Received: 8 November 2020 | Accepted: 9 November 2020

DOI: 10.1002/sim.8830

COMMENTARY

Statistics in Medicine WILEY

Reflecting on “A Statistician in Medicine” in 2020

Walter Dempsey[1,2] | **Bhramar Mukherjee**[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

[2]Institute of Social Research, University of Michigan, Ann Arbor, Michigan

Correspondence
Bhramar Mukherjee, Department of Biostatistics, University of Michigan, Ann Arbor, MI.
Email: bhramar@umich.edu

In this commentary, we revisit Sir Austin Bradford Hill’s seminal Alfred Watson Memorial Lecture in 1962 through the eyes of two practicing biostatisticians of the current era. We summarize some eternal takeaway messages from Hill’s lecture regarding observations and experiments translated through the modern lexicon of causal inference. Finally, we pose a series of questions that we would have liked to pose to Sir Austin Bradford Hill if he were to deliver the lecture in 2020.

KEYWORDS

counterfactual, evidence synthesis, propensity score, *P*-value, randomization, reproducibility

1 | INTRODUCTION

> I returned, and saw under the sun, that the race is not to the swift, not the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favour to men of skill; but time and chance happeneth to them all.
>
> Ecclesiastes 9:11, King James Version

As we reflect on Sir Austin Bradford Hill’s marvelous Alfred Watson Memorial Lecture delivered in 1962 as two practicing biostatisticians 48 years later, we are reminded of the above biblical text that George Orwell admired for its clarity and style of writing. Orwell tried to rewrite this piece from good old English to what he calls the modern English of the worst sort. In his essay *Politics and the English Language*,[1] Orwell criticized the “ugly and inaccurate” written English of his time and “translated” the above original to an ambiguous version emblematic of that period:

> Objective consideration of contemporary phenomena compels the conclusion that success or failure in competitive activities exhibits no tendency to be commensurate with innate capacity, but that a considerable element of the unpredictable must invariably be taken into account.

Orwell points out that this “translation” contains many more syllables but gives no concrete illustrations, as the original did, nor does it contain any vivid, arresting images or phrases. What made the original so good? Of the 49 words in the biblical original, 41 are of one syllable, including sturdy Anglo-Saxon words such as sun, race, swift, strong, bread, wise, skill, and time. Clark 2 notes how the passage moves from the human attributes to things that we cannot control, “Time and Chance.” As we began reading the transcript of Hill’s lecture, we appreciated the beauty and simplicity of his writing that drives home fundamental and eternal statistical messages with powerful examples without getting lost in the “mathematistry” and complexity of modern Statistics.[3] Throughout his lecture, Hill focuses on learning and teaching epistemic statistical “values” instead of learning a particular methodology. Like the biblical passage, Hill’s lecture pivots around the dichotomy of experimental variables that we, as scientists, can control and recognizes unmeasured random noises as “time and chance happeneth to all.”

Standing in 2020, when we are facing a global pandemic ravaging our society[4] and underscoring how fragile our technologically advanced civilization is, we, as statisticians, appreciated reading and writing about this lecture even more. Hill describes medicine as a primary union of two fields: public health focusing on etiology and prevention, prior to contraction of a disease, and, healthcare focusing on treatment of a patient with a disease. There is constant knowledge flow between these two fields: the study of the populations and that of the individual patients. Even in the era of big data and precision medicine, this dichotomy and exchange remains true. In our battle against the COVID-19 pandemic, we are witnessing collaborations between these two fields of public health and healthcare. Nonpharmaceutical interventions have as much role to play as vaccines and treatments for this highly contagious viral infection.

Hill classifies three broad classes of experiments to answer questions that arise in this unified field of medicine

(a) Animal experiments conducted in laboratories.
(b) Designed experiments with humans as units.
(c) Natural experiments and observational studies in the real world.

Hill's emphasis on the theory of experiments is somewhat unexpected for a current graduate student in Biostatistics. Most Biostatistics departments are in the process of phasing out a course on design of experiments from their graduate program, which we believe is detrimental to our profession. A statistician in medicine has to engage from the nascent design phase of a study as true partners in science. In the next few sections of the lecture, Hill advocates for "the permeation of the statistical research with experimental spirit."[5]

2 | ON EXPERIMENTATION AND *DOING*

Spirtes[6] and Pearl[7] emphasize the important distinction between *S*eeing and *D*oing. Experimentation centers around drawing conclusions about how systems respond to external intervention when the intervention is controlled by the experimenter. Hill recognized experiments as *D*oing, altering the system of interest by introducing different stimuli/treatments. Hill intuited, from a methodology-free perspective, that for causal inference the *ideal trial* in many settings is first choosing a randomly selected set of individuals from the target population and then randomly allocating treatment. Hill's ideal trial notionally protects against treatment assignment being correlated with latent health status and biased selection of individuals from the population. Hill, always aware of practical issues when implementing theoretically correct criteria, acknowledged limitations in the ability to recruit a random sample but rightly emphasized treatment randomization.

While Hill's ideal trial is correct, such informal statements can sometimes mask the difficulty in formal description of causal effects even in well-designed experiments. Indeed, modern causal inference has focused on linking Fisher's statistical testing with Hill's intuition by formalizing these notions either through counterfactuals[8-10] or directed acyclic graphs.[7] Often, emphasis is placed on translation of the scientific question into a nonparametrically defined estimand. In many settings, the scientific question is equivalent to estimating the *average treatment effect* (ATE). Consider, for example, a variation on Hill's famous randomized controlled trial[11] where patients with tuberculosis are randomized to either receive streptomycin ($Z = 1$) or not ($Z = 0$) with a known probability $\pi(Z|X)$ that is a function of baseline covariate information X. Then, the ATE can be defined using the potential outcome framework as

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[(I[Z = 1] - I[Z = 0])\frac{Y(Z)}{\pi(Z|X)}\right],$$

where $Y(z)$ is the potential outcome under treatment ($z = 1$) and control ($z = 0$) respectively and the expectation is over the distribution of potential outcomes in the population. The left-hand size defines the target estimand, while the right-hand reexpresses this quantity using the randomized treatment Z. Modern causal inference[10] demonstrated that if the randomized trial satisfies (a) positivity, that is, $0 < \epsilon \leq \pi(z|x) \leq 1 - \epsilon < 1$ for all x and (b) consistency, that is, an individual's potential outcomes under the observed exposure is the outcome observed for that person, then an estimator inversely weighted by the likelihood of treatment assignment is an unbiased estimator of the ATE. In Hill's trial, patients with tuberculosis were all equally likely to receive and not to receive the treatment, that is, $\pi(Z|X) \equiv 1/2$. Moreover, administering treatment to one tuberculosis patient is unlikely to impact another patient. Under these assumptions, the right-hand side simplifies and the ATE of streptomycin compared with control is the difference in means of the treatment and control group.

Above, the *target estimand* was formally defined and, under certain assumptions, a weighted estimator is shown to be consistent for the causal effect of interest. Modern causal inference translated Hill's "trifle bit of danger" to "making assumptions necessary to go from association to causality." This formal language also helps researchers build potentially more efficient estimators than simple mean comparisons. In Hill's streptomycin study, for example, a researcher may have had a prior model for the outcome given treatment and covariate information, denoted $\mu_z(x)$. Then, Hill may incorporate this model by estimating the average treatment effect using a *model-assisted estimator*

$$\sum_{z \in \{0,1\}} (-1)^{z+1} \left(\mathbb{E}[\mu_z(X)] + \mathbb{E}\left[\frac{1[Z=z]}{\pi(Z|X)} (Y(z) - \mu_z(X)) \right] \right).$$

Beyond improving efficiency, formal methods allow us to recognize settings where data from a randomized control trial cannot be used directly to estimate the target estimand of interest without additional assumptions or auxiliary information. For example, suppose patients cannot be blinded to treatment assignment. Then, upon being assigned to the treatment arm, an individual may decide to not comply and forgo taking the treatment. If the scientist is interested in effectiveness, then the mean comparison of treatment and control groups is adequate, that is, the ATE is equivalent to an intention-to-treat analysis. If, however, the scientist's interest is in drug efficacy, that is, assessing the effect of treatment within a population, then one must account for noncompliance. Here, let $W(z)$ denote compliance under exposure status $z \in \{0, 1\}$. Under one-sided compliance, then the population can be split into compliers ($W(z) = z$ for $z \in \{0, 1\}$) and never-takers ($W(z) \equiv 0$). Let binary T denote complier status. Then, the ATE can be expressed as

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{t \in \{0,1\}} \mathbb{E}[Y(1) - Y(0) | T = t] P(T = t).$$

Again, assumptions are needed to proceed. If one is willing to live a trifle dangerously and make the assumption that the noncomplier *ATE* is zero (ie, the exclusion-restriction assumption), then the complier ATE can be calculated as the ratio of the observed ATE and the rate of compliance in the treatment arm.

Randomized trials with compliance and no missing data can yield internally valid causal estimates. Hill's final criteria of a random sample from the population alludes to the fact that often the investigator has a target population of interest. Randomized trial samples, however, are often not representative of target populations of interest.[12] The above expectations are implicitly with respect to the target population; however, the observed data may come from a different distribution. In such settings, one may rely on the assumption that the scientist has observed the factors that moderate treatment effects and differ between sample and population.[13] Armed with this information, one can account for selection bias. Again, causal inference relies on assumptions; living dangerously but explicitly stating how, which allows researchers to go from association to causality.

3 | ON OBSERVATIONAL STUDIES AND *SEEING*

The brevity of "Nature in the Raw" is quite astounding considering that by 1962, Hill was a leader in the two decade debate of the relationship between smoking and lung cancer. In March 1962 (only 6 months prior to the Watson lecture), the Royal College of Physicians had released their report that "clearly indicted cigarette smoking as a cause of lung cancer and bronchitis." For background, in 1948, Richard Doll and Hill had run a case-control study, studying patients who have been diagnosed with lung cancer (cases) and patients who had not been diagnosed with lung cancer (control). Interestingly, all but two cases reported having been smokers in their past.

But "nature is tricky." Doll and Hill were concerned that certain unobserved differences among the cases and controls was contributing to the stark observed difference. Holland and Rubin (1988) formalized this by demonstrating that the assumption of strong ignorability is crucial for causal inferences from such retrospective studies, that is, smoking status is independent of the potential outcomes of lung cancer status given a set of covariates. Modern causal inference has only expanded the list of tools including instrumental variables,[14,15] Mendelian randomization,[16,17] and negative controls.[18]

Hill's final paragraph suggests that his work with retrospective studies along with a 1951 prospective study provided sufficient evidence such that the "most reasonable" explanation was a causal one between smoking and lung cancer. Skeptics, however, remained, most prominently R.A. Fisher. Fisher argued the evidence was "only statistical," while simultaneously arguing that a potential genetic link between smoking and lung cancer could not be refuted. Hill's lecture implicitly acknowledges the foolishness and unscientific nature of such a claim. Indeed, to the genetic comment, Hill tips his hat to the lead architect of the attack against it, Jerome Cornfield. In 1959, Cornfield demonstrated that the genetic link would need to be biologically implausibly strong to account for the difference in risk between smokers and nonsmokers. Cornfield's argument has blossomed into an important aspect of causal inference, *sensitivity analysis* in which one "quantifies how one's inference concerning an outcome of interest varies as a function of the magnitude of nonidentifiable selection bias."[19] Interestingly, sharp upper bounds on causal relative risk using data from case-control studies are only being discovered this year.[20]

4 | CAUSAL CRITERIA AND STATISTICAL EVIDENCE

While Hill's 1962 lecture predates his famous causal criteria[21] by a few years, it is clear that, even in 1962, Hill would not view any checklist as either necessary or sufficient for assessing causality. Hill's 1965 criteria act as a roadmap, helping researchers build complex narratives to help answer causal questions. Hill emphasized the context-specific nature of causal questions and reaching the "most reasonable explanation of a particular set of facts."

Modern causal inference has focused heavily on (a) formal definitions of causal effects of interest, (b) criteria for nonparametric identification, and (c) efficient methods for causal estimation. While our toolkit continues to grow, nothing replaces a simple compass for navigation. Indeed, Hill's lecture reads as a warning to 21st century researchers against automated causal inference. Hill's narrative approach emphasizes causal triangulation. To live up to this, the rich literature of context-free causal methods needs to be married to context-specific reasoning.

Hill may agree with recent work[22] which emphasizes "inference to the best explanation" approaches to causal inference. While Hill's criteria may be "an early rough cut,"[23] it is clear that the statistician's goal when considering causal questions is not to present dichotomized statements of significance, but to provide useful and adequate information to decision makers. To the question "Should we provide streptomycin to all tuberculosis patients?" the statistician should not present a t-table based on a single study and say to the scientific team "the rest is up to you."

As mentioned before, Doll and Hill had tremendous influence and impact in the scientific and policy discourse around smoking and lung cancer.[24] Their persistent arguments of a causal association, along with those of Jerome Cornfield[25] lay the foundation of what is known as sensitivity analysis in modern causal inference. Hill's 1962 lecture mentions related notions of common causes, lurking variables and alternative explanations in the process of establishing a causal association. He cites association of smoking or occupational exposures with increased cancer incidence as an example of such a conceptual framework. The lecture is almost a preamble to his seminal paper in 1965 where he formally introduces Hill's criteria for association versus causation.[21] We found the narrative that Hill shares in his lecture where a purely statistical observation of association led to an ultimate causal conclusion to be quite compelling. This example is about incidence of cataract in infants whose mothers suffered from German measles during pregnancy. Since the initial observation by clinician Sir Norman Gregg, supporting evidence regarding the effects of the rubella virus upon the eyes, ears, and heart of the fetus during its first trimester has evolved without any dispute of alternative explanations. The case-study argues that persuasive and careful *seeing* or observation is important as it may generate plausible hypotheses and finally lead to *doing* or intervention.

5 | CLOSING THOUGHTS

Hill begins his lecture with a reference to all the strange things that Alice saw "Through the Looking Glass" and argues that as statisticians we need to be exploring the unknown more often, live a trifle more dangerously and learn more from related disciplines (in this specific case the field of actuarial sciences, where most attendees of the lecture belonged to). To conclude our discussion we return to the prequel of "Through the Looking Glass," namely, to Lewis Carroll's 1865 classic "Alice's Adventures in Wonderland."

> 'Would you tell me, please, which way I ought to go from here?'
>
> 'That depends a good deal on where you want to get to,' said the Cat.
>
> 'I don't much care where–' said Alice.

> ‘Then it doesn’t matter which way you go,’ said the Cat.
>
> ‘—so long as I get SOMEWHERE,’ Alice added as an explanation.
>
> ‘Oh, you’re sure to do that,’ said the Cat, ’if you only walk long enough.’
>
> Alice felt that this could not be denied, so she tried another question. ’What sort of people live about here?’
>
> ‘In THAT direction,’ the Cat said, waving its right paw round, ’lives a Hatter: and in THAT direction,’ waving the other paw, ’lives a March Hare. Visit either you like: they’re both mad.’
>
> ‘But I don’t want to go among mad people,’ Alice remarked.
>
> ‘Oh, you can’t help that,’ said the Cat: ’we’re all mad here. I’m mad. You’re mad.’
>
> ‘How do you know I’m mad?’ said Alice.
>
> ‘You must be,’ said the Cat, ‘or you wouldn’t have come here.’
>
> Alice’s Adventures in Wonderland, Lewis Caroll

Alice’s quandary exactly reflects our own, with Hill’s body of work replacing the Cheshire Cat. Interested in the current state of statistics in medicine, we turn to Hill and ask which way we, as a profession, ought to go from here? Indeed, reflecting upon Hill’s criteria for causality and his focus on evidence synthesis, we kept wondering how Hill would respond to the following 10 questions, had he given this lecture in 2020 and if we had a chance to be in the audience and raise our hands.

1. Where would Hill fall within the raging discussion and debate around redefining statistical significance?[26,27]
2. How would Hill alter our thinking about replicability and reproducibility in today’s science?[28]
3. Would Hill take sides in the causality debate? Would he lean toward directed acyclic graphs[7] or potential outcomes,[29] or alternatives?[30,31]
4. Would Hill appreciate the rise of Bayesian statistics in medical applications?
5. Would the cross-fertilization of ideas in statistics and computer science and the emergence of the hybrid field of data science excite Hill?
6. Would Hill caution or embrace machine learning techniques in causal inference and decision-making?
7. Would Hill agree that the union of multiple cultures of modeling, stochastic or algorithmic enriches our discipline?[32]
8. Would Hill applaud the general computational advances that enable inference and prediction using large data sets?
9. What would Hill say about the decline of experimental design in biostatistics and statistics training programs?
10. What does Hill see as the core set of questions one must always ask when starting a medical study in 2020, either observational or experimental?

Hill’s lecture leads us to some educated guesses to some of these questions. To question 1, Hill’s observations on decision-making based on evidence for your own self or loved ones vs decision-making regarding a general conceptual population brings us right back to recent papers where the same observation is noted.[33] Dichotomization and calibration of statistical evidence largely depends on the scientific and clinical context. Hill emphasizes the need to take the substantive context and supporting documentation into account. He recommends adopting a holistic approach toward reporting evidence instead of applying a magic threshold of 0.05. To question 4, his Watson memorial lecture suggests Hill is philosophically inclined toward reporting $P(H_0 \mid \text{data})$ as opposed to $P(\text{Data} \mid H_0)$, thus a Bayes factor will possibly be more appealing to him than reporting P-values. To question 8, Hill would probably appreciate the recent computational advances but encourage us to retain our focus on causally interpretable estimands and policy-relevant deliverables at the end of the day when all our complex and esoteric machinery have been put to task. However, he will most likely push us to fully understand the mathematical underpinnings of data-recursive procedures and to appreciate the associated uncertainty of predictions derived from modern algorithmic tools such as random forests, neural nets and support vector machines as “time and chance happeneth to all.” Finally, to question 9, Hill will likely encourage us to embrace the experimental spirit of statistics as a discipline and think hard about design, data collection and information gathering before embarking on an elaborate inferential journey. In modern medicine, as we are increasingly using data from electronic health records, medical claims, smart devices, and social media, to question 10, Sir Austin Bradford Hill will very likely ask two primary questions about the sampling frame: Who is in your study? What is the target population of inference? Foundational statistical principles of representativeness, generalizability, and transportability cannot be forgotten while advancing cutting-edge biomedical science with big data and artificial intelligence.

Hill’s Watson lecture reminds us that at the end of the day, we have to make a practical difference in the domain science in order to demand respect and stature as “A Statistician in Medicine.” This impact requires a zealous blend of practical knowledge, formal mathematical and computational training, true collaborative spirit, and communication skills that calls for a bit of a creative mad mind. Just like Alice, we must be a bit mad, or we would not have ended up in this discipline at this momentous time. Here is to the statistician extraordinaire Sir Austin Bradford Hill, to a touch of madness, and to the methods behind the madness!

ORCID

Walter Dempsey https://orcid.org/0000-0002-7852-2269
Bhramar Mukherjee https://orcid.org/0000-0003-0118-4561

REFERENCES

1. Orwell G. *Politics and the English Language*. London, England: Penguin Classics; 2013.
2. Clark RP. One great moment; 2018. [Online] Accessed March 01, 2018.
3. Little RJ. In praise of simplicity not mathematistry! ten simple powerful ideas for the statistical scientist. *J Am Stat Assoc*. 2013;108(502):359-369.
4. Adhanom T. WHO director-general’s opening remarks at the media briefing on COVID-19 - 11 March 2020; 2018. [Online]. Accessed March 11, 2020.
5. Hill AB. Observation and experiment. *N Engl J Med*. 1953;248(24):995-1001. https://doi.org/10.1056/NEJM195306112482401.
6. Spirtes P, Glymour C, Scheines R. *Causation Prediction and Search*. 2nd ed. New York, NY: Springer-Verlag; 2000.
7. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge, MA: Cambridge University Press; 2009.
8. Rubin D. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6(1):34-58.
9. Imbens G, Rubin D. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press; 2015.
10. Hernán M, Robins J. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC; 2020.
11. Hill AB. Suspended judgment. memories of the British streptomycin trial in tuberculosis. the first randomized clinical. *Trial*. 1990;11(2):77-79.
12. Rothwell P. External validity of randomised controlled trials: “To whom do the results of this trial apply?”. *Lancet*. 2005;365(9453):82-93.
13. Dahabreh I, Robertson S, Steingrimsson J, Stuart E, Hernán M. Extending inferences from randomized trial to a new target population. *Stat Med*. 2020;39(14):1999-2014.
14. Martens E, Pestman W, Boer dA, Belitser S, Klungel O. Instrumental variables: application and limitations. *Epidemiology*. 2006;17(3):260-267.
15. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722-729. https://doi.org/10.1093/ije/29.4.722.
16. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018;362:k601. https://doi.org/10.1136/bmj.k601.
17. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27(8):1133-1163. https://doi.org/10.1002/sim.3034.
18. Shi X, Miao W, Tchetgen ET. A selective review of negative control methods in epidemiology. *Curr Epidemiol Rep*. 2020:1-13.
19. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trial*. New York, NY: Springer; 2000:1-94.
20. Jun SJ, Lee SS. Causal inference in case-control studies. papers; 2020. arXiv.org.
21. Hill SAB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58(5):295-300.
22. Krieger N, Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol*. 2016;45(6):1787-1808.
23. Phillips C, Goodman K. The missed lessons of Sir Bradford Hill. *Epidemiol Perspect Innovat*. 2004;1(3):1–5.
24. Gail MH. Statistics in action. *J Am Stat Assoc*. 1996;91(433):1–13.
25. Greenhouse SW, Greenhouse JB. Cornfield Jerome. *Encyclopedia of Biostatistics*. 2005:1–5.
26. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2:6-10.

27. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2):129-133.
28. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press; 2019.
29. Imbens G. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *National Bureau of Economic Research (Working Paper)*. 2019.
30. Dawid AP. Causal inference without counterfactuals. *J Am Stat Assoc*. 2000;95(450):407-424. https://doi.org/10.1080/01621459.2000.10474210.
31. Gelman A. Long discussion about causal inference and the use of hierarchical models to bridge between different inferential settings; 2012. [Online] Accessed July 16, 2012.
32. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3):199-231. https://doi.org/10.1214/ss/1009213726.
33. McShane B, Gal D. Statistical significance and the dichotomization of evidence. *J Am Stat Assoc*. 2017;112(519):885-908.

How to cite this article: Dempsey W, Mukherjee B. Reflecting on "A Statistician in Medicine" in 2020. *Statistics in Medicine*. 2021;40:42–48. https://doi.org/10.1002/sim.8830