ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - 2026, Volume 21, Issue 2

Will Generative AI Replace Biostatisticians? Opportunities, Challenges, and Professional Responsibility in the Era of Large Language Models

Alessandro Marcon[(1)], Valentina Panetta[(2)], Giuseppe Maglietta[(3)], Lorenza Scotti[(4)], Vittorio Simeon[(5)], Giovanni Veronesi[(6)]; on behalf of the Directory Board of the Italian Society for Medical Statistics and Clinical Epidemiology (SISMEC)

(1) Unit of Epidemiology and Medical Statistics, Department of Diagnostics and Public Health, University of Verona, Verona, Italy (ROR: 039bp8j42)
(2) L'altrastatistica s.r.l., Rome, Italy
(3) Clinical and Epidemiological Research Unit, University Hospital of Parma, Parma, Italy (ROR: 01savwy26)
(4) Department of Translational Medicine, University of Piemonte Orientale, Novara, Italy (ROR: 04387x656)
(5) Medical Statistics Unit, University of Campania "L. Vanvitelli", Naples, Italy (ROR: 02kqnpp86)
(6) Research center in Epidemiology and Preventive Medicine (EPIMED), Department of Medicine and Surgery, University of Insubria, Varese, Italy (ROR: 00s409261)

CORRESPONDING AUTHOR: Alessandro Marcon, Unit of Epidemiology and Medical Statistics, Department of Diagnostics and Public Health, University of Verona, Strada Le Grazie 8, 37134, Verona, Italy. E-mail: alessandro.marcon@univr.it

LANDSCAPE

Artificial Intelligence (AI) indicates a broad range of computational systems that exhibit intelligent behaviour by analysing their environment and acting with some degree of autonomy to achieve specific goals [1]. Large Language Models (LLMs) are a recent and influential development within this field. They are large-scale probabilistic models based on the transformer architecture, a novel class of neural networks designed to model sequential data [2]. LLMs are pre-trained on massive amounts of text content derived from articles, books, and other internet-based sources, with varying degrees of human supervision [3].

LLMs are becoming increasingly popular in many disciplines, including biostatistics (a term we use throughout the commentary to encompass also medical statistics and health data science). A survey conducted within biostatistics units at two academic medical centres in the United States found that LLMs were already quite widespread in late 2024, a period when consumer-grade LLMs were just becoming accessible to biostatisticians [4]. The survey identified three main benefits of LLMs: enhancing communication, clinical knowledge, and quantitative skills of biostatisticians.

The rapid diffusion of large language models has triggered an active debate about the evolving roles of statisticians and biostatisticians in the age of generative AI, as reflected, for example, in the Town Hall discussion held at the 2024 Joint Statistical Meetings [5]. A tutorial on LLMs published in *Statistics in Medicine*, accompanied by an editorial commentary, has further stimulated discussion within the biostatistical community about the potential implications of these technologies for professional practice [6,7]. One focus of the tutorial was to assess LLM performance in conducting statistical analysis under a biostatistician's guidance, rather than merely providing coding assistance. Using a formal experimental design to examine different prompting methods and assess reproducibility, the authors showed that LLMs can perform impressively well on some tasks (e.g., latent class analysis) but also fail significantly on others (e.g., meta-analysis) [6]. This underscores both the potential and limitations of their use in biostatistics [7].

As a Scientific Society with a strong focus on biostatistics, we considered it important to fuel the debate on the evolving role of our discipline in the LLM-centric present and future within our community first, and more broadly within the field of public health. This commentary is organised as follows. First, we address two complementary perspectives in the next two paragraphs: the role of LLMs in a biostatistician's professional practice and the legacy of biostatistics in the era of LLMs. Then, we discuss these questions from a data protection and regulatory perspective, and ground them in the broader scientific, professional, and educational context. Finally, we provide a vision on the role that Scientific Societies, including ours, can play in guiding this transition and contributing as

DOI: 10.54103/2282-0930/31293
Accepted: 12th May 2026

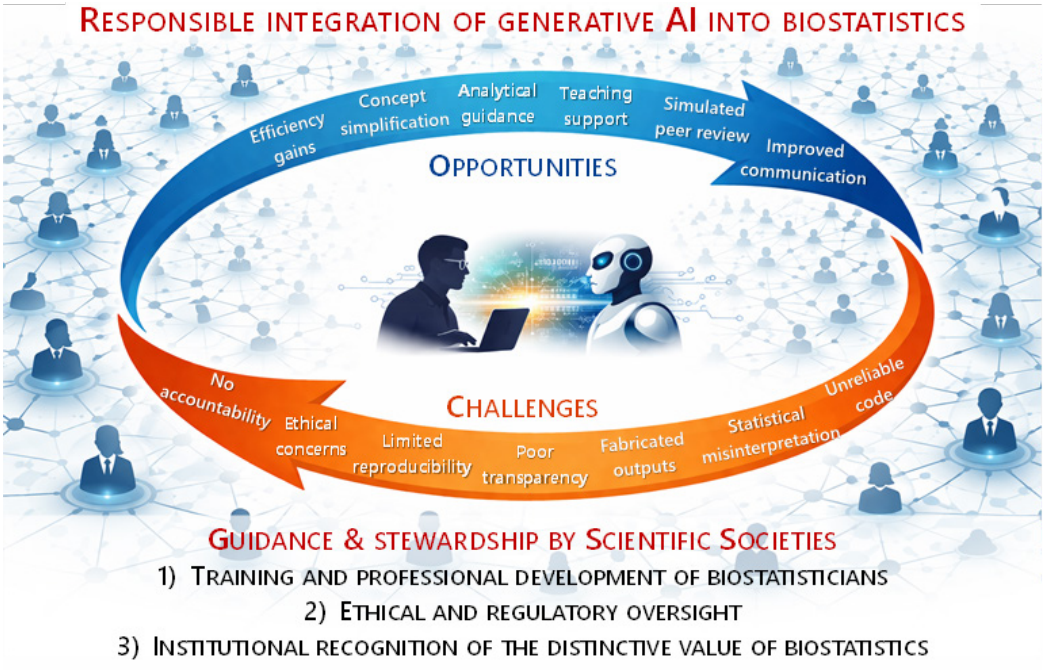

Figure 1. Graphical illustration emphasising the role of scientific societies in guiding the responsible integration of LLMs into biostatistics.

actors – rather than spectators – to writing the next blank chapters (Figure 1).

HOW CAN GENERATIVE AI ASSIST BIOSTATISTICIANS IN THEIR WORK?

What becomes immediately evident to people incorporating LLMs into their professional workflows is the substantial (sometimes striking) efficiency gains, especially in automating routine, repetitive tasks. These include specialised biostatistical skills, such as writing and debugging code, translating code across languages, and automating parts of systematic reviews and simulations [4,6]. Not less importantly, LLMs can assist with general activities, such as composing emails and reports, improving writing quality, and performing administrative and documentation tasks [3,4]. In addition, their ability to adjust tone and phrasing can be particularly useful in high-pressure contexts, where clarity and measured communication are essential.

LLMs can assist in complex analytical thinking and statistical modelling, particularly by outlining modelling strategies, articulating assumptions, and generating structured, step-by-step methodological guidance [4,7]. For example, they can help choose between alternative analytic strategies in contexts such as meta-analysis, latent class models, and causal inference [6]. This capacity can also support self-directed learning, for instance, by prompting LLMs to act as interactive tutors to explain new modelling techniques [3,4,8]. LLMs may also assist with study design and methodological evaluation – core activities of biostatistical practice – including peer review and participation in ethics committees. In these settings, they can serve as simulated peers to "stress-test" analytical strategies by systematically exploring potential weaknesses and methodological pitfalls. Given a complex question, LLMs can mimic human reasoning by searching the internet, extracting related information, exploring and scoring solutions, and iteratively refining their output [7].

This ability to summarise and simplify complex concepts makes LLMs a useful tool for teaching. Indeed, LLMs can provide coherent explanations, justify answers step by step, and contextualise knowledge using examples [8]. They are quite effective at improving communication across disciplinary boundaries, for example, by helping biostatisticians explain methods and results to people without a quantitative background. On the other hand, they can summarise and illustrate complex medical concepts to clinical non-specialists as well [3].

WILL GENERATIVE AI REPLACE BIOSTATISTICIANS?

The very studies highlighting the impressive capabilities of LLMs consistently document substantive limitations, including reasoning errors, unstable conclusions, fabricated references, and inappropriate methodological choices [4,6]. These findings underscore that the outputs of generative AI remain highly dependent on rigorous, domain-specific human oversight [7].

Empirical evaluations have documented instability

in advanced statistical tasks. In structured experiments, identical prompts have produced divergent modelling choices, inconsistent parameter estimates, and even implausible results [6]. Survey data from biostatisticians further report frequent statistical misinterpretations, incorrect code generation, and the fabrication of analytical functions, such as non-existent packages or commands in R or Stata [4].

These limitations are structural, since there are many sources of variability in LLM outputs. LLMs are periodically updated without full transparency, making version control impossible, and their output can vary over time as more data is used to train the models. Moreover, it is sensitive to prompt formulation. This has been systematically examined by Dobler et al. (2025), who applied two alternative strategies: a single comprehensive prompt and an iterative stepwise prompting approach [6]. Finally, even when the model version is identical, LLMs can generate different outputs for the same input prompt [6].

This happens because LLMs are probabilistic models. At each step of text generation, they compute a probability distribution over possible next tokens and, rather than deterministically choosing the single most likely token, they sample from that distribution. Even if low sampling randomness is set (e.g., through the "temperature" model parameter), small differences in early token selection can propagate into substantially different outputs.

From a biostatistical perspective, this raises a fundamental issue of reproducibility. Statistical analysis requires inspectable, version-controlled code and transparent documentation of modelling choices. While LLMs may assist in generating such code, the analytical results themselves must not depend on a non-deterministic system acting as a semi-autonomous analyst. It is worth noting that heterogeneity in modelling decisions is not unique to LLMs; human analysts may also reach different conclusions when analysing the same dataset [9]. The crucial difference lies in methodological accountability: human analysts are expected to justify their choices, document assumptions, and align their analyses with established statistical principles.

LLM outputs may be influenced by socio-demographic bias. Analysing both real and synthetic emergency department cases, Omar et al. (2025) found that when identical clinical scenarios were labelled with certain race and ethnicity, gender identity, sexual orientation, or socioeconomic status, models were more likely to recommend urgent triage, invasive interventions, or mental health assessments than for otherwise identical unlabelled control cases [10]. The study concluded that the magnitude of these differences was not supported by clinical reasoning or established guidelines. This does not imply that LLMs are inherently good or bad. As "stochastic parrots" [11], they reproduce patterns embedded in their training data, rather than apply fair clinical judgement [3].

DATA PROTECTION AND REGULATORY CONSIDERATIONS

Beyond methodological concerns, regulatory and legal considerations introduce additional constraints. Biomedical research relies on individual-level data, much of which qualifies as sensitive under the General Data Protection Regulation (GDPR). The deployment of LLMs raises concerns about data privacy, confidentiality, and the risk of unintended disclosure [3]. Among the principles underlying the GDPR, purpose limitation, data minimisation, storage limitation, and accountability are particularly difficult to operationalise in systems trained through large-scale data aggregation and characterised by limited transparency regarding data provenance and internal model representations. Inadvertent disclosure may occur through interactions with external AI services. Providing identifiable or even indirectly identifiable data to third-party LLM services may therefore create confidentiality and security risks, potentially exposing institutions and researchers to legal liability. For this reason, decisions about whether and how LLMs may be used in data-analysis workflows cannot be treated as purely technical choices but must be embedded within institutional governance and professional standards.

Regulatory oversight has become particularly stringent in Italy, where the national data protection authority (Garante per la protezione dei dati personali) has adopted an assertive approach, imposing substantial fines on providers such as OpenAI for non-compliant data processing practices [12]. In this context, maintaining meaningful human oversight is essential. Safeguards include strict avoidance of uploading identifiable or sensitive data to publicly hosted models, the implementation of locally hosted or institutionally controlled LLMs that operate within secure data environments, and the use of synthetic or de-identified data. Finally, explicit institutional policies and data processing agreements must clearly define data flows and responsibilities.

IMPLICATIONS FOR BIOSTATISTICAL PRACTICE

The adoption of LLMs in biostatistics has raised concerns about potential deterioration in core competencies. However, the notion of skill loss presupposes that relevant competencies were ever fully acquired in the first place. As with the introduction of calculators and statistical software, technology tends not to eliminate expertise but to reshape it, creating a need for reskilling rather than deskilling [7]. Nonetheless, it should be acknowledged that increasing accessibility of generative AI tools may lower the threshold for producing statistical analyses, raising the risk of methodological misuse and the generation of poorly justified results.

In this context, the appropriate benchmark remains the methodological standards and consensus guidelines developed by the statistical community. Biostatistics carries a long-standing methodological legacy that must be preserved and actively leveraged to ensure these new tools are used within a sound statistical framework. LLMs can be tools that streamline routine tasks, thereby freeing biostatisticians for higher-level responsibilities such as defining research questions, contributing to the development of research protocols to safeguard internal validity, and critically interpreting results. Communication is a core professional skill that LLMs can augment but not replace [4].

While LLMs are increasingly integrated into professional workflows – whether through deliberate institutional strategies or individual initiative [13] – formal training opportunities remain limited [4]. The effective and responsible use of these tools requires structured education on their capabilities and limitations, supported by well-designed prompts and systematic verification methods. In practice, this means not only technical instruction but also training in methodological scepticism, quality control, and output validation, to counterbalance LLMs' intrinsic tendency toward over-confident output. Reporting frameworks such as SPIRIT-AI and CONSORT-AI provide a valuable reference point, as they explicitly require a transparent description of AI components, documentation of human–AI interaction, and systematic analysis of errors [14,15]. The need for secure data environments will prompt institutional agreements, which will inevitably shape which LLM tools are adopted. With this premise, a careful evaluation of the pros and cons of different LLM services should inform implementation.

Contrary to early marketing promises, emerging evidence suggests that AI tools may intensify workloads rather than reduce them. Recent studies report that generative AI often leads employees to take on more tasks, work faster, and extend their work hours, thereby increasing cognitive burden despite automating certain routine functions [13]. This intensification effect reinforces the need for structured governance rather than uncritical adoption. In this sociotechnical transition, the role of biostatisticians is also to incorporate new tools in ways that preserve human judgement and well-being.

ROLE OF SCIENTIFIC SOCIETIES

Biostatisticians need not fear professional displacement by LLMs, but should embrace the opportunity to integrate them into their workflows [7]. The distinctive value of biostatisticians lies in their capacity to steward the methodological and professional integrity of scientific research.

Generative AI tools cannot and should not be considered co-authors of scientific work: ultimate responsibility for study design, bias control, data interpretation, and reporting remains unequivocally human [3,16]. Accountability, ethical judgment, and regulatory compliance cannot be delegated to probabilistic systems. Biostatisticians must therefore cultivate scepticism and retain the ability to recognise when LLM-generated outputs are incorrect or misleading [7].

In this evolving landscape, Scientific Societies, including ours, have both the opportunity and the responsibility to guide this transition. Rather than passively adapting to technological change, our Society can actively contribute to shaping the integration of generative AI into medical statistics and clinical epidemiology, ensuring that LLMs are used as instruments of methodological enhancement grounded in ethical responsibility.

In our view, the critical priorities are threefold: strengthening biostatisticians' methodological competences through training and professional development; promoting recognition of their distinctive role in institutional and interdisciplinary settings; and reinforcing regulatory awareness and ethical vigilance, in collaboration with institutions (Figure 1).

Concretely, this entails promoting structured training initiatives and critical discussion within our community first, while also developing position statements and practical guidance to be shared with the broader public health community. Only by doing so can biostatistics confirm its role as a reference point for methodological excellence and responsible innovation in the era of AI.

ACKNOWLEDGEMENTS

ChatGPT (OpenAI) was used to support drafting, language refinement, and structural editing of this commentary. The authors retained full responsibility for the conceptual content, critical interpretation, and final approval of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

FUNDING

The authors received no funding for this work.

REFERENCES

1. European Parliament, Directorate General for Parliamentary Research Services. Artificial intelligence: how does it work, why does it matter, and what we can do about it? LU: Publications Office; 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf.
2. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023;381:eadk6139. https://doi.org/10.1126/science.adk6139.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930–40. https://doi.org/10.1038/s41591-023-02448-8.
4. Grambow SC, Desai M, Weinfurt KP et al. Integrating large language models in biostatistical workflows for clinical and translational research. J Clin Trans Sci 2025;9:e131. https://doi.org/10.1017/cts.2025.10064.
5. Donoho DL, Kang J, Lin X et al. "Rebuilding" Statistics in the Age of AI: A Town Hall Discussion on Culture, Infrastructure, and Training 2026. https://doi.org/10.48550/arXiv.2601.17510.
6. Dobler D, Binder H, Boulesteix A et al. ChatGPT as a Tool for Biostatisticians: A Tutorial on Applications, Opportunities, and Limitations. Statistics in Medicine 2025;44:e70263. https://doi.org/10.1002/sim.70263.
7. Zhu B. Biostatisticians Meet AI : Navigating Shifts While Preserving Principles. Statistics in Medicine 2025;44:e70271. https://doi.org/10.1002/sim.70271.
8. Gilson A, Safranek CW, Huang T et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312. https://doi.org/10.2196/45312.
9. Gould E, Fraser HS, Parker TH et al. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. BMC Biol 2025;23:35. https://doi.org/10.1186/s12915-024-02101-x.
10. Omar M, Soffer S, Agbareia R et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025;31:1873–81. https://doi.org/10.1038/s41591-025-03626-6.
11. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM; 2021, p. 610–23. https://doi.org/10.1145/3442188.3445922.
12. ChatGPT, il Garante privacy chiude l'istruttoria. OpenAI dovrà realizzare una campagna informativa di sei mesi e pagare una sanzione di 15 milioni di euro (press release) 2024. https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/10085432 (accessed February 24, 2026).
13. Ranganathan A, Ye XM. AI Doesn't Reduce Work—It Intensifies It. Harvard Business Review 2026. https://hbr.org/2026/02/ai-doesnt-reduce-work-it-intensifies-it.
14. Liu X, Cruz Rivera S, Moher D et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health 2020;2:e537–48. https://doi.org/10.1016/S2589-7500(20)30218-1.
15. Cruz Rivera S, Liu X, Chan A-W et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.
16. Hamm B, Marti-Bonmati L, Sardanelli F. ESR Journals editors' joint statement on Guidelines for the Use of Large Language Models by Authors, Reviewers, and Editors. Eur Radiol Exp 2024;8:7. https://doi.org/10.1186/s41747-023-00420-2.