The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2022, Vol. 115(4) 145–149
DOI: 10.1177/01410768221082392

Carl Liebermeister and the emergence of modern medical statistics. Part 3: 19th-century criticism and a paradigm lost

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

The critical response to Liebermeister's work

The previous two papers[1,2] have described how the 19th-century German physician Carl Liebermeister made an extraordinary contribution to the debate over the use of statistical methods in medicine. In 1877 he published a method for extracting useful insight even from small studies normally dismissed as worthless by physicians. The method was centred around a formula giving the probability that one treatment was better than another using data from a study of any size. Liebermeister's demonstration that such a formula was even possible put him half a century ahead of the professional statistics community. Moreover, his approach has important advantages even over today's textbook methods.

Yet while his 1877 paper[3] *Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik* (''On Probability Theory Applied to Therapeutic Statistics''; henceforth *Über Wkt*.) was exceptional, the response to it was not. Very soon after its publication, critical reviews appeared revealing yet again the gulf separating traditional physicians from those comfortable with probabilistic reasoning. The nature of these criticisms and the fate of Liebermeister's approach is the focus of this final paper.

Drawing on his unusual combination of mathematical ability and medical experience, the German ophthalmologist Julius Hirschberg (1843–1925) focused on Liebermeister's most impressive claim: that his formula could extract inferential value from even small trials[4] (see supplementary material for both the original German text of this critique and an English translation). Hirschberg's concern was that physicians would be bamboozled by Liebermeister's arcane formulas into over-interpreting their output. As evidence, he cited worked examples in *Über Wkt*. in which the formulas gave huge odds against the outcome of a clinical study (of an antipyretic) being merely a chance effect, in one case exceeding trillions to one. Hirschberg pointed out that this is vastly greater than estimates of the probability the sun will rise tomorrow, given it has done so for at least the last 5000 years. 'Are we really to believe', he asks, 'that the superiority of [the antipyretic therapy] has a tremendously greater degree of probability than the return of daylight?' After pointing to real-world constraints that can undermine such apparent certainty, such as accurate diagnosis and clear end-points, Hirschberg makes clear his belief that probability theory can best serve clinicians as a guide to plausible levels of efficacy, not as dichotomous proof. He argued that a confidence interval ('Territorium der Chance') of the kind advocated by the French clinician Gavarret[5] could help clinicians make reasonable decisions about efficacy. While these standards might be somewhat arbitrary, they suggest Hirschberg envisaged a classification system for strength of evidence, based on which (if any) of these standards are met by a specific study outcome.

He then turns to Liebermeister's remarkable claim to have found probability formulas whose reliability does not depend on trial size. Hirschberg expresses no doubts about their mathematical validity, and emphasises that he does not dispute the value of small studies, or even individual cases. Rather, his concern was that when used in such cases, the formulas are sensitive to practical issues such as the choice of clinical end-point and the impact of stopping a study early, in what would be now called interim analyses. In summary, Hirschberg's critique reflects a concern that the application of apparently sophisticated probability theory can be undermined by inadequate data: '[I]t would not be desirable to clothe the cautious empirical groping by calculation with a shining semblance of

higher certainty'. Such concerns still resonate almost 150 years later.

Similar sentiments appear in a more vituperative critique of *Über Wkt*. published in the same year by Johannes Korteweg (1851–1930), a young Dutch surgeon based in Leiden.[6] In a footnote, Korteweg acknowledges help from his older brother Diederik, a gifted mathematician later celebrated for his work in fluid dynamics. Like Hirschberg, however, the focus of Korteweg's concerns is not the formal derivation of Liebermeister's formulas, but their reliability as a source of clinical insight.

Like Hirschberg, Korteweg points out that reliable conclusions about genuine efficacy based on probability theory require that '[T]he series of numbers to be compared are also comparable in all respects – that is, that the disease character of the cases in both series was the same, the cause of the disease was equally powerful, etc'. While Liebermeister acknowledges this constraint in *Über Wkt*., Korteweg insists it is far from trivial, describing it as one of '... the sources of errors which even someone like Liebermeister cannot avoid when applying the method of the exact sciences to a subject which is not yet entitled to the name ''exact''.'

This leads Korteweg to take issue with Liebermeister's starting premise: that the outcome of clinical studies – with all their potential biases – can be modelled via the random selection of coloured balls from urns. It is striking – especially given the involvement of his mathematically adept brother – that Korteweg does not identify the use of *randomisation* as the key lacuna in Liebermeister's argument, and one that could, moreover, be applied in clinical studies. Instead, he focuses on another, nonetheless important, difference between the probabilistic model and clinical reality: the former assumes nothing is known about the relative proportions of the different colours in each urn. Korteweg points out that many clinical studies involve diseases about which much is known concerning mortality rates and treatment efficacy, at least approximately. Liebermeister's formulas have no place for such insights, including those based on what Korteweg calls 'common sense' about treatments which 'cannot be brought under a mathematical calculation'. Korteweg is here alluding to the problem of the incorporation of prior insight – especially of a subjective nature – in the Bayesian approach to assessing new evidence. As we have seen, however, he was wrong to imply Liebermeister was unaware of it: he explicitly deals with it, albeit discursively, in Supplemental Note 1 to *Über Wkt*. Korteweg is nevertheless correct to warn that such prior evidence can outweigh that produced by single clinical studies, especially those involving small numbers – precisely the kind Liebermeister had brought within reach of probabilistic analysis. Used in isolation, the apparently impressive odds of effectiveness produced by the formulas could lead to what Korteweg calls the 'smothering' of insight from other sources.

He then turns to what he believes is Liebermeister's selective use of published studies of new therapies to demonstrate the value of his formulas, warning of the dangers of what is now termed publication bias:

> *'[C]ommon sense teaches that 100 new drugs are tried without results when 10 have been announced with results. The truth calculus shows that of 100 new but indifferent drugs, 50 happen to be worse, of the latter 50, some, e.g. 10, will happen to give a result that seems worth mentioning. These 10 will be made known and will be subject to further investigation. Liebermeister selects from these 10 the one that seems to be the best by chance and bets 100 to 1 that the good result of this drug is not due to chance alone ...*

Korteweg ends his critique with a stinging – and prophetic – final paragraph:

> It is to be hoped that Liebermeister's treatise will be one of the most useful editions of the Sammlung Klinischer Vorträge and that it will achieve exactly the opposite of what it is intended to do. Its purpose was to push the often erroneous subjective judgment of the physician into the background and to replace it with numbers that would guide, if not overwhelm, the so easily mistaken mind. After the failed attempt made by Liebermeister, everyone will consider it more desirable for the time being to place common sense *above* scholarship, *above* facts, but above all *above* faith in numbers. [emphasis in original]

The last substantive contemporary critique of *Über Wkt*. appears to have come from a young German military physician, Friedrich Martius (1850–1923), and forms part of his review of the role of probability theory in clinical medicine.[7] This had already been the focus of an earlier work[8] in which he argued that probabilistic methods are a 'makeshift necessity' lacking the robustness needed to cut through the complexities of clinical observation and achieve the goal of making medicine 'scientific'. Instead, Martius claimed that simple, empirical induction based on experiment and observation is the surer path to reliable insight. In his follow-up critique Martius sought to clarify 'the still very much mistaken logical foundations of statistics and probability calculus' applied in medicine, including the work of Liebermeister.

The result is a curate's egg of a review combining both familiar and novel criticisms with mathematical errors. Martius starts by conceding that statistics – i.e. raw data such as patient numbers – and probability *per se* are essential tools in the drive to make medicine scientific. The dispute, he states, is about 'the scope of this source of knowledge' (p. 337). He takes issue with Liebermeister's claim that the chief barrier to the use of probabilistic methods is just the practical difficulty of applying the formulas – including the need for large numbers of patients. Instead, he reiterates his belief in the superiority of experimental induction, and then repeats the now-familiar doubts about modelling patient outcomes via the extraction of balls from urns, claiming that if the arguments for this approach were valid, physicians could only reject its adoption on the grounds of 'their mathematical ineptitude' (which Martius himself inadvertently demonstrates by bungling simple probability calculations (see pp. 352, 362)).

Martius then turns to what he sees as compelling arguments against those claiming probabilistic methods can make medicine 'scientific'. First, he picks up on the different probability thresholds advocated by Poisson/Gavarret and Hirschberg for ruling out chance effects.[1] While Fick[9] had already noted that the Poisson/Gavarett threshold of 99.53% was based principally on practical constraints, Martius went further, arguing that Hirschberg's suggestion of a lower probability threshold revealed the 'arbitrariness and uncertainty of the whole procedure'. Nor was he impressed by having a means of extracting insight even from very small clinical trials. He declares (p. 376): 'In fact, I believe that one cannot ignore the conviction that the Liebermeister formulas, though more practically applicable, are more unscientific than those of Poisson'.

Part of the reason seems to be Liebermeister's elimination of the need for *any* threshold for deciding when an effective therapy has been identified, such as the 99.53% adopted by Poisson/Gavarret. Liebermeister believed this would encourage the accumulation of evidence from any well-conducted trial ('silver coins'), including those that might otherwise be discarded as 'failures'. Martius, in contrast, claimed this would encourage the practice of 'subjective preference' over whether a therapy had worked or not.

Martius's objections also seem to reflect a belief that Liebermeister's formulas were based on (unspecified) 'other assumptions' to those used by Poisson, and 'completely neglect the Law of Large Numbers'. This law, first formalised by the Swiss mathematician Jacob Bernoulli (1655–1705), states that the probability of an event can be estimated ever more precisely from its observed frequency as the number of observations increases. This somewhat vague statement was later made mathematically precise and underpins the 'confidence interval' formulas of Poisson and Gavarret. Martius seems to imply, however, that Liebermeister's formulas only work with very small numbers because they have ignored this Law, thus violating a basic tenet of inference incorporated into Poisson's formulas. This is understandable, given Poisson's own peremptory (but incorrect) statement that the Law is 'the basis of all applications of the calculus of probabilities' (Matthews,[10] p. 26). Given Martius's lack of mathematical training, it is also unsurprising that he then fails to appreciate that the Law (of which he gives a somewhat muddled explanation on p. 365) has not been ignored, but circumvented by Liebermeister's inspired use of a mathematical model that obviates the use of approximations reliable only for large sample sizes (see supplementary Appendix 3).

Martius ends by stating that his 'rather negative' views should not be taken as implying rejection of the use of statistics in medicine. Rather, his aim was to counter 'the repeated attempts to exceed the competence of the method on which statistics and probability calculations are based'. Progress in medicine, he insists, 'lies in experimental induction, not in the numerical method'.

Martius's 15,000 word critique appears to mark the end of substantive public discussion of *Über Wkt*. during its author's lifetime. It is perhaps significant that the principal critics were all relatively young and raised similar concerns. This suggests they represented the nature and depth of the scepticism of probabilistic methods which would prevail for decades to come.

The relevance of Über Wkt. for clinical biostatistics in the 21st century

The descent of *Über Wkt*. into obscurity was rapid. Soon after Martius's critique, what may be the first discussion in English of Liebermeister's methodology appeared in the form of a puzzle sent to a British educational journal. It had been posed by the distinguished Scottish mathematician and physician Donald MacAlister (1854–1934), and concerned a small clinical trial of treatments for blood poisoning.[11] After stating the relative efficacy of the two treatments under test, MacAlister asked readers for the probability of the apparently higher efficacy of one approach being merely a fluke. MacAlister credited Liebermeister with providing the means of answering such questions, noting that they showed that some clinical insight could be obtained even

from small samples. Yet after some correspondence about their underlying assumptions, Liebermeister's formulas failed to make any further impact. By the 1940s, the American biostatistician Charles Winsor (1895–1951) praised Liebermeister's 'clear understanding of what statistical methods could and could not do for the practitioner' but declared 'it is clear that few of us today would use the Liebermeister solution'.[12]

This suggests that *Über Wkt*. had already been rendered obsolete by more sophisticated inferential methods. This was not the case: Winsor's remarks merely reflect the fact that by the 1940s other inferential paradigms had pushed aside Bayesianism, along with methods based upon it such as Liebermeister's formulas. Fisher's Exact Test, their equivalent in the so-called frequentist paradigm, had by then been published, apparently in complete ignorance of Liebermeister's work over half a century earlier.[13] Even after his method was independently rediscovered and began to appear in textbooks during the Bayesian renaissance of the 1980s, Liebermeister himself remained forgotten.

Those few who did encounter *Über Wkt*. admired its sophistication and practicality. Zabell[14] described its formulas as 'impressive . . . rigorously derived [with] several examples involving actual clinical data'. Similar praise appeared in Ineichen[15] (in German), Seneta,[16] and Seneta and Phipps.[17] Flatly contradicting Winsor, Seneta et al.[18] argued that 'There is little reason today for using less powerful exact tests in preference to Liebermeister's now that it has been brought out of oblivion'.

Ironically, Winsor's comments appeared just as the most serious objection to Liebermeister's methods was losing its power. Critics had repeatedly warned of the naïveté of interpreting the outcome of clinical studies in terms of the random extraction of balls from urns. From the mid-20th century onwards, clinical trials increasingly featured random allocation of patients to the treatment or control arms. Introduced to counter-biased allocation by triallists,[18] this also effectively turned each patient into a ball in Liebermeister's urn model. The resulting randomised controlled trial cuts through the inferential difficulties that Liebermeister himself recognised as hampering the adoption of his methods.

In truth, there were many other barriers. Some had deep cultural roots, harking back to ancient disputes over whether medicine could be regarded as a 'science' and the scope for quantification in decision-making.[20] Many 19th-century clinicians would have shared Martius's distaste for a mathematical model which represents patients as coloured balls. Those willing to set aside such qualms would have found Liebermeister's formulas unfamiliar and possibly intimidating. Even if they succeeded in extracting the probability of a finding not being a mere fluke, they would still have faced the ineluctable problem of making sense of the answer. Unless all other causes had been eliminated, how reliable was this apparent probability of efficacy – and how high should it be? In the absence of randomisation, the former is very hard to assess; even in its presence, the latter remains a deeply controversial question.[21]

Nevertheless, Liebermeister's concern about the abuse of arbitrary thresholds in deciding 'what works', his insistence that all well-conducted trials have inferential value and his use of Bayesian techniques for its extraction would today put him at the cutting edge of the debate over the use of statistics in medicine. Had the perspicacity of *Über Wkt*. been recognised when it was published, it might have led to more questions being asked about the inferential methods now causing so much concern within the medical research community.[22] As the drive to find more reliable methods continues, we believe Liebermeister should be recognised as one of the founding fathers of 21st-century evidence-based medicine.

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.

Supplemental material: The original German text of *Über Wkt*., the first English translation and a detailed technical commentary are available online at https://osf.io/4dvpk/

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.

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 Roy Soc Med* 2022; 115: 69–72.
2. Held L and Matthews RAJ. Carl Liebermeister and the emergence of modern medical statistics Part 2: The origin, use and limitations of his method. *J Roy Soc Med* 2022; 115: 112–115.
3. Liebermeister C. Über Wahrscheinlichkeitsrechnung in Anwendung auf Therapeutische Statistik. *Sammlung klinischer Vorträge (Innere Medicin No. 31–64)* 1877; 110: 935–962.
4. Hirschberg J. Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Berliner Klinische Wochenschrift* 1877; 21: 297–299.
5. Tröhler U. The French road to Gavarret's clinical application of probabilistic thinking Part 2: Louis-Denis-Jules Gavarret. *J R Soc Med* 2020; 113: 360–366.
6. Korteweg JA. Over de Toepassing der Statistiek op Medische Wetenschappen. *Weekblad van het Nederlandsch Tijdschrift voor Geneeskunde* 1877; 21: 553–560.
7. Martius F. Die Numerische Methode (Statistik und Wahrscheinlichkeitsrechnung) mit besonderer Berücksichtigung ihrer Anwendung auf die Medicin. *Virchows Archiv für Pathologische Anatomie und Physiologie und für Klinische Medicin* 1881; 83: 336–377.
8. Martius F. Die Principien der wissenschaftlichen Forschung in der Therapie. *Volkmanns Sammlung Klinischer Vorträge* 1878; 139: 1169–1188.
9. Fick A. *Die Medicinische Physik*. Braunschweig: F Vieweg und Sohn, 1866.
10. Matthews JR. *Quantification and the Quest For Medical Certainty*. Princeton, NJ: Princeton University Press, 1995.
11. MacAlister D. Probability and Listerism. *Educational Times Reprints* 1882; 37: 40–42.
12. Winsor CP. Probability and Listerism. *Human Biol* 1948; 20: 161–169.
13. Fisher RA. *Statistical Methods for Research Workers*. 5th ed. Edinburgh: Oliver & Boyd, 1934.
14. Zabell S. Discussion of Robin L. Plackett: Fisher's history of inverse probability. *Stat Sci* 1989; 4: 261–263.
15. Ineichen R. Der ierfeldertest von Carl Liebermeister (Bemerkungen zur Entwicklung der medizinischen Statistik im 19. Jahrhundert). *Hist Math* 1994; 21: 28–38.
16. Seneta E. Carl Liebermeister's hypergeometric tails. *Hist Math* 1994; 21: 453–462.
17. Seneta E and Phipps MC. On the comparison of two observed frequencies. *Biomet J* 2001; 43: 23–43.
18. Seneta E, Seif FJ, Liebermeister H and Dietz K. Carl Liebermeister (1833–1901): a pioneer of the investigation and treatment of fever and the developer of a statistical test. *J Med Biogr* 2004; 12: 215–221.
19. Chalmers I. Why the 1948 MRC trial of streptomycin used treatment allocation based on random numbers. *J R Soc Med* 2011; 104: 383–386.
20. Matthews RAJ. The origins of the treatment of uncertainty in clinical medicine. Part 1: ancient roots, familiar disputes. *J R Soc Med* 2020; 113: 193–196.
21. Gibson EW. The role of *p*-values in judging the strength of evidence and realistic replication expectations. *Stat Biopharm Res* 2021; 13: 6–18.
22. Goodman SN. Aligning statistical and scientific reasoning. *Science* 2016; 352: 1180–1181.