The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2022, Vol. 115(3) 112–115
DOI: 10.1177/01410768221082390

Carl Liebermeister and the emergence of modern medical statistics Part 2: the origin, use and limitations of his method

Leonhard Held[1] and Robert AJ Matthews[2]
[1]Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, 8001 Zurich, Switzerland
[2]Department of Mathematics, Aston University, Birmingham B4 7ET, UK
Corresponding author: Leonhard Held. Email: leonhard.held@uzh.ch

Introduction

In the first of these articles[1], we described how in 1877 the German physician Carl Liebermeister published an approach to applying probability theory to clinical trials which promised to transform the use of statistics in medicine.[2] At its heart was a formula which gave researchers what had seemed impossible: a way of calculating the probability that one treatment was better than another using data from a study of any size. At the time, advocates of statistical methods in medicine could only offer formulas showing if the probability of the reality of an effect exceeded some arbitrary threshold, and even then only for studies involving many hundreds of patients. Liebermeister applied his formula to real-life data sets, showing that even small trials – often dismissed as worthless by physicians – could be a source of valuable insight. Unsurprisingly, Liebermeister's approach provoked considerable controversy among his contemporaries, which we shall describe in the concluding third part of this account.

The purpose of the current article is to provide a non-technical explanation of the origins, applications and limitations of Liebermeister's remarkable achievement. This takes the form of a 28-page paper entitled *Über Wahrscheinlichkeits rechnung in Anwendung auf Therapeutische Statistik* ('On Probability Theory Applied to Therapeutic Statistics'; henceforth *Über Wkt.*); the original text in German and an English translation can be found in the online supplementary material as Appendix 1 and 2, respectively. The technical and historical basis of Liebermeister's approach are the subject of Appendix 3. Despite their considerable complexity, a brief and non-technical description is vital to a proper appreciation of both Liebermeister's achievement and its implications.

The essential idea behind the formula

First, in keeping with the practice of the time, Liebermeister worked within the so-called Bayesian inference paradigm, in which data are transformed into insight via Bayes's Theorem. Published posthumously by the eponymous English clergyman-mathematician in 1764, the theorem allows an initial level of belief in a hypothesis – its so-called *prior probability* – to be updated in the light of newly-acquired data, resulting in a new level of belief, or *posterior probability* (for a non-technical explanation see Matthews,[3] pp. 135–146). Liebermeister was thus seeking a means of allowing even small amounts of data to update an initial level of belief about a treatment. The means of doing this is today known as a *likelihood ratio,* and it gives the relative chances of observing the results obtained on the basis of each hypothesis under test. In Liebermeister's case, there were just two hypotheses – that the treatment was genuinely effective or not – with the evidence taking the form of the relative numbers of treated and untreated patients who recovered in their respective groups. The greater the difference in these relative numbers, the less likely mere chance could have been responsible. But how much less likely? Liebermeister needed a means of capturing the effect of chance in such a comparison, and turned to the time-honoured mathematical model of black and white balls in urns. This allowed him to arrive at a formula for the probability that mere chance could have led to more white balls being plucked at random from the 'treatment' urn compared with the 'control' urn.

The derivation is a demanding exercise in advanced mathematics based on Bayes's Theorem (see supplementary Appendix 3 and references therein), but leads to the desired outcome: a formula giving the probability that the treatment under test is

effective, given the data from a clinical study. It should be stressed that Liebermeister's use of Bayes's Theorem ensures that this (posterior) probability is free of the counterintuitive interpretation of *p*-values that remain widely used in the assessment of study outcomes. Contrary to common perception, *p*-values bear no simple relationship to the probability of chance accounting for the outcome, and are notoriously misleading (see, e.g. Wasserstein and Lazar[4]). In contrast, the Bayesian posterior probability is exactly what it appears to be: the probability that the treatment is effective, given the observed outcomes. Moreover, Liebermeister's derivation has the remarkable feature of leading to a formula that is applicable to all sample sizes. Up until this point, attempts to apply probability theory to the outcome of clinical studies had relied on a mathematical approximation that requires substantial sample sizes (see supplementary material, Appendix 3). As such, Liebermeister had preceded by over 50 years the work of the celebrated statistician Ronald Aylmer Fisher (1890–1962), whose well-known Exact Test is widely used to assess differences between small samples – albeit via the problematic concept of *p*-values (Fisher,[5] Section 21.02).

Liebermeister had no reason to mention the interpretational benefits of his approach over *p*-values, as the latter are part of an inferential paradigm that came to prominence after his death. Known as Null Hypothesis Significance Testing, it supplanted the Bayesian approach for reasons beyond the scope of this article (see, e.g. McGrayne,[6] chapter 3). Instead, Liebermeister stressed what was, at the time, the critical advantage of his probability formulas: their probabilistic reliability regardless of sample size. He proceeded to demonstrate this with a set of worked examples, including some based on real clinical studies. These allowed him to highlight the ability of the formulas to extract valuable insight even from small studies. They include the case of quinine and malarial patients, where Liebermeister reports a study (probably his own) involving two groups of 12 patients, 10 of which had become free of fever three days after treatment with quinine, compared with just 2 of those left untreated. Applying his formula, Liebermeister calculated the odds against so large a difference being a fluke as 1666 to 1 against (99.94%) – thus confirming his claim that small studies can nevertheless produce compelling evidence if the effect size is sufficiently large.

In another telling example, Liebermeister examines the assertion that small differences in relative proportions provide no evidential weight of an effect, his aim being '*... to show in what striking way the meaning of the formulas used by representatives of medical statistics has been misunderstood by them ...*'.

This appears to be a direct criticism of the work of the German ophthalmologist Julius Hirschberg (1843–1925) who in 1874 had examined the case of two groups of 300 patients with the same disease, where the mortality in one group is 22%, compared with 16% in the other.[7] Hirschberg argued that given that the difference is just 6%, the true mortality rate is very likely to be the same. Liebermeister disputes this, insisting that 'the general practitioner' would undoubtedly regard the difference as real. He then shows that the theory used by advocates of the probabilistic analysis of clinical studies – which he states is 'not very exact for such small numbers' – implies the chances of the difference not being a fluke are 93.97%, or odds of over 15 to 1. While this fails to pass the (arbitrary) 99.53% probability threshold promoted by advocates of probabilistic analysis, Liebermeister argues that odds of 15 to 1 are

> *... certainly not meaningless. It will depend to a large extent on other circumstances and considerations [such as] whether one wishes to consider them sufficient to take an important decision in relation to future treatment or anything similar.*

Liebermeister then uses his own formula, and finds a probability of 96.91%, adding that this gives 'an even somewhat larger significance' than the standard theory. For technical reasons, this is not quite correct; for an explanation, see example 6 in the Applications section of supplementary Appendix 3. Nevertheless, Liebermeister is making several important points here. First, he is highlighting the dangers inherent in using thresholds to decide whether to accept a specific finding as genuine. In the real world, such decisions are rarely clear cut, but depend on context. As such, they require a more nuanced approach than simply 'pass/fail', and this is what Liebermeister's formula provides: the probability that effectiveness has been demonstrated by the study in question, thus allowing it to be assessed on its own merits. This, in turn, has clear practical value: the ability to base actions on probabilities. Liebermeister illustrates this by assuming that the higher mortality rate was observed in an ordinary hospital, while the lower rate came from identical conditions in a barrack hospital. He then asks whether the evidence is large enough to justify build another barrack hospital, arguing that this ultimately depends on the associated cost:

> *Where the construction of the barracks would be easy to carry out, one would probably proceed without*

> *question to that result. Where, on the other hand, there would be particular difficulties and inconveniences connected with it, and there would be no urgent need for a careful decision, it would be preferable to wait and see whether further observations would increase or decrease the probability.*

This is an example of an informal decision analysis combining the cost of different decisions with the available probabilistic evidence. It is remarkable that Liebermeister considers the possibility of collecting additional data to obtain more reliable evidence on the existence of a difference between the two groups. This highlights another key feature of Liebermeister's approach to inference: recognition of the limits of inference based on probability theory. Even if the observed mortality rates had ruled out chance to better than the 99.53% standard adopted by advocates of probabilistic analysis, this means merely that chance is highly unlikely to account for the difference. The true cause remains to be identified – a key point often overlooked to this day.

Limitations of Liebermeister's method

Liebermeister was well aware that the validity of his method depends on additional assumptions. He states that it is important to make sure that the groups to be compared do not differ with respect to important characteristics at the beginning of his article:

> *Certainly, with the accomplishment of this mathematical and formal part, our task is far from being completed. Rather, the question then arises as to whether the two series of observations, in which the difference in success occurred with different treatment, can really be regarded as comparable in every other respect. There might have also been a decisive change in the character of the disease, in the intensity of the cause of the disease, whether a change in the various other moments, on which the outcome of the disease may depend, has not caused the differences in the observed success.*

Concerns about non–comparability of groups have been common in the 19th century according to Morabia.[8] Vandenbroucke[9] describes the work by Mill[10] and by Claude Bernard[11] as the first accounts discussing the problem of comparability. It is noteworthy how precisely Liebermeister describes the possibility that potential confounders may have caused an observed difference between groups.

In example 5 he compares mortality rates among patients with acute pneumonia in a hospital in Basle, Switzerland. He compares patients treated with antipyretic methods to historical controls without that treatment. He remarks that

> *through precise clinical analysis it was established that the two series of observations were comparable in every other regard.*

Similarly, in example 6 he compares the mortality rates 66/300 and 48/300, an example taken from Hirschberg.[7] Liebermeister argues that there is moderate, but not overwhelming evidence for a true difference in mortality between the two groups. He then asks what factors may have contributed to the reduction in mortality:

> *The question, of course, as to what is the cause of the reduction in mortality, whether a possible difference in treatment or a change in the nature of the epidemic or any other change in circumstances, is not a matter for mathematical analysis, but for clinical analysis.*

These examples show that Liebermeister was well aware of the problem of confounding and that there remains a role for clinical analysis in identifying the true cause of any difference in efficacy. The solution is now known to lie not in 'clinical analysis' but in experimental design, specifically the use of randomised allocation. The power of this methodology to counteract bias was only starting to be recognised by the time of Liebermeister's death in 1901.[12]

Liebermeister concludes the substantive part of *Über Wkt.* by pointing out that the formulas he has derived 'are not only applicable to therapeutic statistics, but also to a large number of other problems in probability calculus'. History records that while both statements are true, Liebermeister's remarkable achievement was destined to be forgotten even within clinical medicine until long after his death. Potential explanations for this will be explored in the next section.

Über Wkt. also includes two appendices covering technical points and giving a more detailed derivation of the formulas. The first appendix deals with a key issue confronting anyone using Bayes's Theorem to turn data into insight. In essence, the theorem shows how a prior level of belief expressed as a probability should be updated in the light of data, producing a posterior probability. But how should that prior level of belief be set? This question has dogged Bayesian inference since its emergence over 250 years ago. Liebermeister's solution was to use a convention widely applied at the time (and since), and which assumed a complete lack of prior insight about the possibility that a finding could be due to chance.

Known technically as a 'non-informative' uniform prior probability distribution, this assumption greatly simplifies the derivation (see supplementary Appendix 3). However, Liebermeister was well aware that other choices could be made:

> *When dealing with tasks concerning the so-called posterior probability, it is not uncommon to be under the illusion that one is approaching the observations without any preconditions. In reality, this is never the case and naturally cannot be the case.*

For example, there may be pre-existing studies that suggest a given treatment is effective; Bayesian inference allows these to be taken into account using prior probabilities. Liebermeister was also aware that prior beliefs can greatly influence the outcome of a Bayesian calculation. This is especially true with the levels of evidence typical of small studies – the focus of Liebermeister's attention – as their relatively weak evidential weight may barely alter prior beliefs. Liebermeister mentions that the balls-in-urns model can accommodate other priors, but gives no mathematical details.

As we shall see in the next paper in this series, this led to criticism of his entire approach, on the grounds that the assumption of complete prior ignorance would often lead to conclusions inconsistent with those based on 'common sense' beliefs of physicians.[13] Yet it must also be admitted that the use of 'common sense' has a chequered record in the history of medicine. It is unclear whether Liebermeister developed the necessary mathematical detail, or rebutted the criticism in general terms. What is known is that the addition of 'informative' priors to Liebermeister's model is far from trivial. The full theory was only developed long after his death by Altham,[14] unaware of the existence of his pioneering work (Altham PME, pers. comm. 2020, to LH).

Declarations

Competing Interests: None declared.

Funding: None declared.

Ethics approval: Not applicable.

Guarantor: RAJM.

Contributorship: LH conceived the project, typeset and translated the original text of *Über Wkt*. and wrote early drafts. RAJM contributed historical context and later drafts.

Provenance: Invited article from the James Lind Library.

Acknowledgements: LH is grateful to Flurin Condrau for useful discussions, to Valentina Held for support in translating Liebermeister (1877) into English, Patricia Altham for comments on the Exact Test, Stephen Senn for helpful comments on an early version of this manuscript and Klaus Dietz for bringing Liebermeister to his attention. RAJM thanks Iain Chalmers for his enthusiasm for this project, Ulrich Tröhler, and Wolfram Liebermeister and Klaus Dietz for assistance with relevant literature. Both authors are grateful to Peter Diggle and Håvard Rue for comments on drafts.

Supplemental material: Supplemental material for this article is available at: https://osf.io/4dvpk/

References

1. Held L and Matthews RAJ. Carl Liebermeister and the emergence of modern medical statistics Part 1: his remarkable work in historical context. *J R Soc Med* 2022; 115: 69–72.
2. Liebermeister C. Über Wahrscheinlichkeitsrechnung in Anwendung auf Therapeutische Statistik. *Sammlung klinischer Vorträge (Innere Medicin No. 31-64)* 1877; 110: 935–962.
3. Matthews RAJ. *Chancing It*. London: Profile Books, 2017, pp.135–146.
4. Wasserstein RL and Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat* 2016; 70: 129–133.
5. Fisher RA. *Statistical Methods for Research Workers*. 5th ed. Edinburgh: Oliver & Boyd, 1934.
6. McGrayne SB. *The Theory That Would Not Die*. Yale: University Press, 2011.
7. Hirschberg J. *Die Mathematischen Grundlagen der Medizinischen Statistik*. Leipzig: Verlag von Veit & Comp, 1874.
8. Morabia A. History of the modern epidemiological concept of confounding. *J Epidemiol Commun Health* 2011; 65: 297–300.
9. Vandenbroucke JP. The history of confounding. *Sozial- Und Praventivmedizin* 2002; 47: 216–224.
10. Mill JS. *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker, 1843.
11. Bernard CI. *Introduction à l Étude de la Médecine Expérimentale*. Paris: JB Bailliere et fils, 1865.
12. Hróbjartsson A, Gøtzsche PC and Gluud C. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. *BMJ* 1998; 317: 1243–1245.
13. Korteweg JA. Over de Toepassing der Statistiek op Medische Wetenschappen. *Weekblad van het Nederlandsch Tijdschrift voor Geneeskunde* 1877; 21: 553–560.
14. Altham PME. Exact Bayesian analysis of a 2×2 contingency table and Fisher's 'Exact' significance test. *J R Stat Soc B* 1969; 31: 261–269.