International Journal of Epidemiology, 2026, **55(1)**, dyaf210
https://doi.org/10.1093/ije/dyaf210
Opinion

International Epidemiological Association
OXFORD

Opinion

Why can't epidemiology be automated (yet)?

David Bann[1,*], Ed Lowther[2], Liam Wright[1] and Yevgeniya Kovalchuk[2]

[1]Centre for Longitudinal Studies, University College London, London, United Kingdom
[2]Centre for Advanced Research Computing, University College London, London, United Kingdom

*Corresponding author. Centre for Longitudinal Studies, University College London, 55–59 Gordon Square, London WC1H 0NU, United Kingdom. E-mail: david.bann@ucl.ac.uk.

Introduction

Epidemiology is concerned with understanding the distribution and determinants of health in the population. A sizable fraction of epidemiological research involves secondary data analysis: statistically analysing data collected from cohorts, cross-sectional studies, or other data sources. Such research comprises a series of cognitive tasks currently conducted, or at least overseen, by humans.

Historically, conducting epidemiological research was a slow, manual endeavor: scanning library shelves, reading physical papers, and manually collecting, coding, and analysing data (Fig. 1). Technological progress has now led to much of this work being done electronically, yet actual scientific progress arguably remains slow; e.g. despite the surge in large cohorts and ballooning data volumes—omics, wearables, administrative linkages, etc.—progress in identifying modifiable causes of disease has proved elusive [1, 2].

Artificial intelligence (AI) represents the next step in the technological evolution of epidemiology (Fig. 1); it can accelerate—or even automate—cognitive tasks, boosting the efficiency of current practice and creating new opportunities for discovery. Epidemiologists often use AI-based tools—sometimes without explicitly knowing it—such as Google Scholar for paper discovery, spell-checkers for writing, and GitHub Copilot for coding.

Despite previous AI "winters", its current era of development, built around the transformer deep-learning architecture [3] that powers modern large language models (LLMs), has generated remarkable progress. LLMs shot into public consciousness in November 2022 with the release of ChatGPT, reportedly the fastest-growing consumer product of all time. Many scientists, particularly younger researchers, now use ChatGPT [4]. Other LLMs have since become publicly available and widely used, and billions of dollars are invested in their training. LLMs predict the next token (typically, a small piece of text) in a sequence and, when developed at a massive scale, have surprisingly useful properties. Productivity increases in tasks relevant to epidemiology have recently been suggested—writing [5], cognitive tasks [6], debating/reasoning [7], and coding [8].

Here, we map the landscape of epidemiological tasks that rely on existing datasets—from literature review through to idea generation, data access, analysis, write-up, and dissemination. We provide a snapshot of AI tools and a repository containing examples of AI-generated epidemiological output, along with prompt and model details (https://github.com/edlowther/automated-epidemiology). The tools were chosen to present an illustration of what, at the time of writing, frontier AI models were capable of. (We use the term AI to refer to computational systems that can perform cognitive tasks relevant to epidemiological research.) Finally, we discuss barriers to deeper AI integration and broader implications for the field, including how epidemiologists can contribute to AI development, addressing recent calls [9]. We note that the issues discussed apply to other fields, e.g. social sciences [10, 11].

Conducting literature reviews

Systematic reviews generally take ≥1 year to undertake [12], with much of this time spent on pain-staking and often tedious screening—manually removing the vast majority of irrelevant articles from the selected pool by comparing them against the same inclusion/exclusion criteria. At face value, this breaks the software engineering principle to "automate repetitive tasks," but an additional motivation for automation is to reduce errors: humans do not screen without error [13].

In recent years, machine-learning tools have become available to speed up screening: authors manually screen a smaller subset of abstracts to train models, which then automatically screen the remainder [14]. Such tools appear to increase efficiency [15, 16], widening the scope to produce or update reviews more rapidly (possibly continuously) and undertake more ambitious evaluations. LLMs can help in other review-related tasks, such as creating synonym/search-term lists and extracting data. Could the entire process of reviewing be automated? Using an agentic AI system (Otto-SR), a 2025 study claimed to have reproduced and updated a Cochrane issue in 2 days—the equivalent of 12 work-years of traditional systematic review work (assuming 1 year per review) [17].

For more ad-hoc literature searches, systematic reviews are typically prohibitively costly, e.g. when informing introductions or discussion sections in original research articles. Researchers are increasingly using AI-augmented search

Received: 21 July 2025; **Accepted:** 30 October 2025

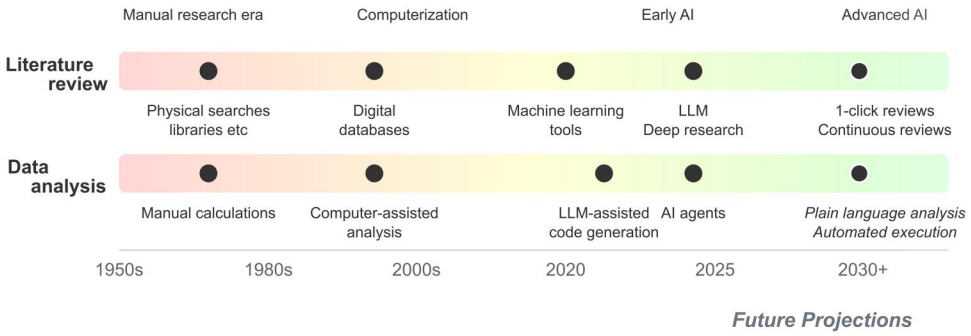

Figure 1. Technological progress in two key epidemiological tasks (literature reviews and data analysis): from manual work and computerization to artificial intelligence (AI)-augmented research. Note that, in some senses, the final tasks listed (e.g. plain-language analysis) are already possible with current AI systems, yet the quality of the outputs is of mixed or as-yet undetermined quality.

tools such as Google Scholar to undertake literature searches—unlike PubMed, it indexes non-health publications (e.g. economics articles), as well as grey literature. Nevertheless, both tools require conversion of the search question (e.g. "What effect does X have on Y?") into terms that are more likely effective for such databases (e.g. "associations between X and Y," a "randomized controlled trial of X on Y," etc.), in addition to a continued manual search of "cited by" articles.

LLMs can answer such single questions directly and recent reasoning LLMs enable AI "sense checks" before a response is produced. Hallucinations—which raised considerable concern in early models—appear to have been reduced [18]. "DeepResearch" capabilities, made available in several leading LLM tools in recent months, enable more extended searches of scholarly literature; users can check the sources provided via links to the full text.

The generality of LLMs means they can usefully sift, connect, and summarize evidence from far-flung disciplines—a task that has otherwise become progressively harder as scientific output has surged [19]. For epidemiologists, such sources range from mechanistic studies in cells, animal models, human autopsy studies, and the social sciences (e.g. psychology, economics, sociology). AI tools could thus make cross-disciplinary triangulation more feasible.

Hallucinations remain a barrier to the trustworthiness of LLMs, but a human barrier exists in accessing research articles. Since the 1970s, the five largest for-profit publishers have steadily increased their market share, accounting for more than half of all papers by 2013 [20]. Papers—and even their abstracts—are copyrighted. This creates a particular barrier for open-source AI systems [21].

Partial access to research articles, limited performance with longer "context windows" (the amount of data the LLM uses from memory), and the capacity of LLMs to provide highly compelling but unsupported narratives mean LLMs may mislead [22]. Ongoing evaluation of such systems is required: empirical study of their sensitivity and specificity in searches, for example. This is a challenge given their rapid development—closed-source frontier LLMs can be rapidly updated or decommissioned.

Creativity and generating hypotheses

It is often assumed that AI systems (particularly LLMs) simply interpolate between data points contained within their training set and are thus not capable of being creative or generating novel ideas—i.e. they are "stochastic parrots" [23]. Setting aside the "incremental" nature of modern science, such claims are at least partly empirically testable: an emerging literature suggests that the creative capability of frontier models may match those of humans in discrete small-scale creative tasks [24]. Their abilities in real-life scientific creativity remain uncertain, as do the comparisons of humans alone versus human–AI collaborations in (i) forming hypotheses that advance epidemiology or (ii) selecting hypotheses that are tractable and falsifiable [25] given the existing data. In other disciplines, such as drug discovery, new scientific findings are seemingly being discovered via AI systems [26].

We prompted a recently developed AI tool (the AI Scientist [27]) to suggest novel hypotheses across two topics: (i) the links between birthweight and subsequent body mass index (BMI) and (ii) social inequalities in mental health (see github.com/edlowther/automated-epidemiology). Many hypotheses appear to have face validity, e.g. suggesting generally underutilized approaches to causal inference (sibling comparison studies and natural experiments). We note that such suggestions were created in "one shot" and are thus the equivalent of a human's first draft. Even if only a fraction of AI-suggested hypotheses are promising, the number that can be created quickly is large and may be especially valuable with discerning humans "in the loop" to select them: an AI-augmented process akin to human brainstorming.

Identifying and accessing data

A common approach in epidemiological research is that groups running specific epidemiological studies (e.g. cohorts or health surveys) publish research focused on using that specific dataset. In this scenario, multiple publications in the literature from different research groups address the same question; yet, subsequently synthesizing such evidence (e.g. via meta-analysis) is not always possible due to methodological differences. Consortia integrating multiple studies are one manual approach to circumventing this, but they are typically set up for specific research questions and are hard to maintain in the long run; when their funding ends, they may cease to operate.

AI may enable a bolder default for epidemiological research, enabling us to ascertain, for each research question, the possible available datasets that could contribute evidence. Of these, which have harmonizable data? And what does that evidence collectively show?

Current barriers to this include the high fixed costs of becoming familiar with datasets and the fragmented approaches to data discovery and access. Platforms to aid cohort

discovery, e.g. the recent Atlas of Longitudinal Data (https://atlaslongitudinaldatasets.ac.uk), are a step forwards in helping to identify datasets; yet, using them highlights our barriers: 10 different cohorts may involve 10 separate access systems, with considerable overlap in the information requested.

The challenge for data providers is whether a single point of entry can be provided—a cohesive streamlined data-access system with interoperable data and necessary safeguards. ORCID provides a centralized and broadly accepted system for verifying researcher identity—could existing centralized systems for data documentation and access be expanded (e.g. UK Data Service for UK cohorts; or the Gateway to Global Aging Data, for older adults) or newly created? Within such systems, AI tools can also facilitate the historically slow and manual process of harmonizing data across different datasets [28] (e.g. the Harmony tool [29]).

Finally, AI tools can aid in the creation of new epidemiological data. In existing cohorts, for example, data held in historic non-electronic form (e.g. paper questionnaires or microfiche) can be digitized by using automated optical character recognition tools. Such tools can also be used to create new retrospective cohorts: many hundreds of papers have now cited the cohort profiles that arose from the discovery and digitization of records, which formed the basis for the Hertfordshire [30] and Lothian [31] cohort studies. AI tools could also improve existing metadata (e.g. annotating questionnaires with associated variable names).

Analysing data

Much like in the literature reviews, epidemiologists are increasingly supported by AI when analysing data. Rather than manually typing out each letter when coding, AI autocompleters such as GitHub Copilot can speed up code writing. For a guide, see https://www.ncrm.ac.uk/resources/online/all/?id=20859. Frontier LLMs are now able to create a complete draft of code in response to a plain-language prompt and then execute this code. The promise is that the rapid, autonomous generation of research code will enable human researchers to spend more time at higher levels of abstraction, e.g. thinking carefully about designing research strategies.

We prompted an agentic AI framework (Data Analysis Crow) to address two research questions and provided simulated data. The responses yielded an analytical plan, analytical code, execution of this code, and visualizations—see github.com/edlowther/automated-epidemiology for full workbooks and Table 1 for a summary. While the outputs were (in our view) impressive, they did contain errors and, in some cases, failed entirely, depending on the underlying LLM used. This suggests that (i) the choice of LLM is important and (ii) code review remains essential.

Often, epidemiologists specialize in one piece of software or programming language (e.g. SAS, SPSS, Stata, R). In one sense, specialization is increasingly not needed, as the barrier to entry lowers to code in multiple languages. What will remain important is the clear articulation of the goals in plain language and code review. The fact that AI systems provide analytical syntax also aids in reproducibility: something that <2% of health researchers currently do [32].

The more mundane aspects of data analysis could also be accelerated by AI. Data cleaning, for example, is often a highly manual and time-intensive task that is required even for well-used datasets, leading to considerable duplication of work. Assuming that data cleaning involves 1 month of unnecessary work (cleaning data that should have otherwise been centrally cleaned)—a task ordinarily repeated across 1000 papers—1000 months (83 years) of scientists' time could be saved in the future. A cursory look at the literature suggests at least six distinct AI data-cleaning tools from 2024 onwards that claim varying levels of accuracy in data cleaning [33–38]. Whether such tools are useful in epidemiological applications remains to be seen. A challenge for epidemiologists will be to make sense of the bewildering numbers of tools released in AI-related fields: the creation and curation of epidemiological benchmarks could provide objective criteria by which they can be continually evaluated.

Barriers to the use of AI in data analysis include the current frequent need for uploading data to cloud providers: this is not possible for many health-related datasets held in sandboxed secure computing environments. Researchers could instead use local open-source models—such models have historically been weaker than the closed-source models, yet, in recent months, the gap has narrowed considerably [39].

Table 1. Evaluating AI-generated analysis: illustrative results from the Data Analysis Crow

		LLM evaluated	
Association	**Challenge**	**GPT-4.1**	**Claude Sonnet 4**
Birthweight → BMI	Data cleaning	✓ Derived BMI ✓ Removed implausible values	✓ Derived BMI ⚠☐ Removed some but not all implausible values
	Designing and executing analytical plan	✓ Created analytical plan ⚠☐ Errors (e.g. <1.2 metre cases excluded, not <1 metre) ✓ Executed results: tables/figures	✓ Created analytical plan ⚠☐ Partial results (API crash)
	Analysis outcome	✓ Interpreted regression output	⚠☐Analysis incomplete
Income → mental health	Data cleaning	⚠☐ Identified sex, assumed value labels ✓ Log-transformed income	⚠☐ Identified sex, assumed value labels ✓ Rescaled income
	Designing and executing analytical plan	✓ Created analytical plan ✓ Executed results: tables/figures	✓ Created analytical plan ⚠☐ Partial results (API crash)
	Analysis outcome	✓ Interpreted regression output	⚠☐Analysis incomplete

Each item was evaluated as follows: ✓: correct or plausible output; ⚠☐: error or concern identified. A simulated dataset was provided, available on the accompanying repository: https://github.com/edlowther/automated-epidemiology. The Data Analysis Crow is available at https://github.com/Future-House/data-analysis-crow. API, application programming interface.

Alternatively, the AI tools could be restricted to accessing metadata (e.g. variable names, labels, and result output) rather than the raw data, or data owners could release synthetic versions of their data.

Epidemiologists will, as ever, need to balance two sets of competing risks. The first is a risk of high-profile data leaks if AI tools are used irresponsibly; the second is a risk that scientific discovery is restricted if such tools are not used. The latter risk is often overlooked in our view, despite threats to the continued existence of epidemiological studies (e.g. funding uncertainty and declining response rates) and the ever-increasing volumes of data collected, which often remain under-researched.

Writing up

Remarkably, given a simple plain-language prompt, frontier LLMs can produce entire epidemiological research papers; github.com/edlowther/automated-epidemiology shows examples of this by using multiple LLMs, each instructed to write a paper on the association between birthweight and adult BMI by using a simulated dataset that we provided.

Where do such papers sit in the current distribution of human-created epidemiological research papers? Despite being given very little contextual background, the highest quality amongst our AI-created outputs (produced by ChatGPT's o3 model) at face value appeared to satisfy the most commonly used consensus criteria for the reporting of epidemiology studies (STROBE guidelines [40], available on the accompanying repository). It also showed signs of reasoning: it identified a sex interaction (associations differed in direction by sex) that we had introduced into the simulated data, despite the prompt instructing the LLM to analyse "adjusted by sex."

In other respects, the AI-produced papers are of low quality, e.g. incorrect referencing and the omission of result items. Yet, as demonstrated in the "Analysing data" section, LLMs can (with tool calling) produce compelling figures and tables. Thus, current barriers to producing high-quality AI-generated manuscripts may partly reflect limitations in how effectively the model is prompted or integrated with tools.

Full (end-to-end) automation

Full automation of epidemiological research papers—from generating the idea all the way through to write-up—is a logical consequence of the capability of AI in each component chained together. Such tools as AI Scientist [27] and data-to-paper [41] are recent open-source examples of this. A fruitful avenue of future research is the evaluation of such systems tailored for epidemiology, e.g. relative to human-generated and AI-with-human-in-the-loop-generated papers.

Dissemination

Researchers are increasingly encouraged to share their research with non-specialist audiences such as the public and government policymakers, and more generally engage in continued public discourse. From the perspective of scientists who are already struggling with a "mountain of small things" [42] (including recent requirements on the reporting of AI use), such tasks may be unwelcome—yet, if public discourse is dominated by a small, vocal, and unrepresentative minority of scientists, then evidence-based policy may suffer [43]. LLMs can speed up the creation of blogs, lay summaries, and social media content if provided with prompts and context (e.g. the research paper), though human oversight may be necessary to ensure accuracy and appropriate nuance. Entire podcasts can now be completely automated; an AI-generated podcast based on this article can be found at github.com/edlowther/automated-epidemiology. If AI increases efficiency, then researchers may be able to move towards a deeper engagement with evidence-based policy—rather than simply advocating that their own work should change policy, creating unbiased evidence across the entire evidence landscape, for example, and carefully considering policy trade-offs [44].

Overall utility

In our judgment, the current capability of AI suggests a promising future to accelerate epidemiology. This is the case whether AI is used for narrow tasks under close human supervision; as a research assistant or collaborator [45, 46] with human oversight; as an expert; or—more controversially—as a semi- or fully autonomous research agent [47]. Each may bring benefits to epidemiology, with further integration an evolving combination of both human-system and AI-capability barriers.

A promise of AI in the short to medium term is that it could enable more time to be spent on high-level tasks (e.g. designing new research questions or data collections) rather than on low-level tasks that are often undesired (e.g. repetitive admin tasks) or uninspiring (e.g. writing code to recode variables) (see Fig. 2). The blend is at our discretion: some investment in

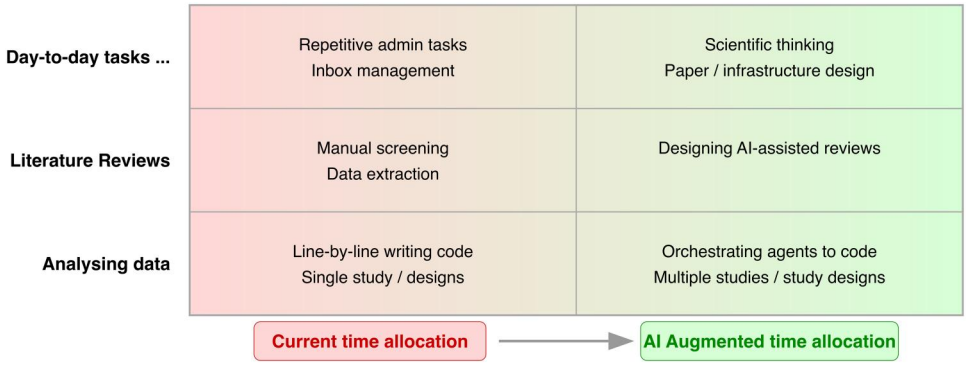

Figure 2. Could AI help to liberate epidemiologists to focus on higher-level tasks? Simplified illustration of our suboptimal current time allocation (left) versus idealized AI-augmented allocation (right).

low-level tasks is likely to be helpful or even necessary to learn (e.g. to deeply understand data, to comprehend statistical methods) and build the foundations needed for higher-level tasks. Future training should seek to optimally balance this and avoid an overreliance on AI, which could enervate the skills required to evaluate and use their outputs judiciously.

If the quality rather than the quantity of our outputs is incentivized, then the net result could be higher-quality science and a bolstering of our discipline. We live in an era that incentivizes scientists to produce masses of papers of questionable quality, including their direct purchase online [48]; this is a human, not an AI, problem.

Existential risks and concluding thoughts

AI developments are rapidly lowering the costs of cognitive tasks and may ultimately lower demand for human epidemiologists—particularly junior epidemiologists, who traditionally lead on writing and analysis tasks, overseen by a senior colleague. If unchecked, this trend could damage the training pipeline, leading to fewer epidemiologists across all levels and thus a collapse in the discipline.

Scientific careers are already uncertain, with rates of pay for epidemiologists generally lower than those in other technical sectors (e.g. tech/pharma/finance). Will our brightest minds wish to become epidemiologists in the future? Addressing structural problems (pay, security) is one route for attracting talent. Another is the appeal of working on interesting and important problems—the integration of AI with epidemiology is one. For example, can AI accelerate or automate epidemiology? How can we use AI to improve epidemiology and avoid a vast expansion of "AI slop"? Can AI benchmarks be tailored/newly created for epidemiology? What biases and risks can AI systems introduce? Can the methods proposed in the AI literature be used to improve prediction [49] and inference [50] in epidemiology? How will AI influence population health?

AI could assist, augment, and automate aspects of epidemiology in the future. If AI in its current iteration were to take over human intelligence entirely, our existential role could be temporary: to produce new "tokens" (data, papers), which vast multibillion-dollar companies use to train AI systems without our consent. Whether such scenarios are good or bad for scientific discovery or humanity at large remains an open question, which epidemiologists can and should contribute to. Realizing the potential of AI for epidemiology will require two-way engagement between epidemiologists and engineers.

Author contributions

Wrote the first draft: D.B. Conducted analysis: E.L., D.B. All authors contributed to generating ideas, compiling material, reviewing and revising the text, as well as the accompanying repository.

Conflict of interest

None declared.

Funding

D.B. and L.W. are funded by the Economic and Social Research Council (ES/W013142/1) and the UKRI Digital Research Infrastructure (DRI) Programme (UKRI/ST/B000295/1).

Use of artificial intelligence (AI) tools

This paper explores the use of AI in epidemiology. Thus, a range of tools were used as noted in the paper and Github Repository. LLMs such as ChatGPT were also used to suggest means of reducing the text length for some paragraphs. The human authors drafted all sections of the text and carefully reviewed any suggestions made. The lead author discloses a proclivity for using em dashes pre-ChatGPT release.

References

1. Brennan P, Davey-Smith G. Identifying novel causes of cancers to enhance cancer prevention: new strategies are needed. *J Natl Cancer Inst* 2022;**114**:353–60.
2. Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol* 2011;**40**:537–62.
3. Vaswani A. Attention is all you need. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, 6000–10.
4. Van Noorden R, Perkel JM. AI and science: what 1,600 researchers think. *Nature* 2023;**621**:672–5.
5. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 2023; **381**:187–92.
6. Haslberger M, Gingrich J, Bhatia J. No great equalizer: experimental evidence on AI in the UK labor market. 2024. Available at SSRN 4594466. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4594466 (5 December 2025, date last accessed).
7. Roldán-Monés T. When GenAI increases inequality: evidence from a university debating competition. 2024. https://poid.lse.ac.uk/PUBLICATIONS/abstract.asp?index=10951
8. Cui ZK, Demirer M, Jaffe S *et al.* The effects of generative AI on high skilled work: evidence from three field experiments with software developers. Available at SSRN 4945566. 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4945566
9. Sung J, Hopper JL. Co-evolution of epidemiology and artificial intelligence: challenges and opportunities. *Int J Epidemiol* 2023; **52**:969–73.
10. Bann D, Wright L, Artificial intelligence. NCRM working paper national centre for research methods. 2025. https://eprints.ncrm.ac.uk/id/eprint/4983/
11. Bail CA. Can generative AI improve social science? *Proc Natl Acad Sci USA* 2024;**121**:e2314021121.
12. Borah R, Brown AW, Capers PL *et al.* Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; **7**:e012545.
13. Wang Z, Nayfeh T, Tetzlaff J *et al.* Error rates of human reviewers during abstract screening in systematic reviews. *PLoS One* 2020; **15**:e0227742.
14. Van De Schoot R, De Bruin J, Schram R *et al.* An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell* 2021;**3**:125–33.
15. Abogunrin S, Muir JM, Zerbini C *et al.* How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. *Front Pharmacol* 2025;**16**:1454245.
16. van Dijk SH, Brusse-Keizer MG, Bucsán CC *et al.* Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open* 2023;**13**:e072254.
17. Cao C, Arora R, Cento P *et al.* Automation of systematic reviews with large language models. medRxiv 2025:2025.06.13.25329541, preprint: not peer reviewed.

18. Haman M, Školník M. Fake no more: The redemption of ChatGPT in literature reviews. *Account Res* 2025;1–3.
19. Hanson MA, Barreiro PG, Crosetto P *et al.* The strain on scientific publishing. *Quant Sci Stud* 2024;**5**:1–21.
20. Larivière V, Haustein S, Mongeon P. The oligopoly of academic publishers in the digital era. *PLoS One* 2015;**10**:e0127502.
21. Culbert J, Hobert A, Jahn N *et al.* Reference coverage analysis of openalex compared to web of science and Scopus. arXiv preprint arXiv: 240116359, 2024, preprint: not peer reviewed.
22. Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. *R Soc Open Sci* 2025; **12**:241776.
23. Bender EM, Gebru T, McMillan-Major A, Mitchell S. 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://dl.acm.org/doi/10.1145/3442188.3445922
24. Sun L, Yuan Y, Yao Y *et al.* Large Language Models show both individual and collective creativity comparable to humans. *Think Skills Creat* 2025;**57**:101870.
25. Huang K, Jin Y, Li R *et al.* Automated hypothesis validation with agentic sequential falsifications. arXiv preprint arXiv: 250209858, 2025, preprint: not peer reviewed.
26. Ghareeb AE, Chang B, Mitchener L *et al.* Robin: a multi-agent system for automating scientific discovery. arXiv preprint arXiv: 250513400, 2025, preprint: not peer reviewed.
27. Lu C, Lu C, Lange RT *et al.* The AI scientist: towards fully automated open-ended scientific discovery. arXiv preprint arXiv: 240806292, 2024, preprint: not peer reviewed.
28. Fortier I, Raina P, Van den Heuvel ER *et al.* Maelstrom research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol* 2017;**46**:103–5.
29. Moltrecht B, McElroy E, Hoffmann MS. Software Profile: Harmony: a web-tool for retrospective, multilingual harmonisation of questionnaire items using natural language processing. 2023.
30. Syddall H, Aihie Sayer A, Dennison E *et al.* Cohort profile: the Hertfordshire cohort study. *Int J Epidemiol* 2005;**34**:1234–42.
31. Deary IJ, Gow AJ, Taylor MD *et al.* The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr* 2007;**7**:28–12.
32. Hamilton DG, Hong K, Fraser H, *et al.* Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. *Bmj* 2023; **382**: e075767. https://doi.org/10.1136/bmj-2023-075767
33. Zhang S, Huang Z, Wu E, Data cleaning using large language models. arXiv preprint arXiv: 241015547 2024, preprint: not peer reviewed.
34. Ni W, Zhang K, Miao X *et al.* IterClean: an iterative data cleaning framework with large language models. In: *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, 2024. https://dl.acm.org/doi/abs/10.1145/3674399.3674436
35. Chen Z, Cao L, Madden S *et al.* SEED: Domain-specific data curation with large language models. arXiv preprint arXiv: 231000749 2023, preprint: not peer reviewed.
36. Biester F, Abdelaal M, Del Gaudio D. Llmclean: context-aware tabular data cleaning via LLM-generated OFDS. In: *European Conference on Advances in Databases and Information Systems*. Springer, 2024. https://link.springer.com/chapter/10.1007/978-3-031-70421-5_7
37. Ahmad MS, Naeem ZA, Eltabakh M *et al.* Retclean: retrieval-based data cleaning using foundation models and data lakes. arXiv preprint arXiv: 230316909, 2023, preprint: not peer reviewed.
38. Abdelaal M, Ktitarev T, Städtler D, Schöning H. SAGED: few-shot meta learning for tabular data error detection. In: *EDBT*, 2024. https://openproceedings.org/2024/conf/edbt/paper-95.pdf
39. Guo D, Yang D, Zhang H *et al.* Deepseek-r1: incentivizing reasoning capability in LLMS via reinforcement learning. *Nature* 2025; **645**:633–8.
40. Von Elm E, Altman DG, Egger M *et al.*; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007;**370**:1453–7.
41. Ifargan T, Hafner L, Kern M *et al.* Autonomous LLM-driven research—from data to human-verifiable research papers. *NEJM AI* 2025;**2**:AIoa2400555.
42. Husain M. A Mountain of Small Things. *Brain* 2024;**3**:739.
43. Smith GD, Blastland M, Munafò M. Covid-19's known unknowns. *BMJ* 2020;**371**:m3979.
44. Bann D, Courtin E, Davies NM *et al.* Dialling back 'impact'claims: researchers should not be compelled to make policy claims based on single studies. *Int J Epidemiol* 2024;**53**:dyad181.
45. Schmidgall S, Su Y, Wang Z *et al.* Agent laboratory: using LLM agents as research assistants. arXiv preprint arXiv: 250104227. 2025, preprint: not peer reviewed.
46. Gottweis J, Weng W-H, Daryin A *et al.* Towards an AI co-scientist. arXiv preprint arXiv: 250218864, 2025, preprint: not peer reviewed.
47. Morris MR, Sohl-Dickstein J, Fiedel N *et al.* Levels of AGI: operationalizing progress on the path to AGI. arXiv preprint arXiv: 231102462, 2023, preprint: not peer reviewed.
48. Richardson RA, Hong SS, Byrne JA *et al.* The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proc Natl Acad Sci USA* 2025;**122**:e2420092122.
49. Savcisens G, Eliassi-Rad T, Hansen LK *et al.* Using sequences of life-events to predict human lives. *Nat Comput Sci* 2024;**4**:43–56.
50. Petersen AH, Osler M, Ekstrøm CT. Data-driven model building for life-course epidemiology. *Am J Epidemiol* 2021;**190**: 1898–907.

International Journal of Epidemiology, 2025, 55, 1–6
https://doi.org/10.1093/ije/dyaf210
Opinion