Health Sciences & Technology 190
Syllabus for *Introduction to Biostatistics in Medicine*
Summer 2025: class meets MWF 09:00AM–12:00PM in
HMS Tosteson Medical Education Center (TMEC) 209

Faculty Instructor: Nima Hejazi, PhD
E-mail: nhejazi@hsph.harvard.edu
Webpage: https://nimahejazi.org
Office Hours Location: TBA
Office Hours: Mon. 12:00–1:00PM or by appt.

Teaching Fellow: Nana Okada
E-mail: nokada@hms.harvard.edu
Office Hours Location: TMEC 338
Office Hours: Fri. 01:15–02:30PM

The instructional staff reserve the right to make changes to this syllabus at any time.

Course Description:

Evidence-based medicine—and evidence-based innovation in the health and medical sciences—depends critically on the *evaluation and interpretation* of data. Not only does this require an understanding of the scientific subject matter (i.e., a medical specialty), but it also necessitates a working knowledge of the *research methodology* used to distill experimental data into evidence. Biostatistics serves as a foundational pillar of research methodology in the health and medical sciences, ideally acting as an engine for evidence evaluation and as a safeguard against drawing incorrect inferences from data. This course provides training on how to comprehend, critique and communicate findings from the health and medical sciences literature, focusing especially on how to assess the importance of chance in the interpretation of experimental data. Major topics to be addressed include elementary probability theory, hypothesis testing, analysis of variance, the general linear model and generalized linear models (e.g., logistic regression), survival analysis and causal inference, and reproducible statistical data analysis. These topics will be reviewed together with the critical reading of studies published in the health and medical sciences literature. Emphasis will be placed on understanding how statistical thinking can help to prevent mistakes that commonly arise from relying upon intuition alone and how even the rote application of statistical techniques and tools can lead to logical mistakes that invalidate an analysis and undermine its scientific conclusions.

Prerequisites and Recommendations:

The course is required for medical (MD) and doctoral (PhD) students in the MIT–Harvard *Health Sciences & Technology* doctoral program. It is generally not open to students in other educational programs, whether at Harvard University or the Massachusetts Institute of Technology, though exceptions may be made on a case-by-case basis with prior approval of the instructor.

- *Pre-requisites:* Intermediate calculus and basic linear algebra.
- *Recommended:* Prior introduction to numerical/statistical programming.

Students are expected to have some prior training in mathematics, including intermediate calculus (partial differentiation, integration) and elementary linear algebra (matrices, determinants, vector spaces). Although students are not expected to be fluent in a high-level statistical or numerical programming language (e.g., `R`, `Julia`), modern biostatistics is a computing-intensive discipline; thus, familiarity with the basics of numerical/statistical programming will be helpful in wading through the material at a fast pace. Students are expected to have the motivation and intellectual maturity to acquire any technical skills and tools they might lack to succeed in the course—without significant help, as that is what is required of professionals.

Requirements and Materials:

The course primarily follows a single textbook: Vu and Harrington (2020)'s *Introductory Statistics for the Life and Biomedical Sciences*, digital copies of which are freely available at `https://openintro.org/book/biostat`; readings assigned from this textbook are required, and some of the material will not be explicitly reviewed in the course lectures. Although lecture notes will also be distributed, these are intended to supplement readings from the textbook. A non-exhaustive list of other possibly helpful and well-written monographs, given in reverse chronological order of their publication, appears below (n.b., some on this list are more advanced† than others). Some material will be drawn from these texts as well.

- Fay and Brittain (2022), *Statistical Hypothesis Testing in Context*
- Brumback (2021), *Fundamentals of Causal Inference with R*
- James et al. (2021), *An Introduction to Statistical Learning with Applications in R*
- Westreich (2019), *Epidemiology by Design: A Causal Approach to the Health Sciences*
- Pearl et al. (2016), *Causal Inference in Statistics: A Primer*
- Friedman et al. (2015), *Fundamentals of Clinical Trials*
- Freedman (2009), *Statistical Models: Theory and Practice*†
- Dalgaard (2008), *Introductory Statistics with R*
- Jewell (2004), *Statistics for Epidemiology*†
- Altman (1990), *Practical Statistics for Medical Research*

Textbook readings and lectures notes will be supplemented by primary research articles and short tutorials chosen from the health and medical sciences literature or the (bio)statistics literature. *The BMJ* has also published a series of "Statistics Notes" (`https://www.bmj.com/specialties/statistics-notes`), which are intended for applied scientists and medical researchers—these make for excellent overviews of many topics, so we will review a selection of these as well. To develop an appreciation for statistical science, you may also *optionally* read Senn (2022)'s *Dicing with Death: Living by Data*, a recently updated popular science book on health and medical statistics.

Lecture notes, classroom activity materials, homework assignments, upcoming due dates, and this syllabus will be made available via the course's Canvas intranet website at `https://canvas.harvard.edu/courses/159381`. Materials from the previous iteration of the course are archived at `https://hst190.nimahejazi.org`.

Pedagogic note: This is an introductory course intended to equip students with the basic skills and necessary background to evaluate statistical evidence as it is presented in the health and medical sciences literature. The emphasis is on *critical thinking* and *statistical rethinking*—not on performing mathematical manipulations by rote. A key skill in thinking as a (bio)statistician is to learn to translate scientific questions into statistical queries, which may then be addressed by the application of statistical methods to carefully collected experimental data, and to appreciate the fundamental limits of what it is possible to learn from a given experimental or observational study. Necessarily, the focus of this course will *not* be on *how to apply* various common statistical methods (e.g., to develop a "statistics toolbox") but instead on understanding *when* and *why* a chosen methodological approach may be appropriate and *how to interpret* the results of the application of appropriately selected statistical procedures in relevant scientific context.

Course Policies, Expectations, and Grading Scheme

Accommodations: Please speak with the instructional staff as soon as possible if you require any particular accommodations, and we will work out any necessary arrangements as best as possible.

Scheduling Conflicts: Notify the instructional team *by the second class session* about any known or potential conflicts (e.g., religious observances, job interviews, professional conferences). Due to the compressed and interactive nature of the class, you should not enroll in the course if you foresee the possibility of missing the equivalent of two or more class meetings.

Collaboration and Independence: All homework assignments should clearly list collaborators and references. Homework assignments will not be considered for credit if they are a replicate of another submitted assignment. While collaboration is highly encouraged, assignment submissions should be written up entirely independently.

Grading Scheme: Per HST policy, there are two components to the final grade:

- A ***Content*** *grade* based on your mastery of the course didactic material as evidenced through various assessments. This will be evaluated primarily by way of
  - *Course project*: A final project will be assigned in lieu of a final exam. Groups of 2-4 students will work on these projects together, learning from each other in the process. The project provides a hands-on opportunity to draw on the methods discussed in class to develop a statistical analysis plan tied to a motivating scientific question and dataset, while allowing for the receipt of feedback from peers and the instructional team. Project requirements include a brief presentation during the final class session and a short write-up describing the scientific question and recommended statistical analysis plan. *Project presentations will take place during the final class session—Friday, September 5th.* A handout outlining specific expectations will be provided two weeks prior to this date.
  - *Assignments*: Each week, a homework problem set, to be handed in via Canvas, will be distributed. Each student must submit their own solutions to these assignments. Answer keys to these assessments will *not* be distributed; however, following each deadline, small groups of students will be asked to revisit each problem set and to develop a new set of solutions "in committee"; groups of students will be randomly selected to present these solutions. Given the compressed timeline of the course, no late assignments will be accepted for any reason. Understanding of the content of these assessments will also be evaluated via short, open-note "concept checks" administered at random during the class sessions.
- A ***Professionalism*** *grade* based on attendance, engagement, effort made to seek assistance, and the approach taken to homework solutions.
  - *Attendance*: The levels of background and prior statistical experience vary considerably within the HST cohort: Some students will have already encountered (and mastered) some of the technical ideas covered in the course, while others may be encountering the ideas for the first time. The course is not intended merely to review technical ideas but is designed to encourage the development of skills and intuition required to appreciate when and how statistical methods can be brought to bear on problems in the health and medical sciences—an endeavor best undertaken carefully and collaboratively. The goal is for individual students to contribute to the entire class's collective learning, which requires in-the-classroom attendance. While attendance will not be formally recorded, it

is clear to the instructional team who is absent from the room. If it is necessary to miss a class meeting for any reason, please inform the course instructor well in advance.

- *Engagement*: The in-class group exercises are intended to provide opportunities for small groups of students to grapple with a problem—that is, to *work together* on solutions and to push each other to understand different approaches and perspectives. For those with prior experience, it may seem "easier" to quickly solve a problem on their own—but these exercises are also about *working with peers*, comparing ideas and teaching others where needed. Successful completion of the tasks is less important that the group experience; thus, it is necessary for students to remain fully engaged during these exercises. As the instructional team will circulate between groups during these exercises—checking in with groups and helping guide them (as needed), and posing new questions when appropriate—they will also be assessing the level of engagement between peers.

- *Homework solutions*: Students are encouraged to (orally) discuss problem sets with their peers, but each student must write up solutions separately. It is imperative to have worked through each problem individually and to make sure that the answers submitted, inclusive of computer code, are the results of one's own understanding of the problem (*the goal is learning, not accruing "points"*). In situations in which the level of collaboration may be questionable (rather than clearly appropriate), it is the students' responsibility to check in with the instructor. Students are also required to acknowledge all collaborators when submitting completed assignments. Furthermore, the use of `ChatGPT` or similar tools is entirely forbidden. Note that beyond the grading of homework, the completion of assignments will be assessed via in-class "concept checks" (administered at random), and neither `ChatGPT` nor peer collaborators will be of much help in such situations.

- *Seeking assistance*: Both the course instructor and the teaching fellow will have office hours each week. Students should also feel free to reach out to the instructional staff to set up individual meetings to discuss the course material. Our goal is to let everyone feel that they have learned something in the course (even if they have great statistical skills coming in) and for everyone to finish the course with a sense of general comfort in approaching statistical problems and questions in the health and medical sciences.

Course Evaluation

Constructive feedback from students is a valuable resource for improving the teaching and learning experience. The feedback should be specific, focused, and respectful. It should address aspects of the course and teaching that are positive, as well as those which need improvement.

Course Outline

The course is divided into a series of modules. The exact weekly coverage of topics is subject to change, as it will depend on the progress of the class. Anticipated time for the completion of each module is given below. Since the course consists in roughly 10 3-hour meetings, each module will be covered in just one or two class sessions. (n.b., The use of ⚕ below denotes *required* clinically oriented reading assignments, on which in-class group-based learning activities will be based.)

1. **Biostatistics and data science in medicine.** Brief history of biostatistics; the role of (bio)statistics in evaluating evidence; principles of data collection; numerical and categorical data; exploratory data analysis; principles of reproducible statistical data science; overview of observational studies and randomized controlled trials (RCTs).

- Time anticipated: 1 meeting (Mon., Aug. 11$^{\text{th}}$)
- Books and tutorials: Vu and Harrington (2020, Sec. 1.2, 1.3.5–1.3.7, 1.4–1.6), Altman (1990, Ch. 1–3), Jewell (2004, Ch. 1–2), Freedman (2009, Ch. 1), Senn (2022, Ch. 1–2)
- Notes from *The BMJ*: Altman and Bland (1999b), Altman and Bland (1994), Altman and Bland (1998a), Altman and Bland (1999a), Bland and Altman (2011)
- Literature and commentaries: Kluytmans-van Den Bergh et al. (2020, presenting data) ⚕, Altman (1998), van der Laan (2015), Stark and Saltelli (2018), Wainer (1984), Broman and Woo (2018), Peng and Hicks (2021), Jordan (2019)

2. **Concepts of probability.** Basic axioms and laws of probability and conditional probability; Bayes' theorem; random variables and their essential properties; expectation, conditional expectation, and covariance; common families of random variables (Normal, Binomial, Poisson); marginal, joint, and conditional distributions; Bernoulli trials.
   - Time anticipated: 1 meeting (Wed., Aug. 13$^{\text{th}}$)
   - Books and tutorials: Vu and Harrington (2020, Sec. 2.1–2.2, 3.1.1–3.1.3, 3.2–3.4), Altman (1990, Ch. 4), Jewell (2004, Ch. 3), Senn (2022, Ch. 3–4)
   - Notes from *The BMJ*: None assigned
   - Literature and commentaries: Spiegelhalter (2024, on probability), Freedman and Stark (2001, on probability), Altamirano et al. (2020, measuring incidence) ⚕, Milne and Whitty (1995, using approximations) ⚕

3. **Concepts of statistics.** Point estimation and variability of point estimates; confidence intervals and hypothesis testing (and their duality); the Normal and t-distributions (as sampling distributions of test statistics); one-sample inference (means and proportions); two-sample inference for paired and independent data (difference-of-means and difference-of-proportions); analysis of variance (ANOVA); contingency tables, χ^2 tests of association, and Fisher's exact test; overview of permutation testing.
   - Time anticipated: 2 meetings (Fri., Aug. 15$^{\text{th}}$; Mon., Aug. 18$^{\text{th}}$)
   - Books and tutorials: Vu and Harrington (2020, Sec. 4.1–4.3, 5.1–5.5, 8.1–8.5), Altman (1990, Ch. 5, 8–10), Jewell (2004, Ch. 4–7, 11), Senn (2022, Ch. 5)
   - Notes from *The BMJ*: Bland and Altman (1994a), Bland and Altman (1995), Altman and Bland (1996), Altman and Bland (2005), Altman and Bland (2014a)
   - Literature and commentaries: Cranston et al. (2020, inference with one sample) ⚕, Chowdhary et al. (2020, inference with two samples) ⚕, Cole et al. (2021)

4. **Linear regression and its generalizations.** Regression to the mean; the general linear model and interpretation of model parameters; estimating linear regression parameters via least squares; point estimation of and inference for regression parameters; multiple linear regression and statistical interaction; ANOVA (revisited) and ANCOVA; model evaluation and model selection; introduction to generalized linear models (GLMs) via logistic regression.
   - Time anticipated: 3 meetings (Wed., Aug. 20$^{\text{th}}$; Fri., Aug. 22$^{\text{nd}}$; Mon., Aug. 25$^{\text{th}}$)
   - Books and tutorials: Vu and Harrington (2020, Sec. 6.1–6.5, 7.1–7.8), LaValley (2008), Altman (1990, Ch. 11–12, 15), Jewell (2004, Ch. 12–13), Freedman (2009, Ch. 2–5), Senn (2022, Ch. 6, 8)
   - Notes from *The BMJ*: Altman and Royston (2006), Bland and Altman (1994b), Bland and Altman (1994c), Bland and Altman (2000)
   - Literature and commentaries: Turner et al. (2020, linear regression) ⚕, Avadhanula et al. (2020, logistic regression) ⚕, Freedman (2008)

5. **Survival analysis, causal inference, and clinical trials.** Time-to-event outcomes and right-censoring; the Kaplan-Meier estimator of the survival curve, Greenwood's formula, and the log-rank test; overview of the Cox proportional hazards model; RCTs, randomization, and confounding; the Neyman-Rubin potential outcomes framework, the fundamental problem of causal inference, and target trial emulation; the estimands framework for clinical trials.
   - Time anticipated: 3 meetings (Wed., Aug. 27^{th}; Fri., Aug. 29^{th}; Wed., Sep. 3^{rd})
   - Books and tutorials: Clark et al. (2003), Bradburn et al. (2003), Dalgaard (2008, Ch. 14), Altman (1990, Ch. 13), Jewell (2004, Ch. 8, 17), Senn (2022, Ch. 7, 9)
   - Notes from *The BMJ*: Altman and Bland (1998b), Bland and Altman (1998), Bland and Altman (2004), Altman and Bland (2007), Altman and Bland (2014b)
   - Literature and commentaries: Nguyen et al. (2019, survival analysis) ⚕, Hernán et al. (2008, causal inference) ⚕, Weir et al. (2024, estimands framework) ⚕, Kahan et al. (2024), Hernán and Robins (2006), Hernán and Robins (2016), Hernán (2018), US Food and Drug Administration (2021), Kahan et al. (2023)

Class Session Structure

The class sessions focus on *synchronous learning*: Each of the 10 learning sessions will be 3 hours, and will be divided unevenly into three parts:

- Part I (1.25 hours) begins with a brief 3-5 question, open-note worksheet intended to cover the textbook-based reading material. After a 10-minute period to complete said worksheet, a random selection of students will present their answers orally as a "concept check." After this, approximately 1 hour will be dedicated to a lecture delivered by the faculty instructor. A 10-minute break will be provided before Part II.
- Part II (0.5 hours) is devoted to a guided discussion of the assigned reading from the primary literature—on health and medical sciences or applied (bio)statistics—focusing on aspects of the reading that pertain specifically to the material discussed in the course. Each discussion will focus on a clinically oriented paper (marked with the ⚕ symbol above) and will be led by a small group of students assigned in advance of the given class session.
- Part III (1.0 hour) is dedicated to students working in small groups (of 2-4) for approximately 45 minutes. During this portion, students will work as a group to answer a problem set consisting of 2-3 problems. One student in each group will be assigned the role of note-taker. The remaining time will consist of a guided discussion of the problem sets, with the assigned note-taker serving as the representative of each group.

Throughout each class session, Parts I and II will be separated by a 5-10 minute break, as will Parts II and III. During these breaks, the instructional team will remain available for questions.

References

Jonathan Altamirano, Prasanthi Govindarajan, Andra L Blomkalns, Lauren E Kushner, Bryan Andrew Stevens, Benjamin A Pinsky, and Yvonne Maldonado. Assessment of sensitivity and specificity of patient-collected lower nasal specimens for Severe Acute Respiratory Syndrome Coronavirus 2 testing. *JAMA Network Open*, 3(6):e2012005, 2020. doi:10.1001/jamanetworkopen.2020.12005.

Douglas G Altman. *Practical Statistics for Medical Research*. CRC Press, 1990.

Douglas G Altman. Statistical reviewing for medical journals. *Statistics in Medicine*, 17(23): 2661–2674, 1998. doi:10.1002/(SICI)1097-0258(19981215)17:23<2661::AID-SIM33>3.0.CO;2-B.

Douglas G Altman and J Martin Bland. Statistics notes: Quartiles, quintiles, centiles, and other quantiles. *The BMJ*, 309(6960):996, 1994. doi:10.1136/bmj.309.6960.996.

Douglas G Altman and J Martin Bland. Statistics notes: Comparing several groups using analysis of variance. *The BMJ*, 312(7044):1472–1473, 1996. doi:10.1136/bmj.312.7044.1472.

Douglas G Altman and J Martin Bland. Generalisation and extrapolation. *The BMJ*, 317(7155): 409, 1998a. doi:10.1136/bmj.317.7155.409.

Douglas G Altman and J Martin Bland. Time to event (survival) data. *The BMJ*, 317(7156): 468–469, 1998b. doi:10.1136/bmj.317.7156.468.

Douglas G Altman and J Martin Bland. Treatment allocation in controlled trials: Why randomise? *The BMJ*, 318(7192):1209, 1999a. doi:10.1136/bmj.318.7192.1209.

Douglas G Altman and J Martin Bland. Statistics notes: Variables and parameters. *The BMJ*, 318 (7199):1667, 1999b. doi:10.1136/bmj.318.7199.1667.

Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *The BMJ*, 331 (7521):903, 2005. doi:10.1136/bmj.331.7521.903.

Douglas G Altman and J Martin Bland. Missing data. *The BMJ*, 334(7590):424, 2007. doi:10.1136/bmj.38977.682025.2C.

Douglas G Altman and J Martin Bland. Uncertainty and sampling error. *The BMJ*, 349, 2014a. doi:10.1136/bmj.g7064.

Douglas G Altman and J Martin Bland. Uncertainty beyond sampling error. *The BMJ*, 349, 2014b. doi:10.1136/bmj.g7065.

Douglas G Altman and Patrick Royston. The cost of dichotomising continuous variables. *The BMJ*, 332(7549):1080, 2006. doi:10.1136/bmj.332.7549.1080.

Shirisha Avadhanula, Wendy J Introne, Sungyoung Auh, Steven J Soldin, Brian Stolze, Debra Regier, Carla Ciccone, Fady Hannah-Shmouni, Armando C Filie, Kenneth D Burman, et al. Assessment of thyroid function in patients with alkaptonuria. *JAMA Network Open*, 3(3):e201357, 2020. doi:10.1001/jamanetworkopen.2020.1357.

J Martin Bland and Douglas G Altman. Statistics notes: One- and two-sided tests of significance. *The BMJ*, 309(6949):248, 1994a. doi:10.1136/bmj.309.6949.248.

J Martin Bland and Douglas G Altman. Statistics notes: Regression towards the mean. *The BMJ*, 308(6942):1499, 1994b. doi:10.1136/bmj.308.6942.1499.

J Martin Bland and Douglas G Altman. Statistics notes: Some examples of regression towards the mean. *The BMJ*, 309(6957):780, 1994c. doi:10.1136/bmj.309.6957.780.

J Martin Bland and Douglas G Altman. Statistics notes: The normal distribution. *The BMJ*, 310 (6975):298, 1995. doi:10.1136/bmj.310.6975.298.

J Martin Bland and Douglas G Altman. Survival probabilities (the Kaplan-Meier method). *The BMJ*, 317(7172):1572, 1998. doi:10.1136/bmj.317.7172.1572.

J Martin Bland and Douglas G Altman. The odds ratio. *The BMJ*, 320(7247):1468, 2000. doi:10.1136/bmj.320.7247.1468.

J Martin Bland and Douglas G Altman. The logrank test. *The BMJ*, 328(7447):1073, 2004. doi:10.1136/bmj.328.7447.1073.

J Martin Bland and Douglas G Altman. Comparisons within randomised groups can be very misleading. *The BMJ*, 342, 2011. doi:10.1136/bmj.d561.

Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas Graham Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British Journal of Cancer*, 89(3):431–436, 2003. doi:10.1038/sj.bjc.6601119.

Karl W Broman and Kara H Woo. Data organization in spreadsheets. *The American Statistician*, 72(1):2–10, 2018. doi:10.1080/00031305.2017.1375989.

Babette A Brumback. *Fundamentals of Causal Inference with R*. CRC Press, 2021. doi:10.1201/9781003146674.

Mudit Chowdhary, Akansha Chowdhary, Trevor J Royce, Kirtesh R Patel, Arpit M Chhabra, Shikha Jain, Miriam A Knoll, Neha Vapiwala, Barbara Pro, and Gaurav Marwaha. Women's representation in leadership positions in academic medical oncology, radiation oncology, and surgical oncology programs. *JAMA Network Open*, 3(3):e200708, 2020. doi:10.1001/jamanetworkopen.2020.0708.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, 2003. doi:10.1038/sj.bjc.6601118.

Stephen R Cole, Jessie K Edwards, and Sander Greenland. Surprise! *American Journal of Epidemiology*, 190(2):191–193, 2021. doi:10.1093/aje/kwaa136.

Jessica S Cranston, Sophia Finn Tiene, Karin Nielsen-Saines, Zilton Vasconcelos, Marcos V Pone, Sheila Pone, Andrea Zin, Tania Saad Salles, Jose Paulo Pereira, Dulce Orofino, Patricia Basil, Tara Kerin, Kristina Adachi, Fernanda Mendes Soares, Andrea Dunshee de Abranches, Ana Carolina C da Costa, and Maria Elisabeth Lopes Moreira. Association between antenatal exposure to Zika virus and anatomical and neurodevelopmental abnormalities in children. *JAMA Network Open*, 3(7):e209303, 2020. doi:10.1001/jamanetworkopen.2020.9303.

Peter Dalgaard. *Introductory Statistics with R*. Springer, 2008. doi:10.1007/978-0-387-79054-1.

Michael P Fay and Erica H Brittain. *Statistical Hypothesis Testing in Context: Reproducibility, Inference, and Science*. Cambridge University Press, 2022. doi:10.1017/9781108528825.

David A Freedman. Randomization does not justify logistic regression. *Statistical Science*, 23(2): 237–249, 2008. doi:10.1214/08-STS262.

David A Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009. doi:10.1017/CBO9780511815867.

David A Freedman and Philip B Stark. What is the chance of an earthquake? Technical Report 611, University of California, Berkeley, September 2001. URL https://statistics.berkeley.edu/tech-reports/611.

Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of Clinical Trials*. Springer, 2015. doi:10.1007/978-3-319-18539-2.

Miguel A Hernán. The c-word: scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 108(5):616–619, 2018. doi:10.2105/AJPH.2018.304337.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006. doi:10.1136/jech.2004.029496.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016. doi:10.1093/aje/kwv254.

Miguel A Hernán, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, 19(6):766–779, 2008. doi:10.1097/EDE.0b013e3181875e61.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2$^{\mathrm{nd}}$ edition, 2021. doi:10.1007/978-1-0716-1418-1.

Nicholas P Jewell. *Statistics for Epidemiology*. CRC Press, 2004. doi:10.1201/9781482286014.

Michael I Jordan. Artificial intelligence—The revolution hasn't happened yet. *Harvard Data Science Review*, 1(1), 2019. URL https://hdsr.mitpress.mit.edu/pub/wot7mkc1.

Brennan C Kahan, Suzie Cro, Fan Li, and Michael O Harhay. Eliminating ambiguous treatment effects using estimands. *American Journal of Epidemiology*, 192(6):987–994, 2023. doi:10.1093/aje/kwad036.

Brennan C Kahan, Joanna Hindley, Mark Edwards, Suzie Cro, and Tim P Morris. The estimands framework: A primer on the ICH E9 (R1) addendum. *The BMJ*, 384:e076316, 2024. doi:10.1136/bmj-2023-076316.

Marjolein F Q Kluytmans-van Den Bergh, Anton G M Buiting, Suzan D Pas, Robbert G Bentvelsen, Wouter Van Den Bijllaardt, Anne J G Van Oudheusden, Miranda M L Van Rijen, Jaco J Verweij, Marion P G Koopmans, and Jan A J W Kluytmans. Prevalence and clinical presentation of health care workers with symptoms of Coronavirus Disease 2019 in two Dutch hospitals during an early phase of the pandemic. *JAMA Network Open*, 3(5):e209673, 2020. doi:10.1001/jamanetworkopen.2020.9673.

Michael P LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008. doi:10.1161/CIRCULATIONAHA.106.682658.

Eugene Milne and Paula Whitty. Calculation of the need for paediatric intensive care beds. *Archives of Disease in Childhood*, 73(6):505–507, 1995. doi:10.1136/adc.73.6.505.

Vy T Nguyen, Ross D Zafonte, Jarvis T Chen, Kalé Z Kponee-Shovein, Sabrina Paganoni, Alvaro Pascual-Leone, Frank E Speizer, Aaron L Baggish, Herman A Taylor, Lee M Nadler, et al. Mortality among professional American-style football players and professional American baseball players. *JAMA Network Open*, 2(5):e194223, 2019. doi:10.1001/jamanetworkopen.2019.4223.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

Roger D Peng and Stephanie C Hicks. Reproducible research: A retrospective. *Annual Review of Public Health*, 42:79–93, 2021. doi:10.1146/annurev-publhealth-012420-105110.

Stephen Senn. *Dicing with Death: Living by Data*. Cambridge University Press, 2022. doi:10.1017/9781009000185.

David Spiegelhalter. Why probability probably doesn't exist (but it is useful to act like it does). *Nature*, 636(8043):560–563, 2024. doi:10.1038/d41586-024-04096-5.

Philip B Stark and Andrea Saltelli. Cargo-cult statistics and scientific crisis. *Significance*, 15(4): 40–43, 2018. doi:10.1111/j.1740-9713.2018.01174.x.

Lindsey Turner, Julien Leider, Elizabeth Piekarz-Porter, and Jamie F Chriqui. Association of state laws regarding snacks in US schools with students' consumption of solid fats and added sugars. *JAMA Network Open*, 3(1):e1918436, 2020. doi:10.1001/jamanetworkopen.2019.18436.

US Food and Drug Administration. Statistical Principles for Clinical Trials E9 (R1): Addendum: Estimands and Sensitivity Analysis in Clinical Trials, 2021. URL `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clin`

Mark J van der Laan. Statistics as a science, not an art: The way to survive in data science. *Amstat News*, 2015. URL `https://magazine.amstat.org/blog/2015/02/01/statscience_feb2015/`.

Julie Vu and David Harrington. *Introductory Statistics for the Life and Biomedical Sciences*. OpenIntro, 2020. URL `https://openintro.org/book/biostat`.

Howard Wainer. How to display data badly. *The American Statistician*, 38(2):137–147, 1984.

Isabelle R Weir, Suzanne M Dufault, and Patrick P J Phillips. Estimands for clinical endpoints in tuberculosis treatment randomized controlled trials: A retrospective application in a completed trial. *Trials*, 25(1), 2024. doi:10.1186/s13063-024-07999-w.

Daniel Westreich. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press, 2019.