OPEN ACCESS

EDITORIAL

The Innovation Medicine

Escaping the data misuse maze: Reorienting medical research toward clinical needs

Guoshuang Feng,[1,*] Yang Zhao,[2] Fangrong Yan,[3] Feng Chen,[2,*] and Chen Yao[4,*]

[1]Big Data Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China

[2]School of Public Health, Nanjing Medical University, Nanjing 211166, China

[3]Department of Biostatistics, China Pharmaceutical University, Nanjing 210009, China

[4]Clinical Research Institute, Institute of Advanced Clinical Medicine, Peking University, Beijing 100191, China

*Correspondence: glxfgsh@163.com (G.F.); fengchen@njmu.edu.cn (F.C.); yaoc301@qq.com (C.Y.)

Received: July 21, 2025; Accepted: August 14, 2025; Published Online: August 15, 2025; https://doi.org/10.59717/j.xinn-med.2025.100153

Citation: Feng G., Zhao Y., Yan F., et al. (2025). Escaping the data misuse maze: Reorienting medical research toward clinical needs. *The Innovation Medicine* **3**:100153.

Over the past decade, the global volume of medical research publications has surged, particularly those relying on public database mining. Spick et al. reported that publications utilizing National Health and Nutrition Examination Survey (NHANES) data rose from an average of 4 per year between 2014 and 2021 to 190 in 2024 (as of October).[1] However, despite this apparent growth, the proportion of transformative scientific advancements with clear clinical translational value has markedly declined.[2] Instead, many studies are technically sophisticated but clinically irrelevant. Spick et al. noted that NHANES-based papers often follow a formulaic pattern: correlating a health condition with nearly all conceivable environmental or physiological factors, generating a new paper with each new metric. This reveals a dangerous trend: medical research is shifting from "problem-driven" to "data-driven." When researchers prioritize "What can I do with this data?" over "What clinical problem needs addressing?", the essence of scientific inquiry risks dangerous distortion. This article mainly focuses on this phenomenon, discussing the consequences and causes of data misuse in medical research and proposing possible suggestions (Figure 1).

The definition of "data misuse" remains challenging. In this article, we

Figure 1. Escaping the data misuse maze: reorienting medical research toward clinical needs.

define data misuse as: (1) over-exploitation of data; (2) analytical errors (e.g., inappropriate model selection); (3) technical misuse (e.g., manipulation of models or variables); (4) misrepresentation (e.g., sampling bias); and (5) misinterpretation. Data misuse has spawned a proliferation of "invalid research"—studies disconnected from clinical problems, reliant on mechanical data mining, and unable to guide clinical practice or improve patient outcomes. These studies consume vast research resources and manifest in several ways: repetitive mining of the same public databases yielding highly homogeneous conclusions; manipulating variables, models, or datasets to produce positive results; fixating on marginal statistical associations while sidestepping clinical relevance; or applying complex algorithms to large datasets, yielding findings that fail to inform clinical decision-making. The consequences of this trend are severe:

Academic Pollution Chain: A systemic cycle of bias—from data bias to

analytical bias to publication bias—produces seemingly scientific but hollow studies. At the data source, public databases often harbor inherent flaws, such as sampling bias or missing data, which researchers frequently overlook. During analysis, some employ tactics like increasing model complexity or repetitive subgroup analyses to generate publishable positive results, often false positives. At the publication stage, positive results are disproportionately favored due to publication bias, perpetuating a vicious cycle of systemic bias.

Inversion of Research Hypotheses: When conclusions stem primarily from data mining, the scientific process deviates from formulating and testing hypotheses to retrofitting hypotheses after data exploration. In some fields, formulaic templates for data-driven papers have emerged, leading to unreliable published results.

P-hacking and False Positives: To secure publication, researchers may manipulate p-values (e.g., through data trimming, selective reporting, or variable manipulation) to achieve statistical significance (e.g., $p < 0.05$), even when results lack practical meaning. In public database analyses, cherry-picking favorable data or repeating analyses until significant results emerge is common, misleading subsequent research.

Misguiding Clinical Decision-Making: Data arbitrage creates an academic "bubble" that generates voluminous literature but fails to deliver meaningful medical value. This overwhelms clinicians, who struggle to distinguish valuable insights from noise, hindering evidence-based practice.

Wasted Research Resources: Chalmers et al. estimated that approximately 85% of global research funding is wasted,[3] largely because studies are not rooted in genuine clinical needs or patient/physician priorities. Research often misaligns with clinical concerns—for example, researchers focus on pharmacological interventions while clinicians and patients prioritize non-pharmacological approaches.[4] When studies are not designed to address pressing clinical questions, they often yield papers without practical utility.

Stifled Innovation: In evaluation systems that prioritize publication quantity, high-output data pipelines dominate, marginalizing researchers tackling complex, time-intensive clinical challenges. The focus on "hot" topics generates redundant papers, crowding out practical research addressing grass-roots medical issues, such as postoperative infection control or medication adherence, thus stifling innovation.

Why has rigorous science veered toward data misuse?

Treating Publication as the End Goal: Many researchers prioritize publication over clinical relevance, treating papers as ends rather than tools for improving patient care.

Distorted Academic Incentives: The "publish-or-perish" pressure in academia and healthcare ties promotions, tenure, funding, and institutional rankings to publication metrics, incentivizing rapid, low-value output over clinical impact.

Clinician-Statistician Disconnect: Superficial collaborations between clinicians and statisticians—often limited to clinicians providing data and statisticians performing analyses—contribute to methodological flaws, overinterpretation, and low-value research. Some statisticians overemphasize advanced methods without clarifying their applicability, fostering clinicians' overreliance on complex tools (e.g., using machine learning for linear relationships).

Data Accessibility Fuels Academic Opportunism: The low-barrier accessibility of public datasets (e.g., NHANES) enables rapid paper production and incentivizes "fishing expeditions"—statistical significance chasing without prior clinical hypotheses. Minimal data acquisition costs make public database mining a "low-risk, high-return" tactic, disincentivizing resource-intensive investigations of pressing clinical questions. Compounding the problem, paper mills exploit these datasets to mass-produce assembly-line publications.

Simplified Tools Enable Methodological Misuse: User-friendly statistical packages and "one-click" complex algorithms (e.g., machine learning modules) dramatically lower technical barriers. Methods requiring rigorous statistical training are now routinely misapplied by insufficiently trained users, leading to model misspecification, overfitting, and clinically opaque outputs. Tool developers rarely emphasize applicability constraints, while users prioritize technical novelty over clinical utility, yielding statistically significant but clinically unactionable results.

To escape the data mining maze, medical research must realign with clinical needs and embrace transparency in data analysis, review, and publication processes:

Move Beyond Publication Metrics: Reduce the emphasis on paper quantity in evaluations, adopting systems like the "representative work" model, which prioritizes originality, methodological rigor, clinical/public health impact, and translational potential. Long-term clinical and translational research should be granted extended evaluation timelines to encourage deep, impactful work. Diverse metrics should be incorporated, such as those recommended by the San Francisco Declaration on Research Assessment (DORA), to recognize data sharing, software tool development, clinical guideline contributions, patent translation, and real-world patient benefit.

Strengthen Journal Oversight: Journals must rigorously evaluate studies' clinical relevance and methodological soundness, particularly for public database-driven research. Authors should be required to disclose data limitations (e.g., population bias, measurement errors) and demonstrate result robustness. Sharing negative or null results should be encouraged to counter publication bias and provide a more complete evidence base. Hypothesis-driven studies must prioritize pre-registered protocols to ensure scientific integrity, and journals should reject repetitive, low-value papers while enforcing strict penalties for academic misconduct.

Guide Research Toward Clinical Priorities: Authoritative bodies should steer researchers toward high-impact areas using tools like Evidence Gap Maps (EGMs) to identify research gaps and redundancies. For instance, Saran et al. utilized EGMs to analyze the research gaps concerning children's well-being in low- and middle-income countries.[5] The findings revealed significant research gaps in areas such as child abuse and neglect, as well as traffic accidents. This study also had practical implications, such as influencing the focus of foundations.

Foster Deep Clinician-Statistician Collaboration: Medical statisticians should serve as co-definers of scientific questions, core decision-makers in study design, and gatekeepers of result interpretation, not merely data analysts. They must guide clinicians on appropriate method application, caution against misuse and misinterpretation. They should prioritize suitable, simple, robust models unless complex clinical problems necessitate advanced approaches, ensuring interpretability over technical showmanship.

CONCLUSION

Every landmark medical advance—from the serendipitous observation of penicillin's antibacterial properties to the life-saving translation of insulin therapy—originated in a profound understanding of clinical challenges and unwavering focus on patient needs, not blind data mining. Amid the deluge of data, researchers must hold fast to clinical needs as their "Ariadne's thread," guiding them through the data maze to find true value and impact. To achieve this, we must fundamentally recalibrate our focus: let us measure impact in patient outcomes, not p-values.

REFERENCES

1. Suchak T., Aliu A. E., Harrison C., et al. (2025). Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the NHANES US national health database. *PLoS Biol.* **23**:e3003152. DOI:10.1371/journal.pbio.3003152
2. Park M., Leahey E. and Funk R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature* **613**:138–144. DOI:10.1038/s41586-022-05543-x
3. Chalmers I. and Glasziou P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet* **374**:86–89. DOI:10.1016/S0140-6736(09)60329-9
4. Crowe S., Fenton M., Hall M., et al. (2015). Patients', clinicians' and the research communities' priorities for treatment research: There is an important mismatch. *Res. Invol. Engagem.* **1**:2. DOI:10.1186/s40900-015-0003-x
5. Saran A., White H., Albright K., et al. (2020). Mega-map of systematic reviews and evidence and gap maps on the interventions to improve child well-being in low- and middle-income countries. *Campbell Syst. Rev.* **16**:e1116. DOI:10.1002/cl2.1116

DECLARATION OF INTERESTS

Guoshuang Feng is an Editorial Board member of The Innovation Medicine. He was blinded from reviewing or making final decisions on the manuscript. The article was subject to the journal's standard procedures, with peer review handled independently of this Editorial Board member and their research groups. The other authors declare no conflict of interest.