stems from a decoupling of the model from the physical reality it is supposed to represent. Our analysis is specifically designed to detect this decoupling, serving as a reality check against the tendency of large-scale models to prioritize internal statistical consistency over external biological truth.

## 5. Discussion: reclaiming common sense in the age of AI and big data

The evidence presented in this investigation leads to a definitive conclusion: to move from epistemic exclusion to epistemic inclusion, we must treat medical AI as a mediator, not a master. Our present study demonstrates that opening the black box does not require complex new bureaucracies, but rather a return to the best practices of auditing algorithmic outputs against gold standards and the plausibility of clinical experience.

We have argued throughout this work that the scientific community must have the courage to declare that a study is not better simply because its database is larger. Size does not excuse a violation of biological plausibility. By confronting the math with the same rigor we apply to ethical principles, we reclaim the clinician's role in the meaning-making process.

We acknowledge certain limitations in our study. As a critical investigation based on case-audit, this work does not aim to provide a comprehensive statistical re-analysis of all existing medical Big Data, nor does it dismiss the undeniable potential of large-scale datasets in medicine for retrospective studies. Rather, its scope is limited to exposing specific, high-profile instances where the decoupling of statistical significance from biological plausibility has led to evident clinical paradoxes. Furthermore, while we rely on national gold standards as a benchmark for truth, we recognize that these registries themselves are subject to reporting delays and bureaucratic constraints. However, the magnitude of the discrepancies identified, such as the vanishing of entire patient populations, suggests a structural instability that transcends simple reporting lag.

Nonetheless, our analysis highlights the pervasive, everyday risk of becoming lost in the labyrinth of medical Big Data, where countless inferences are algorithmically deduced but remain unvalidated by a confrontation with reality [19]. We must remain vigilant in recognizing that biological reality does not yet, and perhaps never will, coincide 100% with a database, or even a specific fragment of it, regardless of its scale. To mistake the digital sample for the clinical whole is to succumb to an anomaly that replaces the living patient with a statistical parrot.

Finally, we contend and reiterate that the international scientific community must demand that these databases be made public and accessible for independent auditing. Only through such transparency can we transform the black box of Big Data from a tool of statistical over-inference into a shared resource for human flourishing, grounded in the common sense of clinical reality. Reclaiming epistemic agency is not just a philosophical necessity; it is a clinical obligation to prevent the definitive erosion of medical truth.

## 6. Conclusion: from statistical mirage to epistemic integrity

This investigation has demonstrated that the current crisis in computational medicine is, at its core, a crisis of biological anchoring. By auditing three high-profile cases, we have exposed how the epistemic obligation toward Big Data has allowed a bureaucratized truth to supersede the historical and empirical reality of human pathology.

Our analysis of the investigated datasets reveals that the sheer scale of modern databases acts as a double-edged sword: while it provides unprecedented statistical power, it simultaneously creates a black box where entire patient populations can bureaucratically vanish, and where biological miracles, such as an over 75% reduction in schizophrenia, are manufactured by data-dredging rather than clinical reality. This represents the ultimate epistemic injustice: the systematic marginalization of human expertise in favor of an algorithm that, despite its millions of records, functions as a stochastic parrot devoid of clinical common sense.

We conclude that the proof of causality cannot be replaced by the fury of the new or the index of surprise provided by the p-value. To reclaim epistemic agency, the scientific community must enforce a return to biological plausibility as the final arbiter of truth. A statistical coefficient must never again be allowed to override the national gold standards and the lived experience of clinicians.

Finally, the restoration of medical truth demands transparency. We reiterate our call for the public accessibility of the underlying databases to allow for independent international auditing. Only by grounding our computational tools in the common sense of clinical reality can we prevent the definitive erosion of medicine into a mere branch of administrative accounting. In closing, it is crucial to posit that, while our analysis has utilized high-profile COVID-19 case studies, it is crucial to recognize that these examples serve as an extraordinary stress-test for the entire ecosystem of medical AI. The unprecedented scale of pharmacological surveillance during the pandemic simply acted as a magnifying glass, exposing pre-existing structural instabilities in how large-scale databases are curated and audited. The principles established here, most notably our anchoring to biological reality, are universal. They apply to any medical AI operation, from oncology predictive models to chronic disease management. Our findings suggest that without a formalized auditing protocol that transcends specific clinical domains, medical AI risks creating situations where the statistical power is systematically mistaken for pathological truth. In the end, the very duty of the researcher is not to govern the data narrative, but to ensure that the data still speaks for the patient.

## Consent to participate

Not applicable.

## Data availabilty statement

All data analyzed in this study are aggregate statistics extracted from the published peer-reviewed articles cited in the References section. No individual-level raw data were used, and no new datasets were generated; therefore, data sharing is not applicable. All other reasonable requests for information can be addressed to: marco.roccetti@unibo.it.

## Software availability

Not applicable: no specific software code was generated for this study.

## Author contribution

The author (MR) conceived, designed, performed the analysis, wrote, and revised this manuscript. The author has read and agreed to the published version of the manuscript.

## Ethics approval

Not applicable: neither humans nor animals nor plants nor personal data were involved in this study.

## Statement on the use of Artificial Intelligence

This work represents a transparent attempt to specify the role of AI in scientific reporting. Gemini (Google) was used for AI-assisted copy editing limited to error detection and correction, as well as formatting to adhere to the journal's style requirements. All substantive content, concepts, case studies, relative analysis, and critical interpretations are the author's own.

## Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Not applicable.

## References

[1] Herzog C, Branford J. Relational ethics and structural epistemic injustice of AI in medicine. Philos Technol 2025;38(160). https://doi.org/10.1007/s13347-025-00987-1.
[2] Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.
[3] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Phil Trans Roy Soc Lond 1933;231(694–706):289–337. https://doi.org/10.1098/rsta.1933.0009.
[4] Popper K. The logic of scientific discovery. London: Hutchinson; 1959.
[5] Roccetti M, Delnevo GJ, 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 1919;6(1):70. https://doi.org/10.1186/s40537-019-0235-y.
[6] Kim HJ, Kim M-H, Choi MG, et al. 1-year risks of cancers associated with COVID-19 vaccination. Biomark Res 2025;13:114. https://doi.org/10.1186/s40364-025-00831-w.
[7] Kang MJ, Jung K-W, Bang SH, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2020. Cancer Res Treat 2023;55(2):385–99. https://doi.org/10.4143/crt.2023.447.
[8] Park EH, Jung K-W, Park NJ, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2021. Cancer Res Treat 2024;56(2):357–71. https://doi.org/10.4143/crt.2024.253.
[9] Park EH, Jung K-W, Park NJ, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2022. Cancer Res Treat 2025;57(2):312–30. https://doi.org/10.4143/crt.2025.264.
[10] Roccetti M. A critical note on contradictions in South Korean cancer incidence rates showing concurrent increases in the vaccinated and overall decrease. Preprints.org. 2025. https://doi.org/10.20944/preprints202510.0883.v1.
[11] Kuperwasser C, El-Deiry WS. COVID vaccination and post-infection cancer signals: evaluating patterns and potential biological mechanisms. Oncotarget 2026;17:1–29. https://doi.org/10.18632/oncotarget.28824.
[12] Kim HJ, Kim M-H, Park SJ, et al. Autoimmune adverse events after COVID-19 vaccination: a nationwide population-based cohort study in Korea. J Allergy Clin Immunol 2024;153(6):1711–20. https://doi.org/10.1016/j.jaci.2024.01.025.
[13] Kang H, Lee S. Prevalence and incidence of vitiligo and associated comorbidities: a nationwide population-based study in Korea. Clin Exp Dermatol 2023;48:484–9. https://doi.org/10.1093/ced/llad028.
[14] 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.
[15] Kim HJ, Kim MH, Choi MG, et al. Psychiatric adverse events following COVID-19 vaccination: a population-based cohort study in Seoul, South Korea. Mol Psychiatr 2024. https://doi.org/10.1038/s41380-024-02627-0.
[16] Cho SJ, Kim J, Kang YJ, et al. Annual prevalence and incidence of schizophrenia and similar psychotic disorders in the Republic of Korea: a national health insurance Data- based study. Psychiatry Investig 2020;17(1):61–70. https://doi.org/10.30773/pi.2019.0041.
[17] Roccetti M. Before the algorithm: an exemplar case of the necessity of statistical testing for epidemiological consistency in public health data. AIMS Public Health 2026;13(1):121–34. https://doi.org/10.3934/publichealth.2026008.
[18] Bender EM, Gebru T, McMillan Major A, et al. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021. https://doi.org/10.1145/3442188.34459.
[19] Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318(6):517–8. https://doi.org/10.1001/jama.2017.7797.