

Review Article

The Rebirth of Ecological Study

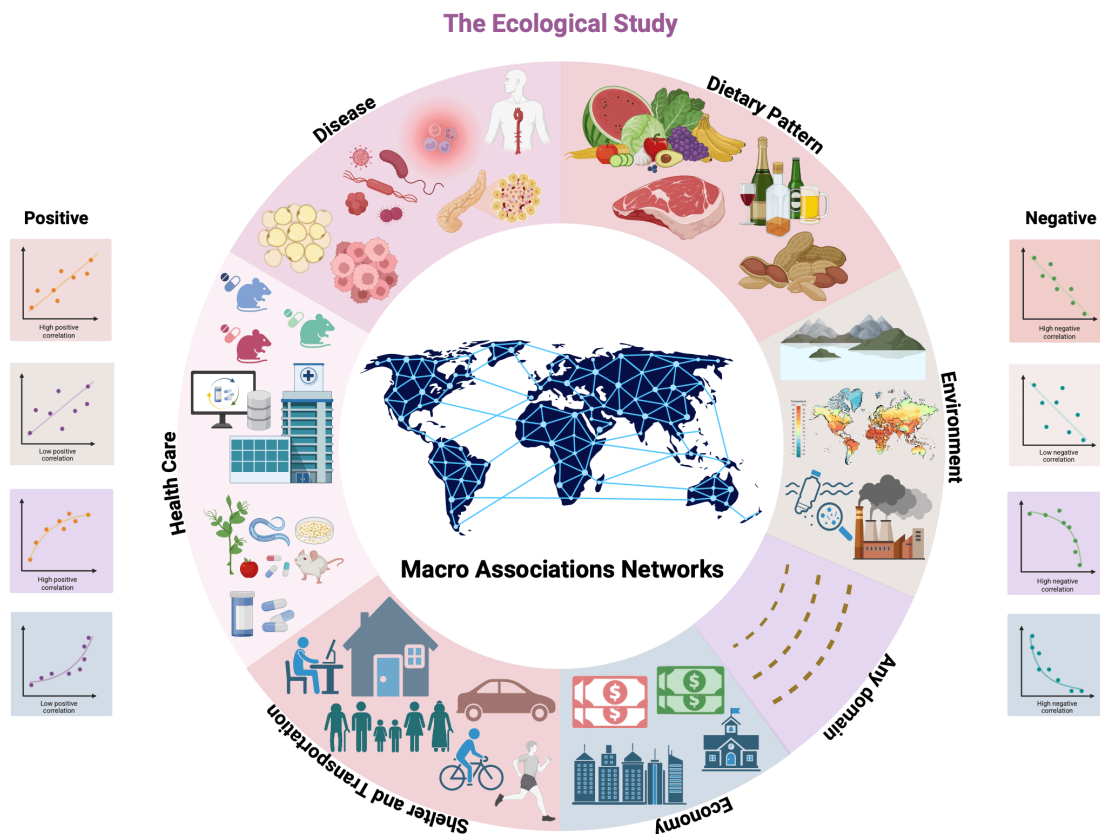
Hongyue Ma^{1,2}

1. School of Life Sciences, Westlake University, Hangzhou, China; 2. Haide College, Ocean University of China, China

Ecological study, historically criticized for its vulnerability to confounding and its reliance on simple group-level correlations, is now undergoing a conceptual and methodological renewal. The advent of large-scale open datasets, advances in computation, and the development of macro association analysis (MA-analysis) provide the foundation for this transformation. MA-analysis reframes ecological study from isolated pairwise associations toward the construction of scalable, multi-level networks that capture complex relationships across domains. Through the Three Stars Global Macro Association-analysis (TSGMA) framework, associations can be progressively classified, moving from exploratory correlations through confounder adjusted links and on to temporally validated links, which enhances both the rigor and the interpretability of the findings. This shift enables ecological study to evolve into a robust platform for mapping macro-level determinants of health, sustainability, and societal change. Ultimately, the construction of pan-domain MA networks offers not only a means of integrating fragmented knowledge, but also a pathway toward understanding the systemic forces that have shaped human development and that will guide planetary futures.

Corresponding author: Hongyue Ma, mahongyue@westlake.edu.cn

Graphical Abstract



1. Introduction

An ecological study is an observational study in which the units of analysis are populations or groups, rather than individuals^[1]. Historically, ecological studies played a foundational role in epidemiology, particularly when individual-level data were scarce or unobtainable^[2]. Landmark findings such as national dietary patterns and population-level disease burdens emerged from such studies, shaping early public health strategy^[3].

However, the method fell out of favor. In the late 20th century, limited data availability, lack of statistical standardization, and the absence of computational infrastructure undermined the reproducibility and perceived rigor of ecological studies^{[4][5]}. As evidence-based medicine advanced, so too did the emphasis on internal validity and individual-level inference. Increasing awareness of confounding and the ecological fallacy led to ecological studies being marginalized in the modern hierarchy of evidence.

Large-scale exploratory studies using population-level data became rare, and ecological designs were often dismissed as outdated or weak.

Yet the landscape has shifted dramatically. In the 21st century, the confluence of high-performance computation, algorithmic advances, and standardized global data infrastructure has ushered in a new era^{[6][7][8][9][10][11][12]}. Public datasets from hundreds of countries, covering health, diet, agriculture, policy, economics, and demography, are now openly available and increasingly harmonized across time and space. Paradoxically, even as data for population level analysis have flourished, ecological study has remained overlooked, constrained by historical stigma and by the absence of methodological modernization.

It's a missed opportunity. In the present era, however, the unprecedented availability of large-scale datasets on population and group, together with the conceptual advance of macro associations (MA), has given rise to an entirely new paradigm of ecological study^[13]. This paradigm leverages the simplest analytical approaches to interrogate the most complex systems: assembling diverse datasets defined at the population level, applying ecological study designs to explore broad patterns of potential associations, and systematically identifying novel linkages that may reshape prevailing research frameworks. By structuring and interconnecting previously fragmented datasets into MA networks, this approach enables the exploration of large-scale, high-dimensional associations and the discovery of emergent regularities governing human health, environment, and societal development. In doing so, the meaning of ecological study is elevated from narrow and often inaccurate pairwise correlation analyses to a comprehensive, scalable, and more reliable exploration of macro level associations across disciplines, and ultimately across domains. Such a shift marks the beginning of a new trajectory in which ecological study evolves into a critical tool for mapping the interconnected dynamics of human survival and planetary sustainability.

2. The rise, decline, and rebirth of ecological study

Ecological study, defined by its use of population or group level units as the basis of analysis, emerged as a pragmatic response to data scarcity in early epidemiology (Figure 1)^{[1][4]}. In the mid-20th century, when individual-level surveillance systems were rare or nonexistent, researchers turned to national aggregates to interrogate public health patterns^{[3][14][15]}. This era laid the theoretical groundwork for population-level thinking, with early contributions such as Frost's (1934) tuberculosis analyses and Kuller's proposal

of “macro epidemiology” exemplifying the use of aggregate data to generate population-wide insights^{[16][17][18][19][20][21]}.

The method flourished during the 1960s to 1980s, finding broad application in nutrition, environmental health, chronic disease epidemiology, and social determinants of health^{[2][22][23]}. Studies linking dietary fat to cardiovascular mortality or air pollution to respiratory disease provided compelling, although tentative, evidence of macro level exposures^{[24][25][26]}. However, this period also marked the rise of conceptual critiques. The “ecological fallacy,” formally articulated by Robinson in 1950, warned of the dangers of misattributing group-level correlations to individual-level mechanisms^[27]. This was followed by methodologically grounded concerns about uncontrolled confounding, effect modification, and the misinterpretation of spatial autocorrelation, challenges that were difficult to address given the computational and statistical tools available at the time^{[5][28]}.

By the 1990s and early 2000s, ecological study had undergone a clear decline in status^{[4][29][30]}. The ascendance of evidence-based medicine (EBM), bolstered by randomized trials and large cohort designs, codified a new hierarchy of evidence in which ecological designs were demoted to a very low tier^{[31][32][33]}. This marginalization was reinforced by the rise of meta-analytic techniques and causal inference frameworks that emphasized internal validity^[33]. Despite the growing availability of national and global statistics from agencies like WHO, FAO, and the World Bank, these were rarely leveraged systematically. As a result, ecological studies, although theoretically promising, remained methodologically fragmented, inconsistently reported, and increasingly viewed as exploratory relics of a pre-modern data era.

The COVID-19 pandemic, however, disrupted this trajectory in meaningful ways. Faced with a rapidly evolving global crisis, policymakers and scientists turned to population-level data to generate real-time situational awareness^{[34][35][36]}. From early cross-country comparisons of mortality and intervention timing to global dashboards mapping testing and vaccination coverage, ecological logic, including aggregation, comparison, and temporal mapping, re-entered the epidemiological mainstream^{[37][38][39]}. While limitations of such analyses were widely acknowledged, COVID-19 underscored the utility of rapid, population-scale insight generation in the absence of granular cohort data^{[40][41]}. Moreover, it accelerated the institutional standardization of global health metrics, catalyzed open data infrastructures, and validated the importance of population-level reasoning in times of uncertainty.

Today, a convergence of factors, including standardized longitudinal data, widespread open access, and maturing computational methods, suggests that the conditions are in place for a renewed form of

ecological inquiry. Yet the question remains: can ecological study evolve beyond its historical limitations to become a robust, reproducible, and scalable scientific framework? This is not only possible, but necessary. Through the introduction of macro association analysis (MA-analysis), the use of Three Stars Global Macro Association-analysis (TSGMA), and the construction of MA networks, ecological study can be transformed from a loose design category into a rigorous engine of exploratory discovery, capable of scanning structured macro-data for hypothesis generation and policy prioritization^[13].

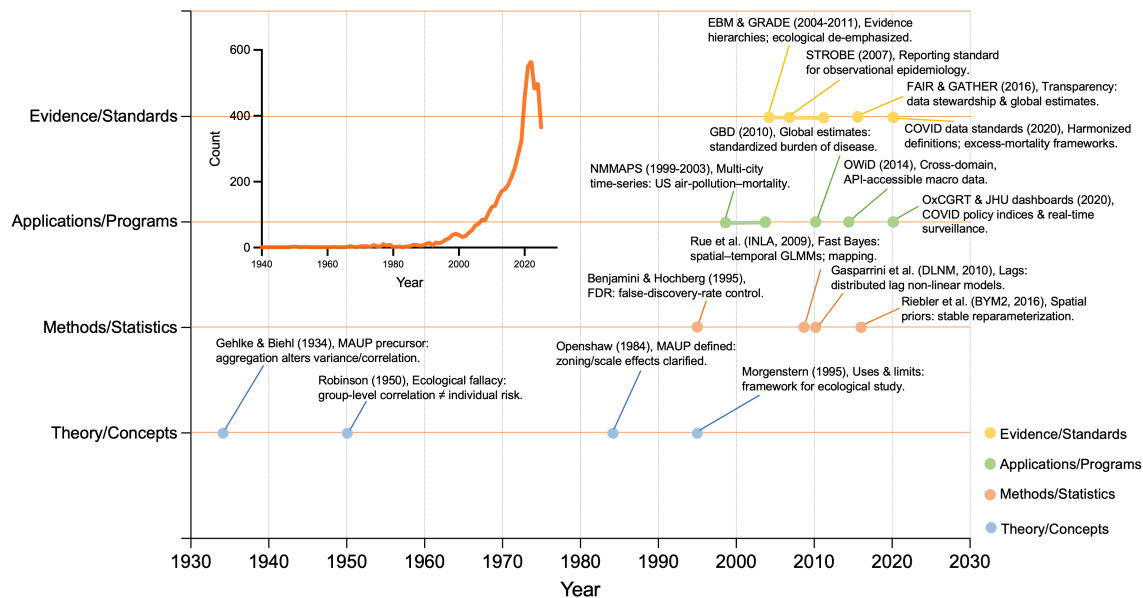


Figure 1. The development of ecological study. Publication counts are based on a rough PubMed search using “ecological study” as the title term. Since its introduction, the design has not generated a large volume of publications. Although the number of studies has shown a steady increase since the 1990s, the overall output has remained relatively modest. The emergence of COVID-19 prompted a surge of ecological studies that relied on population-level data to examine the pandemic, yet this expansion was largely confined to COVID-19 related applications rather than reflecting a broader methodological adoption. Following the peak of the pandemic, the number of publications has again shown a declining trend.

3. Unprecedented foundations: data, computation, and the new horizon for ecological study

3.1. Data explosion and infrastructural maturity

Data has always been the limiting substrate of ecological analysis. As a method relying on group-level statistics, ecological study is fundamentally constrained by the availability, coverage, granularity, and standardization of population-level data^{[1][4]}. For decades, such data was sparse, fragmented, and inconsistent across regions and time, which rendered ecological analysis narrow in scope and low in reproducibility^{[1][4]}.

The past decade, however, has witnessed a profound reversal. National statistical agencies, international organizations, and open-science platforms have converged toward standardized, high-resolution, longitudinal data systems. Massive, structured repositories, such as Our World in Data (<https://ourworldindata.org/>), FAOSTAT (<https://www.fao.org/statistics/en>), WHO Global Health Observatory (<https://www.who.int/data/gho>), and the World Bank Open Data (<https://databank.worldbank.org/databases>), now offer harmonized population-level variables across domains: health, diet, economics, climate, education, agriculture, governance, and beyond.

Moreover, these data are no longer scattered across PDF reports or proprietary silos. APIs, bulk CSV downloads, and machine-readable metadata have enabled automated, reproducible ingestion pipelines at scale. For ecological study, this constitutes a foundational shift: what was once a data-starved method is now data-rich. Yet paradoxically, most of these databases are limited to relatively simple displays of their contents, with few offering built-in systematic analyses or dynamically updating their analytical frameworks as new data are added. Only a small fraction is routinely analyzed, and even then, typically in domain-specific silos. The potential for cross-domain, cross-national, time-aware macro analysis remains largely untapped. This abundance of data alone does not guarantee analytic insight, but it does make ecological study technically feasible, globally scalable, and potentially transformative if supported by appropriate analytical frameworks.

3.2. Computational and algorithmic advances

Historically, ecological analysis depended on basic bivariate correlation and linear regression, often with minimal adjustment for confounding or structure^{[3][5][22][42]}. This simplicity was both a strength, as it

permitted accessibility and interpretability, and a fatal weakness, as ecological studies were frequently dismissed as crude, over confounded, and methodologically fragile^{[2][4]}.

Today, that limitation is far less constraining. The computational landscape has been radically transformed. With the rise of cloud computing, GPU acceleration, distributed storage, and scalable statistical libraries, researchers can now process and analyze tens of thousands of variable pairs across countries and timepoints in minutes rather than months^{[43][44][45]}. This enables exhaustive screening, bootstrapped replication, rolling window validation, and lag-aware testing at population scale.

Crucially, advances in compositional data analysis (for example, CLR transformation), robust estimation, partial correlation networks, and causal graph theory provide tools to overcome many of ecological study's classical pitfalls^{[46][47][48][49]}. Mixed models, hierarchical regression, and Bayesian priors allow structured control of multilevel confounding and regional heterogeneity^[50]. Ecological analysis, once methodologically static, is now technically modular, extensible, and statistically defensible.

Perhaps most transformative is the incorporation of Artificial Intelligence (AI)-driven workflows. Large language models and automated code synthesis now accelerate data wrangling, documentation, and sensitivity analysis^{[51][52][53][54]}. AutoML platforms and explainable AI can prioritize association candidates, detect non-linear structures, and expose latent interaction patterns, while at the same time preserving interpretability constraints that are critical for ecological reasoning^{[47][55]}.

The combination of scalable computation and intelligent automation does not merely improve existing ecological methods; it redefines them. In this context, the barrier to performing macro scale, high dimensional, and robust ecological analysis is no longer technical but conceptual. What is needed is a clear framework that governs discovery without overstating causality. This is the role assigned to TSGMA and MA networks, which are introduced in the next section.

4. The MA-analysis, MA networks, and ecological study

Recently, a methodological innovation has been proposed that fundamentally reshapes the analytical paradigm of ecological study: the MA-analysis^[13]. This approach is centered on incorporating as many variables as possible in order to perform large-scale, cross-field exploratory analyses and to detect potential association networks at a macro scale (Figure 2). Its primary goal is to construct networks of associations within and across domains, to uncover hidden linkages, and to provide a broad framework that can guide future research and disciplinary development. Ultimately, MA-analysis aims to build MA

networks and even pan-domain MA networks, connecting macro-level events across diverse dimensions of human society, the environment, and the economy, thereby enabling a deeper understanding of both societal progress and planetary evolution.

By complementing variables across dimensions, MA-analysis helps overcome one of the most persistent limitations of traditional ecological study: the distortion of results by unmeasured or dominant confounders in point-to-point analyses^{[4][5][28]}. Constructing single-domain MA networks transforms ecological study from the analysis of simple linear associations into multidimensional mapping within a domain. Extending this framework to cross-domain MA networks elevates ecological study from two-dimensional surfaces to multidimensional structures, fundamentally altering its scope and significance. For ecological study to achieve true conceptual breakthroughs, it must be integrated with the perspective of MA, thereby situating findings within higher-dimensional contexts.

The conceptual reach of MA-analysis should extend beyond population-level data. In principle, it could be applied at multiple scales, including the molecular level, for example by quantifying associations between the human genome and planetary ecology or between species level genomes and ecosystem development. By building such comprehensive relational networks, MA-analysis could ultimately support the creation of a “virtual Earth system”, analogous to the concept of a “virtual cell,” where diverse events are computationally integrated to reveal emergent laws of interaction^{[56][57][58][59]}. Just as genomics has enabled massively complex computations to quantify patterns across billions of bases, an Earth-scale MA framework could quantify virtually all recorded events, including human health, ecological, economic, and social data, and compute their interdependencies. The resulting networks would not only allow retrospective insights into human and planetary history, but also provide a foundation for forecasting future trajectories.

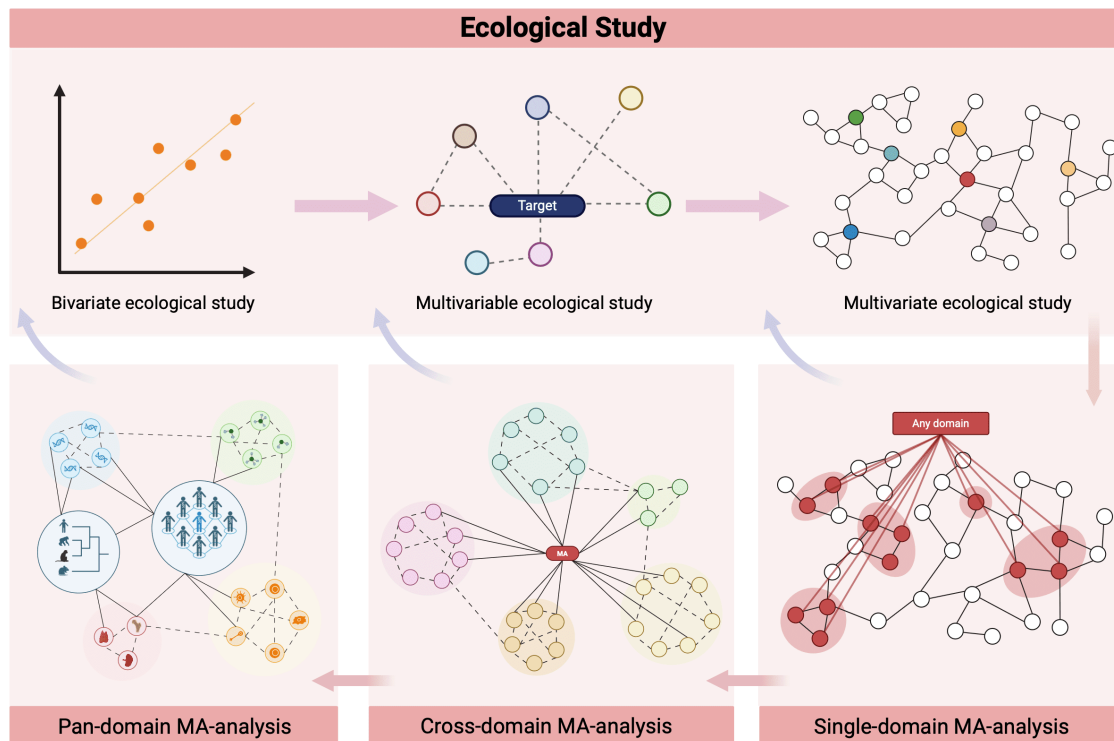


Figure 2. From traditional ecological study to MA-analysis. Conventional ecological studies, typically limited to pairwise associations, can be progressively extended to multivariable and multivariate analyses that capture multiple associations simultaneously, and further to the construction of small-scale local networks. When these localized analyses are expanded to encompass an entire disciplinary domain, they form the basis for single-domain MA networks. With further integration, such frameworks can evolve into cross-domain MA networks that span multiple fields, and ultimately into pan-domain MA networks that capture associations across the full spectrum of knowledge. At each stage of this progression, the constructed networks can be revisited from a higher level of abstraction, allowing earlier results to be re-examined with greater precision and contextual understanding.

Moreover, the development of the TSGMA framework offers ecological study a systematic approach for conducting MA-analysis^{[13][60]}. TSGMA extends conventional correlation analysis into three tiers of evaluation, classifying associations by their robustness and interpretability. One-star analysis functions much like traditional ecological study, identifying preliminary correlations between variables. Two-star analysis incorporates key representative confounders, such as Gross Domestic Product (GDP) per capita in economic analyses, through partial correlation, thereby testing whether associations remain stable after adjustment. Because over-controlling for multiple variables can reduce interpretability and

precision, extracting the most influential confounders is essential. Three-star analysis introduces the temporal dimension, examining whether variables co-vary over time and whether trends in one variable mirror fluctuations in another. Associations that remain consistent across these tiers can then be considered progressively stronger and more reliable.

In network construction, these three levels correspond to one-star, two-star, and three-star MA networks. For relatively modest sets of variables, such as those confined to a specific field with less variables, TSGMA provides accurate results and robust inference. However, as the dimensionality of MA-analysis expands to encompass entire single-domain or cross domains (for example, complete lifestyle profiles against all disease groups), applying full TSGMA at scale becomes computationally prohibitive. Thus, in practice, MA-analysis and TSGMA must be applied in a complementary and staged manner. When the primary aim is to determine accurate associations for a limited number of variables, TSGMA is appropriate. Conversely, when constructing large-scale MA networks, initial analyses may rely primarily on one-star TSGMA to build the preliminary network framework. Such a network, though exploratory, is meaningful at the macro level and inherently complementary across variables. Higher-order refinements can then be introduced gradually by incorporating two-star and three-star analyses, rather than attempting to complete full TSGMA prior to network construction (Figure 3).

The resulting preliminary MA networks provide a foundational scaffold encompassing all variables. Any subset of this network can subsequently be extracted for more targeted TSGMA or for domain-specific, fine-grained investigations. In this way, MA-analysis and TSGMA together offer a scalable, iterative pathway: from broad exploratory mapping of MA to progressively refined, confounder-adjusted, and temporally validated networks.

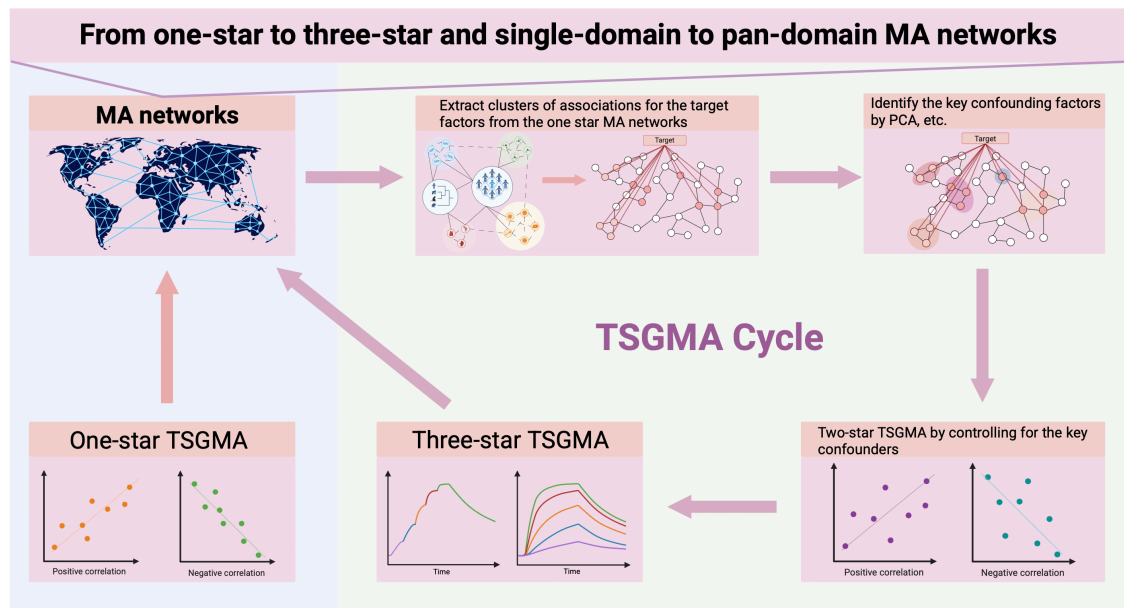


Figure 3. Construction of MA networks. Using one-star TSGMA, it is possible to construct MA networks ranging from single-domain to pan-domain scales. Once such networks are established, either specific target variables or the entire set of variables can be extracted for subsequent analyses. After extracting a subnetwork centered on the target variable, we can further analyze it to identify the key confounding factors influencing this subnetwork, and subsequently apply two-star TSGMA with adjustment for these confounders. Further application of three-star TSGMA introduces temporal validation, thereby establishing higher-level and more robust associations. These enhanced results are then reintegrated into the MA networks, progressively advancing them into two- to three-star MA networks. This iterative process constitutes an MA cycle, through which networks are continuously refined and ultimately consolidated into a stable and reliable framework of MA.

5. Challenges in ecological study and MA network construction

As outlined above, confounding remains one of the most intractable limitations of ecological study and is a key factor constraining its development. Even in the two-star and three-star tiers of TSGMA, confounders can only be partially controlled. For example, in analyses of diet and non-high-density lipoprotein cholesterol (non-HDL-C) level previously, the observed positive association between vegetable intake and non-HDL-C level persist across one-star and two-star analyses, which likely reflects the influence of residual and unmeasured confounding^[13]. In conventional epidemiology as well, complete control of confounders is rarely feasible; most studies adjust only for a limited set of general

covariates such as body mass index, smoking, alcohol consumption, or educational level, while many latent confounders remain unknown^{[61][62][63][64]}. Although these standard variables may capture a large portion of confounding, they cannot fully eliminate bias^{[61][62]}. Nevertheless, as systematic one-star MA networks are constructed and iteratively refined through the MA cycle, the influence of confounders can be progressively diluted, improving the quality and reliability of the resulting associations. In large-scale MA settings, as the number and diversity of variables increase, the accuracy of derived relationships can approach levels unattainable in smaller-scale analyses.

As noted earlier, ecological study is also vulnerable to spurious associations driven by collinearity and other structural factors^{[4][5]}. Yet an exclusive focus on direct associations may obscure the value of these spurious links. Even apparent false associations can signal underlying, meaningful processes; they represent connections between variables that, while not directly causal, may still point to latent common determinants. By aggregating both true and spurious associations into large-scale networks, events can be grouped into meaningful clusters and examined collectively. Further internal analysis can then reveal which linkages are artifacts and which reflect genuine causal relationships. Importantly, once a sufficiently large one-star MA network is assembled, it begins to exhibit a self-correcting property: through clustering and internal comparisons, the network itself can help identify the factors responsible for spurious associations, highlight candidate causal pathways, and iteratively refine its own structure.

Because the construction of MA networks involves enormous data volumes, computational demands, and disciplinary breadth, it constitutes a major undertaking requiring extensive collaboration across fields and substantial investment of human and financial resources. Yet advances in computer science, the growing availability of open data, and the inherent simplicity of MA-analysis make this endeavor feasible. With the support of artificial intelligence, both the labor and cost required for such large-scale integration can be markedly reduced. At the same time, decisions regarding how special variables are defined, quantified, and incorporated into these networks, as well as how molecular level data will eventually be linked with macro level data, require systematic planning and standardization. Population units can also be progressively refined: from national or regional aggregates down to subnational units or even villages, enabling subgroup and diversity analyses. Such finer-grained stratification increases the number of groups, the diversity of comparisons, and the potential for nuanced inference, but it also requires balancing subgroup size against the precision and stability of results.

Within the TSGMA framework itself, further challenges arise. In two-star analyses, the identification of confounders must be precise. In the context of large MA networks, there will inevitably be numerous

correlated potential confounders. Simply including all of them indiscriminately would obscure the original associations^{[13][65][66][67][68]}. Instead, it is necessary to identify the small subset of variables that exert the greatest influence and account for the majority of confounding. Some variables may display atypical patterns, appearing uncorrelated under one-star analysis but becoming associated after proper adjustment, which highlights the need to reconsider how such variables are analyzed and incorporated. Without this nuance, traditional TSGMA could overlook variables that are intrinsically related to the outcomes of interest. In three-star analyses, additional complexity arises from time-series effects, unpredictable fluctuations, and subtle lagged responses. Because each population unit requires its own time-series computation, the computational burden can be substantial. Developing rapid and accurate time series matching methods, potentially supported by artificial intelligence, will therefore be essential for scaling TSGMA to the three-star level.

6. Future directions of ecological study

When ecological study is reconsidered through the lens of MA networks, it can, for the first time, become truly “ecological.” Just as ecosystems are composed of populations yet operate not through isolated events but through interlinked processes that together sustain system dynamics, so too should ecological study move beyond the analysis of single associations. Instead, it should be conceived as the study of interconnected networks of events that collectively shape population- and planetary-level outcomes.

Only when the number of variables and associations reaches a sufficiently large scale can ecological study genuinely reflect the complexity implied by its name. The future of ecological study, therefore, lies in its transformation: from simple analyses of isolated associations, or its use as a subsidiary component within other study designs, toward systematic MA-analysis within specific domains; from there, to integrative frameworks spanning multiple domains; and ultimately, to the construction of pan-domain networks. At its core, the ultimate goal of MA-analysis and MA networks is to quantify the interconnections among all events in the world, from the molecular level to humanity and even the planetary scale, thereby revealing the full spectrum of linkages and constructing a comprehensive network of all events. Such networks hold the potential not only to illuminate the pathways by which human societies and ecosystems have developed, but also to provide a framework for anticipating future trajectories, thereby helping us to understand both where we come from and where we may be heading.

7. Conclusion

Ecological study, once marginalized as a method of limited precision, is now poised for a profound transformation. The convergence of standardized macro-level datasets, advances in computation, and the conceptual innovation of MA-analysis together provide the foundation for a new era. By scaling from one-star exploratory associations to multi-tiered, confounder-adjusted, and temporally validated networks, ecological study can evolve from simple pairwise comparisons into a robust framework for mapping the interconnected systems that shape human societies and the planet. The construction of single-domain or cross-domain MA networks, and ultimately pan-domain MA networks, offers not only a means of revisiting long-standing scientific questions with unprecedented depth but also a platform for identifying the key determinants of collective health, sustainability, and resilience. In doing so, ecological study regains its relevance as a truly ecological science: one capable of integrating diverse layers of information, from molecular to planetary, into a coherent architecture for understanding both our past and our possible futures.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Author Contributions

Hongyue Ma performed all analyses and wrote the manuscript.

Resource and Data Availability

This study did not generate any new samples or unique reagents and does not report original code. The data presented in this paper are available from the lead contact upon reasonable request. Requests for additional information, resources, or reagents should be directed to Hongyue Ma at mahongyue@westlake.edu.cn and will be fulfilled where appropriate.

References

1. ^{a, b, c, d}Morgenstern H (1995). "Ecologic Studies in Epidemiology: Concepts, Principles, and Methods." *Annu Rev Public Health*. 16:61–81. doi:[10.1146/annurev.pu.16.050195.000425](https://doi.org/10.1146/annurev.pu.16.050195.000425).
2. ^{a, b, c}Morgenstern H (1982). "Uses of Ecologic Analysis in Epidemiologic Research." *Am J Public Health*. 72:1336–1344. doi:[10.2105/ajph.72.12.1336](https://doi.org/10.2105/ajph.72.12.1336).
3. ^{a, b, c}Keys A (1957). "Diet and the Epidemiology of Coronary Heart Disease." *J Am Med Assoc*. 164:1912–1919. doi:[10.1001/jama.1957.62980170024007e](https://doi.org/10.1001/jama.1957.62980170024007e).
4. ^{a, b, c, d, e, f, g, h}Wakefield J (2008). "Ecologic Studies Revisited." *Annu Rev Public Health*. 29:75–90. doi:[10.1146/annurev.publhealth.29.020907.090821](https://doi.org/10.1146/annurev.publhealth.29.020907.090821).
5. ^{a, b, c, d, e}Greenland S, Morgenstern H (1989). "Ecological Bias, Confounding, and Effect Modification." *Int J Epidemiol*. 18:269–274. doi:[10.1093/ije/18.1.269](https://doi.org/10.1093/ije/18.1.269).
6. [^]LeCun Y, Bengio Y, Hinton G (2015). "Deep Learning." *Nature*. 521:436–444. doi:[10.1038/nature14539](https://doi.org/10.1038/nature14539).
7. [^]Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data*. 3:160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
8. [^]Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017). "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*. 202:18–27. doi:[10.1016/j.rse.2017.06.031](https://doi.org/10.1016/j.rse.2017.06.031).
9. [^]Bycroft C, Freeman C, Petkova D, Band, G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. (2018). "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature*. 562:203–209. doi:[10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z).
10. [^]Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C, Rod s-Guirao L (2021). "A Global Database of COVID-19 Vaccinations." *Nature Human Behaviour*. 5:947–953. doi:[10.1038/s41562-021-01122-8](https://doi.org/10.1038/s41562-021-01122-8).
11. [^]Tatem AJ (2017). "WorldPop, Open Data for Spatial Demography." *Sci Data*. 4:170004. doi:[10.1038/sdata.2017.174](https://doi.org/10.1038/sdata.2017.174).
12. [^]Anonymous (2020). "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019." *Lancet*. 396:1204–1222. doi:[10.1016/S0140-6736\(20\)30925-9](https://doi.org/10.1016/S0140-6736(20)30925-9).

13. ^{a, b, c, d, e, f}Ma H (2025). "TSGMA: Identification of Macro Associations from Global Data to Build Global MA Networks." *medRxiv*. 2025.2006.2009.25329257. doi:[10.1101/2025.06.09.25329257](https://doi.org/10.1101/2025.06.09.25329257).
14. ^ΔLangmuir AD (1963). "The Surveillance of Communicable Diseases of National Importance." *N Engl J Med*. 268:182–192. doi:[10.1056/nejm196301242680405](https://doi.org/10.1056/nejm196301242680405).
15. ^ΔRobinson W (2009). "Ecological Correlations and the Behavior of Individuals*." *Int J Epidemiol*. 38:337–341. doi:[10.1093/ije/dyn357](https://doi.org/10.1093/ije/dyn357).
16. ^ΔFrost WH (1995). "The Age Selection of Mortality from Tuberculosis in Successive Decades. 1939." *Am J Epidemiol*. 141:4–9; discussion 3. doi:[10.1093/oxfordjournals.aje.a117343](https://doi.org/10.1093/oxfordjournals.aje.a117343).
17. ^ΔComstock GW (2001). "Cohort Analysis: W.H. Frost's Contributions to the Epidemiology of Tuberculosis and Chronic Disease." *Soz Praventivmed*. 46:7–12. doi:[10.1007/bf01318793](https://doi.org/10.1007/bf01318793).
18. ^ΔRose G (1985). "Sick Individuals and Sick Populations." *Int J Epidemiol*. 14:32–38. doi:[10.1093/ije/14.1.32](https://doi.org/10.1093/ije/14.1.32).
19. ^ΔOmran AR (2005). "The Epidemiologic Transition: A Theory of the Epidemiology of Population Change. 1971." *Milbank Q*. 83:731–757. doi:[10.1111/j.1468-0009.2005.00398.x](https://doi.org/10.1111/j.1468-0009.2005.00398.x).
20. ^ΔKuller LH (1999). "Circular Epidemiology." *Am J Epidemiol*. 150:897–903. doi:[10.1093/oxfordjournals.aje.a010097](https://doi.org/10.1093/oxfordjournals.aje.a010097).
21. ^ΔAnonymous (2005). "The Macro-Epidemiology of Influenza Vaccination in 56 Countries, 1997–2003." *Vaccine*. 23:5133–5143. doi:[10.1016/j.vaccine.2005.06.010](https://doi.org/10.1016/j.vaccine.2005.06.010).
22. ^{a, b}Armstrong B, Doll R (1975). "Environmental Factors and Cancer Incidence and Mortality in Different Countries, with Special Reference to Dietary Practices." *Int J Cancer*. 15:617–631. doi:[10.1002/ijc.2910150411](https://doi.org/10.1002/ijc.2910150411).
23. ^ΔGray AM (1982). "Inequalities in Health. The Black Report: A Summary and Comment." *Int J Health Serv*. 12:349–380. doi:[10.2190/xxmm-jmqu-2a7y-hx1e](https://doi.org/10.2190/xxmm-jmqu-2a7y-hx1e).
24. ^ΔLave LB, Seskin EP (1970). "Air Pollution and Human Health." *Science*. 169:723–733. doi:[10.1126/science.169.3947.723](https://doi.org/10.1126/science.169.3947.723).
25. ^ΔYerushalmy J, Hilleboe HE (1957). "Fat in the Diet and Mortality from Heart Disease; A Methodologic Note." *NY State J Med*. 57:2343–2354.
26. ^ΔPope CA 3rd (1991). "Respiratory Hospital Admissions Associated with PM10 Pollution in Utah, Salt Lake, and Cache Valleys." *Arch Environ Health*. 46:90–97. doi:[10.1080/00039896.1991.9937434](https://doi.org/10.1080/00039896.1991.9937434).
27. ^ΔRobinson WS (1950). "Ecological Correlations and the Behavior of Individuals." *Int J Epidemiol*. 15:351–357.
28. ^{a, b}Piantadosi S, Byar DP, Green SB (1988). "The Ecological Fallacy." *Am J Epidemiol*. 127:893–904. doi:[10.1093/oxfordjournals.aje.a114892](https://doi.org/10.1093/oxfordjournals.aje.a114892).

29. [△]Pekkanen J, Pearce N (2001). "Environmental Epidemiology: Challenges and Opportunities." *Environ Health Perspect.* **109**:1–5. doi:[10.1289/ehp.011091](https://doi.org/10.1289/ehp.011091).
30. [△]Subramanian SV, Jones K, Kaddour A, Krieger N (2009). "Revisiting Robinson: The Perils of Individualistic and Ecologic Fallacy." *Int J Epidemiol.* **38**:342–360; author reply 370–343. doi:[10.1093/ije/dyn359](https://doi.org/10.1093/ije/dyn359).
31. [△]Burns PB, Rohrich RJ, Chung KC (2011). "The Levels of Evidence and Their Role in Evidence-Based Medicine." *Plast Reconstr Surg.* **128**:305–310. doi:[10.1097/PRS.0b013e318219c171](https://doi.org/10.1097/PRS.0b013e318219c171).
32. [△]Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH (2011). "GRADE Guidelines: 3. Rating the Quality of Evidence." *J Clin Epidemiol.* **64**:401–406. doi:[10.1016/j.jclinepi.2010.07.015](https://doi.org/10.1016/j.jclinepi.2010.07.015).
33. [△]Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (1996). "Evidence Based Medicine: What It Is and What It Isn't." *Bmj.* **312**:71–72. doi:[10.1136/bmj.312.7023.71](https://doi.org/10.1136/bmj.312.7023.71).
34. [△]Dong E, Du H, Gardner L (2020). "An Interactive Web-Based Dashboard to Track COVID-19 in Real Time." *Lancet Infect Dis.* **20**:533–534. doi:[10.1016/s1473-3099\(20\)30120-1](https://doi.org/10.1016/s1473-3099(20)30120-1).
35. [△]Hasell J, Mathieu E, Beltekian D, Macdonald B, Giattino C, Ortiz-Ospina E, Roser M, Ritchie H (2020). "A Cross-Country Database of COVID-19 Testing." *Sci Data.* **7**:345. doi:[10.1038/s41597-020-00688-8](https://doi.org/10.1038/s41597-020-00688-8).
36. [△]Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C, Rodés-Guirao L (2021). "A Global Database of COVID-19 Vaccinations." *Nat Hum Behav.* **5**:947–953. doi:[10.1038/s41562-021-01122-8](https://doi.org/10.1038/s41562-021-01122-8).
37. [△]Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, Whittaker C, Zhu H, Berah T, Eaton JW, et al. (2020). "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe." *Nature.* **584**:257–261. doi:[10.1038/s41586-020-2405-7](https://doi.org/10.1038/s41586-020-2405-7).
38. [△]Hsiang S, Allen D, Annan-Phan S, Bell K, Bolliger I, Chong T, Druckenmiller H, Huang LY, Hultgren A, Krasovich E, et al. (2020). "The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic." *Nature.* **584**:262–267. doi:[10.1038/s41586-020-2404-8](https://doi.org/10.1038/s41586-020-2404-8).
39. [△]Brauner JM, Mindermann S, Sharma M, Johnston D, Salvatier J, Gavenčiak T, Stephenson AB, Leech G, Altman G, Mikulík V, et al. (2021). "Inferring the Effectiveness of Government Interventions against COVID-19." *Science.* **371**. doi:[10.1126/science.abd9338](https://doi.org/10.1126/science.abd9338).
40. [△]Björk J, Modig K, Kahn F, Ahlbom A (2021). "Revival of Ecological Studies during the COVID-19 Pandemic." *Eur J Epidemiol.* **36**:1225–1229. doi:[10.1007/s10654-021-00830-9](https://doi.org/10.1007/s10654-021-00830-9).
41. [△]Chen YC, Chow-In Ko P, Lee WC, Lien WC (2021). "Ecological Fallacy in COVID-19 Epidemiological Inference: Influenza Vaccination Rate as an Example." *J Formos Med Assoc.* **120**:1655–1656. doi:[10.1016/j.jfma.2021.03.011](https://doi.org/10.1016/j.jfma.2021.03.011).

42. [△]Dufault B, Klar N (2011). "The Quality of Modern Cross-Sectional Ecologic Studies: A Bibliometric Review." *Am J Epidemiol*. 174:1101–1107. doi:[10.1093/aje/kwr241](https://doi.org/10.1093/aje/kwr241).
43. [△]Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Moore R (2017). "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment*. 202.
44. [△]Meng X, Bradley J, Yavuz B, Sparks E, Talwalkar A (2015). "MLlib: Machine Learning in Apache Spark." *Journal of Machine Learning Research*. 17:1235–1241.
45. [△]Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, Vassilakis T (2010). "Dremel: Interactive Analysis of Web-Scale Datasets." *Communications of the Acm*. 3:114–123.
46. [△]Aitchison J (1982). "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B (Methodological)*. 44:139–160. doi:[10.1111/j.2517-6161.1982.tb01195.x](https://doi.org/10.1111/j.2517-6161.1982.tb01195.x).
47. [△]^bFriedman J (2008). "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics*. 9:432–441.
48. [△]Epskamp S, Fried EI (2018). "A Tutorial on Regularized Partial Correlation Networks." *Psychological Methods*.
49. [△]Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*. 35:279–300.
50. [△]Diez-Roux AV (2000). "Multilevel Analysis in Public Health Research." *Annu Rev Public Health*. 21:171–192. doi:[10.1146/annurev.publhealth.21.1.171](https://doi.org/10.1146/annurev.publhealth.21.1.171).
51. [△]Liu L, Hasegawa S, Sampat SK, Xenochristou M, Chen W-P, Kato T, Kakibuchi T, Asai T (2024). "AutoDW: Automatic Data Wrangling Leveraging Large Language Models." *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. Association for Computing Machinery.
52. [△]Dvivedi SS, Vijay, V, Pujari SLR, Lodh S, Kumar D (2024). "A Comparative Analysis of Large Language Models for Code Documentation Generation." *Proceedings of the 1st ACM International Conference on AI-Powered Software*. Association for Computing Machinery.
53. [△]Xu W, Luo G, Meng W, Zhai X, Zheng K, Wu J, Li Y, Xing A, Li J, Li Z, et al. (2025). "MRAgent: An LLM-Based Automated Agent for Causal Knowledge Discovery in Disease via Mendelian Randomization." *Brief Bioinform*. 26. doi:[10.1093/bib/bbaf140](https://doi.org/10.1093/bib/bbaf140).
54. [△]Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B (2024). "Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models." *Pharmacoecon Open*. 8:205–220. doi:[10.1007/s41669-024-00476-9](https://doi.org/10.1007/s41669-024-00476-9).

55. [△]Friedman JH (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. 29:1189–1232.
56. [△]Loew LM, Schaff JC (2001). "The Virtual Cell: A Software Environment for Computational Cell Biology." *Trends Biotechnol.* 19:401–406. doi:[10.1016/s0167-7799\(01\)01740-1](https://doi.org/10.1016/s0167-7799(01)01740-1).
57. [△]Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr, Assad-Garcia N, Glass JI, Covert MW (2012). "A Whole-Cell Computational Model Predicts Phenotype from Genotype." *Cell*. 150:389–401. doi:[10.1016/j.cell.2012.05.044](https://doi.org/10.1016/j.cell.2012.05.044).
58. [△]Hazeleger W, Aerts JPM, Bauer P, Bierkens MFP, Camps-Valls G, Dekker MM, Doblas-Reyes FJ, Eyring V, Finkenauer C, Grundner A, et al. (2024). "Digital Twins of the Earth with and for Humans." *Communications Earth & Environment*. 5:463. doi:[10.1038/s43247-024-01626-x](https://doi.org/10.1038/s43247-024-01626-x).
59. [△]Bauer P, Stevens B, Hazeleger W (2021). "A Digital Twin of Earth for the Green Transition." *Nature Climate Change*. 11:80–83. doi:[10.1038/s41558-021-00986-y](https://doi.org/10.1038/s41558-021-00986-y).
60. [△]Ma H, Qi X (2023). "Red Meat Consumption and Cancer Risk: A Systematic Analysis of Global Data." *Foods*. 12. doi:[10.3390/foods12224164](https://doi.org/10.3390/foods12224164).
61. [△]VanderWeele TJ, Ding P (2017). "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Ann Intern Med*. 167:268–274. doi:[10.7326/m16-2607](https://doi.org/10.7326/m16-2607).
62. [△]Fewell Z, Davey Smith G, Sterne JA (2007). "The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study." *Am J Epidemiol*. 166:646–655. doi:[10.1093/aje/kwm165](https://doi.org/10.1093/aje/kwm165).
63. [△]Nørgaard M, Ehrenstein V, Vandenbroucke JP (2017). "Confounding in Observational Studies Based on Large Health Care Databases: Problems and Potential Solutions – A Primer for the Clinician." *Clin Epidemiol*. 9: 185–193. doi:[10.2147/clep.S129879](https://doi.org/10.2147/clep.S129879).
64. [△]Wallach JD, Serghiou S, Chu L, Egilman AC, Vasiliou V, Ross JS, Ioannidis JPA (2020). "Evaluation of Confounding in Epidemiologic Studies Assessing Alcohol Consumption on the Risk of Ischemic Heart Disease." *BMJ Medical Research Methodology*. 20:64. doi:[10.1186/s12874-020-0914-6](https://doi.org/10.1186/s12874-020-0914-6).
65. [△]Schisterman EF, Cole SR, Platt RW (2009). "Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies." *Epidemiology*. 20:488–495. doi:[10.1097/EDE.0b013e3181a819a1](https://doi.org/10.1097/EDE.0b013e3181a819a1).
66. [△]VanderWeele TJ (2019). "Principles of Confounder Selection." *Eur J Epidemiol*. 34:211–219. doi:[10.1007/s10654-019-00494-6](https://doi.org/10.1007/s10654-019-00494-6).
67. [△]Greenland S, Pearl J, Robins JM (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology*. 10:37–48.

68. [△]Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006). "Variable Selection for Propensity Score Models." *Am J Epidemiol.* **163**:1149–1156. doi:[10.1093/aje/kwj149](https://doi.org/10.1093/aje/kwj149).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.