Preventive Medicine 48 (2009) 600–603

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Commentary

Povl Heiberg's 1897 methodological study on the statistical method as an aid in therapeutic trials

Christian Gluud[a,*], Jørgen Hilden[b]

[a] Copenhagen Trial Unit, Center for Clinical Intervention Research, Department 3344, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK 2100 Copenhagen, Denmark
[b] Department of Biostatistics, Institute of Public Health Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

Detection and evaluation of the effects of treatments was given impetus in the late 1800s by the arrival of some effective medical treatments, such as salicylic medicines, phenacetin, and diphtheria antitoxins (Morse, 1878; Roux et al., 1894; Sørensen, 1896; Fibiger, 1898; Haas, 1983; Hróbjartsson et al., 1998; Lafont, 2007).

One remarkable commentator during this important period was the Danish physician Povl Heiberg (Gluud, 2008; Gluud and Hilden, 2008). In an article published in 1897 Heiberg comments on the fruitlessness of arguments among specialists whose opinions are based on weak evidence (Heiberg, 1897). Instead Heiberg proposes that scientific evidence about the effects of interventions should be sought using "statistical-therapeutic experiments, with the real aim of finding a safe and common therapy that every unbiased physician is obliged to give" to the patient (Heiberg, 1897). The clarity of Heiberg's 1897 article is very impressive, and it is difficult to imagine that it was written more than a century ago.

Who was Povl Heiberg?

Povl Heiberg was born in Sorø, Denmark, in 1868 and died in Denmark in 1963 (Gluud, 2008). He was the son of a priest and a mother, who ran the vicarage farm including a dairy. Povl Heiberg received a good deal of his school education at a Danish boarding school, and became a mathematical student in 1886.

Heiberg studied medicine at Copenhagen University from 1886 to 1893. After holding a number of clinical positions in hospitals and general practice between 1893 and 1906, he became a district medical officer in 1906, and held this position until he retired in 1938.

Heiberg published more than 60 peer reviewed articles (several of which were in English or German), a number of books, and many reports on health statistics. He played an active part in public debate at meetings and in newspapers, and was editor of Ugeskrift for Læger (the Danish Medical Journal) for four years.

Heiberg's 1897 masterpiece

Heiberg acknowledges that statistical-therapeutic experiments cannot be applied to every clinical question; that physicians have difficulties in dealing with statistics; and that statistical methods can be misused (Heiberg, 1897). But he advocates strongly the use of clinical trials and statistical tests before deciding whether interventions should be introduced in clinical practice, and he warns against embarking on a trial if one does not have access to a sufficient number of patients.

Heiberg makes it clear that he appreciates the dangers of systematic errors (resulting from 'bias') as well as random errors (resulting from 'the play of chance') in clinical research (Heiberg, 1897). Heiberg also realizes that diseases may be subject to periodic changes, and that, when attempting to evaluate therapeutic effects, errors or erroneous judgements may result from insufficient diagnostic precision, and variations in age and other prognostic variables. Heiberg (1897) stresses the need for homogeneous patient populations and notes that the circumstances of hospital admission may affect what we would nowadays call 'the severity spectrum'. He even recognizes that bias may be introduced when rumors that a new intervention is successful are responsible for attracting, differentially, the interest of patients with less severe disease. When relying on historical controls, this leads to the false inference that the new intervention is responsible for the apparently improved outcome.

To avoid false conclusions resulting from such lopsided comparisons, Heiberg proposes that experimental interventions should be used in alternate patients, although he actually recommends treatment allocation by date of admission to hospital as a way of reducing selection bias. He also recognizes that observer bias may arise from unblinded experiments, especially if a participant can observe the reaction of the patient next to him.

Heiberg is particularly forceful in arguing that patients assigned to standard treatment (control group) cannot be considered deprived of a chance of cure. He recognizes that use of the specious thinking underlying such notions short-circuits the empirical process necessary for protecting patients from unrecognized adverse effects of experimental interventions, and hinders therapeutic progress. Accordingly, alternate allocation to new or standard treatments poses no ethical problem for him.

Heiberg's historical interest

As well as covering Danish discussions on medical statistical thinking during the 1800s, Heiberg reviews its international historical development, referring to Bernoulli (1713), LaPlace (1812), Poisson (1837), Bouillaud (1840), Gavarret (1840), Hirschberg (1874), Westergaard (1882), and Thiele (1889) (for references, see

* Corresponding author.
E-mail addresses: cgluud@ctu.rh.dk (C. Gluud).

0091-7435/$ – see front matter
doi:10.1016/j.ypmed.2009.04.001

original or translation of Heiberg's 1897 article). In particular, Heiberg stresses Gavarret's pioneering role in introducing statistics into medicine (Gavarret, 1840). As a result, Heiberg recognizes that random error (from 'the play of chance') in small data sets can lead to mistaken inferences. Using numerous tables, he shows how this can happen, and why it is necessary to clarify how likely it is that differences between treatment comparison groups can be explained by play of chance. An aspect that is striking is his clear conception of the implications of fluctuation greater than that predicted by binomial or Poisson variability (overdispersion). He notes that, whenever this occurs, it is necessary to examine the data for possible sources of heterogeneity, such as seasonal variation in the severity or incidence of a disease.

Who were Heiberg's teachers?

Like Gavarret two generations earlier, Heiberg refrains from explaining the mathematical tools he presents and it is unclear what he has invented himself and how much he has adopted from his teachers.

The inspiration for Heiberg's clear understanding of the risks of systematic errors ('biases') and random errors (resulting from 'the play of chance') in therapeutic experiments is likely to have come from the Danish mathematical statistician TN Thiele (1838–1910). Thiele deals with both types of errors in his 1889 textbook 'Forelæsninger over almindelig Iagttagelseslære' ('Lectures on general observation theory') (Thiele, 1889; Lauritzen, 1999, 2007). Heiberg had also read Westergaard's 'Statistikkens Theori i Grundrids' ('Fundamental Theory of Statistics') (Westergaard, 1890), as well as the guidelines for medical statistics in Westergaard's 'Die Lehre von der Mortalität und Morbilität' ('The Theory of Mortality and Morbidity') (Westergaard, 1882). In addition, Heiberg also knew Sørensen's 'Ledetraad for Læger ved statistiske Undersøgelser' ('Guidelines for Doctors in Statistical Analyses') (Sørensen 1889), which was acclaimed because of its clear and easy-to-understand presentation. Thiele was an internationally renowned expert in mathematical statistics, who was well ahead of his time, and his books contain early formulations of analysis of variance and other novel statistical techniques (Schweder, 1999; Lauritzen, 1999, 2007; Norberg, 2008).

Heiberg's mathematical tools

Heiberg's mathematical tools are essentially those that follow from the use of the Gaussian approximation to calculate how frequently deviations of a certain magnitude can be expected, given standard binomial and Poisson situations. A warning is needed about Heiberg's terminology at this point. When he speaks about "the law of large numbers", he means precisely how frequently deviations of a certain magnitude can be expected, given standard binomial and Poisson situations (Heiberg, 1897).

Heiberg takes it as an empirical matter whether the law holds, the envisaged alternative being fluctuations that are greater than predicted by binomial or Poisson variability (overdispersion). (In modern texts 'the laws of large numbers' are theorems about the mathematical conditions under which the distribution of an average has the usual asymptotic properties. This has nothing to do with the empirical question of whether a given data source is overdispersed).

A sophisticated statistical analysis in Heiberg's paper is one that involves an age-adjusted mortality comparison between two successive calendar years. First, each pair of age classes gives rise to a 2-by-2 table (year versus outcome) (our present Table 1). In the table, comparisons are based on 'expected number of deaths' – indirect standardization – using the proportion dying in the first year to calculate the expected mortality in the second. This technique and terminology was introduced in the 1700s and is prominent in Westergaard's writings (Keiding, 1987). Keiding is unclear from where Westergaard came by the standardization technique and conjectures "that he picked it up during his visits to England" (around 1880) (Keiding, 1987).

The calculated standard error that Heiberg presents in the 4th column of the calculation table (Table 1) involves the added sophistication of taking into account that the 'expected numbers' themselves propagate an associated binomial uncertainty. So in the end, each age class is dealt with by its own 2-by-2 table analysis, summarized using a z statistic ($=\sqrt{\chi^2}$) in the final column. Furthermore, the deviations and their variances are summed (note that $14.22 = 3.42 + 11.52 + \ldots$), and a summary z of 6.4 is calculated as evidence of a difference in mortality. The inspiration for this is likely to have come from Westergaard's book, but we have only found hints there as to how such an analysis should be done, and no fully worked example. Neither did we find any hints of such an analysis in Sørensen's 'Guidelines for Doctors in Statistical Analyses' (Sørensen, 1889). Inspiration from the visionary Danish mathematical statistician Thiele could at most be indirect, as we have found no similar models or hints to such models in his books produced before (Thiele, 1889; Lauritzen, 2007) and after (Thiele, 1903) Heiberg's remarkable article.

Anyhow, Heiberg's procedure is analogous to that proposed on somewhat intuitive grounds by Mantel and Haenszel more than 60 years later (Mantel and Haenszel, 1959). This procedure has subsequently been given theoretic underpinning by many researchers (Kuritz et al., 1988). Mantel and Haenszel do not refer to Heiberg's work or to Westergaard's work for that matter. We will be interested to see whether there is an earlier example than Heiberg's of a confounder-adjusted 2-by-2 analysis equivalent to the Mantel–Haenszel procedure. Incidentally, Heiberg does not comment on the evidential impact of his $z = 6.4$ (we find 6.0), possibly because he did not have a Gaussian distribution table at hand. The observed z is actually equivalent to $P \approx 10^{-9}$.

Table 1
Translated text and tables extracted from Heiberg's 1897 masterpiece on clinical research (Heiberg, 1897).

"A study in a limited area, like a town or a major health insurance society, showed that under constant 'external circumstances' the fluctuations in risk of a specific disease fell within the limits of the law of large numbers. Then the following development occurred in 2 successive years:

	Year (age)	0–1	1–5	5–15	Adult men and women
First year	Diseased	65	818	1196	1260
First year	Dead	19	191	89	23
Second year	Diseased	36	559	681	639
Second year	Dead	7	75	26	4

If the above numbers are ordered in such a way that comparison becomes possible, we obtain the following table:

Age	Dead in second year	Calculated dead in the second year according to first year's experience	Difference	Standard error less than	Prespecified range of fluctuation, as multiple of standard error
0–1 year	7	11	−4	3.4	1.2
1–5 years	75	130	−55	11.5	4.8
5–15 years	26	51	−25	7.1	3.5
Adults	4	12	−8	3.5	2.3
	112	204	−92	14.2	6.4

Fluctuations in the same direction appear in all 4 age groups: in 2 of them they are quite substantial, and for all groups, after having eliminated the age-factor, a fluctuation of 6 times the average error is observed. These numbers clearly show that a new causal factor has appeared, but whether this is a result of spontaneous decline of the risk of the disease or a decline caused by a new treatment, no analyses of numbers can tell us."

After age-standardization, a lethality difference between two successive calendar years is documented by calculating a summary $z = 6.4$. This standard-normal z statistic is the famous one proposed by Mantel and Haenszel (1959) 60 years later.

Heiberg's vocabulary was not like present day 'statistical test' vocabulary. For example, confidence limits were presented under ad hoc names, such as Gavarret's 'limits of oscillation' (Gavarret, 1840). Comparing Heiberg's text with modern texts on clinical trials one also notes a number of methodological caveats of which Heiberg was probably unaware. Heiberg did not seem to recognize the dangers of treatment allocation by date or simple alternation, where foreknowledge of upcoming allocations can result in selection biases. These dangers were not noticed until the 1930s. Bradford Hill, for example, became aware that an alternate allocation scheme had not been strictly observed in a Medical Research Council trial conducted in the early 1930s and that selection bias had thus probably undermined the validity of the comparisons made in the study (Chalmers, 2003; Chalmers, 2008). The full consequences of the dangers of not adhering to adequate generation of the allocation sequence and adequate allocation concealment first became fully recognized about one century later (Schulz et al., 1995; Moher et al., 1998; Kjaergard et al., 2001; Wood et al., 2008). However, from a modern-day perspective, Heiberg nowhere erred, except when writing towards the end of his essay about calculating "the chances that this [observed] difference is real." In this he makes the all too common slip of interpreting a *P*-value as the probability that the null hypothesis is false. Heiberg may well be excused for this, as the application of probabilistic reasoning about chance errors in medicine was still in its infancy.

Heiberg's assessment of his 1897 article

It seems very likely that Heiberg himself was aware of the central importance of his 1897 article. In an article in Ugeskrift for Læger published in 1940 entitled 'Reflections on numeric studies', he stressed the importance of mathematical thinking and the avoidance of systematic errors and random errors in evaluating therapies (Heiberg, 1940). In his autobiography he also repeatedly referred to the mathematical education he had received at home, in school, and later as a student in addition to having a shortened version of his remarkable 1897 article republished 61 years after its first appearance (Heiberg, 1958).

The impact of Heiberg's 1897 article

To the best of our knowledge Heiberg's 1897 article had limited impact. The only author referring to Heiberg's article we are aware of is Johannes Fibiger, who refers to it in his quasi-randomized study from 1898 (Fibiger, 1898; Hróbjartsson et al., 1998). Heiberg's republication just mentioned may well reflect how disappointed Heiberg was that his ideas had not caught on. At a time when casuistry and case series dominated clinical journals, few clinical investigators were probably prepared to having their eyes opened to methodological rigor, and those few who studied it may perhaps just have seen it as a concoction of ideas from existing sources, which – superficially regarded – it was. One may also conjecture that methodological treatises tended to be dismissed by clinicians by the same argument as is sometimes used today: Very good – in theory – but in real life is too multi-facetted and no two patients are alike.

As stated, Heiberg's advice was cited and followed by his younger colleague Fibiger in his diphtheria antiserum study (Fibiger, 1898; Hróbjartsson et al., 1998). Ironically, Fibiger's work was also largely forgotten, and when he won the Nobel prize in 1926, it was for a different – and flawed – line of research in experimental oncology (Hróbjartsson et al., 1998; Modlin et al., 2001).

Most of the modern clinical and statistical research methodology that came to be used in therapeutic trials after the Second World War (Gluud and Nikolova, 2007) was developed by and around Ronald Fisher and Bradford Hill (Chalmers, 2003). Thereafter, the methodological snowball began rolling, quickly merging with snowballs from theoretical statistics, from other empirical disciplines and from the growing formalization of clinical decision-making, clinical trials, research ethics, epidemiology and, most recently, systematic reviews with meta-analyses (Thorlund et al., 2009).

Conflict of interest statement

There are no conflicts of interest in connection with this publication.

Acknowledgments

We thank Iain Chalmers and Jan P Vandenbroucke for helpful comments and suggestions on earlier drafts. This paper is an extended version of a commentary that accompanies our translation of Heiberg's 1897 article at The James Lind Library (Gluud and Hilden, 2008; Heiberg, 1897).

References

Chalmers, I., 2003. Fisher and Bradford Hill: theory and pragmatism? Int. J. Epidemiol. 32, 922–924.
Chalmers, I., 2008. MRC Therapeutic Trials Committee's report on serum treatment of lobar pneumonia. BMJ 1934 The James Lind Library (www.jameslindlibrary.org).
Fibiger, J., 1898. Om Serumbehandling af Difteri [About serum treatment of diphtheria]. Hospitalstidende 6, 309–325 and 337–350.
Gavarret, L.D.J., 1840. Principes généraux de statistique médicale: ou développement des règles qui doivent présider à son emploi. [General Principles of Medical Statistics, or Development of Rules that should Govern their Use]. Bechet Jeune & Labé, Paris. The James Lind Library (www.jameslindlibrary.org).
Gluud, C., 2008. Povl Heiberg (1868–1963). The James Lind Library (www.jameslindlibrary.org).
Gluud, C., Nikolova, D., 2007. Likely country of origin in publications on randomised controlled trials and controlled clinical trials during the last 60 years. Trials 8, 7.
Gluud, C., Hilden, J., 2008. Povl Heiberg's 1897 methodological study on the statistical method as an aid in therapeutic trials. The James Lind Library (www.jameslindlibrary.org).
Haas, H., 1983. History of antipyretic analgesic therapy. Am. J. Med. 75 (5A), 1–3.
Heiberg, P., 1897. Studier over den statistiske undersøgelsesmetode som hjælpemiddel ved terapeutiske undersøgelser [Studies on the statistical study design as an aid in therapeutic trials]. Bibliotek for Læger 89, 1–40 Translated from Danish into English by: Sarah Louise Klingenberg, Mette Hansen, Dimitrinka Nikolova, and Christian Gluud. The James Lind Library (www.jameslindlibrary.org). http://www.jameslindlibrary.org/trial_records/19th_Century/heiberg/heiberg-1897-translation-of-whole-article.pdf.
Heiberg, P., 1940. Causerier over talmæssige undersøgelser [Reflections on numeric studies]. Ugeskrift for Læger 102, 540–550.
Heiberg, P., 1958. Spredte erindringer. Fra en gammel læges liv. [Random memories. From an old physician's life]. Arne Frost-Hansens Forlag, København, pp. 1–157.
Hróbjartsson, A., Gøtzsche, P.C., Gluud, C., 1998. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. BMJ 317, 1243–1245.
Keiding, N., 1987. The method of expected number of deaths, 1786–1886–1986. Int. Stat. Rev. 55, 1–20.
Kjaergard, L.L., Villumsen, J., Gluud, C., 2001. Reported methodological quality and discrepancies between large and small randomised trials in meta-analyses. Ann. Intern. Med. 135, 982–989.
Kuritz, S.J., Landis, J.R., Koch, G.G., 1988. A general overview of Mantel–Haenszel methods: applications and recent developments. Ann. Rev. Public Health 9, 123–160.
Lafont, O., 2007. From the willow to aspirin. Rev. Hist. Pharm. (Paris) 55, 209–216.
Lauritzen, S.L., 1999. Aspects of T.N. Thiele's contributions to statistics. http://www.stat.fi/isi99/proceedings.html. (Accessed 2 May, 2008).
Lauritzen, S.L., 2007. Thiele: Pioneer in Statistics. Oxford University Press, Oxford.
Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22, 719–748.
Modlin, I.M., Kidd, M., Hinoue, T., 2001. Of Fibiger and fables: a cautionary tale of cockroaches and Helicobacter pylori. J. Clin. Gastroenterol. 33 (3), 177–179.
Moher, D., Pham, B., Jones, A., et al., 1998. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352, 609–613.
Morse, N., 1878. Ueber eine neue Darstellungsmethode der Acetylamidophenole. [On a new production method for acetylamidophenole]. Berichte der deutschen chemischen Gesellschaft 11, 232–233 10.1002/cber.18780110151.
Norberg, R., 2008. Thiele, T.N. (1838–1910). http://stats.lse.ac.uk/norberg/links/papers/thi-eas.pdf. (Accessed 21 April, 2008).
Roux, M.E., Martin, M.L., Chaillou, M.A., 1894. Trois cents cas de diphtérie traités par le sérum antidiphtérique. [Three hundred patients treated with anti-diphtheria serum]. Ann. Inst. Pasteur 8, 640–662.
Schulz, K.F., Chalmers, I., Hayes, R.J., Altman, D.G., 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273, 408–412.
Schweder, T., 1999. Early statistics in the Nordic countries — when did the Scandinavians slip behind the British? http://www.stat.fi/isi99/proceedings/arkisto/varasto/schw0844.pdf.
Sørensen, S., 1896. Forsøg med Serumterapi ved Difteritis [Trials on serum therapy for diphtheria]. Hospitalstidende 4, 621–628.

Sørensen, T., 1889. Ledetraad for Læger ved statistiske Undersøgelser. [Guidelines for Doctors in Statistical Analyses]. Published with support from the Danish General Medical Association. Supplement to Ugeskrift for Læger XX;6: 1–62. Bianco Lunos Kgl. Hof-Bogtrykkeri, Kjøbenhavn.

Thiele, T.N., 1889. Forelæsninger over almindelig Iagttagelseslære. [Lectures on the general theory of observation]. Referred to by Heiberg P (1897). Studier over den statistiske undersøgelsesmetode som hjælpemiddel ved terapeutiske undersøgelser [Studies on the statistical study design as an aid in therapeutic trials]. Bibliotek for Læger 1897 (89), 1–40.

Thiele, T.N., 1903. Theory of Observations. Layton, London. (Reprinted in Ann Math Statist 1931; 2:165–308.

Thorlund, K., Devereaux, P.J., Wetterslev, J., et al., 2009. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int. J. Epidemiol. 38, 276–286.

Westergaard, H., 1882. Die Lehre von der Mortalität und Morbilität. [The Theory of Mortality and Morbidity]. Jena, Gustav Fischer Verlag.

Westergaard, H., 1890. Statistikkens Theori i Grundrids. [Fundamental Theory of Statistics]. København. (also published in German 1890. [Die Grundzüge der Theorie der Statistik.]. Gustav Fischer Verlag, Jena.

Wood, L., Egger, M., Gluud, L.L., et al., 2008. Empirical evidence of bias: methodological quality and treatment effect estimates in controlled trials with different interventions and outcomes. Meta-epidemiological study. BMJ 336, 601–605.

Methods, too, divide social scientists. In a very general sense, we can talk of a distinction between hard methods (usually based on a positivist epistemol-ogy and a belief in the reality of social concepts) and soft methods (relying more on interpretation). Yet matters are in practice a great deal more compli-cated, with dierent forms of information being suitable for dierent forms of analysis. There is scope for combining methods through triangulation, but, in order to do this, we need to be clear of the assumptions that underlie each and to ensure that they are not incompatible. Most scienc