Submitted to Statistical Science

Heterogeneity in Statistics: A Conceptual and Methodological Review

Zhanshan (Sam) Ma, Shu Liu, and Aaron M. Ellison

Abstract. Heterogeneity—the presence of meaningful variation across observations, in models, and in inferences—is a foundational concept in statistics that has many meanings. This review synthesizes the evolution of the meanings, methodologies, and interpretations of the four dominant and interconnected types of heterogeneity: (1) *heteroscedasticity* (non-constant variance), historically treated as a nuisance but now modeled as substantive information in fields from finance to ecology; (2) *generalized heterogeneity* (i.e., variation in parameters or effects), addressed via Gaussian graphical models and frailty-based network models that uncover latent subgroup structures; (3) *frailty* (unobserved heterogeneity), whose effects are uniquely captured in survival analysis through frailty and accelerated failure time models. and (4) *covariance and dependence* (i.e., structured relationships among observations), formalized theoretically by Price's Equation and handled practically by mixed models and generalized estimating equations (GEEs). These four ways in which heterogeneity is used in contemporary statistical research illustrate a progression from controlling variation to learning from it, and can be embedded in a broader ontology (hierarchical taxonomy) of types and subtypes of heterogeneity that span observational, model-based, and inferential domains. Mixed-effects models, Bayesian methods, causal forests, and AI-enhanced survival models are unifying platforms for jointly modeling different types of heterogeneity. Examples from applied sciences that use statistics extensively illustrate how heterogeneity has been transformed from a statistical nuisance into a source of scientific discovery. Advances in estimation, diagnostics, and causal interpretation have made meta-analysis into an exemplar for quantifying and investigating between-study heterogeneity. We conclude with practical guidelines for diagnosing, modeling, and reporting heterogeneity, and identify future challenges for dealing with heterogeneity in causal attribution, high-dimensional data, interpretability, and interdisciplinary integration. Embracing heterogeneity as a fundamental feature of complex systems represents a maturation of statistical science whose application from generalizable models to personalized medicine can provide more nuanced insights into the interpretation of complex datasets.

Key words and phrases: AI and Machine Learning, Covariance and dependence, Frailty modeling, Heterogeneity, Heteroscedasticity, Meta-analysis, Survival Analysis.

Zhanshan (Sam) Ma is a Professor and Principal Investigator at the Computational Biology and Medical Ecology, Biostatistics and Image Genetics Lab, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China and in the Microbiome Medicine and Advanced AI Lab, Kunming, China (email: ma@vandals.uidaho.edu). Shu Liu is a Professor and Principal Investigator at the Biostatistics and Image Genetics Lab, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China (email: liushu@mail.kiz.ac.cn) Aaron M. Ellison (ORCID: 0000-0003-4151-6081) is the Senior Research Fellow, Emeritus, at Harvard University, Harvard Forest, Petersham, Massachusetts 01366, USA, and the Founding Principal of Sound Solutions for Sustainable Science, Boston, Massachusetts 02135, USA (email: aaron@ssforss.com).

1. INTRODUCTION: THE MULTIFACETED NATURE OF HETEROGENEITY

Heterogeneity is a foundational concept in statistics that in general refers to the presence of meaningful variation among observations or groups in models and inferences. In practice, however, heterogeneity has taken on many different meanings that can be systematically organized into three conceptual domains: *heterogeneity in observations* (data), which includes variation in measurements, distributions, and spatiotemporal structures; *heterogeneity in models*, which includes differences in parameters, functional forms, unobserved frailty, and error variances; and *heterogeneity in inferences*, which arises from variability in results across studies or analytical methods. A simplified ontology (hierarchical taxonomy) of the different domains, types, and subtypes of heterogeneity (Table 1) differentiates the meanings of heterogeneity and the different methods that statisticians have developed to deal with them.

Four interrelated types of heterogeneity are paramount in statistical methodology and inference: (i) *Heteroscedasticity* a.k.a. heterogeneity in variances (the second moment of the error distribution; Pagan and Breusch, 1979; White, 1980); (ii) *"Generalized" Heterogeneity* in model parameters or effects (e.g., varying treatment effects across subgroups; Fisher, 1925; Raudenbush and Bryk, 2002); (iii) *Frailty* or unobserved heterogeneity that arises from unobserved individual susceptibilities or latent traits, and which is often modeled with random effects, particularly in survival and network analyses (Muthén, 2004; Vaupel, Manton and Stallard, 1979); and (iv) *Covariance and Dependence* that results in heterogeneity in the relationships between observations (e.g., non-zero off-diagonals in the variance-covariance matrix; Cressie, 1993; Laird and Ware, 1982).

Although there are other types of heterogeneity (Table 1), we focus our review of heterogeneity on these four types for three reasons. First, they represent active areas of methodological developments aimed at shifting heterogeneity from a nuisance parameter to a source of insight (e.g., Bland and Altman, 1986; Higgins and Thompson, 2002). Second, subtypes of heterogeneity in observations (e.g., spatial, temporal, distributional shape; Table 1) often are interpreted differently in different basic and applied scientific fields, such as ecology, climatology, economics, and epidemiology, that extensively use and often develop statistical methods. In those fields (and many others), heterogeneity is used most frequently to mean or measure difference or diversity (Shavit and Ellison, 2021). Examples include genetic variation in biology, socioeconomic disparity in the social sciences, historical contingency in the humanities, and metaphysical diversity in philosophy (Box 1). Many basic and applied research questions involving "heterogeneity" that originated in these fields have been addressed using well-established statistical methods (salient examples include Hill numbers and Réenyi entropy, kriging, and time-series models in Cressie, 1993; Gotelli and Ellison, 2012; Chao et al., 2014; Box et al., 2015; Dormann and Ellison, 2025).

Finally, in regression diagnostics, mixed-effects models, network analysis, survival modeling, and through recent advances in computational and AI-driven modeling, these four types of heterogeneity also have anchored major themes in the historical and contemporary development of statistical inference and models. These four types also come together in meta-analysis, in which quantifying and investigating heterogeneity plays a central role. Taken as a whole, a more nuanced understanding of heterogeneity can be used to trace a key evolutionary change in the development of statistical inference: from controlling variation to learning from it.

Box 1. The philosophical distinction between heterogeneity and diversity or difference, and its consequences for measuring them.

Shavit and Ellison (2021) identified a fundamental distinction between heterogeneity and measures of difference and diversity. They summarized the distinction between diversity (or difference) and heterogeneity in the aphorism "a zoo is diverse, whereas an ecosystem is heterogeneous." This framing distinguishes a *population*—a diverse aggregate of discrete, *non-interacting* objects—from a *collective*—a group characterized by joint processes and a common structure that arises from the integration of different constituent entities through their specific interactions *and* the resulting network of relationships within the collective and between the collective and its environment.

Measures of diversity or difference describe composition of the population (how many individuals of different kinds), whereas measures of heterogeneity should describe the emergent properties of a system's relational structure. Quantifying heterogeneity (*sensu* Shavit and Ellison, 2021) thus requires a methodological shift from counting entities or assessing indices of diversity (review and synthesis in Chao et al., 2014) to mapping and quantifying the structure and strength of connections between them. The latter necessitates analyses of weighted networks (see §3.2 and e.g., Box-Steffensmeier, Christenson and Morgan, 2018; Ma and Ellison, 2025).

Table 1: An ontology or hierarchical taxonomy for classifying different types and subtypes of statistical heterogeneity. We distinguish heterogeneity in data, models, and inference (**Domain**), **Types** and **Subtypes** of heterogeneity within each domain, and the commonly-used meanings (**Definitions**) and common **Methods** used (and key references for them) for each of the types and subtypes.

Domain	Type	Subtype	Definitions	Methods
Heterogeneity in Observations (Data)	**Sample / Population Heterogeneity**	*Between-Unit Heterogeneity*	Variation between different observational units (subjects, groups, clusters); quantified via between-subjects variance in ANOVA (Fisher, 1925; Scheffé, 1959)	1. Mixed / multilevel models (Raudenbush and Bryk, 2002) 2. ANOVA (Fisher, 1925) 3. ICC (Shrout and Fleiss, 1979)
		Within-Unit Heterogeneity	Variation within the same unit over time or conditions; reflects intra-individual variability or measurement error (Bland and Altman, 1999; Fisher, 1925)	1. Repeated measures ANOVA (Maxwell and Delaney, 2004) 2. Mixed models (Laird and Ware, 1982) 3. Bland-Altman analysis (Bland and Altman, 1999)
	Distributional Heterogeneity	*Location Heterogeneity*	Differences in central tendency (mean, median) across groups or distributions (Casella and Berger, 2024).	1. *t*-test / ANOVA (Student, 1908; Fisher, 1925) 2. Mann-Whitney / Kruskal-Wallis (Kruskal and Wallis, 1952) 3. Regression with group indicators (Gelman and Hill, 2006)
		Scale Heterogeneity	Differences in dispersion or variance across groups or distributions. (Lehmann and Romano, 2005).	1. Levene's test (Levene, 1960) 2. Bartlett's test (Bartlett, 1937) 3. Brown-Forsythe test (Brown and Forsythe, 1974)
		Shape Heterogeneity	Differences in distributional form: skewness, kurtosis, or modality (Casella and Berger, 2024).	1. Q-Q plots Wilk and Gnanadesikan (1968) 2. Kolmogorov-Smirnov test (Kolmogoroff, 1933) 3. Anderson-Darling test (Anderson and Darling, 1954)
	Spatiotemporal Heterogeneity	*Spatial Heterogeneity*	Variation structured across geographical or spatial coordinates (Cressie, 1993).	1. Moran's I or Geary's C (Moran, 1950; Cliff and Ord, 1973) 2. Kriging (Cressie, 1993) 3. Spatial regression (SAR, CAR) (Anselin, 1988)
		Temporal Heterogeneity	Variation structured across time; includes trends, cycles, and non-stationarity (Box et al., 2015).	1. Time series decomposition (STL) (Cleveland and Cleveland, 1990) 2. ARIMA models (Box et al., 2015) 3. Generalized additive models (GAMs) (Hastie and Tibshirani, 1990)

Continued on next page

Table 1 – *continued from previous page*

Domain	Type	Subtype	Definitions	Methods
		Spatiotemporal Heterogeneity	Combined variation across both space and time (Cressie and Wikle, 2011).	1. Spatiotemporal kriging (Cressie and Wikle, 2011) 2. Bayesian hierarchical spatiotemporal models (Banerjeee, Carlin and Gelfand, 2025) 3. INLA (Rue, Martino and Chopin, 2009)
Heterogeneity in Models	**Parameter Heterogeneity**	*Fixed Parameter Heterogeneity*	Systematic differences in model parameters across known, observed groups (Gelman and Hill, 2006)	1. ANCOVA / regression with interactions (Maxwell and Delaney, 2004) 2. Stratified analysis (Cochran, 1968) 3. Subgroup analysis (Wang et al., 2007)
		Random Parameter Heterogeneity	Parameters vary randomly across units according to a latent distribution (Raudenbush and Bryk, 2002).	1. Random coefficients models (Laird and Ware, 1982) 2. Hierarchical linear models (Raudenbush and Bryk, 2002) 3. Generalized estimating equations (GEEs) (Liang and Zeger, 1986)
	Structural Heterogeneity	*Functional Form Heterogeneity*	Different mathematical relationships apply to different subsets of data (Hansen, 2001).	1. Regime-switching models (Hamilton, 1989) 2. Piecewise and segmented regression (Muggeo, 2003) 3. Generalized additive mixed models (GAMMs) Wood (2017)
		Latent Class Heterogeneity	Data arise from a mixture of unobserved subpopulations with distinct models (McLachlan and Peel, 2000).	1. Finite mixture models (McLachlan and Peel, 2000) 2. Latent class analysis (LCA) (Lazarsfeld and Henry, 1968) 3. Model-based clustering (mclust) Fraley and Raftery (2002)
	Unobserved Heterogeneity	*Frailty Models*	Unmeasured factors causing dependence in survival or clustered data; shared latent frailty (Vaupel, Manton and Stallard, 1979).	1. Shared frailty models (Vaupel, Manton and Stallard, 1979) 2. Correlated frailty models (Yashin and Iachine, 1999) 3. Additive frailty models (Petersen, Andersen and Gill, 1996)
		Random Effects (General)	Unobserved unit-specific effects in panel or longitudinal data (Raudenbush and Bryk, 2002).	1. Panel data models (Fixed or random effects) (Wooldridge, 2010) 2. Mixed models for repeated measures (Laird and Ware, 1982) 3. Multilevel models Goldstein (2011)

Continued on next page

Table 1 – *continued from previous page*

Domain	Type	Subtype	Definitions	Methods
	Heteroscedasticity		Non-constant variance of errors across observations in a regression model (White, 1980).	1. Breusch-Pagan test (Pagan and Breusch, 1979) 2. White Test (White, 1980) 3. Weighted least squares (WLS) (Carroll and Ruppert, 1988)
Heterogeneity in Inference	**Meta-Analytic Heterogeneity**	*Statistical Heterogeneity*	Variability in effect sizes across studies beyond chance (Higgins and Thompson, 2002).	1. Cochran's Q (Cochran, 1954) 2. I^2 Statistic (Higgins and Thompson, 2002) 3. τ^2 DerSimonian and Laird (1986)
		Methodological Heterogeneity	Variability due to differences in study designs, populations, or interventions (Borenstein et al., 2009; Gurevitch et al., 2018)	1. Subgroup analysis by study design (Higgins and Thomas, 2024) 2. Meta-regression (Thompson and Higgins, 2002) 3. Sensitivity analysis (leave-one-out) (Viechtbauer and Cheung, 2010)
	Estimator Heterogeneity	*Estimator / Algorithm Heterogeneity*	Different statistical methods or algorithms yield different results from the same data (Ho et al., 2007).	1. Sensitivity analysis (Leamer, 1983) 2. Method (Ho et al., 2007) 3. Ensemble and consensus methods Breiman (1996)

2. HETEROSCEDASTICITY: FROM VIOLATION OF MODEL ASSUMPTIONS TO A FEATURE TO BE MODELED

Heteroscedasticity refers specifically to the variation in finite variance among random variables (i.e., observations) or sequences of random variables (i.e., groups of observations). Heteroscedasticity violates a key assumption of classical linear models such as the analysis of variance (ANOVA) and ordinary least-squares regression. Historically, heteroscedasticity was treated as a statistical nuisance to be tested for and corrected (e.g., by various transformations prior to analysis Box and Cox, 1964). However, the development of robust standard errors (White, 1980) and variance-function modeling within mixed-effects frameworks (Pinheiro and Bates, 2000) transformed heteroscedasticity from a problem to be fixed into a pattern to be understood and modeled. As a result, heterogeneity is now considered to be a substantive feature of interest in, for example, fields ranging from finance, where it is used to model and forecast market volatility (e.g., Bhambu et al., 2025), to ecology, where it reflects underlying "patchiness" or variability in environmental conditions (e.g., Dutilleul and Legendre, 1993; Veum et al., 2025). This transformed perspective enables has changed the types of questions that domain scientists ask, from "what is the average effect?" or "are there significant differences among groups?" to "does the intervention affect outcome consistency?" or "What variables predict differential variability?" The latter questions are particularly valuable in precision medicine, educational interventions, and making and evaluating evidence-based policies, where understanding not only central tendency and its expected variance, but also variation (and uncertainty) at different hierarchical scales (e.g., individuals, groups, populations) leads to better decisions.

2.1 Historical Development and Early Diagnostics

The problem of heteroscedasticity was first identified by Fisher (1925). His formulation of ANOVA required equality of variances across treatment groups (i.e., homoscedasticity) as a prerequisite for valid inference. Just over a decade after the introduction of ANOVA, Bartlett (1937) provided the first formal diagnostic test for homogeneity of variance (the eponymous Bartlett's test), although its computational complexity initially limited its use. Bartlett and Kendall (1946) subsequently introduced the logarithmic transformation of sample variances, enabling ANOVA-like decomposition of heterogeneous variances into structured components (e.g., group, time, residual effects). That paper represented an important conceptual shift: from merely testing for heterogeneity to analyzing its structure. Hartley (1950) addressed the practical barrier of computational complexity with his maximum F-ratio statistic ($F_{max} = s^2_{max}/s^2_{min}$), that could be compared to the critical value of the F_{max} distribution with (ν, ν) degrees of freedom (for which he provided appropriate look-up tables). Hartley (1950) also showed that his F_{max} test had only minimally lower power than Bartlett's test.

2.2 Using Heteroscedasticity to Improve Modeling

As statistical practice expanded beyond controlled settings in agricultural fields and laboratories, the limitations of statistical tests and inferences that assumed not only homoscedasticity but also Gaussian (normal) distributions of residuals became apparent. Levene (1960) recognized the circularity inherent in Bartlett's test—normally-distributed data were required to test the assumptions of methods that assumed normally-distributed data—and developed a robust alternative to test for normality using absolute deviations from group means. Brown and Forsythe (1974) further refined Levene's approach by using deviations from group medians, creating a test that was robust both to non-normality and to outliers. These developments reflected a philosophical shift: rather than expecting data to fit mathematical ideals, statisticians were adapting methods to handle data as it actually existed. White's (1980) heteroscedasticity-consistent covariance matrix estimator represented a further paradigm shift. His "sandwich" estimator provided standard errors that remained valid even when variances differed among groups, enabling researches to make reliable and trustworthy inferences without transforming or otherwise manipulating their data. Mixed-effects models further allowed researchers to parameterize variance structures explicitly, modeling how variance changes across groups, time, or conditions (e.g., Laird and Ware, 1982; Pinheiro and Bates, 2000).

2.3 Applications in Other Disciplines

The evolution from seeing heteroscedasticity as a violation of assumptions to treating it as meaningful information to be modeled led to significant innovations in a range of fields that use and develop statistical methods. For example, in financial and econometric modeling, ARCH/GARCH models explicitly model volatility clustering in time-series data, treating heteroscedasticity not as noise but as a meaningful feature of financial markets that reflects changing regimes of stability and turbulence (e.g., Bhambu et al., 2025). The results of such analyses convey crucial information for risk management and derivative pricing. In ecology, patchy variance can indicate environmental heterogeneity that influences species distributions, community dynamics, resource availability, and measures of environmental "health" (e.g., Dutilleul and Legendre, 1993; Veum et al., 2025). In medical research, treatments might affect not only average outcomes but also their predictability; some interventions might make clinical responses more consistent across patients whereas others introduce greater variability that has pushed clinical practice in the direction of personalized medicine (e.g., Liu, 2025). This perspective has also permeated into the social sciences, where heteroscedasticity in educational or economic outcomes can imply differential treatment effects or unequal resource allocation (e.g., Adom, Adams and Quagrainie, 2024). In sum, accounting directly for heteroscedasticity and using it effectively in statistical models has become a powerful tool for uncovering problems, mechanisms, and solutions that would be obscured by focusing solely on mean effects.

3. FROM HETEROSCEDASTICITY TO GENERALIZED HETEROGENEITY

In all scientific disciplines, including statistics, heterogeneity has come to mean much more than simply heteroscedasticity. The distinction between what we, for convenience only,

refer to here as "generalized heterogeneity" and heteroscedasticity was articulated most clearly in the early 1990s by ecologists (Kolasa and Rollo, 1991; McIntosh, 1991; Dutilleul and Legendre, 1993; Li and Reynolds, 1995). Although McIntosh (1991) asserted that "[t]he terms of an ideal scientific language and the concepts they express are clear and precise, [and] ambiguity is a barrier to progress", he did little more than illustrate the wide range of different and often contradictory ways that ecologists and evolutionary biologists used the term heterogeneity. Two years later, Dutilleul and Legendre (1993) argued that heterogeneity (at least in ecology) represented a broad scientific "paradigm" (albeit a paradigm simply being a distinct set of concepts or thought patterns, not the Kuhnian sense of a paradigm as the concepts and practices that define a scientific discipline at any particular period of time; Kuhn 1996). Dutilleul and Legendre's (1993) paradigm encompassed differences in means, frequencies, distributions, and interactions, whereas heteroscedasticity was (correctly) seen as merely one specific statistical manifestation of heterogeneity (Table 1). In practice, however, Dutilleul and Legendre (1993) dealt only with what they called "spatial heterogeneity," by which they meant aggregation or overdispersion in spatial density of organisms or environmental characteristics relative to the expectation that the objects of interest were distributed spatially at random (Ripley, 1981, 1987). Li and Reynolds (1995) formalized this concept of ecological heterogeneity, defining it as "the complexity and/or variability of a system property in space and/or time" (the temporal dimension reflected parts of the broader concept of heterogeneity articulated earlier by Kolasa and Rollo, 1991). This use of heterogeneity, along with blocked or spatially-balanced experimental designs that acknowledge rather than ignore spatial heterogeneity and ensure that subsequent modeling includes covariance structures that reflect underlying ecological processes, continues to dominate the ecological literature on the topic. Relevant statistical methods are found in the pages of Cressie (1993) and Cressie and Wikle (2011).

Similar issues arise in the so-called "unit heterogeneity" seen in observational studies (Rosenbaum, 2005). In experimental studies where the investigator controls which individuals are assigned (usually randomly) to specific treatments, increasing the sample size or decreasing the variance (heterogeneity) among experimental units within treatments reduce bias in estimates. In contrast, in observational studies where the investigator has no control over treatment assignments, reducing heterogeneity reduces both variability among samples and bias, whereas increasing sample size may reduce variability among samples but does not affect unobserved bias (Rosenbaum, 2005), In modern causal inference, causal forests can be used to estimate heterogeneous treatment effects across subpopulations, providing valid inference through what Wager and Athey (2018) call "honesty." However, new types of data derived from genomics and analysis of social and other complex networks, have expanded the range of heterogeneity considered in ecology and other disciplines.

3.1 Gaussian Graphical Models for Heterogeneity Analysis

The advent of high-dimensional data in genomics and other fields has created new challenges and opportunities for analyzing heterogeneity. For example, Ren et al. (2021a) developed a penalized fusion approach within the Gaussian Graphical Model (GGM) framework (Box 2) that automatically detects subgroups with distinct network structures. Rather than requiring an *a priori* specification of the number of subgroups, their GGM-based heterogeneity analysis (incorporated in the HeteroGGM package Ren et al., 2021b) instead uses fusion penalties to merge similar subgroups during estimation (Ren et al., 2021a). This methodology represents a significant advance over simple clustering of observations based on mean values. It identifies clusters based on underlying relationships among variables and quantifies a deeper form of heterogeneity in complex systems. The ability to uncover latent subgroup structures without specifying clusters *a priori* has proven useful in exploratory precision medicine. For example, in lung cancer, tumor subtypes often exhibit distinct ("heterogeneous") gene regulatory networks with different network connectivity patterns that corresponded to differences in survival and responses to therapeutic interventions.

Box 2. Core concepts and methods of Gaussian graphical models.

Gaussian graphical models (GGMs) represent the conditional dependence structure among a set of normally distributed variables. They are used to model networks in which nodes are variables and edges represent direct relationships between the nodes. GGMs are particularly valuable for detecting *relational heterogeneity*: differences in network structure across unknown subgroups.

In a GGM, an edge between two variables is absent if and only if they are conditionally independent given all other variables. Conditional independence is inferred from the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ (that is, the precision matrix is the inverse of the covariance matrix). Specifically, if an element of $\mathbf{\Omega}$, $\omega_{ij} = 0$, the variables i and j are inferred to be conditionally independent. The sparse precision matrix can be estimated using penalized likelihood methods (e.g., graphical lasso; Friedman, Hastie and Tibshirani, 2007). The use and interpretation of the estimated precision matrix in a GGM is related to, but different from, those of the covariance matrix in mixed-effects models. In a GGM model, conditional dependence occurs among many variables at a single level, whereas in mixed-effects models (see §5.2 and Box 3) $\mathbf{\Sigma}$ typically specifies marginal covariance structures for random effects across hierarchical levels.

The HeteroGGM R package (Ren et al., 2021a) implements subgroup detection without *a priori* specification of groups. The FASTGGM algorithm (Wang et al., 2016) enables scalable inference to the high-dimensional data characteristic of genomic studies; He et al. (2019) provided a statistical test for determining whether structured networks differed between two or more conditions (e.g., healthy vs. diseased).

GGMs excel at quantifying relational or structural heterogeneity. They can be used to determine how a network of condi-

tional dependencies among a large set of variables differs across unobserved subgroups (Ren et al., 2021a). GGMs also provide a powerful framework for moving beyond marginal correlations to conditional dependence: the direct relationships between variables after controlling for all others. The FASTGGM algorithm addressed the computational challenges of applying GGMs to high-dimensional biological data (Wang et al., 2016), making it feasible to infer gene regulatory networks with thousands of variables. He et al. (2019) further extended the GGM framework to differential network analysis, developing statistical tests for comparing network structures between conditions (e.g., healthy vs. diseased tissue). Their method provides point estimates, measures of uncertainty, and control of false discovery rates; the latter is of particular importance in systems biology because researchers in that field often want to identify not just differential gene expression but also how and which interactions change under different conditions.

3.2 Heterogeneity in Networks

3.2.1 Network Models with Unobserved Heterogeneity Social networks are more complicated than genomic networks, and early methods to study them, such as exponential random graph models (ERGMs) assumed complete model specification (i.e., all relevant attributes of nodes that influenced the formation of edges between them were included in the model). This assumption is clearly unrealistic, as researchers rarely measure all characteristics that influence social ties (e.g., Box-Steffensmeier, Christenson and Morgan, 2018; Box-Steffensmeier et al., 2019). To deal with this problem Box-Steffensmeier, Christenson and Morgan (2018) proposed a frailty ERGM (FERGM) that introduced node-level random effects to capture unobserved heterogeneity, much like frailty terms in survival analysis (see §4, below). In applications to legislative collaboration networks and friendship networks, Box-Steffensmeier et al. (2019) showed how ignoring unobserved heterogeneity could lead to overconfidence in effects like homophily (the "birds of a feather effect," in which people tend to form connections with others who are similar to them in socioeconomic status, values, beliefs, attitudes, or other such characteristics). When frailty terms were included, some previously significant effects disappeared, demonstrating that what appeared to be strong social patterns (e.g., preferential attachment based on ideology) might instead reflect unmeasured individual predispositions (e.g., general sociability or institutional position). The FERGM framework thus provides a crucial correction for network inference by ensuring that identified structural effects are not confounded by latent, node-specific propensities.

3.2.2 Heterogeneity in weighted multilevel networks Unlike social networks that include multiple individuals of a single species (generally humans, but birds, mammals, and other species also form social networks; see, e.g., the papers in a special feature published in *Methods in Ecology and Evolution* and *Journal of Animal Ecology*: Sosa et al., 2021), food webs and other multilevel networks include multiple individuals of multiple species linked by "who-eats-whom" interactions (McCann, 2011). Classical descriptions and quantification of food webs for the most part have been based on their species diversity (i.e., number of nodes) and topological properties (e.g., number of unweighted feeding links between them [edges]), but a more accurate analysis of their structure should also include the relative size (i.e., abundance) of each node and the weights of their edges (i.e., interaction strength) in the network (e.g., Emmerson, Montoya and Woodward, 2005; Girardin et al., 2023), which naturally leads to measuring its heterogeneity (Box 1; see also Shavit and Ellison 2021; Ma and Ellison 2025).

Shavit and Ellison (2021) suggested that the heterogeneity of a network with weighted nodes and edges could be measured by its ascendency A, defined by Ulanowicz (2004) as:

$$A = \sum_{i,j} T_{i,j} \ln\left(\frac{T_{ij}T_{..}}{T_{i.}T_{.j}}\right) \tag{1}$$

where $T_{i,j}$ is a measure of the strength of the edge between nodes i and j and the dot subscript indicates the summation over the missing index (e.g., for N total nodes, $T_{.j} = \sum_{i=1}^{N} T_{i,j}$). Although directly measuring node weights (e.g., number of individuals) is relatively straightforward, directly measuring interaction strength is not, and despite their importance for policy-making and management (e.g., Fath et al., 2019), comparatively few food webs with species weights and interaction strengths needed to compute heterogeneity as network ascendency (Eqn. 1) have been published (e.g. Baird, Asmus and Asmus, 2007). However, with the increasing availability of large datasets derived from sampling environmental DNA that provide measures of relative abundance (i.e., node weights) of species (or, for microbes, "operational taxonomical units"; see, e.g., Thompson et al., 2017), edge weights are estimated using the weighted network algorithm in the igraph package (Csárdi et al., 2026) in the R software system (R Core Team, 2025).

An alternative approach, which quantifies heterogeneity of nodes within networks and the overall heterogeneity of the entire network was proposed by Ma and Ellison (2025). They first estimated the weighted connectedness of node i as:

$$C_i^W = C_i \cdot [R_{i1}, R_{i2}, \ldots, R_{iD_i}] \tag{2}$$

where R_{ij} represents the vector of edge weights between nodes i and $j \in \{1, 2, \ldots, D_i\}$, D_i is the degree of node i (number of edges to or from node i), and C_i^W is, itself, a vector (the scalar product of C_i and $[R_{ij}]$. Then, two measures of the heterogeneity of the network can be computed:

1. Node heterogeneity

$$H_i = \frac{V_i}{M_i} \tag{3}$$

where M_i and V_i are, respectively, the mean and variance of the vector of C_i^W (Eqn. 2) and $H_i \in [0, \infty)$; and

2. Network heterogeneity

$$H = \left(\sum_{i=1}^{N} h_i^q\right)^{1/1-q} \qquad (q \neq 1)$$
$$H = \exp\left(-\sum_{i=1}^{N} h_i \log(h_i)\right) \qquad (q = 1) \tag{4}$$

where h_i is the relative node heterogeneity

$$h_i = \frac{H_i}{\sum_{i=1}^{N} H_i},$$

H_i is defined in Eqn. 3, and q is the order of the Rényi entropy (Rényi, 1961; Chao et al., 2014).

Alternatively, network heterogeneity can be modeled probabilistically from power-law distributions of node degree or abundance distributions (Ma, 2025), or from distributions of node heterogeneities H_i (Eqn. 3). Ma and Ellison (2025) apply these methods to the Earth Microbiome dataset (Thompson et al., 2017), and also present ways to test statistically for differences in heterogeneity between nodes within networks and between entire networks.

4. SURVIVAL ANALYSIS AND THE TEMPORAL DIMENSION OF HETEROGENEITY

Survival analysis provides a perspective on heterogeneity that is distinct from the explicit focus of ecologists on spatial heterogeneity and the implicit spatial emphasis of network analysis (but see Ferreira et al. 2020 for a review of spatiotemporal network analysis). Heterogeneity enters survival analysis through its estimation of variation in unobserved individual susceptibilities ("frailties") that affect hazard rates. This type of heterogeneity has real-world consequences: individuals with higher frailty experience events earlier, progressively changing the composition of the risk set over time (Vaupel, Manton and Stallard, 1979; Aalen, 1988). Demographers, epidemiologists, and reliability engineers also distinguish between individual-level hazards and population-level hazards (Ma and Krings, 2008a,b,c, 2011).

4.1 Definition of the Hazard Function

For a survival model, we define the hazard function $\lambda(t)$ as the instantaneous risk of an event at time t, given that an individual (or population) survives until t. For a given unobserved (latent) individual, the frailty, $(Z) : Z > 0$ and $E[Z] = 1$, is a multiplicative random variable representing an unobserved individual's susceptibility. Higher values of Z imply greater inherent risk. As "frail" individuals (those individuals i with higher values of Z_i) experience events earlier, the mean frailty among survivors $E[Z|T > t]$ decreases over time. This causes the observed population hazard to diverge from individual hazards, often leading to deceleration or plateaus in the aggregate hazard curve.

The different methods for individual- and population-level hazards reflect classic trade-offs in statistical modeling. Frailty models explicitly parameterize heterogeneity through random effects on the hazard scale (Vaupel, Manton and Stallard, 1979; Aalen, 1992), and use Gamma or inverse Gaussian distributions to represent unobserved susceptibility. These models offer rich interpretability but face significant identifiability challenges (Keiding, Andersen and Klein, 1997); the baseline hazard and frailty distributions cannot be jointly estimated from standard survival data. In contrast, accelerated failure time (AFT) models estimate survival time directly, but relegate all heterogeneity (variation in frailty) to the error term. AFT models sacrifice direct interpretation of hazard dynamics but give more stable and invariant estimates of covariate effects, particularly when the true frailty distribution is unknown.

Survival analysis also uniquely addresses what might be termed "compositional heterogeneity:" the systematic change in population characteristics over time due to selective attrition. Zajacova, Goldman and Rodríguez (2009) demonstrated how compositional heterogeneity can lead to the misleading conclusion that specific causes of mortality continue into old age as opposed to the more general pattern that the differential effects of specific causes decrease with age (see also Vaupel, Manton and Stallard, 1979; Zarulli, 2016).

4.2 Estimating Selection with the Proportional Frailty Model

The core of frailty theory is in formalizing the selection (i.e., mortality) process through conditional expectations. The standard proportional frailty model specifies the individual (conditional) hazard as

$$\lambda(t|Z) = Z \cdot \lambda_0(t) \cdot exp(\beta^T X), \tag{5}$$

where Z represents the individual's frailty, $\lambda_0(t)$ is the baseline hazard, and X are observed covariates. The population (marginal) hazard $\mu(t)$, which is what is actually observed in aggregate data, emerges as an average over the surviving population:

$$\mu(t) = E[Z|T > t] \cdot \lambda_0 \cdot exp(\beta^T X). \tag{6}$$

The mean frailty among survivors, $\mu(t)$, defines the selection effect.

Because individuals with higher Z values experience events (hazards) earlier, Eqn. 5 decreases monotonically with time. Under the commonly used assumption that $Z \sim \Gamma(\alpha, \theta)$,

$$E[Z|T > t] = 1/[1 + \theta\Lambda_0(t)],$$

where $\Lambda_0(t)$ is the cumulative baseline hazard. Substitution yields the population hazard:

$$\mu(t) = \lambda_0(t) \cdot exp(\beta^T X)/[1 + \theta\Lambda_0(t)]. \tag{7}$$

Eqn. 7 makes the "dragging down" effect explicit: even if every individual's hazard increases indefinitely with time (as in a Gompertz model), the population hazard $\mu(t)$ will eventually decelerate, plateau, or even decline as the surviving population becomes increasingly selected for robustness (Vaupel, Manton and Stallard, 1979). This mathematical relationship explains numerous empirical phenomena, from mortality plateaus at advanced ages to converging hazard ratios between demographic groups, as artifacts of selection rather than reflections of fundamental biological or social processes (Vaupel, Manton and Stallard, 1979; Zarulli, 2016).

The choice of the distribution to use to model frailty with a proportional hazard model asserts a specific assumption or hypothesis about the nature of heterogeneity in a given studied population (Ma and Krings, 2008c). The Gamma distribution, which is commonly used because it is mathematically tractable and conjugate with the Weibull distribution used to model hazards, assumes continuous, moderately skewed heterogeneity in frailty. The compound Poisson distribution, introduced to frailty models by Aalen (1992), includes a point mass at zero, implying the existence of an immune subpopulation (or at least one

that is not susceptible to the hazard under study). This compound Poisson distribution is used to model not only diseases for which immunity exists but also consumer behavior. The positive stable distribution possesses the unique property that proportional hazards are preserved at the population level, albeit with attenuated coefficients, and suggests the hypothesis that heterogeneity in frailty manifests primarily through "early dependence" patterns (Balan and Putter, 2020). Inverse Gaussian and lognormal distributions are used to model different tail behaviors, and allow tests to determine whether latent susceptibility follows heavier-tailed or log-symmetric patterns.

4.3 Methodological Advances in Modeling Frailty: Shared, Conditional, Hierarchical and Accelerated Failure Time Models

4.3.1 Shared Frailty Models A major advancement was extending frailty models to clustered or correlated event data (e.g., families, litter-mates, batches of manufactured items). In such datasets, all members of a cluster share a common frailty value Z_i, which simultaneously quantifies within-cluster heterogeneity and induces stochastic dependence among the event times of members of any given cluster. The resulting joint survival function takes the form of an Archimedean copula, directly linking the frailty variance θ to association measures like Kendall's τ (Aalen, Borgan and Gjessing, 2008).

4.3.2 Conditional Frailty Models For recurrent event data, (Box-Steffensmeier, De Boef and Joyce, 2017) developed a model that disentangles two distinct sources of within-subject correlation: event dependence (where the occurrence of one event alters the risk of subsequent events, modeled through stratification of the baseline hazard by event number) and persistent heterogeneity, modeled through a shared frailty term. This method has proven to be valuable in social science, where it is used to distinguish true contagion from stable unobserved propensities.

4.3.3 Hierarchical (Multilevel) Frailty Models These models extend the shared frailty concept to nested data structures (e.g., patients within hospitals within regions) by including multiple, crossed random effects. They formally decompose the total heterogeneity variance into level-specific components (e.g., $\theta_{hospital}$, $\theta_{patient}$), allowing researchers to partition latent variation into contextual and individual factors. This method is widely used in genetic epidemiology for separating genetic variance from components of shared environmental variance (e.g., Aalen et al., 2015).

4.3.4 Accelerated failure time (AFT) models AFT models estimate survival time, T, directly:

$$log(T) = \beta^T X + \sigma.$$

Heterogeneity (i.e., variation in frailty) inflates the scale parameter σ in the AFT model. An important advantage that comes with AFTs is that covariate effects β are often invariant with respect to the unknown frailty distribution.

4.4 Integrating Deep Learning and AI into Frailty Models Increases Explainability

Deep learning and AI have fundamentally expanded the scope of heterogeneity modeling by moving beyond low-dimensional latent structures to higher-dimensional, non-linear representations learned directly from complex, unstructured data (Wiegrebe et al., 2024). This shift has addressed a key limitation of the ML and Bayesian frailty models described in the preceding subsections that rely on strong parametric assumptions and simple random-effect structures. Deep-learning models have led to the discovery of intricate interaction patterns and time-varying effects.

For example, the non-proportional Cox-Time model parameterizes the log-hazard as a neural network function of both covariates and time, $\eta(x,t) = g(x,t)$, thereby learning how covariate effects evolve non-linearly over time and providing estimates of complex time-varying heterogeneity in risk (Kvamme, Ørnulf Borgan and Scheel, 2019). For more complex datasets, such as those that combine histopathology images with genomic profiles, multimodal learning architectures integrate modality-specific subnetworks whose outputs are fused for joint prediction. This allows quantification of heterogeneity arising from disparate data sources.

Other recent frameworks build heterogeneity modeling directly into their architecture. OTSurv (Ren et al., 2025) frames survival prediction from whole-slide images as an optimal transport problem, ensuring that prognostically critical but rare morphological features are not overwhelmed by common patterns and preserves heterogeneity in pathological representation. Similarly, temporal and heterogeneous graph neural networks for predictive maintenance use feature-wise linear modulation (FiLM) layers to dynamically adapt global models to specific sensor contexts and temporal regimes, explicitly modeling heterogeneity inherent in multi-sensor systems (THGNN; Wen et al., 2025).

Although deep-learning models do not solve the fundamental identifiability problem of classical frailty models, they mitigate the problem through architectural regularization and constraints (e.g., the optimal transport constraints in OTSurv). These act as implicit priors guiding the model away from arbitrary, non-identifiable solutions. Furthermore, these models can identify complex interaction structures between observed features and latent susceptibility, moving beyond the standard multiplicative frailty assumption (Eqns. 5, 7) to scenarios where covariates may modify heterogeneity itself.

However, the complex output of deep-learning models can be difficult to interpret. Unlike classical frailty models where heterogeneity is explicitly parameterized by θ and the distribution of Z (Eqns. 5, 7), the heterogeneity estimated by deep neural networks is distributed across millions of parameters in non-linear transformations. This interpretability challenge has spurred the development of explainable AI (XAI) methods for survival analysis. Krzyziński et al. (2023) introduced SurvSHAP(t), which extended Shapley values (derived from a game-theoretic approach to feature attribution) to functional outputs. SurvSHAP(t) produces time-dependent Shapley values, $\phi_t(x^*)$, which quantify the contribution of each feature to the predicted survival probability at each time point t. This allows researchers to more clearly answer questions such as "does a genetic variant confer early risk that diminishes over time?" or "do certain histopathological features indicate constant frailty or time-varying susceptibility?"

Furthermore, most current deep survival models lack explicit, interpretable parameters for unobserved heterogeneity. They do not provide variance decompositions, heritability estimates, or random-effect predictions that have made classical frailty models useful and scientifically valuable. This suggests a new research challenge: the development of hybrid models that integrate deep-learning architectures with statistically principled frailty components. Such models would combine deep-feature extractors for complex data, explicit frailty structures with interpretable variance parameters, and joint estimation, thus allowing deep features to inform frailty estimation and vice versa. Initial steps in addressing this challenge could include adding random intercepts to the penultimate layer of a deep survival network or using variational auto-encoders to learn low-dimensional latent representations that serve as interpretable frailty variables.

5. COVARIANCE AND DEPENDENCE: MODELING STRUCTURED RELATIONSHIPS

5.1 The Price Equation

What is now known as the Price Equation (Price, 1970),

$$\Delta\bar{z} = \frac{1}{\bar{w}}Cov(w,z) + \frac{1}{\bar{w}}E(w\Delta z) \tag{8}$$

was developed as a general mathematical treatment for the by-then well-known and well-accepted biological process of evolution through natural selection (Fisher, 1930; Mayr, 1942). The Price equation decomposes evolutionary change into two components: selection (estimated as the covariance between fitness w and traits z) and transmission (estimated as the expected changes in trait values Δz across generations). Even though Price (1970) was focused on differential selection as the driving force of evolution in nature, Eqn. 8 has turned out to apply to any system in which success or outcome depends on multiple traits or factors; the covariance structure among those components determines how the system changes through time. As with the importance of link weights in the dynamics of networks (§4), interactions between traits (in this case, estimated as correlations among components) can matter much more to the outcome than it does to their individual values. In essence, the Price Equation serves as a unifying lens: understanding and modeling heterogeneity as non-random, structured covariance can be used to model and predict changes in complex systems.

5.2 Modeling Correlated Data: From GMMs to Mixed-Effect Models and General Estimating Equations

The practical challenge of analyzing correlated data, common in longitudinal, clustered, or spatial studies, led to a number of important methodological innovations. For example, networks of conditional dependencies modeled by Gaussian graphical models (§3.1) can help to identify predictors of the marginal covariances that the Price Equation highlights. More generally, the covariance structures modeled in mixed-effects models (MEMs; Box 3) can be interpreted as parameterizing the ''selection'' term (the covariance term in Eqn. 8) for clustered data. MEMs model several forms of heterogeneity: heteroscedasticity and changing residual variance (σ^2 in Eqn. 9) across groups or over time; aspects of more general heterogeneity, including group differences (the random intercepts μ_{0j} estimate baseline differences between groups) and differential effects (the random slopes μ_{1j} model how the effect of a covariate varies across groups). The covariance matrix $\mathbf{\Sigma}$ estimates covariance within groups and helps model within-group dependence.

The development of flexible covariance structures, such as autoregressive (AR) functions for temporal data, the Matérn covariance function for spatial data (e.g., Gelfand and Schliep, 2016), and compound symmetry for clustered data, enabled researchers to tailor their models to the specific dependence patterns in their data. If $\mathbf{\Sigma}$ is constrained to reflect a graphical model, it estimates the same network of conditional dependencies modeled with GGMs. In sum, GGMs ask whether interaction networks differ between groups, whereas MEMs ask by how much does the outcome vary between groups, and how is that variation structured. Together, GGMs and MEMs represent powerful tools for modeling different dimensions of structured heterogeneity. The former identifies heterogeneity in relationships among variables, the latter, the heterogeneity of units and contexts.

Box 3. A basic mixed-effects model.

Mixed-effects models (MEMs), also known as multilevel or hierarchical models, are used extensively for data with inherent groupings or clusters (e.g., students in schools, repeated measurements per patient). Their power lies in explicitly modeling multiple sources of variation.

The basic linear MEM can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \mu_{0j} + \mu_{1j} x_{xj} + \epsilon_{ij} \tag{9}$$

where: y_{ij} is the outcome for observation i in group j; β_0 and β_1 are, respectively, the average intercept and slope ("fixed" effects); μ_{0j} and μ_{1j} are the deviations of each group from β_i ("random" effects); $\mu \sim \mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the covariance matrix for random effects); and the residual error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Parameters are typically estimated via maximum likelihood (ML), restricted maximum likelihood (REML), or Bayesian methods in the `R` packages lme4 (Bates et al., 2015), nlme4 (Pinheiro and Bates, 2000) , or blme (Chung et al., 2013).

Finally, Generalized Estimating Equations (GEEs) provide a semi-parametric framework that maintains model validity even when the assumed correlation structure is misspecified (Liang and Zeger, 1986). This robustness property, which yields consistent "population-average" estimates, has made GEEs indispensable in clinical trials, epidemiology, and meta-analysis, which we turn to in the next section (§6).

6. THE CENTRAL ROLE OF HETEROGENEITY IN META-ANALYSIS

Meta-analysis occupies an important position with respect to heterogeneity. It provides the most explicit and formalized

framework for quantifying, testing, and explaining between-study heterogeneity, which in meta-analysis is defined as the variability in estimates of effect sizes across studies that exceed what would be expected from sampling error alone (Deeks, Altman and Bradburn, 2001). Meta-analysis serves as a conceptual bridge between statistical theory and empirical synthesis, and its focus on heterogeneity illustrates that it is a mature field of inquiry in which explaining variation is more important than estimating a single pooled effect (Gurevitch et al., 2018). Furthermore, the problem of estimating explanatory covariates for between-study variance (τ^2; §6.2) is conceptually and methodologically similar to the problem of identifying traits (z) that co-vary with fitness (w) in the Price Equation (Eqn. 8).

Meta-analysis illustrates clearly a primary theme of this review: heterogeneity is not a statistical nuisance but rather reflects a central question: *Why do effects differ in different contexts*? Through meta-regression and subgroup analysis, heterogeneity is used as a source of discovery rather than a problem to be eliminated (Deeks, Altman and Bradburn, 2001). This is particularly true for meta-analyses aimed at broad generalizations where identifying sources of heterogeneity is central to understanding the overall phenomenon (Gurevitch et al., 2018). The meta-analytic framework also illustrates the integration of different types of heterogeneity. Although it starts with questions about general heterogeneity (varying effect sizes), it must also address potential heteroscedasticity (differing within-study variances) and dependence (correlated effect estimates in multivariate or network meta-analysis). Key components of meta-analysis and standard methods of estimation used by meta-analysts are summarized in Box 4.

6.1 Detection and Quantification of Heterogeneity in Meta-analysis

Early methodological work established Cochran's Q as a standard tool for estimating heterogeneity in a meta-analysis. The Q statistic is calculated as the weighted sum of squared differences between effects θ_i observed in individual studies i and the pooled effect $\bar{\theta}$:

$$Q = \sum w_i(\theta_i - \bar{\theta})^2.$$

Q approximately follows a χ^2 distribution with $k-1$ degrees of freedom under the null hypothesis of homogeneity. However, the χ^2 test of Q has low statistical power, especially in meta-analyses with few studies or unbalanced weights; the latter occurs when when one large study dominates the analysis (Hardy and Thompson, 1998). Thus, the critical value for tests of significance using Cochran's Q normally is set at $\alpha = 0.1$ and a "non-significant" result should not be interpreted as evidence *for* homogeneity among studies (Hardy and Thompson, 1998). Rather, random-effects models should be used to identify and test for heterogeneity between studies (Gurevitch et al., 2018).

Higgins and Thompson (2002) introduced the I^2 statistic,

$$I^2 = max(0\%, \frac{Q-(k-1)}{Q} \times 100\%)$$

as an alternative to Q. Conventionally, I^2 values of 25%, 50%, and 75% are taken to characterize "low," "moderate," and "high" heterogeneity. However, I^2 can be biased when the number of studies used in the meta-analysis is small; thus, reporting confidence intervals for I^2 is recommended (von Hippel, 2015). Alternative measures of heterogeneity for correlated outcomes include the univariate R^2 statistic (the ratio of the variance of the treatment effect estimated from random- and fixed-effects models), a multivariate statistic H^2 (the ratio of a generalization of Cochran's Q and its associated degrees of freedom), and a multivariate analog of I^2 (Jackson, White and Riley, 2012). All these measures express heterogeneity as a percentage of total variability, making the concept of heterogeneity comparable across different meta-analyses (Higgins and Thompson, 2002).

Box 4. Components of, and estimation in, meta-analysis.

The primary goal of a meta-analysis is to use data from multiple, comparable studies to estimate an overall effect size Θ and its between-study variance between studies τ^2 (Gurevitch et al., 2018). To get to this end-point, the meta-analyst first needs to know (or calculate) the effect size θ_i, within-group standard deviation s_i or variance s_i^2, and sample size N_i of each study i. Common measures of effect size include the sample mean, correlation or regression coefficient, or, for experiments with "control" and "treatment" groups with continuous outcomes, the mean or standardized difference (e.g., Cohen's $d = \frac{\bar{Y}_1 - \bar{Y}_2}{s}$ or the less-biased Hedges' g; see, e.g., Hedges, 1981; Gurevitch et al., 2018; Dormann and Ellison, 2025). For experiments resulting in binary data, standardized effect sizes are estimated by odds ratios or risk ratios. For time-to-event data, the hazard ratio can be used as a standardized effect size.

Estimation is done either with a fixed-effects model that assumes a single true effect θ across all studies (each weighted equivalently with weights $w_i = 1/v_i$) or a random-effects model that assumes that the true effects $\theta_i \sim \mathcal{N}(\mu, \tau^2)$ with weights $w_i^* = 1/(v_i + \hat{\tau}^2)$ (DerSimonian and Laird, 1986). The random-effects model is preferred when clinical or methodological diversity is suspected or known.

The pooled effect under the random-effects model is calculated as $\bar{\theta}^* = \sum w_i^* \theta_i / \sum w_i^*$, with variance $Var(\bar{\theta}^*) = 1/\sum w_i^*$. The 95% confidence interval is constructed as $\bar{\theta}^* \pm t_{0.975} \times Var(\bar{\theta}^*)$, where the Hartung-Knapp-Sidik-Jonkman adjustment (using a t-distribution with $k-1$ degrees of freedom) is recommended to maintain better coverage, especially when τ^2 is estimated with uncertainty (IntHout, Ioannidis and Borm, 2014).

6.2 Estimating the Between-Study Variance (τ^2)

The random-effects meta-analysis model (Box 4) uses τ^2 as a direct, estimable parameter representing between-study heterogeneity (DerSimonian and Laird, 1986). The value of using τ^2 as opposed to Q, I^2, or other measures used in binary tests for homogeneity *vs.* heterogeneity is that τ^2 is a continuous variable to be estimated, interpreted, and explained.

6.2.1 Estimating τ^2 The choice of estimator for τ^2 significantly influences inference (Langan et al., 2019). Common ones include the DerSimonian and Laird estimator (DerSimonian and Laird, 1986), which is simple but often negatively biased; the Paule-Mandel estimator (Jackson et al., 2017), which is more accurate but can show positive bias with variable study sizes; and REML estimates, which are more robust and have minimal bias. Recent large-scale simulations recommend REML as the preferred estimator, with the Paule-Mandel estimate as a practical non-iterative alternative (IntHout, Ioannidis and Borm, 2014). However, all τ^2 estimators are imprecise and unreliable in meta-analyses with fewer than 10 studies; using the Hartung-Knapp-Sidik-Jonkman method (IntHout, Ioannidis and Borm, 2014) to construct confidence intervals maintains near-nominal coverage even when τ^2 is poorly estimated (Langan et al., 2019).

6.2.2 Specific challenges in estimating τ^2 There are three specific cases where estimating τ^2 presents unique challenges.

First, when combining results from only two studies, the Hartung-Knapp-Sidik-Jonkman method (IntHout, Ioannidis and Borm, 2014) controls Type I error but has extremely low power (<15% if the the effects sizes in the two studies really are homogeneous). In contrast the other classical methods of examining heterogeneity (Q, I^2 and its multivariate analog, R^2, and H^2 all have greatly inflated Type I error probabilities. There is currently no statistically sound method for confirmatory meta-analysis of two studies when heterogeneity cannot be ruled out (Gonnermann et al., 2015).

Second, when effects are measured as proportions or prevalence (so-called "prevalence meta-analysis"), extremely high I^2 values (median $\approx 97\%$) are typical but they generally are a mathematical artifact of large samples and variability in proportional data (Migliavaca et al., 2022). Rather than estimating variance or other measures of heterogeneity, Migliavaca et al. (2022) recommended that meta-analysts estimate prediction intervals for pre-planned (a priori) analyses of subgroups.

Finally, high heterogeneity (median $I^2 \approx 85\%$) is the norm in meta-analyses of ecological and evolutionary data. Random-effects or multi-level models should always be used when synthesizing studies in ecology and evolutionary biology. Although careful attention should be paid to accurate and meaningful reporting of heterogeneity statistics in ecology and evolution (Senior et al., 2016), the more important research goal in these fields is identifying causal predictors of heterogeneity, such as phylogenetic processes or environmental covariates (Gurevitch et al., 2018).

6.3 Explaining and Predicting Heterogeneity

Quantifying heterogeneity is only the first step in understanding it. Identifying the causes of heterogeneity and modeling or predicting it are the crucial next steps that make meta-analysis a powerful engine for generating new scientific insights (Petitti, 2001; Gurevitch et al., 2018). For example, subgroup analysis and meta-regression can be used to test whether and how study-level characteristics cause or modify observed effects (Gurevitch et al., 2018). Graphical diagnostics are invaluable adjuncts to explaining results of meta-analyses. Forest plots can be used to display study estimates, funnel plots can be used to assess or highlight publication bias; and Baujat plots can identify studies that disproportionately contribute to estimates of pooled effect sizes or observed heterogeneity (Baujat et al., 2002). Sensitivity analyses by sequential leave-one-out analyses can further assess the robustness of the results. In general, reporting prediction intervals supplement confidence intervals on τ^2; the former better illustrate the expected range of effect sizes in subsequent (i.e., out-of-sample) studies (Higgins, Thompson and Spiegelhalter, 2008; Dormann and Ellison, 2025).

Finally, recent work has highlighted distinctions among three types of heterogeneity that can affect conclusions derived from meta-analysis: population heterogeneity that results from different samples; design heterogeneity, a consequence of the use of different protocols in synthesized studies; and analytical heterogeneity, which can arise because different studies combined in a meta-analysis used different analytical methods (Holzmeister et al., 2024; Krefeld-Schwalb, Hua and Johnson, 2025). Large syntheses have revealed that design and analytical heterogeneity can be substantial, and simulations have illustrated that conclusions drawn from meta-analyses of studies with either design or analytical heterogeneity (or both) can be misleading (see Chapter 10 in Dormann and Ellison, 2025). Preregistration of prospective meta-analyses and improved standards of reporting, statistical testing, and prediction can help ameliorate these effects and increase the reliability of conclusions drawn from meta-analyses and research syntheses (Gurevitch et al., 2018). An alternative approach, currently used most frequently in research with human subjects, is federated analysis (Hallock et al., 2021; Dormann and Ellison, 2025).

7. INTEGRATING MULTIPLE TYPES OF HETEROGENEITY

Rather than treating each different type of heterogeneity in Table 1 as a separate problem requiring a separate solution, MEMs (Box 3, above), GEEs (§5.2), machine- and deep-learning models (§4.4), and other methods can be used to jointly model multiple types of heterogeneity within a mathematically coherent and robust framework. This kind of integration reflects a more mature statistical approach; the ability to specify complex variance-covariance structures enables researchers to closely align their statistical model with the hypothesized data-generating process, leading to more accurate inference and deeper understanding. Consider, for example, a longitudinal study of student achievement across multiple schools. With MEMs or GEEs, fixed effects estimate average growth trajectories; random intercepts estimate baseline differences between schools (general heterogeneity); random slopes allow growth rates to vary across schools (general heterogeneity); covariance structures (e.g., autoregressive, compound symmetry) model the dependence among repeated measures within students; and variance functions allow measurement error to change over time or across groups (heteroscedasticity). Similarly, OTSurv applies optimal transport principles within a multiple-instance learning framework to handle heterogeneity in survival prediction from histopathology images (Ren et al., 2025), ensuring that rare but prognostically important tissue patterns are represented. Temporal and heterogeneous graph neural networks

(THGNN Wen et al., 2025) explicitly model multiple sources of heterogeneity (equipment-specific, temporal, relational) in predictive maintenance applications. These approaches address longstanding challenges: they maintain flexibility without overfitting through sophisticated regularization, provide stable estimates even with complex interaction structures, and, with the aid of explainability tools like SurvSHAP(t) (Krzyziński et al., 2023), can yield more interpretable insights than classical "black-box" machine-learning models in high-dimensional settings.

7.1 Meta-Analysis of Multi-Arm Studies

Advanced meta-analytic methods also analyze multiple types of heterogeneity. In particular, network meta-analysis of multiple clinical trials, each of which are designed to assess the efficacy of multiple therapeutic interventions in a single study (so-called "multi-arm" studies) use both direct and indirect evidence to account for correlations among multiple simultaneous treatment effects. Meta-analysis of such studies is done using a multivariate random-effects model with a structured covariance matrix (Axon, Dwan and Richardson, 2023). This kind of meta-analysis handles multiple correlated outcomes per study by borrowing strength across outcomes and properly accounting for their dependence (Rücker and Schwarzer, 2025). These methods use sophisticated covariance structures to model dependencies while simultaneously estimating the core heterogeneity parameters (τ^2) for each treatment comparison or outcome (Rücker and Schwarzer, 2025). The practical implementation of meta-analysis in the `R` packages metafor (Viechtbauer and Cheung, 2010) and netmeta (Balduzzi et al., 2023) provide accessible implementations that allow applied researchers in medicine and social sciences to easily run and report results from these complex meta-analyses.

7.2 Bayesian Approaches for Propagating Uncertainty in Estimating Heterogeneity

Bayesian methods provide a natural framework for handling and propagating the uncertainty inherent in heterogeneity estimation, which is often substantial, especially for variance components (e.g., Banerjeee, Carlin and Gelfand, 2025). Using inverse-Wishart priors for covariance matrices or the more flexible Lewandowski-Kurowicka-Joe (LKJ) distribution (Lewandowski, Kurowicka and Joe, 2009) as a prior for correlation matrices, Bayesian models can incorporate prior knowledge about plausible heterogeneity structures by regularizing estimates away from extreme correlation values (Barnard, McCulloch and Meng, 2000; Gelman et al., 2013). This is particularly valuable in high-dimensional settings or with sparse data. Stan (Carpenter et al., 2017) and other probabilistic programming languages have dramatically lowered barriers to implementing these complex models. Researchers can now specify intricate multi-level models with crossed random effects and structured covariance matrices with relatively concise code. Hamiltonian Monte Carlo (HMC) methods handle run the computations for sampling from the posterior distribution more quickly than other methods (e.g., Monnahan, Thorson and Branch, 2017). This allows for full uncertainty quantification for all parameters, including those used to estimate heterogeneity.

8. GUIDELINES FOR USING STATISTICAL METHODS FOR MEASURING AND MODELING HETEROGENEITY

Researchers in many fields use methods developed by statisticians for measuring, modeling, and interpreting heterogeneity. In this section we present a set of guidelines for researchers and practitioners in such fields who may not be well-versed in the theoretical intricacies of the methods and models presented above. Before proceeding through these guidelines, it is critical to be clear about exactly the type and subtype of heterogeneity that is of interest and being studied (Table 1.

8.1 Visualize, Test, Estimate

Assessing, measuring, and testing for heterogeneity starts with visualizing the data. Residual plots (e.g., residuals vs. fitted values) can reveal patterns of heteroscedasticity; variograms or spatial correlograms can unveil temporal or spatial dependencies; and forest plots can clearly illustrate between-study heterogeneity. Even if visualization suggests substantial heterogeneity, formal testing should be employed judiciously. Heteroscedasticity can be tested with Levene's or Breusch-Pagan tests. Intraclass correlation coefficients can test for significant clustering in the data, and Moran's I can be used to test for spatial dependence. In meta-analysis, Cochran's Q, Higgins and Thompson's I^2, or likelihood ratio tests can be used to comparing heterogeneity in models with and without random effects to test for more general types of heterogeneity. In most cases, simulation can be more useful than many of these classical statistical tests (e.g., Dormann and Ellison, 2025) Finally, it is not enough to simply report results of these various statistical tests. Reporting effect sizes, such as the magnitude and proportion of variance components in mixed models or τ^2 and its confidence (or prediction) interval in meta-analysis, provides much more useful information than the results of the significance tests alone.

8.2 Choose an Appropriate Model

The choice of modeling strategy should be guided by the type of heterogeneity and the research question of interest. Although there are no optimal approaches for modeling each type or subtype of heterogeneity, we suggest the following:

1. To model heteroscedasticity when it is not the focus of the research, use robust ("sandwich") standard errors to obtain valid inference for mean effects without modeling the variance structure itself (White, 1980).
2. To model heteroscedasticity when it is of direct interest, employ explicit variance modeling using mixed-effects models with variance functions (e.g., `varIdent()` or `varPower()` in nlme4 Pinheiro and Bates, 2000) to predict how variability changes across groups or through time.
3. To model general heterogeneity with a known grouping structure, use random effects (intercepts, slopes) in a mixed model or use a stratified analysis.
4. To model general heterogeneity when the underlying structure (e.g., groups) is unknown, consider using machine- or deep-learning models such as causal forests to discover subpopulations or subgroups without the

need for *a priori* specification (Wager and Athey, 2018), or use clustering methods like the penalized fusion in in HeteroGMM (Ren et al., 2021a).
5. To model dependence and heterogeneity in clustered or longitudinal data, use generalized estimating equations for population-average inference that is robust to misspecification of the correlation structure, or use mixed-effects models for subject-specific inference and prediction.
6. To model more complex spatiotemporal heterogeneity, use specialized covariance structures (e.g., spatial Matérn, temporal ARIMA) within mixed models or dedicated geostatistical methods (kriging) for prediction.
7. To model unknown heterogeneity in hazards, use proportional, shared, conditional, or hierarchical frailty models; the choice depends on the type of data at hand.

8.3 Clearly Report and Carefully Interpret the Results

Always distinguish between statistical significance and substantive importance. A significant Q-test or a large τ^2 may imply "significant" variability (i.e., beyond chance alone), but its practical importance depends on the context and the magnitude of the observed effect. To clearly illustrate the context and magnitude, always report estimates of heterogeneity along with estimates of main effects. Examples include estimates and proportions of variance components in mixed-effect models or estimates of τ^2 and I^2 (with confidence intervals) in meta-analyses. Do not ignore potentially important implications of any observed heterogeneity. Does it suggest the presence of moderator variables? Does it limit the generalizability of an average effect? Could it reflect measurement error or bias? Finally, take advantage of clear figures and other methods of visualizing the results. For example, network diagrams can illustrate heterogeneous structures modeled by GGMs, and time-dependent Shapley plots from SurvSHAP(t) can illustrate how importance of specific features changes over time in predicting future survival of individuals subject to particular hazards. Forest plots are standard ways to illustrate heterogeneity among studies used in meta-analysis and caterpillar plots of random effects can show the distribution of deviations from expectations of particular clusters.

9. FUTURE DIRECTIONS AND OPEN CHALLENGES

9.1 Causal Explanations of Heterogeneity

An open challenge in all disciplines is to move beyond describing and (statistically) modeling heterogeneity to explaining its causes. Although methods like causal forests can estimate heterogeneous treatment effects, a deeper challenge is causal attribution: determining *why* observed effects vary. Causal attribution—understanding the underlying mechanisms that drive differential responses—requires more than simply identifying moderators or effect modifiers. Rather, causal attribution will require integrating causal discovery methods with estimation of heterogeneous treatment effects along with methods that account for heterogeneity among populations. Developing such frameworks and methods will be particularly critical for further advancing personalized medicine and evidence-based policy, for which it is most important to understand not just *if* an intervention or policy works but *for whom and why* it works.

9.2 High-Dimensional, Multimodal, and Complex Data

Many new data sources, including genomics, neuroimaging, digital phenotyping, and ecological sensor networks, present unprecedented challenges for all types of statistical analysis. Modeling heterogeneity in these Big datasets must continue to be interpretable while scaling to settings where there are relatively few observations for millions of variables (i.e., $p \gg n$). Integrative analyses that combine multiple data types (e.g., genomic, transcriptomic, imaging, clinical) may require identification of new types of heterogeneity and models that can handle dependencies across modalities and scales. Methods are needed to disentangle technical batch effects, biological heterogeneity, and measurement error in these large, complex, and integrated datasets. Multi-view and multi-task learning with shared and modality-specific latent factors are promising directions (e.g., Guo et al., 2025).

9.3 Interpretability, Fairness, and Communication

As models for heterogeneity become more sophisticated (e.g., complex Bayesian hierarchal models, hybrid and multi-view deep learning models), communicating their implications to diverse audiences of scientists in other fields, clinicians, policymakers, and interested non-specialists becomes both more important and more challenging. There is a pressing need for informative methods and advanced visualizations that can convey uncertainty in estimates of heterogeneity (and other model outputs) and their consequences for particular policies, interventions, or decisions. Creating ethical models of heterogeneity and linking them with equity is also critical: models that identify subgroup effects can be used to ensure equitable outcomes, but they also can perpetuate or exacerbate biases if heterogeneity is conflated with other important attributes or goals (Shavit and Ellison, 2021).

9.4 Software, Accessibility, and Reproducibility

Continued development of open-source, well-documented software (not only in `R`, but also in Python and Julia) will serve to make methods for estimating, modeling, and visualizing heterogeneity more accessible. User-friendly interfaces (e.g., Shiny apps in `R` that maintain reproducibility), comprehensive tutorials and vignettes, and other educational resources are needed to bridge the gap between methodological innovations and practical applications. Further, promoting reproducible research practices is essential, especially given the flexibility and potential for "forking paths" in exploration of heterogeneity (e.g., choosing different frailty distributions, τ^2 estimators, or ML hyperparameters). As with any research plan, preregistration of plans for studying heterogeneity in single observational or experimental studies and meta-analyses can help mitigate researcher degrees of freedom (Simmons, Nelson and Simonsohn, 2011; Nosek et al., 2018).

9.5 Integration Across Disciplinary Silos

Finally, as should be apparent from Table 1 and the different domain-specific examples in the previous sections, dif-

ferent fields have developed, usually in isolation, parallel approaches to estimating, modeling, and interpreting heterogeneity. For example, statistics emphasizes formal parameterization and inference (e.g., variance components, random effects); machine learning focuses on prediction and pattern discovery (clustering, representation learning); and other fields have developed domain-specific concepts (e.g., spatial heterogeneity in ecology, frailty in demography, unobserved confounding in econometrics). Greater cross-disciplinary dialogue could more quickly rationalize and synthesize these concepts and types of heterogeneity, leading to more integrated and general approaches applicable to many fields.

10. CONCLUSION

The statistical understanding and treatment of heterogeneity have undergone a profound transformation from early efforts to test and eliminate "unwanted" heteroscedasticity. Statisticians have developed robust methods to accommodate heterogeneity and actively model, explain, and learn from its many types. This evolution mirrors broader shifts in scientific thinking: from seeking universal laws and average effects to embracing context-dependence and individual differences as fundamental aspects of reality.

Heteroscedasticity is now recognized not as a mere violation of assumptions of simple ANOVA and ordinary least-squares regression models but rather as an indicator of meaningful scientific information about differential variability. Models of general heterogeneity have challenged simplistic narratives about average effects and led to context-sensitive inference in fields ranging from genomics to ecology and economics. Survival analysis that incorporates heterogeneity has shown how unobserved frailty and selection processes shape population dynamics, with value to diverse fields, including aging research and engineering reliability. With the exponential growth in research output across all scientific fields, meta-analysis has become an exemplar for studies whose primary goals are to explicitly estimate and identify causal mechanisms of heterogeneity.

Heteroscedasticity, general heterogeneity, covariance and dependence, and frailty increasingly are being integrated within unified modeling frameworks. Mixed-effects models provide a foundational platform for joint specification. Survival analysis, with its unique focus on dynamic selection, provides a powerful lens on temporal heterogeneity. Bayesian hierarchical models reveal how uncertainty propagates through analysis of heterogeneity. Meta-analysis formalizes the synthesis of heterogeneous evidence. Machine learning, deep learning, and AI are being used to extend these methods to high-dimensional, complex Big datasets.

Future progress in understanding heterogeneity will depend on continued methodological innovation, accessible tools, and clear communication among statisticians, computer scientists, and researchers in fields where heterogeneity is now seen as meaningful and rich source of scientific insight. By embracing heterogeneity, researchers can build models that better reflect reality, ask more nuanced questions, and generate more meaningful, generalizable, and actionable knowledge in all scientific fields.

ACKNOWLEDGMENTS

We thank Ronny Vallejos for useful comments on an early draft of the manuscript.

FUNDING

Z(S)M's work on this review was supported in part by a 2023–2024 Charles Bullard Fellowship in Forest Research at Harvard University.

REFERENCES

AALEN, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* **7** 1121–1137. https://doi.org/10.1002/sim.4780071105

AALEN, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability* **2** 951–972. https://doi.org/10.1214/aoap/1177005583

AALEN, O., BORGAN, Ø. and GJESSING, H. (2008). *Survival and event history analysis: a process point of view*. Springer, New York. https://doi.org/10.1007/978-0-387-68560-1

AALEN, O. O., VALBERG, M., GROTMOL, T. and TRETLI, S. (2015). Understanding variation in disease risk: the elusive concept of frailty. *International Journal of Epidemiology* **44** 1408–1421. https://doi.org/10.1093/ije/dyu192

ADOM, P. K., ADAMS, S. and QUAGRAINIE, F. A. (2024). Parental financial inclusion and child proficiency in low-income settings: Family effects. *Development and Sustainability in Economics and Finance* **2–4** 100009. https://doi.org/10.1016/j.dsef.2024.100009

ANDERSON, T. W. and DARLING, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association* **49** 765–769.

ANSELIN, L. (1988). *Spatial Econometrics: Methods and Models*. Springer, Dordrecht.

AXON, E., DWAN, K. and RICHARDSON, R. (2023). Multiarm studies and how to handle them in a meta-analysis: A tutorial. *Cochrane Evidence Synthesis and Methods* **1** e12033. https://doi.org/10.1002/cesm.12033

BAIRD, D., ASMUS, H. and ASMUS, R. (2007). Trophic dynamics of eight intertidal communities of the Sylt-Rømø Bight ecosystem, northern Wadden Sea. *Marine Ecology Progress Series* **351** 25–41. https://doi.org/10.3354/meps07137

BALAN, T. A. and PUTTER, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research* **29** 3424–3454. https://doi.org/10.1177/0962280220921889

BALDUZZI, S., RÜCKER, G., NIKOLAKOPOULOU, A., PAPAKONSTANTINOU, T., SALANTI, G., EFTHIMIOU, O. and SCHWARZER, G. (2023). netmeta: An R Package for Network Meta-Analysis Using Frequentist Methods. *Journal of Statistical Software* **106** 1—40. https://doi.org/10.18637/jss.v106.i02

BANERJEEE, S., CARLIN, B. P. and GELFAND, A. E. (2025). *Hierarchical Modeling and Analysis for Spatial Data*, 3rd ed. Chapman and Hall/CRC, New York.

BARNARD, J., MCCULLOCH, R. and MENG, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10** 1281–1311. https://www.jstor.org/stable/24306780.

BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **160** 268–282.

BARTLETT, M. S. and KENDALL, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society* **8** 128–138.

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67** 1–48.

BAUJAT, B., MAHÉ, C., PIGNON, J.-P. and HILL, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine* **21** 2641–2652. https://doi.org/10.1002/sim.1221

BHAMBU, A., BERA, K., NATARAJAN, S. and SUGANTHAN, P. N. (2025). High frequency volatility forecasting and risk assessment using neural networks-based heteroscedasticity model. *Engineering Applications of Artificial Intelligence* **149** 110397. https://doi.org/10.1016/j.engappai.2025.110397

BLAND, J. M. and ALTMAN, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **1** 307–310.

BLAND, J. M. and ALTMAN, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8** 135–160. https://doi.org/10.1177/096228029900800204

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. and ROTHSTEIN, H. R. (2009). *Overview*. John Wiley & Sons, Ltd, Chichester. https://doi.org/10.1002/9780470743386.ch10

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26** 211–252.

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. and LJUNG, G. M. (2015). *Time Series Analysis: Forecasting and Control*, 5th ed. John Wiley & Sons, Ltd.

BOX-STEFFENSMEIER, J. M., CHRISTENSON, D. P. and MORGAN, J. W. (2018). Modeling unobserved heterogeneity in social networks with the frailty exponential random graph model. *Political Analysis* **26** 3–19. https://doi.org/10.1017/pan.2017.23

BOX-STEFFENSMEIER, J. M., DE BOEF, S. and JOYCE, K. A. (2017). Event dependence and heterogeneity in duration models: the conditional frailty model. *Political Analysis* **15** 237–256. https://doi.org/10.1093/pan/mpm013

BOX-STEFFENSMEIER, J. M., CAMPBELL, B. W., CHRISTENSON, D. P. and MORGAN, J. W. (2019). Substantive implications of unobserved heterogeneity: Testing the frailty approach to exponential random graph models. *Social Networks* **59** 141–153. https://doi.org/10.1016/j.socnet.2019.07.002

BREIMAN, L. (1996). Bagging Predictors. *Machine Learning* **24** 123–140. https://doi.org/10.1023/A:1018054314350

BROWN, M. B. and FORSYTHE, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association* **69** 364–367. https://doi.org/10.1080/01621459.1974.10482955

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76** 1–32. https://doi.org/10.18637/jss.v076.i01

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall/CRC, New York.

CASELLA, G. and BERGER, R. L. (2024). *Statistical inference*, 2nd ed. Chapman and Hall/CRC, New York.

CHAO, A., GOTELLI, N. J., HSIEH, T. C., SNADER, E. L., MA, K. H., COLWELL, R. K. and ELLISON, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84** 45–67. https://doi.org/10.1890/13-0133.1

CHUNG, Y., RABE-HESKETH, S., DORIE, V., GELMAN, A. and LIU, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika* **78** 685–709. https://doi.org/10.1007/s11336-013-9328-2

CLEVELAND, R. B. and CLEVELAND, W. S. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* **6** 3–73.

CLIFF, A. D. and ORD, J. K. (1973). *Spatial Autocorrelation*. Pion, London.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129. https://doi.org/10.2307/3001666

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295–313.

CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, New York.

CSÁRDI, G., NEPUSZ, T., TRAAG, V., HORVÁT, S., ZANINI, F., NOOM, D., MÜLLER, K., SCHOCH, D. and SALMON, M. (2026). igraph: Network analysis and visualization in R, version 2.2.2. https://CRAN.R-project.org/package=igraph.

DEEKS, J. J., ALTMAN, D. G. and BRADBURN, M. J. (2001). *Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis* In *Systematic Reviews in Health Care* 15, 285–312. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470693926.ch15

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7** 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

DORMANN, C. F. and ELLISON, A. M. (2025). *Statistics by Simulation: A Synthetic Data Approach*. Princeton University Press, Princeton.

DUTILLEUL, P. and LEGENDRE, P. (1993). Spatial heterogeneity against heteroscedasticity: an ecological paradigm versus a statistical concept. *Oikos* **66** 152–171. https://doi.org/10.2307/3545210

EMMERSON, M. C., MONTOYA, J. M. and WOODWARD, G. (2005). *Body size, interaction strength and food web dynamics* In *Dynamic food webs: multispecies assemblages, ecosystem development and environmental change* 167—178. Elsevier, New York.

FATH, B. D., ASMUS, H., ASMUS, R., BAIRD, D., BORRETT, S. R., DE JONGE, V. N., LUDOVISI, A., NIQUIL, N., SCHARLER, U. M., SCHÜCKEL, U. and WOLFF, M. (2019). Ecological network analysis metrics: The need for an entire ecosystem approach in management and policy. *Ocean & Coastal Management* **174** 1–14. https://doi.org/10.1016/j.ocecoaman.2019.03.007

FERREIRA, L. N., VEGA-OLIVEROS, D. A., COTACALLAPA, M., CARDOSO, M. F., QUILES, M. G., ZHAO, L. and MACAU, E. E. N. (2020). Spatiotemporal data analysis with chronological networks. *Nature Communications* **11** 4036. https://doi.org/10.1038/s41467-020-17634-2

FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1930). *The Genetical Theory of Natural Selection (1999 Complete Variorum Edition, edited with a foreword and notes by J. H. Bennett)*. Oxford University Press, Oxford.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97** 611–631.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

GELFAND, A. E. and SCHLIEP, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics* **18** 86–104. https://doi.org/10.1016/j.spasta.2016.03.006

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton.

GIRARDIN, V., GRENTE, T., NIQUIL, N. and REGNAULT, P. (2023). Analysis of ecological networks: Linear inverse modeling and information theory tools. *Physical Sciences Forum* **9** 24. https://doi.org/10.3390/psf2023009024

GOLDSTEIN, H. (2011). *Multilevel Statistical Models*, 4th ed. Wiley, New York.

GONNERMANN, A., FRAMKE, T., GROSSHENNIG, A. and KOCH, A. (2015). No solution yet for combining two independent studies in the presence of heterogeneity. *Statistics in Medicine* **34** 2476–2480. https://doi.org/10.1002/sim.6473

GOTELLI, N. J. and ELLISON, A. M. (2012). *A Primer of Ecological Statistics*, 2nd ed. Oxford University Press, New York.

GUO, T., SHEN, D., KOU, Y. and NIE, T. (2025). Multi-task self-supervised learning based fusion representation for Multi-view clustering. *Information Sciences* **694** 121705. https://doi.org/10.1016/j.ins.2024.121705

GUREVITCH, J., KORICHEVA, J., NAKAGAWA, S. and STEWART, G. (2018). Meta-analysis and the science of research synthesis. *Nature* **555** 175–182. https://doi.org/10.1038/nature25753

HALLOCK, H., MARSHALL, S. E., 'T HOEN, P. A. C., NYGÅRD, J. F., HOORNE, B., FOX, C. and ALAGARATNAM, S. (2021). Federated networks for distributed analysis of health data. *Frontiers in Public Health* **9** 712569. https://doi.org/10.3389/fpubh.2021.712569

HAMILTON, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57** 357–384.

HANSEN, B. E. (2001). The new econometrics of structural change: Dating breaks in U.S. labour productivity. *Journal of Economic Perspectives* **15** 117–128. https://doi.org/10.1257/jep.15.4.117

HARDY, R. J. and THOMPSON, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* **17** 841–856. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<841::aid-sim781>3.0.co;2-d

HARTLEY, H. O. (1950). The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika* **37** 308–312.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*, 1st ed. Chapman and Hall/CRC, New York.

HE, H., CAO, S., ZHANG, J.-G., SHEN, H., WANG, Y.-P. and DENG, H.-W. (2019). A statistical test for differential network analysis based on inference of Gaussian graphical model. *Scientific Reports* **9** 10863. https://doi.org/10.1038/s41598-019-47362-7

HEDGES, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* **6** 107–128. https://doi.org/10.3102/10769986006002107

HIGGINS, J. P. T. and THOMAS, J., eds. (2024). *Cochrane Handbook for Systematic Reviews of Interventions, Version 6.5*. Cochrane Available from www.chochrane.org/handbook.

HIGGINS, J. P. T. and THOMPSON, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21** 1539–1558. https://doi.org/10.1002/sim.1186

HIGGINS, J. P. T., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2008). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society* **172** 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236. https://doi.org/10.1093/pan/mpl013

HOLZMEISTER, F., JOHANNESSON, M., BÖHM, R., DREBER, A., HUBER, J. and KIRCHLER, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences* **121** e2403490121. https://doi.org/10.1073/pnas.2403490121

INTHOUT, J., IOANNIDIS, J. P. and BORM, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology* **14** 25. https://doi.org/10.1186/1471-2288-14-25

JACKSON, D., WHITE, I. R. and RILEY, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* **31** 3805-20. https://doi.org/10.1002/sim.5453

JACKSON, D., VERONIKI, A. A., LAW, M., TRICCO, A. C. and BAKER, R. (2017). Paule-Mandel estimators for network meta-analysis with random inconsistency effects. *Research Synthesis Methods* **8** 416–434. https://doi.org/10.1002/jrsm.1244

KEIDING, N., ANDERSEN, P. K. and KLEIN, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16** 215–224.

KOLASA, J. and ROLLO, C. D. (1991). *Introduction: The heterogeneity of heterogeneity: A glossary* In *Ecological Heterogeneity* 1–23. Springer, New York. https://doi.org/10.1007/978-1-4612-3062-5_1

KOLMOGOROFF, A. (1933). Sulla determinazione empírica di uma legge di distribuzione. *Giornale dell''Istituto italiano degli attuari* 83–92.

KREFELD-SCHWALB, A., HUA, X. and JOHNSON, E. J. (2025). Measuring population heterogeneity requires heterogeneous populations. *Proceedings of the National Academy of Sciences* **122** e2425536122. https://doi.org/10.1073/pnas.2425536122

KRUSKAL, W. H. and WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47** 583–621.

KRZYZIŃSKI, M., SPYTEK, M., BANIECKI, H. and BIECEK, P. (2023). SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems* **262** 110234. https://doi.org/10.1016/j.knosys.2022.110234

KUHN, T. S. (1996). *The Structure of Scientific Revolutions*, 3rd ed. University of Chicago Press, Chicago.

KVAMME, H., ØRNULF BORGAN and SCHEEL, I. (2019). Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research* **20** 1–30. http://jmlr.org/papers/v20/18-424.html.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LANGAN, D., HIGGINS, J. P. T., JACKSON, D., BOWDEN, J., VERONIKI, A. A., KONTOPANTELIS, E., VIECHTBAUER, W. and SIMMONDS, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods* **10** 83–98. https://doi.org/10.1002/jrsm.1316

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, New York.

LEAMER, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review* **73** 31–43.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York.

LEVENE, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, ed.) 278–292. Stanford University Press.

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100** 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

LI, H. and REYNOLDS, J. F. (1995). On definition and quantification of heterogeneity. *Oikos* **73** 280–284. https://doi.org/10.2307/3545921

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LIU, Z. (2025). Model-based optimal randomization procedure for treatment–covariate interaction tests. *Statistical Methods in Medical Research* **34** 1732–1750. https://doi.org/10.1177/09622802241298703

MA, Z. (2025). Taylor's Power Law meets ecological networks: Introducing TPLoN as a novel framework for heterogeneity analysis. *Oikos* **2025** e10956. https://doi.org/10.1002/oik.10956

MA, Z. S. and ELLISON, A. M. (2025). A novel three-layer methodology for heterogeneity analysis: From metrics, through statistical tests and modeling to applications. *bioRxiv*. https://doi.org/10.64898/2025.12.18.695221

MA, Z. and KRINGS, A. (2008a). Multivariate survival analysis (I): Shared frailty approaches to reliability and dependence modeling. In *2008 IEEE Aerospace Conference* 1–21. https://doi.org/10.1109/AERO.2008.4526618

MA, Z. and KRINGS, A. W. (2008b). Competing risks analysis of reliability, survivability, and prognostics and health management (PHM). In *2008 IEEE Aerospace Conference* 1–21. https://doi.org/10.1109/AERO.2008.4526636

MA, Z. and KRINGS, A. W. (2008c). Survival analysis approach to reliability, survivability and prognostics and health management (PHM). In *2008 IEEE Aerospace Conference* 1–20. https://doi.org/10.1109/AERO.2008.4526634

MA, Z. and KRINGS, A. W. (2011). Dynamic hybrid fault modeling and extended evolutionary game theory for reliability, survivability and fault tolerance analyses. *IEEE Transactions on Reliability* **60** 180–196. https://doi.org/10.1109/TR.2011.2104997

MAXWELL, S. E. and DELANEY, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd ed. Routledge, New York.

MAYR, E. (1942). *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. Harvard University Press, Cambridge.

MCCANN, K. S. (2011). *Food Webs*. Princeton University Press, Princeton.

MCINTOSH, R. P. (1991). *Concept and terminology of homogeneity and heterogeneity in ecology* In *Ecological Heterogeneity* 24–46. Springer, New York. https://doi.org/10.1007/978-1-4612-3062-5_2

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

MIGLIAVACA, C., STEIN, C., COLPANI, V., BARKER, T., ZIEGELMANN, P. and FALAVIGNA, M. (2022). Meta-analysis of prevalence: I^2 statistic and how to deal with heterogeneity. *Research Synthesis Methods* **13** 363–367. https://doi.org/10.1002/jrsm.1547

MONNAHAN, C. C., THORSON, J. T. and BRANCH, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* **8** 339–348. https://doi.org/10.1111/2041-210X.12681

MORAN, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37** 17–23.

MUGGEO, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine* **22** 3055–3071. https://doi.org/10.1002/sim.1545

MUTHÉN, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences* 345–368.

NOSEK, B. A., EBERSOLE, C. R., DEHAVEN, A. C. and MELLOR, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA* **11** 2600–2606. https://doi.org/10.1073/pnas.170827411

PAGAN, A. R. and BREUSCH, T. S. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47** 1287–1294.

PETERSEN, J. H., ANDERSEN, P. K. and GILL, R. D. (1996). Variance components models for survival data. *Statistica Neerlandica* **50** 193–211. https://doi.org/10.1111/j.1467-9574.1996.tb01487.x

PETITTI, D. B. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine* **20** 3625–3633. https://doi.org/10.1002/sim.1091

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*, 1st ed. Springer. https://doi.org/10.1007/b98882

PRICE, G. R. (1970). Selection and covariance. *Nature* **227** 520–521. https://doi.org/10.1038/227520a0

RAUDENBUSH, S. W. and BRYK, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Sage Publications, Thousand Oaks.

REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021a). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics* **78** 524–535. https://doi.org/10.1111/biom.13426

REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021b). HeteroGGM: an R package for Gaussian graphical model-based heterogeneity analysis. *Bioinformatics* **37** 3073–3074. https://doi.org/10.1093/bioinformatics/btab134

REN, Q., WANG, Y., FANG, R., LING, H. and YOU, C. (2025). OTSurv: a novel multiple instance learning framework for survival prediction with heterogeneity-aware optimal transport. *arXiv*. https://doi.org/10.48550/arXiv.2506.20741

RÉNYI, A. (1961). On measures of information and entropy. In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960* 547–561.

RIPLEY, B. D. (1981). *Spatial statistics*. Wiley, New York.

RIPLEY, B. D. (1987). Spatial point pattern analysis in ecology. In *Developments in numerical ecology. NATO ASI Series, Vol. G 14.* (L. P. and L. Legendre, eds.) 407–429. Springer, Berlin.

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician* **59** 147–152.

RÜCKER, G. and SCHWARZER, G. (2025). Trials and triangles: Network meta-analysis of multi-arm trials with correlated arms. *Research Synthesis Methods* **16** 961–974. https://doi.org/10.1017/rsm.2025.10026

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian Models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B* **71** 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

SCHEFFÉ, H. (1959). *The analysis of variance*. Wiley, Oxford.

SENIOR, A. M., GRUEBER, C. E., KAMIYA, T., LAGISZ, M., O'DWYER, K., SANTOS, E. S. A. and NAKAGAWA, S. (2016). Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications. *Ecology* **97** 3293–3299. https://doi.org/10.1002/ecy.1591

SHAVIT, A. and ELLISON, A. M. (2021). Diverse populations are conflated with heterogeneous collectives. *The Journal of Philosophy* **118** 525–548. https://doi.org/10.5840/jphil20211181037

SHROUT, P. E. and FLEISS, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86** 420–428. https://doi.org/10.1037//0033-2909.86.2.420

SIMMONS, J. P., NELSON, L. D. and SIMONSOHN, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22** 1359–1366. https://doi.org/10.1177/0956797611417632

SOSA, S., JACOBY, D. M. P., LIHOREAU, M. and SUEUR, C. (2021). Animal social networks: Towards an integrative framework embedding social interactions, space and time. *Methods in Ecology and Evolution* **12** 4–9. https://doi.org/10.1111/2041-210X.13539

STUDENT (1908). The probable error of a mean. *Biometrika* **6** 1–25.

R CORE TEAM (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

THOMPSON, S. G. and HIGGINS, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* **21** 1559–1573. https://doi.org/10.1002/sim.1187

THOMPSON, L. R., SANDERS, J. G., MCDONALD, D., [...] and EARTH MICROBIOME PROJECT CONSORTIUM (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551** 457—463. https://doi.org/10.1038/nature24621

ULANOWICZ, R. E. (2004). Quantitative methods for ecological network analysis. *Computational Biology and Chemistry* **28** 321–339.

VAUPEL, J. W., MANTON, K. G. and STALLARD, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16** 439–454. https://doi.org/10.2307/2061224

VEUM, K. S., PARKER, P. A., HOLAN, S. H., PAIS, N. V., WILLS, S. A., AMSILI, J. P., NUNES, M. R., VAN ES, H. M., SEYBOLD, C. A. and KARLEN, D. L. (2025). Spatially explicit heteroskedastic modeling for the Soil Health Assessment Protocol and Evaluation version 1.0S. *Soil Science Society of America Journal* **89** e70065. https://doi.org/10.1002/saj2.70065

VIECHTBAUER, W. and CHEUNG, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* **1** 112–125. https://doi.org/10.1002/jrsm.11

VON HIPPEL, P. T. (2015). The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology* **15** 35. https://doi.org/10.1186/s12874-015-0024-z

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

WANG, R., LAGAKOS, S. W., WARE, J. H., HUNTER, D. J. and DRAZEN, J. M. (2007). Statistics in medicine — reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357** 2189–2194. https://doi.org/10.1056/NEJMsr077003

WANG, T., REN, Z., DING, Y., FANG, Z., SUN, Z., MACDONALD, M. L., SWEET, R. A., WANG, J. and CHEN, W. (2016). FastGGM: an efficient algorithm for the inference of Gaussian graphical model in biological networks. *PLoS Computational Biology* **12** e1004755. https://doi.org/10.1371/journal.pcbi.1004755

WEN, Z., FANG, Y., WEI, P., LIU, F., CHEN, Z. and WU, M. (2025). Temporal and heterogeneous graph neural network for remaining useful life prediction. *IEEE Transactions on Neural Networks and Learning Systems* **36** 19748–19761. https://doi.org/10.1109/TNNLS.2025.3592788

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838.

WIEGREBE, S., KOPPER, P., SONABEND, R., BISCHL, B. and BENDER, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review* **57** 65. https://doi.org/10.1007/s10462-023-10681-3

WILK, M. B. and GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55** 1–17.

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315370279

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge.

YASHIN, A. I. and IACHINE, I. A. (1999). What difference does the dependence between durations make? Insights for population studies of aging. *Lifetime Data Analysis* **5** 5–22. https://doi.org/10.1023/a:1009622214567

ZAJACOVA, A., GOLDMAN, N. and RODRÍGUEZ, G. (2009). Unobserved heterogeneity can confound the effect of education on mortality. *Mathematical population studies* **16** 153–173. https://doi.org/10.1080/08898480902790528

ZARULLI, V. (2016). Unobserved heterogeneity of frailty in the analysis of socioeconomic differences in health and mortality. *European Journal of Population* **32** 55–72. https://doi.org/10.1007/s10680-015-9361-1