The Royal Society of Medicine

Journal of the Royal Society of Medicine; 2022, Vol. 115(2) 69–72
DOI: 10.1177/01410768221077205

Carl Liebermeister and the emergence of modern medical statistics, Part 1: his remarkable work in historical context

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

The role of probabilistic reasoning in medicine has been a source of controversy for millennia.[1] This reflects its implications for such vexed questions as the extent to which medicine is an art or science, and the ability of insights from clinical trials to inform the treatment of individuals. By the mid-19th century, these issues had been joined by a third: the role of statistical analysis in assessing the effectiveness of therapies tested in patient studies.[2,3] Ironically, while early advocates of such ‘numerical methods’ claimed they brought objective clarity to clinical decisions, their principal effect was to provoke more debate, frequently fuelled by misunderstandings and ill-founded arguments. As we shall see, a key driver was the reliance of numerical methods on concepts drawn from the theory of probability, a branch of mathematics notoriously prone to misconceptions. When the debate went into abeyance in the 1880s, the sceptics of numerical methods were in the ascendancy, where they would remain until the early decades of the 20th century.[4]

The ceasefire followed criticism of a remarkable contribution to the debate published in 1877 by Carl Liebermeister.[2,4,5] Born in 1833 in Ronsdorf, Germany, Liebermeister studied medicine and held several senior posts, including Professor of Internal Medicine at Basel, until his death in 1901.[2,6,7] Liebermeister wrote many papers, not all on clinical topics; a list with 93 of his publications originally compiled by his daughter Marie in 1919 can be found in Baumberger[6] (pp. 159–169). His best-known contribution to medicine is the eponymous rule relating changes in body temperature with pulse rate and is still invoked today (e.g. Voets[8]).

Liebermeister’s controversial contribution to the debate over the role of statistical methods in medicine[9] is far less well-known – and, we will argue, undeservedly so. It takes the form of a 28-page paper entitled *Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik* (‘On Probability Theory Applied to Therapeutic Statistics’; henceforth *Über Wkt*.). Published in 1877 in *Sammlung Klinischer Vortrage* (‘Collection of Clinical Lectures’) edited by the distinguished German surgeon Professor Richard Volkmann, it attracted considerable attention at the time. Today it is all but forgotten except by medical historians. However, it remains remarkable in many respects, not least its relevance for current concerns about the use of statistical methods in the interpretation of clinical studies. Indeed, we will argue that *Über Wkt*. contains the building blocks of a paradigm shift in the use of statistics in medicine which would have better served clinicians than today’s predominant methodology.

The purpose of this series of papers is to bring Liebermeister’s work and its significance to a wider audience. We describe its contents and discuss their implications, both for medical science as practised at the time and today. We also describe the reaction *Über Wkt*. provoked among Liebermeister’s contemporaries, and in particular the supposedly damning critiques that left sceptics of numerical methods in the ascendancy for decades to come. We will show that these critiques rested on various misconceptions about the theory of probability held by authors lacking Liebermeister’s considerable ability in the subject. Finally, to further assist recognition of Liebermeister’s remarkable contribution to the history of medical science, we provide supplementary material online, including the first English translation of *Über Wkt*. and a detailed technical commentary.

Historical context

Liebermeister’s *Über Wkt*. is his sole contribution to the long-running debate on the role of probability theory in medicine. This is somewhat surprising, given the quantitative turn taken by this debate in the early 19th century and Liebermeister's considerable mathematical abilities, made strikingly apparent in *Über Wkt*. This shift had been prompted by claims concerning the efficacy of certain medical procedures made by French physicians Louis and Civiale[10] in the 1830s. Advocates of the use of probability theory argued that assessments of efficacy required more than just raw statistics such as simple counts of how many patients did or did not benefit, for if the numbers involved were inadequate, any apparent success could simply be the result of chance. Probability theory, it was argued, provided objective tools for assessing the risk of being thus misled.

The quality of the debate was constrained by the inadequate technical background of many of the participants. This was somewhat remedied in 1840 with the publication of *Principes généraux de statistique médicale* by the mathematically trained French physician Jules Gavarret[11] (see Huth[12] and Tröhler[10] for reviews). Gavarret's methods drew heavily on results in probability theory published by the distinguished French mathematician Simeon-Denis Poisson.[13] In essence, these offered a means of assessing whether the difference in proportion of patients benefiting from different treatments was so large that it could not reasonably be ascribed to chance. This was achieved by a kind of 'confidence interval' calculated from data using formulas based on Poisson's work. The resulting *limites d'oscillation*, as Gavarret called them, gave the range of values within which the difference in proportion could be said to lie with a certain probability, P. If this range did not include zero – corresponding to no difference – then the possibility that the difference was merely a chance effect could be ruled out with probability P.

However, in developing his methodology, Gavarret faced the same question that had confronted mathematicians since the emergence of probability theory in the 17th century: what value of P is sufficiently compelling? This question had a direct bearing on whether medicine can be deemed 'scientific', given the long-standing view that this requires that its findings be both certain – i.e. $P = 100\%$ – and objective.[1] Gavarret recognised that the former was unattainable, not least because it would require an infinite number of patients. This still left the problem of determining a value of P above which chance could reasonably be discarded as the explanation of a finding. In the absence of any obvious objective approach, Gavarret adopted the seemingly arbitrary value used by Poisson of 99.53%. This figure has its origins in nothing more profound than computational convenience and practicality (see Appendix 3 of supplementary material). Perhaps recognising – correctly – the potential criticism that this made the choice 'unscientific', Gavarret attempted to put forward a rationale for the adoption of 99.53% as a threshold above which chance effects could safely be ruled out. It rested on what he called a 'general principle' that (Gavarret,[11] p. 258):

> '*from the moment a test observer has arrived at a high degree of probability relative to the existence of a fact, he can use it as if he were* absolutely certain' [emphasis added].

To this he added two criteria to be met by any specific choice of probability threshold: first, that it be 'high enough to leave no doubt', and second, that it not require 'too large a number of observations' in order for findings to exceed the probability threshold. As to its specific value, Gavarret simply resorted to an appeal to authority: 99.53% should be adopted because that was the value chosen by Poisson, 'an authority which doubtless no-one will try to dispute the importance of'. Gavarret concluded by claiming that:

> '*It is conceivable, in fact, that it would be unreasonable to raise the slightest doubt as to the existence of an event, when there is a bet of 212 against 1 whether it has occurred or not'.*

If Gavarret's attempt to give this threshold some objective basis seems unconvincing, it is because there can be no such basis, a fact insufficiently appreciated in relation to today's *de facto* standard of 95%,[14] and indeed the (remarkably similar) threshold of 99.50% now advocated by some in connection with the so-called 'replication crisis'.[15] This was picked up by the German physiologist Adolph Fick,[16] who noted that the 99.53% threshold 'does not at all have intrinsic reasons' (Matthews,[4] p. 53). He did not suggest an alternative, recognizing perhaps both the impossibility of justifying one and the dangers of highlighting this 'unscientific' aspect of Gavarret's approach, about which he was otherwise supportive.

A more practical concern even among supporters of Gavarret's approach was that its probability threshold appeared to demand the use of very large numbers of patients. This prompted the mathematically trained German clinician Julius Hirschberg (1843–1925) to publish revised formulas involving the less demanding 91.6% confidence level.[17] His justification for reducing the threshold was perfectly reasonable: findings failing to meet Gavarret's standard still contained valuable evidence. However, Hirschberg's precise choice of threshold was apparently dictated by the patently subjective desire to lend credibility to a method of removing cataracts developed by his one-time teacher Albrecht von Graefe (Matthews,[4] pp. 53–56).

It was against this background that Liebermeister entered the debate with the publication of *Über Wkt*. in 1877, offering a far more radical solution to the problem of arbitrary thresholds.

Liebermeister's remarkable achievement

Liebermeister's *Über Wkt*. opens with the bold statement that 'practical doctors' such as himself had often shown 'irrefutably' that making conclusive assessments of treatments requires the use of probability theory. The reference to 'practical doctors' is telling, as it seems directed at sceptics who might otherwise dismiss Liebermeister as one of the 'strangers at the bedside' – i.e. mathematicians – seeking to challenge the physician's role.[18] He then sets out the specific goal of his paper:

> *'[...] to examine how large the probability is that the observed differences in success are not simply due to chance. And for this question only probability calculus gives the necessary indication.'*

While insisting that this is a crucial first step in assessing a therapy, Liebermeister is careful not to over-claim, stressing that the theory of probability 'is completely incapable' of determining the *actual* cause of the difference; that remains 'a matter for clinical analysis'. He then goes on to concede that the reason probability theory had failed to become part of the clinician's toolkit is not because its importance had not been recognised, but because the available theory 'has so far been too incomplete and inconvenient'. Liebermeister credits Poisson as having developed the requisite theory, but observes that the simplifying assumptions underpinning the resulting formulas come at a heavy price: the need for 'hundreds and often many thousands of individual observations'. This, he adds, would raise many practical problems. He then makes a key observation: that the need for such large numbers has been routinely accepted by the medical community to the point of becoming

> *'...an unshakeable dogma that series of observations which do not consist of very large numbers* cannot prove anything at all, *that it is unscientific to want to draw conclusions from small numbers'*. [Emphasis added].

But, asks Liebermeister, is this assumption actually valid? He cites the possibility that even a small study of a highly effective treatment – such as quinine with malarial patients – could still produce strong evidence of efficacy. Liebermeister then argues that the failure of the existing theory to cope with small numbers suggests it is still 'highly imperfect', and that mathematicians 'have not yet provided us' with tools for such situations. He points out that effect sizes would obviously need to be substantial for a small trial to rule out chance to a given standard of probability. Even so, there would be situations where even small studies could rule out chance with high confidence. Liebermeister took the strikingly modern view that all well-conducted studies, whatever their outcome, should be allowed to contribute to the cumulative evidence for the reality of an effect:

> *Truly, we are not so rich in gold coins in the empirical foundations of therapy that we could be advised to throw all silver coins into the water! And a handful of silver coins is often worth more than a single gold coin.*

Such gathering of evidence of varying strength is redolent of today's notion of the *systematic review* (e.g. Chalmers and Altman[19]). Liebermeister's prescience is, however, most striking in his identification of what remains arguably the principal barrier to making the most of clinical data: the use of probability thresholds to decide 'what works'. He argued that both Poisson/Gavarret's threshold of a 99.53% probability and Hirschberg's less demanding 91.6% were both arbitrary and squandered the full power of probabilistic reasoning in assessing clinical evidence – an argument that remains valid for today's *de facto* threshold of 95%. Liebermeister called for this simplistic 'pass/fail' approach to be replaced by the calculation for each study of the exact probability that its findings represent a genuine effect:

> *In fact, if the probability calculus is to be applied to the assessment of therapeutic results with benefit and in an extensive way, then it is necessary.... that one can calculate with certainty and accuracy for each available observation material with which degree of probability chance is excluded. Only when this is possible can we use all our series of observations in a scientific way, by giving each of them exactly the value it deserves.*

Liebermeister then proceeds to develop what had eluded such giants of statistical methodology as Poisson: a means of reliably calculating this probability for any size of study. To achieve this remarkable result, Liebermeister used an approach that allowed him to avoid the approximations that had undermined reliable inferences from very small trials. In doing so, he preceded Fisher's development of the celebrated Exact Test (Fisher,[20] Section 21.02) by more than 50 years. Moreover, Liebermeister's approach is free of interpretational issues which bedevil the use of standard statistical methods to this day.

In Part 2 of this study of Liebermeister's contribution to medical statistics, we outline the basis of this remarkable method, and discuss its uses and limitations.

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.

Acknowledgements: LH is grateful to Flurin Condrau for useful discussions, to Valentina Held for support in translating Liebermeister[9] 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.

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/

References

1. 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.
2. Tröhler U. Probabilistic thinking and the evaluation of therapies, 1700–1900. JLL Bulletin: Commentaries on the History of Treatment Evaluation. See www.jameslindlibrary.org/articles/probabilistic-thinking-and-the-evaluation-of-therapies-1700-1900/ (last checked 24 January 2022).
3. Tröhler U. Probabilistic thinking and evaluation of therapies: an introductory overview. *J Roy Soc Med* 2020; 113: 274–277.
4. Matthews JR. *Quantification and the Quest for Medical Certainty*. Princeton NJ: Princeton, University Press, 1995.
5. Ineichen R. Der "Vierfeldertest" von Carl Liebermeister (Bemerkungen zur Entwicklung der medizinischen Statistik im 19. Jahrhundert). *Historia Mathematica* 1994; 21: 28–38.
6. Baumberger HR. Carl Liebermeister (1833–1901). *Züricher Medizingeschichtliche Abhandlungen* New Series No. 137. 1545–1614. Zurich: Juris Druck Verlag, 1980.
7. 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 Biog* 2004; 12: 215–221.
8. Voets PJ. Central line-associated bloodstream infections and catheter dwell-time: a theoretical foundation for a rule of thumb. *J Theor Biol* 2018; 445: 31–32.
9. Liebermeister C. Über Wahrscheinlichkeitsrechnung in Anwendung auf Therapeutische Statistik. *Sammlung Klinischer Vorträge (Innere Medicin No. 31-64)* 1877; 110: 935–962.
10. Tröhler U. The French road to Gavarret's clinical application of probabilistic thinking Part 2: Louis-Denis-Jules Gavarret. *J Roy Soc Med* 2020; 113: 360–366.
11. Gavarret. *Principes de Statistique Médicale: ou développement des règles qui doivent présider à son employ* Béchet Jeune et Labé, Libraires de la faculte de medicine de Paris, 1840. See also 1844 translation into German by Landmann, S. (1844). *Allgemeine Grundsätze der medicinischen Statistik*.
12. Huth E. Jules Gavarret's Principes Généraux de Statistique Médicale. *J R Soc Med* 2008; 101: 205–212.
13. Poisson. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*, Paris: Bachelier, 1837.
14. Kennedy-Shaffer L. Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize p-values and significance testing. *Am Stat* 2019; 73: 82–90.
15. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Human Behav* 2018; 2: 6–10.
16. Fick A. *Die Medicinische Physik*. Braunschweig: F Vieweg und Sohn, 1866.
17. Hirschberg J. *Die Mathematischen Grundlagen der Medizinischen Statistik*. Leipzig: Verlag von Veit & Comp, 1874.
18. Tröhler U. Conclusions and perspectives, part II: social, national, and long-term perspectives. *J R Soc Med* 2020; 114: 132–139.
19. Chalmers I and Altman DG (eds) *Systematic Reviews*. London: BMJ Publishing, 1995.
20. Fisher RA. *Statistical Methods for Research Workers*. 5th ed. Edinburgh: Oliver & Boyd, 1934.