

Standardization and control for confounding in observational studies: a historical perspective

Niels Keiding[†] and David Clayton[‡]

University of Copenhagen and University of Cambridge

Abstract. Control for confounders in observational studies was generally handled through stratification and standardization until the 1960s. Standardization typically reweights the stratum-specific rates so that exposure categories become comparable. With the development first of loglinear models, soon also of nonlinear regression techniques (logistic regression, failure time regression) that the emerging computers could handle, regression modelling became the preferred approach, just as was already the case with multiple regression analysis for continuous outcomes. Since the mid 1990s it has become increasingly obvious that weighting methods are still often useful, sometimes even necessary. On this background we aim at describing the emergence of the modelling approach and the refinement of the weighting approach for confounder control.

Key words and phrases: $2 \times 2 \times K$ table, Causality, Decomposition of rates, Epidemiology, Expected number of deaths, Log-linear model, Marginal structural model, National Halothane Study, Odds ratio, Rate ratio, Transportability, H.Westergaard, G.U.Yule.

1. INTRODUCTION: CONFOUNDING AND STANDARDIZATION

In this paper we survey the development of modern methods for controlling for confounding in observational studies, with a primary focus on discrete responses in demography, epidemiology and social science. The forerunners of these methods are the methods of *standardization of rates*, which go back at least to the

Department of Biostatistics University of Copenhagen (e-mail: nike@sund.ku.dk)
Emeritus Honorary Professor of Biostatistics University of Cambridge (e-mail: dc208@cam.ac.uk)

*This work grew out of an invited paper by David Clayton in the session “Meaningful Parametrizations” at the International Biometric Conference in Freiburg in 2002. The authors are grateful to the reviewers for comments and detailed suggestions on earlier drafts.

[†]Important parts of Niels Keiding’s work were done on study leaves at the MRC Biostatistics Unit, Cambridge, October-November 2009 and as Visiting Senior Research Fellow at Jesus College, Oxford, Hilary Term 2012; he is most grateful for the hospitality at these institutions.

[‡]David Clayton was supported by a Wellcome Trust Principal Research Fellowship until his retirement in March, 2012.

	Study population	Standard population
No. of individuals	$A_1 \dots A_k$	$S_1 \dots S_k$
Age distribution	$a_1 \dots a_k, \sum a_i = 1$	$s_1 \dots s_k, \sum s_i = 1$
Death rates	$\alpha_1 \dots \alpha_k$	$\lambda_1 \dots \lambda_k$
Actual no. of deaths	$\sum A_i \alpha_i$	$\sum S_i \lambda_i$
Crude death rate	$\sum A_i \alpha_i / \sum A_i$	$\sum S_i \lambda_i / \sum S_i$

TABLE 1
Age standardization: some notation

18th century (see Keiding (1987) for a review). These methods tackle the problem of comparing rates between populations with different age structures by applying age-specific rates to a single “target” age structure and, thereafter, comparing predicted *marginal* summaries in this target population. However, over the 20th century, the methodological focus swung towards indices which summarize comparisons of *conditional* (covariate-specific) rates. This difference of approach has, at its heart, the distinction between, for example, a ratio of averages and an average of ratios — a distinction discussed at some length in the important papers by Yule (1934) and Kitagawa (1964), which we shall discuss in Section 4. The change of emphasis from a marginal to conditional focus led eventually to the modern dominance of the regression modelling approach in these fields. Clayton and Hills (1993, p. 135) likened the two approaches to the two paradigms for dealing with extraneous variables in experimental science, namely (a) to make a marginal comparison after ensuring, by *randomization*, that the distributions of such variables are equal, and (b) to fix, or *control*, such influences and make comparisons conditional upon these fixed values. In sections following, we shall chart how, in observational studies, statistical approaches swung from the former to the latter. Finally we note that some recent methodological developments have required a movement in the reverse direction.

We shall start by recalling the basic concepts of direct and indirect standardization in the simplest case where a *study population* is to be compared to a *standard population*. Table 1 introduces some notation, where there are k age groups. In indirect standardization, we apply the age-specific death rates for the standard population to the age distribution of the study, yielding the counterfactual number of deaths in the study population if the rates had been the same as the standard rates. The Standardized Mortality Ratio (SMR) is the ratio between the observed number of deaths in the study population to this “expected” number:

$$\text{SMR} = \sum A_i \alpha_i / \sum A_i \lambda_i$$

Note that the numerator does not require knowledge of the age distribution of deaths in the study group. This property has often been useful.

In *direct standardization* one calculates what the marginal death rate would have been in the study population if its age distribution had been the same as in the standard population

$$\text{(Direct) standardized rate} = \sum s_i \alpha_i = \sum S_i \alpha_i / \sum S_i.$$

This is sometimes expressed relative to the marginal rate in the standard population — the *Comparative Mortality Figure* (CMF):

$$\text{CMF} = \sum S_i \alpha_i / \sum S_i \lambda_i.$$

Sato and Matsuyama (2003) and Hernán and Robins (2006) gave concise and readable accounts of the connection of standardization to modern causal analysis. Assume that, as in the above simple situation, outcome is binary (death) and exposure is binary — individuals are either exposed (study population) or unexposed (standard population). Each individual may be thought of as having a different risk for each exposure state, even though only one state can be observed in practice. In addition to depending on exposure, risks depend on a discrete confounder (age group). The causal effect of the exposure can be defined as the ratio of the marginal risk in a population of individuals had they been exposed to the risk for the same individuals had they not been exposed. Conditional exchangeability is assumed; for a given value of the confounder (in the present case, within each age group) the counterfactual risks for each individual do not depend on the actual exposure status. Then the marginal death rate in the unexposed (standard) population of individuals had they been exposed is estimated by the directly standardized rate, so that the causal risk ratio for the unexposed population is estimated by the CMF. Similarly, the causal risk ratio for the exposed population is estimated by the SMR. We may estimate the death rate of the exposed population had they not been exposed by the *indirectly standardized death rate*, obtained by multiplying the crude rate in the standard population by the SMR:

$$\text{Indirect standardized rate} = \frac{\sum S_i \lambda_i}{\sum S_i} \times \frac{\sum A_i \alpha_i}{\sum A_i \lambda_i}.$$

Both direct and indirect approaches are based on comparison of marginal risks although, as pointed out by Miettinen (1972b), they focus on different “target” populations; indirect standardization may be said to have the study population as its target, while direct standardization has the standard population as its target. Indeed, the CMF is identical to the (reciprocal of) the SMR if “study” and “standard” populations are interchanged.

In many epidemiological and biostatistical contexts it is natural to use the total population (exposed + unexposed) as basis for statements about causal risk ratios. With $N_i = A_i + S_i$, the total population size in age group i , the causal risk ratio in the total population will be

$$\frac{\sum N_i \alpha_i}{\sum N_i \lambda_i} = \frac{\sum A_i (N_i/A_i) \alpha_i}{\sum S_i (N_i/S_i) \lambda_i}.$$

This rearrangement of the formula shows that we may interpret standardization with the total population as target as an inverse probability weighting method in which the weighting compensates for non-observation of the counterfactual exposure state for each subject. In the numerator, the contributions of the A_i exposed subjects are inversely weighted by A_i/N_i , which estimates the probability that a subject in age group i of the total study was observed in the exposed state. Similarly, in the denominator, the S_i unexposed subjects are inversely weighted by the probability that a subject was observed in the unexposed state. The method of inverse probability weighting is an important tool in marginal structural models and other methods in modern causal analysis.

Thus, while there are obvious similarities between direct and indirect standardization, there are also important differences. In particular, when the aim is to compare rates in *several* study populations, reversal of the roles of study and

standard population is no longer possible and Yule (1934) pointed out important faults with the indirect approach in this context. Such considerations will lead us, eventually, to see indirect standardization as dependent on an implicit model and, therefore, as a forerunner of the modern conditional modelling approach.

The plan of this paper is to present selected highlights from the historical development of confounder control with focus on the interplay between marginal or conditional choice of target on one hand, and the role of (parametric or non-parametric) statistical models on the other. Section 2 recalls the development of standardization techniques during the 19th century. Section 3 deals with early 20th century approaches to the problem of causal inference, focusing particularly on the contributions of Yule and Pearson. Section 4 records highlights from the parallel development in the social sciences, focusing on the further development of standardization methods in the 20th century — largely in the social sciences. Section 5 deals with the important developments in the 1950's and early 1960's surrounding the analysis of the $2 \times 2 \times K$ contingency table, and Section 6 briefly summarizes the subsequent rise and dominance of regression models. Section 7 points out that the values of parameters in (conditional) probability models are not always the only focus of analysis, that marginal predictions in different target populations are often important, and that such predictions require careful examination of our assumptions. Finally, Section 8 contains a brief concluding summary.

Here we have used the word ‘rate’ as a synonym for ‘proportion’, reflecting usage at the time. It was later recognized that a distinction should properly be made (Elandt-Johnson, 1975; Miettinen, 1976a) and modern usage reflects this. However for this historical review it has been more convenient to follow the older terminology.

2. STANDARDIZATION OF MORTALITY RATES IN THE 19TH CENTURY

Neison’s sanatory comparison of districts

It is fair to start the description of direct and indirect standardization with the paper by Neison (1844), read to the Statistical Society of London on 15 January 1844, responding to claims made at the previous meeting (18 December 1843) of the Society by Chadwick (1844) about “representing the duration of life”.

Chadwick was concerned with comparing mortality “amongst different classes of the community, and amongst the populations of different districts and countries”. He began his article by quoting the 18th century practice of using “proportions of death” (what we would now call the crude death rate): the simple ratio of number of deaths in a year to the size of the population that year. Under the Enlightenment age assumption of stationary population, it is an elementary demographic fact that the crude death rate is the inverse of the average life time in the population, but as Chadwick pointed out, the stationarity assumption was not valid in England at the time. Instead, Chadwick proposed the average age of death (that is, among those dying in the year studied). Neison responded

That the average age of those who die in one community cannot be taken as a test of the value of life when compared with that in another district is evident from the fact that no two districts or places are under the same distribution of population as to ages

To remedy this Neison proposed to not only calculate the average age at death in

each district, but

also what would have been the average age at death if placed under the same population as the metropolis.

This is what we now call *direct standardization*, referring the age-specific mortality rates in the various districts to the same age distribution. A little later Neison remarked that

Another method of viewing this question would be to apply the same rate of mortality to different populations.

what we today call *indirect standardization*.

Keiding (1987) described the prehistory of indirect standardization in 18th century actuarial contexts; although Neison was himself an actuary, we have found no evidence that this literature was known to Neison, who apparently developed direct as well as indirect standardization over Christmas 1843. Schweber (2001, 2006), cf. Bellhouse (2008), attempted a historical-sociological discussion of the debate between Chadwick and Neison.

A few years later Neison (1851) published an elaborate survey “On the rate of mortality among persons of intemperate habits” in which he wrote in the typical style of the time

From the rate of sixteen upwards, it will be seen that the rate of mortality exceeds that of the general population of England and Wales. In the 6111.5 years of life to which the observations extend, 357 deaths have taken place; but if these lives had been subject to the same rate of mortality as the population generally, the number of deaths would only have been 110, showing a difference of 3.25 times. . . . If there be anything, therefore, in the usages of society calculated to destroy life, the most powerful is certainly the use of strong drink.

In other words, an SMR of 3.25.

Expected numbers of deaths (indirect standardization) were calculated in the English official statistical literature, particularly by W. Farr, e.g. Farr (1859), who chose the standard mortality rates as the annual age-specific death rates for 1849-53 in the “healthy districts”, defined as those with average crude mortality rates of at most 17/1000 (see Keiding (1987), for an example). W. Ogle initiated routine use of (direct) standardization in the Registrar-General’s report of 1883, using the 1881 population census of England and Wales as the standard. In 1883, direct standardization of official mortality statistics was also started in Hamburg by G. Koch. Elaborate discussions on the best choice of an international standard age distribution took place over several biennial sessions of the International Statistical Institute, cf. Körösi (1892–1893), Ogle (1892) and von Bortkiewicz (1904).

Westergaard and indirect standardization

Little methodological refinement of the standardization methods seems to have taken place in the 19th century. One exception is the work by the Danish economist and statistician H. Westergaard, who already in his first major publication, Westergaard (1882) (an extension, in German, of a prize paper that he had submitted to the University of Copenhagen the year before), carefully described what he called *die Methode der erwartungsmässig Gestorbenen* (the method of expected deaths), i.e. indirect standardization. He was well aware of the danger that other factors could distort the result from a standardization by age alone and illustrated in a small introductory example the importance of what we would nowadays call

	Years at risk	Dead	Expected Number of Deaths according to	
			three special districts	whole country
Copenhagen	7127	108	156	98
Provincial towns	9556.5	159	183	143
Rural districts	4213.5	74	53	60
Whole country	20897.0	341	392	301

TABLE 2

Distribution of deaths of Danish medical doctors 1815-1870, as well as the expected number of deaths if the doctors had been subjected to the mortality of the general (male) population, based on age-specific mortality rates for Denmark as a whole as well as on age-specific mortality rates separately for each of the three districts Copenhagen, Provincial Towns, Rural Districts (Westergaard, 1882, p. 40).

confounder control, and how the method of expected number of deaths could be used in this connection.

Table 2 shows that when comparing the mortality of medical doctors with that of the general population, it makes a big difference whether the calculation of expected number of deaths is performed for the country as a whole or specifically (we would say “conditionally”) for each urbanization stratum. In Westergaard’s words, our English translation:

It is seen from this how difficult it is to conduct a scientific statistical calculation. The two methods both look correct, and still yield very different results. According to one method one would conclude that the medical professionals live under very unhealthy conditions, according to the other, that their health is relatively good. The difficulty derives from the fact that there *exist two causes*: the medical profession and the place of residence; both causes have to be taken into account, and if one neglects one of them, the place of residence, and only with the help of the general life table considers the influence of the other, one will make an erroneous conclusion. The safest is to continue the stratification of the material until no further disruptive causes exist; if one has no other proof, then a safe sign that this has been achieved, is that further stratification of the material does not change the results.

This general strategy of stratifying until the theoretical variance had been achieved, eliminating any residual heterogeneity beyond the basic binomial variation, was heavily influenced by the then current attempts by Quetelet and Lexis in identifying homogeneous subgroups in data from social statistics, for which the normal distribution could be used, preferably with the interpretation of an approximation to the binomial (see Stigler (1986) for an exposition on Quetelet and Lexis). In his review of the book Thiele (1881) criticized Westergaard’s account for overinterpreting the role of mathematical results such as the law of large numbers (as the central limit theorem was then termed) in empirical sciences. As we shall see, however, Westergaard remained fascinated by the occurrence of binomially distributed data in social statistics.

Westergaard also outlined a derivation of the standard error of the expected number of deaths, using what we would call a Poisson approximation argument similar to the famous approximation by Yule (1934) fifty years later for the standard error of the SMR. We shall see later that Kilpatrick (1962) had the last word on this matter by justifying Yule’s approximation in the framework of maximum likelihood estimation in a proportional hazards model.

Standard error considerations accompany the many concrete calculations on human mortality throughout Westergaard’s book from 1882, which in our view is original in its efforts to integrate statistical considerations of uncertainty into mortality analysis, with indirect standardization as the central tool. In the sec-

ond edition of the book Westergaard (1901, p. 25) explained that the method of expected number of deaths has (our translation)

the advantage of summarizing many small series of observations with all their random differences without having to abandon the classification according to age or other groupings (e.g. occupation, residence etc.), in other words obtaining the advantage of an extensive material, without having to fear its disadvantages.

When Westergaard (1916) finally presented his views on statistics in English, the printed comments in what we now call JASA were supplemented by a detailed review by Edgeworth (1917) for the Royal Statistical Society. Westergaard (1916, p. 246) had gone as far as to write

In vital or economic statistics most numbers have a much wider margin of deviation than is experienced in games. Thus the death rate, the birth rate, the marriage rate, or the relative frequency of suicide fluctuates within wide limits. But it can be proved that, by dividing the observations, sooner or later a marked tendency to the binomial law is revealed in some parts of the observations. Thus, the birth rate varies greatly from year to year; but every year nearly the same ratio between boys and girls, and the same proportions of stillbirths, and of twins are observed . . .

and (p.248)

. . . there is no difficulty in getting several important results concerning relative numbers. The level of mortality may be very different from year to year, but we can perceive a tendency to the binomial law in the relative numbers, the death rates by age, sex, occupation etc.

Edgeworth questioned that “Westergaard’s panacea” would work as a general remedy in all situations, and continued

It never seems to have occurred to him that the ‘physical’ as distinguished from the ‘combinatorial’ distribution, to use Lexis’ distinction, may be treated by the law of error [the normal distribution].

Edgeworth here referred to the empirical (physical) variance as opposed to the binomial (combinatorial). Lexis (1876), in the context of time series of rates, had defined what we now call the overdispersion ratio between these two.

Indirect standardization does not require the age distribution of the cases Regarding standardization, Westergaard (1916, p. 261 ff.) explained and exemplified the method of expected number of deaths, as usual without quoting Neison or other earlier users of that method, such as Farr, and went on

English statisticians often use a modification of the method just described of calculating expected deaths; viz., the method of ‘standards’ (in fact the method of expected deaths can quite as well claim the name of a ‘standard’ method)

and after having outlined direct standardization concluded

In the present case the two forms of comparison lead to nearly the same result, and this will generally be the case, if the age distribution in the special group is not much different from that of the general population. But on the whole the method described last is a little more complicated than the calculation of expected deaths, and in particular not applicable, if the age distribution of the deaths of the barristers and solicitors is unknown.

This last point (that indirect standardization does not require the breakdown of cases in the study population by age) has often been emphasized as an important advantage of indirect standardization. An interesting application was the study of the emerging *fall of the birth rate* read to the Royal Statistical Society in December 1905 by Newsholme and Stevenson (1906) and Yule (1906). (Yule (1920) later presented a concise popular version of the main findings to the Cambridge

Eugenics Society, still interesting reading). The problem was that English birth statistics did not include the age distribution of the mother, and it was therefore recommended to use some standard age-specific birth rates (here: those of Sweden for 1891) and then indirect standardization.

Westergaard and an early randomised clinical trial

Westergaard (1918) published a lengthy rebuttal (“On the future of statistics”) to Edgeworth’s critique. Westergaard was here mainly concerned with the statistician’s overall ambition of contributing to “find the causality”, and with a main point being his criticism of “correlation based on Bravais’s formula” as not indicating causality. However, he also had an interesting, albeit somewhat cryptic, reference to a topic that was to become absolutely central in the coming years: that simple binomial variation is justified under random sampling. In his 1916 paper, he had advocated (p. 238) that

in many cases it will be practically impossible to do without representative statistics.

(Edgeworth (1917) taught Westergaard that the correct phrase was “sampling”, and Westergaard replied that English was for him a foreign language.) To illustrate this, Westergaard (1916, p. 245) wrote

The same formula in a little more complicated form can be applied to the chief problem in medical statistics; viz., to find whether a particular method of treatment of disease is effective. Let the mortality of patients suffering from the disease be p_2 , when treated with a serum, p_1 , when treated without it, and let the numbers in each case be n_2 and n_1 . Then the mean error of the difference between the frequencies of dying in the two groups will be $\sqrt{p_1q_1/n_1 + p_2q_2/n_2}$ and we can get an approximation by putting the observed relative values instead of p_1 and p_2 .

In his rebuttal, Westergaard (1918, p. 508) revealed that this was not just a hypothetical example:

A very interesting method of sampling was tried several years ago in a Danish hospital for epidemic diseases in order to test the influence of serum on patients suffering from diphtheria. Patients brought into the hospital one day were treated with serum, the next day’s patients got no injection, and so on alternately. Here in all probability the two series of observations were homogeneous.

Westergaard here referred to the experiment by Fibiger (1898), discussed by Hrobjartsson et al. (1998) as “the first randomized clinical trial” and further documented in the James Lind Library:

(http://www.jameslindlibrary.org/trial_records/19th_Century/fibiger/fibiger_kp.html).

3. ASSOCIATION, AND CAUSALITY: YULE, PEARSON, AND FOLLOWING

The topic of causality in the early statistical literature is particularly associated with Yule and with Pearson, although they were far from the first to grapple with the problem. Yule considered the topic mainly in the context of discrete data, while Pearson considered mainly continuous variables. It is perhaps this which led to some dispute between them, particularly in regard to measures of association. For a detailed review of their differences, see Aldrich (1995).

Yule’s measures of association and partial association

For a 2×2 table with entries a, b, c, d , Yule (1900) defined the association measure $Q = (ad - bc)/(ad + bc)$, noting that it equals 0 under independence and 1 or -1 under complete association. There are of course many choices of association measure that fulfil these conditions. Pearson (1900, pp. 14-18) immediately made

Town	Date	Attack Rate under 10		Attack rate over 10		Yule's Q	
		Vaccinated	Unvaccinated	Vaccinated	Unvaccinated	< 10	> 10
Sheffield	1887–88	7.9	67.6	28.3	53.6	.92	.49
Warrington	1892–93	4.4	54.5	29.9	57.6	.93	.52
Dewsbury	1891–92	10.2	50.8	27.7	53.4	.80	.50
Leicester	1892–93	2.5	35.3	22.2	47.0	.91	.51
Gloucester	1895–96	8.8	46.3	32.2	50.0	.80	.36

TABLE 3

Yule's analysis of the association between smallpox vaccination and attack rates (defined as percentage contracting the disease in "invaded household")

strong objections to Yule's choice; he wanted a parameter that agreed well with the correlation if the 2×2 table was generated from an underlying bivariate normal distribution. The discussion between Yule and Pearson and their camps went on for more than a decade. It was chronicled from a historical-sociological viewpoint by MacKenzie (MacKenzie, 1978, 1981).

That he regarded the concrete values of Q meaningful outside of 0 or 1 is illustrated by his analysis of the association between smallpox vaccination and attack, as measured by Q , in several towns (Table 3). The values of Q were much higher for young children than for older people, but did not vary markedly between different towns, despite considerable variation in attack rates. This use of Q is different from an immediately interpretable population summary measure and it is closer to how we use models and parameters today. Indeed, since Q is a simple transformation of the odds ratio, $(ad)/(bc)$, Yule's analyses of association anticipate modern orthodoxy ($Q = 0.9$ corresponds to an odds ratio of 19, and $Q = 0.5$ to an odds ratio of 3).

Yule's view on *causal* association was largely expounded by consideration of its antithesis, which he termed "illusory" or "misleading" association. Chief amongst the reasons for such non-causal association he identified as that due to the direct effect of a third variable on outcome. His discussion of this phenomenon in Yule (1903) (under the heading "On the fallacies that may be caused by the mixing of distinct records") and, later, in his 1911 book (Yule, 1911) came to be termed "Yule's paradox", describing the situation in which two variables are *marginally* associated but not associated when examined in subgroups in which the third, causal, variable is held constant. The idea of measuring the strength of association holding further variables constant, which Yule termed "partial" association, was thus identified as an important protection against fallacious causal explanations. However, he did not formally consider modelling these partial associations. Indeed, he commented (Yule, 1900):

The number of possible partial coefficients becomes very high as soon as we go beyond four or five variables.

Yule did not discuss more parsimonious definitions of partial association, although clearly he regarded the empirical stability of Q over different subgroups of data as a strong point in its favour. Commenting on some data on recovery from smallpox, in Yule (1912) he later wrote

This, as it seems to me, is a most important property ... If you told any man of ordinary intelligence that the association between treatment and recovery was low at the beginning of the experiment, reached a maximum when 50 per cent. of the cases were treated and then fell off again as the proportion of cases treated was further increased, he would, I think, be legitimately puzzled, and would require a good deal of explanation as to what you meant by association. ... The association coefficient Q keeps the same value throughout, quite unaffected by the ratio of cases

treated to cases untreated.

Pearson and Tocher's test for identity of two mortality distributions

Pearson regarded the theory of correlation as of fundamental importance, even to the extent of replacing “the old idea of causality” (Pearson, 1910). Nevertheless, he recognised the existence of “spurious” correlations due to incorrect use of indices or, later, due to a third variable such as race (Pearson et al., 1899).

Although most of Pearson's work concerned correlation between continuous variables, perhaps the most relevant to our present discussion is his work, with J.F. Tocher, on comparing mortality distributions. Pearson and Tocher (1915) posed the question of finding a proper test for comparing two mortality distributions. Having pointed out the problems of comparing crude mortality rates, they considered comparison of standardized rates (or, rather, proportions). In their notation, if we denote the number of deaths in age group s ($= 1, \dots, S$) in the two samples to be compared by d_s, d'_s and the corresponding numbers of persons at risk by a_s, a'_s , then two age-standardised rates can be calculated as

$$M = \frac{1}{A} \sum A_s \frac{d_s}{a_s}, \text{ and } M' = \frac{1}{A} \sum A_s \frac{d'_s}{a'_s}$$

where A_s represent the standard population in age group s and $A = \sum A_s$. Noting that the difference between standardized rates can be expressed as a weighted mean of the differences between age-specific rates,

$$M' - M = \sum \frac{A_s}{A} \left(\frac{d_s}{a_s} - \frac{d'_s}{a'_s} \right),$$

they showed that, under the null hypothesis that the true rates are equal for the two groups to be compared,

$$\text{Var}(M' - M) = \sum \left(\frac{A_s}{A} \right)^2 p_s (1 - p_s) \left(\frac{1}{a_s} + \frac{1}{a'_s} \right)$$

where p_s denote the (common) age-specific binomial probabilities. Finally, for large studies, they advocated estimation of p_s by $(d_s + d'_s)/(a_s + a'_s)$ and treating $(M' - M)$ as approximately normally distributed or, equivalently,

$$Q^2 = \frac{(M' - M)^2}{\widehat{\text{Var}}(M' - M)},$$

as a chi-squared variate on one degree of freedom (note that their Q^2 is not directly related to Yule's Q). However they pointed out a major problem with this approach; that different choices of standard population lead to different answers, and that there would usually be objections to any one choice. In an attempt to resolve this difficulty they proposed choosing the weights A_s/A to maximise the test statistic and showed that the resulting Q^2 is a χ^2 test on S degrees of freedom. This is because, as Fisher (1922) remarked, each age-specific 2×2 -table of districts vs. survival contributes an independent degree of freedom to the χ^2 test.

Pearson and Tocher's derivation of this test anticipates the much later, and more general, derivation of the score test as a “Lagrange multiplier test”. However, the maximized test statistic could sometimes involve negative weights, A_s ,

which they described as “irrational”. This feature of the test makes it sensitive to differences in mortality in different directions at different ages. They discussed the desirability of this feature and noted that it should be possible to carry out the maximization subject to the weights being positive but “could not see how” to do this (the derivation of a test designed to detect differences in the same direction in all age groups was not to be proposed until the work of Cochran, nearly forty years later — see our discussion of the $2 \times 2 \times K$ below). However, they argued that the sensitivity of their test to differences in death rates in different directions in different age groups in fact represented an improvement over the comparison of corrected, or standardized, rates since “that idea is essentially imperfect and does not really distinguish between differences in the manner of dying.”

Further application of the method of expected numbers of deaths

As described in Section 2, Westergaard (1882) from the very beginning emphasized that the method of expected numbers of death could be calculated according to any stratification, not just age. Encouraged by Westergaard (1916)’s survey in English, Woodbury (1922) demonstrated this through the example of infant mortality as related to mother’s age, parity (called here *order of birth*), earnings of father and plural births. For example: the crude death rates by order of births form a clear J-shaped pattern with nadir at third birth; assuming that only age of the mother was a determinant one can calculate the expected rates for each order of birth, and one gets still a J, though somewhat attenuated, showing that a bit of the effect of birth order is explained by mother’s age. Woodbury did not forget to warn:

Since it is an averaging process the method will yield satisfactory results only when an average is appropriate.

Stouffer and Tibbitts (1933) followed up by pointing out that in many situations the calculations of expected numbers for χ^2 tests would coincide with the “Westergaard method”.

4. STANDARDIZATION IN THE 20TH CENTURY

Although, as we have seen, standardization methods were widely used in the 19th century, it was in the 20th century that a more careful examination of the properties of these methods was made. Particularly important are the authoritative reviews by Yule (1934) and, thirty years later, by Kitagawa (Kitagawa, 1964, 1966). Both these authors saw the primary aim as being the construction of what Yule termed “an average ratio of mortalities”, although Yule went on to remark

in Annual Reports and Statistical Reviews the process is always carried a stage further, viz. to the calculation of a “standardized death-rate”. This extension is really superfluous, though it may have its conveniences.

(the standardized rate in the study population being constructed by multiplying the crude rate in the standard population by the standardized ratio of rates for the study population versus the standard population).

Ratio of averages or average of ratios?

Both Yule and Kitagawa noted that central to the discussion was the consideration of two sorts of indices. The first of these, termed a “ratio of averages” by Yule, has the form $\sum w_i x_i / \sum w_i y_i$ while the second, which he termed an “average of ratios”, has the form $\sum w_i^* (x_i / y_i) / \sum w_i^*$. Kitagawa noted that economists would describe the former as an “aggregative index” and the latter as an “average of relatives”.

Both authors pointed out that, although the two types of index seem to be doing rather different things, it is somewhat puzzling that they are algebraically equivalent — we only have to write $w_i^* = w_i y_i$. It is important to note, however, that the algebraic equivalence does not mean that a given index is equally interpretable in either sense. Thus, for the index to be interpretable as a ratio of averages, the weights w_i must reflect some population distribution so that numerator and denominator of the index represent marginal expectations in the same population. Alternatively, to present the average of the age-specific ratios, x_i/y_i , as a single measure of the age-specific effect would be misleading if they were not reasonably homogeneous. Kitagawa concluded

the choice between an aggregative index and an average of relatives in a mortality analysis, for example, should be made on the basis of whether the researcher wants to compare two schedules of death rates in terms of the total number of deaths they would yield in a standard population *or* in terms of the relative (proportionate) differences between corresponding specific rates in the two schedules. Both types of index can be useful when correctly applied and interpreted.

Here Kitagawa very clearly defined the distinction between what we, in the Introduction, termed the *marginal* and the *conditional* targets. Immediately after this definition, she hastened to point out that

It must be recognized at the outset, however, that no single summary statistic can be a substitute for a detailed comparison of the specific rates in two or more schedules of rates.

On the matter of averaging different ratios, Yule (1934) started his paper with the example of comparing the death rates for England and Wales for 1901 and 1931. His Table I contains these for both sexes in 5-year age groups and he commented

...the rates have fallen at all ages up to 75 for males and 85 for females. At the same time the amount of the fall is very different at different ages, apart even from the actual rise in old age. The problem is simply to obtain some satisfactory form of average of all the ratios shown in columns 4 and 7, an average which will measure in summary form the general fall in mortality between the two epochs, just as an index-number measures the general fall or rise in prices.

So far, there is no requirement for these ratios to be similar. However, when describing indirect standardization, Yule (1934, p. 12) pointed out that

if ... all the ratios of sub-rates are the same, no variation of weighting can make any difference

and warned (p. 13)

And perhaps it may be remarked that ... if the ratios m_{ur}/m_{sr} are very different in different age groups, any comparative mortality figure becomes of questionable value.

The issue of constancy of ratios was picked up in the printed discussion of the paper (Yule, 1934, p. 76) by Percy Stokes, seconder of vote of thanks:

Those of us who have taught these methods to students have been accustomed to point out that they lead to identical results when the local rates bear to the standard rates the same proportion at every age.

Comparability of mortality ratios

Yule noted that, particularly in official mortality statistics, standardization is applied to many different study populations so that, as well as the standardized ratio of mortality in each study population to the standard population being meaningful in its own right, the comparison of the indices for two study populations should also be meaningful. He drew attention to the fact that the ratio of two seemingly legitimate indices is not necessarily itself a legitimate index. He

concluded that either type of index could legitimately be used either if the same weights w_i are used across study populations (for ratios of averages) or if the same w_i^* are used (for averages of ratios).

Denoting a standardized ratio for comparing study groups A and B with standard by ${}_sR_a$ and ${}_sR_b$ respectively, Yule suggested that ${}_sR_a/{}_sR_b$ should be a legitimate index of the ratio of mortalities in population A to that in population B. He also suggested that, ideally, ${}_aR_b = {}_sR_a/{}_sR_b$ but noted that, whereas the CMF of direct standardization fulfils the former criterion, no method of standardization hitherto suggested fulfilled this more stringent criterion. Indirect standardization fulfils neither criterion and Yule judged it to be “hardly a method of standardization at all”.

Yule’s paper is also famous for its derivation of standard errors of comparative mortality figures; for the particular case of the SMR we have

$$\text{SMR} = \text{Observed/Expected}, \quad O/E$$

and

$$\text{S.E.}(\text{SMR}) \approx \sqrt{O/E}.$$

As noted earlier, this was already derived by Westergaard (1882), although this was apparently not generally known.

A final matter occupying no less than twelve pages of Yule (1934) is the discussion of a context-free average, termed by Yule his C_3 method or the *equivalent average death rate*, which is just the simple average of all age-specific death rates. This quantity could also be explained as the death rate standardized to a population with equal numbers in each age group. As we shall see below, it was further discussed by Kilpatrick (1962) and rediscovered by Day (1976) in an application to cancer epidemiology. In modern survival analysis it is called the cumulative hazard and estimated nonparametrically by the Nelson–Aalen estimator (Nelson, 1972; Aalen, 1978; Andersen et al., 1993).

Elaboration: Rosenberg’s test factor standardization

During World War II, the United States Army established a Research Branch to investigate problems of morale, soldier preferences and other issues to provide information that would allow the military to make sensible decisions on practical issues involved in army life. To formalize some of the tools used in that generally rather practical research, illustrated with concrete examples from that work, Kendall and Lazarsfeld (1950) introduced and discussed the terminology of elaboration: A statistical relation has been established between two variables, one of which is assumed to be the cause, the other to be the effect. The aim is to further understand that relation by introducing a third variable (called test factor) related to the ‘cause’ as well as the ‘effect’. Kendall and Lazarsfeld carefully distinguished between antecedent and intervening test variables, depending on the temporal order of the ‘cause’ and the test variables. If the population is stratified according to an antecedent test factor, and the partial relationships between the two original variables then vanish, the relation between ‘cause’ and ‘effect’ has been explained through their relations to the test variable, which is then termed spurious. If the association between cause and effect disappears (is reduced) by controlling on the intervening variable, Kendall and Lazarsfeld talk about complete (partial) interpretation of the original two-factor relationship.

We note that interpretation has gone out of use at least in epidemiological applications and in most of modern causal inference where the focus is on obtaining an undiluted measure of the causal effect of the ‘cause’, not diluting this effect by conditioning on variables on the causal pathway from cause to effect (see Pearl (2001) or Petersen et al. (2006)). Instead a general area of Mediation Analysis has grown up, see (MacKinnon, 2008, Sec. 1.8) for a useful historical survey.

Rosenberg (1962) used standardization to obtain a single summary measure from all the partial (i.e. conditional) associations resulting from the stratification in an elaboration. Rosenberg’s famous example was a study of the possible association between religious affiliation and self-esteem for high school students, controlling for (all combinations of) father’s education, social class identification, and high school grades. Thus, this is an example of interpretation by conditioning on variables that might mediate an effect of religious affiliation on sons’ self-esteem. The crude association showed higher self-esteem for Jews than for Catholics and Protestants; by standardizing on the joint distribution of the three covariates in the total population this difference was halved.

Rosenberg emphasized that in survey research the end product of the standardization exercise is not a single rate as in demography, but

In survey research, however, we are interested in *total distributions*. Thus, if we examine the association between X and Y standardizing on Z , we must emerge with a standardized table (of the joint distribution of X and Y) which contains all the cells of the original table.

Rosenberg indicated shortcuts to avoid repeating the same calculations when calculating the entries of this table.

The Peters-Belson approach

This technique (Peters, 1941; Belson, 1956) was developed for comparing an experimental group with a control group in an observational study on some continuous outcome. The proposal is to regress the outcome on covariates only in the control group and use the resulting regression equation to predict the results for the experimental group under the assumption of no difference between the groups. A simple test of no differences concludes the analysis. Cochran (1969) showed that under some assumptions of (much) larger variance in experimental group than control group this technique might yield stronger inference than standard analysis of covariance, and that it will also be robust to certain types of effect modification. The technique has recently been revived by Graubard et al. (2005).

Decomposition of crude rate differences and ratios

Several authors have suggested a decomposition of a contrast between two crude rates into a component due to differences between the age-specific rates, and a component due to differences between the age structures of the two populations.

Kitagawa (1955) proposed an additive decomposition in which the difference in crude rates is expressed as a sum of (a) the difference between the (direct) standardized rates, and (b) a residual due to the difference in age structure. Rather than treating one population as the standard population and the second as the study population, she treated them symmetrically, standardizing both to

the mean of the two population age structures:

$$\begin{aligned} & \text{Crude rate (study)} - \text{Crude rate (standard)} \\ &= \sum a_i \alpha_i - \sum s_i \lambda_i \\ &= \sum (\alpha_i - \lambda_i) \frac{a_i + s_i}{2} + \sum (a_i - s_i) \frac{\alpha_i + \lambda_i}{2}. \end{aligned}$$

The first term contrasts the standardized rates while the second contrasts the age structures.

However, ratio comparisons are more frequently employed when contrasting rates and several authors have considered a multiplicative decomposition in which the ratio of crude rates is expressed as the product of a standardized rate ratio and a factor reflecting the effect of the different age structures. Such a decomposition, in which the age-standardized measure is the SMR, was proposed by Miettinen (1972b):

$$\frac{\text{Crude rate (study)}}{\text{Crude rate (standard)}} = \frac{\sum a_i \alpha_i}{\sum s_i \lambda_i} = \frac{\sum a_i \alpha_i}{\sum a_i \lambda_i} \times \frac{\sum a_i \lambda_i}{\sum s_i \lambda_i}.$$

The first term is the SMR and the second, which reflects the effect of the differing age structures, Miettinen termed the “confounding risk ratio”.

Kitagawa (1955) had also proposed a multiplicative decomposition which, as in her additive decomposition, treated the two populations symmetrically. Here, the standardized ratio measure was inspired by the literature on price indices in economics. If, in a “base” year, the price of commodity i is p_{0i} and the quantity purchased is q_{0i} and, in year t the equivalent values are p_{ti} and q_{ti} , then an overall comparison of prices requires adjustment for differing consumption patterns. Simple relative indices can be constructed by fixing consumption at base or at t . The former is Laspeyres’s index, $\sum p_{ti} q_{0i} / \sum p_{0i} q_{0i}$, and the latter is Paasche’s index, $\sum p_{ti} q_{ti} / \sum p_{ti} q_{ti}$. These are asymmetric with respect to the two time points and this asymmetry is addressed in Fisher’s “ideal” index, defined as the geometric mean of Laspeyres’s and Paasche’s indices. Kitagawa noted that Laspeyres’s and Paasche’s indices are directly analogous to the CMF and SMR respectively and, in her symmetric decomposition,

$$\frac{\sum a_i \alpha_i}{\sum s_i \lambda_i} = \sqrt{\frac{\sum s_i \alpha_i}{\sum s_i \lambda_i} \times \frac{\sum a_i \alpha_i}{\sum a_i \lambda_i}} \times \sqrt{\frac{\sum \lambda_i a_i}{\sum \lambda_i s_i} \times \frac{\sum \alpha_i a_i}{\sum \alpha_i s_i}}$$

the first term is “ideal” index formed by the geometric mean of the CMF and SMR, and the second term is

the geometric mean of two indexes summarizing differences in I -composition; one an aggregative index using the I -specific rates of the base population as weights, and the second an aggregative index using the I -specific rates of the given population as weights.

Kitagawa’s paper concluded with a detailed comparison to the “Westergaard method” as documented by Woodbury (1922). Woodbury’s paper had also inspired Kitagawa’s contemporary R. H. Turner, also Ph.D. from the University of Chicago, to develop an approach to additive decomposition according to several covariates (Turner, 1949), showing how the “non-white–white” differential in labour force participation is associated with marital status, household relationship and age. Kitagawa’s decomposition paper continues to be frequently cited

	Cases	Controls	
Exposed	A	B	$N = A + B + C + D$
Not exposed	C	D	

TABLE 4
Frequencies in a 2×2 contingency table derived from a case-control study

and the technique is still included in current textbooks in demography (e.g. Preston et al. (2001)). There has been a considerable further development of additive decomposition ideas, for recent reviews see Chevan and Sutherland (2009) for the development in demography and Powers and Yun (2009) for decomposition of hazard rate models and some references to developments in econometrics and to some extent in sociology. We return in Section 6 to the connection with the method of “purging” suggested by C.C. Clogg.

5. ODDS RATIOS AND THE $2 \times 2 \times K$ CONTINGENCY TABLE

Case-control studies and the odds ratio

Although the case-control study has a long history, its use to provide quantitative measures of the strength of association is more recent, generally being attributed to Cornfield (1951). Table 4 sets out results from a hypothetical case-control study comparing some exposure in cases of a disease with that in a control group of individuals free of the disease. In this work, he demonstrated that, if the disease is rare, i.e. prevalence of disease in the population, X , is near zero and the proportion of cases and controls exposed are p_1 and p_0 respectively, then the prevalence of disease in exposed subjects is, to a close approximation, Xp_1/p_0 , and $X(1 - p_1)/(1 - p_0)$ in subjects not exposed. Thus the ratio of prevalences is approximated by the *odds ratio*

$$\frac{p_1}{1 - p_1} / \frac{p_0}{1 - p_0}$$

which can be estimated by $(AD)/(BC)$.

In this work, Cornfield discussed the problem of bias due to poor control selection, but did not explicitly address the problem of confounding by a third factor. In later work (Cornfield, 1956) did consider the case of the $2 \times 2 \times K$ table in which the K strata were different case-control studies. However his analysis focussed on the *consistency* of the stratum-specific odds ratios; having excluded outlying studies he, at this stage, ignored Yule’s paradox, simply summing over the remaining studies and calculating the odds ratio in the the marginal 2×2 table.

Interaction and “Simpson’s paradox”

Bartlett (1935) linked consistency of odds ratios in contingency tables with the concept of “interaction”. Specifically, he defined zero second order interaction in the $2 \times 2 \times 2$ contingency table of variables X , Y and Z as occurring when the odds ratios between X and Y conditional upon the level of Z are stable across levels of Z . (Because of the symmetry of the odds ratio measure, the roles of the three variables are interchangeable). In an important and much cited paper, Simpson (1951) discussed interpretation of no interaction in the $2 \times 2 \times 2$ table, noting that “there is considerable scope for paradox”.

If one were to read only the abstract of Simpson’s paper, one could be forgiven for believing that he had simply restated Yule’s paradox in this rather special case:

it is shown by an example that vanishing of this second order interaction does not necessarily justify the mechanical procedure of forming the three component 2×2 tables and testing each of these for significance by standard methods.

(by “component” tables, he meant the marginal tables). Thus, “Simpson’s paradox” is often identified with Yule’s paradox, sometimes being referred to as the Yule-Simpson paradox. However, the body of Simpson’s paper contains a much more subtle point about the nature of confounding.

Simpson’s example is a table in which X and Y are both associated with Z , in which there is no second order interaction, and the conditional odds ratios for X versus Y are 1.2 while the marginal odds ratio is 1.0. He pointed out that if X is a medical treatment, Y an outcome, and Z sex, then there is clearly a treatment effect — the conditional odds ratio provides the “right” answer, the treatment effect having been destroyed in the margin by negative confounding by sex. Simpson compared this with an imaginary experiment concerning a pack of playing cards which have been played with by a baby in such a way that red cards and court cards, being more attractive, have become dirtier. Variables X and Y now denote red/black and court/plain and Z denotes the cleanliness of the cards. In this case, Simpson pointed out that the *marginal* table of X versus Y , “provides what we would call the sensible answer, that there is no such association”. This is, perhaps, the real Simpson’s paradox — the same table demonstrates Yule’s paradox when labelled one way but does not when it is labelled another way. Simpson’s paper pointed out that the causal status of variables is central; one can condition on *causes* when forming conditional estimates of treatment effects, but not upon *effects*. As we shall see in the next section, this point is central to the problem of time-dependent confounding which have inspired much recent methodological advance. A closely related issue is the phenomenon of selection bias, famously discussed by Berkson (1946) in relation to hospital-based studies. There X and Y are observed only when an effect, Z (e.g. attending hospital), takes on a specific value.

A further contribution of Simpson’s paper was to point out the “non-collapsibility” of the odds ratio measure in this zero interaction case; the conditional and marginal odds ratios between X and Y are only the same if either X is conditionally independent of Z given Y , or Y is conditionally independent of Z given X . Note that these conditions may not be satisfied even in randomized studies — another of the paradoxes to which Simpson drew attention. For a more detailed discussion of Simpson’s paper see Hernán et al. (2011).

Cochran’s analyses of the $2 \times 2 \times K$ table

In his important paper on “methods for strengthening the common χ^2 test”, Cochran (1954) proposed a “combined test of significance of the difference in occurrence rates in the two samples” when “the whole procedure is repeated a number of times under somewhat differing environmental conditions”. He pointed out that carrying out the χ^2 test in the marginal table

is legitimate only if the probability p of an occurrence (on the null hypothesis) can be assumed to be the same in all the individual 2×2 tables.

(He did not further qualify this statement in the light of Simpson’s insight discussed above). He proposed three alternative analyses. The first of these was to add up the χ^2 test statistics from each table, and to compare the result with the χ^2 distribution on K degrees of freedom. This, as already noted, is equivalent to Pearson and Tocher’s earlier proposal but Cochran judged it a poor method since

It takes no account of the signs of the differences ($p_1 - p_0$) in the two samples, and consequently lacks power in detecting a difference that shows up consistently in the same direction in all or most of the individual tables.

The second alternative he considered was to calculate the “ χ ” value for each table — the square roots of the χ^2 statistics, with signs equal to those of the corresponding $(p_1 - p_0)$ ’s — and to compare the sum of these values with the normal distribution with mean zero and variance K . He noted, however, that this method would not be appropriate if the sample sizes (the “ N ’s”) vary substantially between tables, since

Tables that have very small N ’s cannot be expected to be of much use in detecting a difference, yet they receive the same weight as tables with large N ’s.

He also noted that variation of the probabilities of outcome between tables would also adversely affect the power of this method:

Further, if the p ’s vary from say 0 to 50%, the difference that we are trying to detect, if present, is unlikely to be constant at all levels of p . A large amount of experience suggests that the difference is more likely to be constant on the probit or logit scale.

It is clear, therefore, that Cochran considered the ideal analysis to be based on a model of “constant effect” across the tables. Indeed, when the data were sufficiently extensive, he advocated use of empirical logit or probit transformation of the observed proportions followed by model fitting by weighted least squares. Such an approach, based on fitting a formal model to a table of proportions had already been pioneered by Dyke and Patterson (1952), and will be discussed in Section 6.

In situations in which the data were not sufficiently extensive to allow an approach based on empirical transforms, Cochran proposed an alternative test “in the original scale”. This involved calculating a weighted mean of the differences $d = (p_1 - p_0)$ over tables. In our notation, comparing the prevalence of exposure between cases and controls,

$$\begin{aligned} d_i &= \frac{A_i}{A_i + C_i} - \frac{B_i}{B_i + D_i} \\ w_i &= \left(\frac{1}{A_i + C_i} + \frac{1}{B_i + D_i} \right)^{-1} \\ \bar{d} &= \sum w_i d_i / \sum w_i \end{aligned}$$

In calculating the variance of \bar{d} , he estimated the variance of the d_i ’s under a binomial model using a plug-in estimate for the expected values of p_{1i}, p_{0i} under the null hypothesis: $(A_i + B_i)/N_i$. Cochran described the resulting test as performing well “under a wide range of variations in the N ’s and p ’s from table to table”.

A point of some interest is Cochran’s choice of weights which, as pointed out by Birch (1964), was “rather heuristic”. If this procedure had truly been, as Cochran described it, an analysis “in the original scale”, one would naturally have weighted the differences inversely by their variance. But this does not lead to Cochran’s weights, and he provided no justification for his alternative choice. A likely possibility is that he noted that weighting inversely by precision leads to two different

tests according to whether we choose to compare the proportions exposed between cases and control, or the proportions of cases between exposed and unexposed groups. Cochran's choice of weights avoided this embarrassment.

Mantel and Haenszel

Seemingly unaware of Cochran's work, Mantel and Haenszel (1959) considered the analysis of the $2 \times 2 \times K$ contingency table. This paper explicitly related the discussion to control for confounding in case-control studies. Before discussing this famous paper, however, it is interesting that the same authors had suggested an alternative approach a year earlier (Haenszel et al., 1958).

As in Cochran's analysis, the idea was based on post-stratification of cases and controls into strata which are as homogeneous as possible. Arguing by analogy with the method of indirect standardization of rates, they suggested that the influence of confounding on the odds ratio could be assessed by calculating, for each stratum, s , the "expected" frequencies in the 2×2 table under the assumption of no partial association within strata and calculating the marginal odds ratio under this assumption. The observed marginal odds ratio was then adjusted by this factor. Thus, denoting the expected frequencies by a_i, b_i, c_i and d_i where $a_i = (A_i + B_i)(A_i + C_i)/N_i$ etc., their proposed index was

$$\frac{\sum A_i \sum D_i / \sum B_i \sum C_i}{\sum a_i \sum d_i / \sum b_i \sum c_i}.$$

The use of the stratum-specific expected frequencies in this way can be regarded as an early attempt, in the case-control setting, to estimate what later became known as the "confounding risk ratio" and which we described in Section 4.

In their later paper, Mantel and Haenszel (1959) themselves criticized this adjusted index which, they stated "can be seen to have a bias towards unity" and does "not yield an appropriate adjusted relative risk". (Somewhat unconvincingly they claimed that they had used the index fully realizing its deficiencies "to present results more nearly comparable with those reported by other investigators using similarly biased estimators"!)

These statements were not formally justified and beg the question as to what, precisely, is the estimand? One can only assume that they were referring to the case in which the stratum-specific odds ratios are equal and provide a single estimand. This is the case in which Yule's Q is stable across subgroups. The alternative estimator they proposed:

$$\frac{\sum A_i D_i / N_i}{\sum B_i C_i / N_i}$$

is a consistent estimator of the stratum specific odds ratio in this circumstance. They also proposed a test for association between exposure and disease within strata. The test statistic is the sum, across strata, of the differences between observed and "expected" frequencies in one cell of each table:

$$\begin{aligned} \sum (A_i - a_i) &= \sum A_i - \frac{(A_i + B_i)(A_i + C_i)}{N_i} \\ &= \sum \frac{1}{N_i} (A_i D_i - B_i C_i) \end{aligned}$$

and its variance under the null hypothesis is

$$\sum \frac{(A_i + B_i)(C_i + D_i)(A_i + C_i)(B_i + D_i)}{N_i^2(N_i - 1)}.$$

Some algebra shows that the Mantel-Haenszel test statistic is identical to Cochran's $\sum w_i d_i$. There is a slight difference between the two procedures in that, in calculating the variance, Mantel and Haenszel used a hypergeometric assumption to avoid the need to estimate a nuisance parameter in each stratum in the "two binomials" formulation. This results in the $(N_i - 1)$ term in the above variance formula instead of N_i — a distinction which can become important when there are a large number of sparsely populated strata.

Whereas considerations of bias and, as later shown, optimal properties of their proposed test depend on the assumption of constancy of the odds ratio across strata, Mantel and Haenszel were at pains to disown such a model. They proposed that any standardized, or corrected, summary odds ratio would be some sort of weighted average of the stratum-specific odds ratios and identified that one might choose weights either by *precision* or by *importance*. On the former:

If one could assume that the increased relative risk associated with a factor was constant over all subclassifications, the estimation problem would reduce to weighting the several subclassification estimates according to their relative precisions. The complex maximum likelihood iterative procedure necessary for obtaining such a weighted estimate would seem to be unjustified, since the assumption of a constant relative risk can be discarded as usually untenable.

They described the weighting scheme used in the Mantel-Haenszel estimator as approximately weighting by precision. Indeed, it turns out that these weights correspond to optimal weighting by precision for odds ratios close to 1.0.

An alternative standardized odds ratio estimate, in the spirit of weighting and mirroring direct standardization was proposed by Miettinen (1972a). This is

$$\frac{\sum W_i A_i / B_i}{\sum W_i C_i / D_i}$$

where the weights reflect the population distribution of the stratifying variable. This index can be unstable when strata are sparse, but Greenland (1982) pointed out that it has clear advantages over the Mantel-Haenszel estimate when the odds ratios differ between strata. This follows from our earlier discussion (Section 4) of the distinction between a ratio of averages and an average of ratios. Since the numerator and denominator of the Mantel-Haenszel estimator do not have an interpretation in terms of the population average of a meaningful quantity, the index must be interpreted as an average of ratios, despite its usual algebraic representation. Thus, despite the protestations of Mantel and Haenszel to the contrary, its usefulness depends on approximate stability of the stratum-specific odds ratios. Greenland pointed out that Miettinen's index has an interpretation as a ratio of marginal expectations of epidemiologically meaningful quantities and, therefore, may be useful even when odds ratios are heterogeneous. He went on to propose some improvements to address its instability.

As was noted earlier, there was a widespread belief that controlling for confounding in case-control studies was largely a matter to be dealt with at the design stage, by appropriate "cross-matching" of controls to cases. Mantel and Haenszel, however, pointed out that such matching nevertheless needed to be taken account of in the analysis:

when matching is made on a large number of factors, not even the fiction of a random sampling of control individuals can be maintained.

They showed that the test and estimate they had proposed were still correct in the setting of closely matched studies. Despite this, misconceptions about matching persisted for more than a decade.

6. THE EMERGENCE OF FORMAL MODELS

Except for linear regression analysis for quantitative data, proper statistical models, in the sense we know today, were slow to appear for the purpose of what we now call confounder control.

We begin this section with the early multiplicative intensity age-cohort model for death rates by Kermack et al. (1934a,b), even though it was strangely isolated as a statistical innovation: no one outside of a narrow circle of cohort analysts seems to have quoted it before 1976. First, we must mention two precursors from the actuarial environment.

Actuarial analyses of cohort life tables

Two papers were read to audiences of actuaries on the same evening: 31 January 1927. Derrick (1927), in the Institute of Actuaries in London, studied mortality in England and Wales 1841-1925, omitting the war (and pandemic) years 1915-20. On a clever graph of age-specific mortality (on a logarithmic scale) against year of birth he generalized the parallelism of these curves to a hypothesis that mortality was given by a constant age structure, a decreasing multiplicative generation effect, and no period effect, and even ventured to extrapolate the mortality for existing cohorts into the future.

Davidson and Reid, in the Faculty of Actuaries in Edinburgh, first gave an exposition of estimating mortality rates in a Bayesian framework (posterior mode), including the maximum likelihood estimator interpretation of the empirical mortality obtained from an uninformative prior (Davidson and Reid, 1926–1927). They proceeded to discuss how the *mortality variation force* might possibly depend on age and calendar year and arrived at a discussion on how to predict future mortality, where they remarked (p. 195) that this would be much easier if

there is in existence a law of mortality which, when applied to *consecutive* human life — that is, when applied to trace individuals born in a particular calendar year throughout the rest of their lives — gives satisfactory results

or, as we would say, if the cohort life table could be modelled. Davidson and Reid also explained their idea through a well-chosen, though purely theoretical, graph.

The multiplicative model of Kermack, McKendrick, and McKinley

Kermack, McKendrick and McKinley published an analysis of death-rates in England and Wales since 1845, in Scotland since 1860 and in Sweden since 1751 in two companion papers. In the substantive presentation in *The Lancet* (Kermack et al., 1934a) - republished by *International Journal of Epidemiology* (2001) with discussion of the epidemiological cohort analysis aspects - they observed and discussed a clear pattern in these rates as a product of a factor only depending on age and a factor only depending on year of birth.

The technical elaboration in *Journal of Hygiene* (Kermack et al., 1934b) started from the partial differential equation describing age-time dependent population growth with $\nu_{t,a}da$ denoting the number of persons at time t with age between a and $a + da$, giving the death rate at time t and age a

$$-\frac{1}{\nu_{t,a}} \left(\frac{\partial \nu_{t,a}}{\partial t} + \frac{\partial \nu_{t,a}}{\partial a} \right) = f(t, a)$$

here quoted from McKendrick (1925–26) (cf. Keiding (2011) for comments on the history of this equation) and postulate at once the multiplicative model for

$$f(t, a) = \alpha(t, a)\beta_a.$$

The paper is largely concerned with estimation of the parameters and of the standard errors of these estimates, some attention is also given to the possibility of fitting the age effect β_a to the Gompertz-Makeham distribution.

This fine statistical paper was quoted very little in the following 45 years and thus does not seem to have influenced the further developments of statistical models in the area. One cannot avoid speculating what would have happened if this paper had appeared in a statistical journal rather than in the *Journal of Hygiene*. 1934 was the year when Yule had his major discussion paper on standardization in the Royal Statistical Society. In all fairness, it should on the other hand be emphasized that Kermack et al. did not connect to the then current discussions of general issues of standardization.

The SMR as maximum likelihood estimator

Kilpatrick (1962), in a paper based on his Ph.D. at Queen's University at Belfast, specified for the first time a mortality index as an estimator of a parameter in a well-specified statistical model — that in which the age-specific relative death rate in each age group estimates a constant, and only differs from it by random variation. Kilpatrick's introduction is a good example of a statistical view on standardization, in some ways rather reminiscent of Westergaard:

The mortality experienced by different groups of individuals is best compared, using specific death rates of sub-groups alike in every respect, apart from the single factor by which the total population is divided. This situation is rarely, if ever, realized and we have to be satisfied with mortality comparisons between groups of individuals alike with regard to two, three or four major factors known to affect the risk of death.

In this paper groups are defined as aggregates of occupations (social classes). It is assumed that age is the only factor related to an individual's mortality within a group. This example may readily be extended to other factors such as sex, marital status, residence, etc. Although the association of social class and age-specific mortality may be evaluated by comparisons between social classes, specific death rates of a social class are more frequently compared with the corresponding rates of the total population. It is this type of comparison which is considered here.

Kilpatrick then narrowed the focus to developing an index I_{us} which

should represent the "average" excess or deficit mortality in group u compared with the standard s

and noted that, with θ_x representing the ratio between the mortality rates in age group x in the study group and the total population,

Recent authors ... have shown that the SMR can be misleading if there is much variation in θ_x over the age range considered. It would, therefore, seem desirable to test the significance of this variation in θ_x before placing confidence on the results of the SMR or any other index. ... This paper proposes a simple test for heterogeneity in θ_x and shows that the SMR is equivalent to the maximum likelihood estimate of a common θ when the θ_x do not differ significantly. It follows therefore that the SMR has a minimum standard error.

Formally, Kilpatrick assumed the observed age-specific rates in the study group to follow Poisson distributions with rate parameters $\theta\lambda_i$. The λ_i 's and the denominators, A_i , were treated as deterministic constants, and the mortality ratio, θ as a parameter to be estimated.

We note that the view of standardization as an estimation problem in a well-specified statistical model was principally different from earlier authors. One could refer to the paper by Kermack et al. (1934b) discussed above (which specifies a similar model), but they did not explicitly see their model as being related to

standardization; their paper has been quoted rarely and it seems that Kilpatrick was unaware of it.

Once standardization is formulated as an estimation problem, the obvious question is to find an *optimal* estimator, and Kilpatrick showed that the standardized mortality ratio (SMR)

$$\hat{\theta} = \frac{\text{Observed number of deaths in the study population}}{\text{Expected number of deaths in the study population}}$$

has minimum variance among all indices, and that it is the maximum likelihood estimator in the model specified by deterministic standard age-specific death rates and a constant age-specific rate ratio.

Kilpatrick noted that while the SMR is, in a sense, optimal for comparing one study group to a standard, the weights change from one study group to the next so that it cannot be directly used for comparing several groups. As we have seen, this point had been made often before, particularly forcefully by Yule (1934). Kilpatrick compared the SMR to the comparative mortality index (CMF) obtained from direct standardization, and to Yule's index (the ratio of "equivalent death rates", i.e. direct standardization using equally large age groups). He concluded

Where appropriate and possible, a test of heterogeneity on age-specific mortality ratios should precede the use of an index. When there is insufficient information to conduct the test of heterogeneity, conclusions based solely on the index value may apply to none of the individuals studied. Caution is strongly urged in the interpretation of mortality indices.

Kalton — statistical view of standardization in survey research

Kalton (1968) surveyed, from a rather mainstream statistical view, standardization as a technique for estimating the contrast between two groups and to test the hypothesis that this contrast vanishes. Kalton emphasized that

... if the estimate is to be meaningful, there must be virtually no 'interaction' effect in the variable under study between the groups and the control variable (that is, there must be a constant difference in the group means of the variable under study for all levels of the control variable), but this requirement may be somewhat relaxed for the significance test.

This distinction implies that the optimal weights are not the same for the estimation problem and the testing problem. Without commenting on the causal structure of 'elaboration' Kalton (1968) also gave further insightful technical statistical comments to Rosenberg's example (see above) and the use of optimum weights for testing no effect of religious group.

Kalton seems to have been unaware of Kilpatrick's paper six years earlier, but took a similar mainstream statistical view of standardization: that presentation of a single summary measure of the within-stratum effect of the study variable implies a model of no interaction between stratum and study variable.

Indirect standardization without external standard

Kilpatrick had opened the way to a fully model-based analysis of rates in lieu of indirect standardization, and authoritative surveys based on this approach were indeed published by Holford (1980), Hobcraft et al. (1982), Breslow et al. (1983), Borgan (1984), Hoem (1987). Still, modified versions of the old technique of indirect standardization remained part of the tool kit for many years.

An interesting example is the attempt by Mantel and Stark (1968) to standardize the incidence of mongolism for both birth order and maternal age. Standardized for one of these factors, the incidence still increased as function of the other, but the authors felt it

desirable to obtain some simple descriptive statistics by which the reader could judge for himself what the data showed. . . . What was required was that we determine simultaneously a set of birth-order category rates which when used as a standard set gave a set of indirect-adjusted maternal-age category rates which in turn, when used as a standard set, implied the original set of birth-order category rates.

The authors achieved that through an iterative procedure, which always converged to “indirect, unconfounded” adjusted rates, where the convergent solutions varied with the initial set of standard rates, although they all preserved the *ratios* of the various birth-order category-adjusted rates and the ratios of the various maternal-age category-adjusted rates. Osborn (1975) and Breslow and Day (1975) formulated multiplicative models for the rates and used the same iterative indirect standardization algorithm for the parameters. Generalizing Kilpatrick’s model to multiple study groups, the age-specific rate in age group i and study group j is assumed to be $\theta_j \lambda_i$. Treating λ_i ’s as known, the θ_j ’s can be estimated by SMRs; the θ_j ’s can then be treated as known and the λ_i ’s estimated by SMRs (although the indeterminacy identified by Mantel and Stark must be resolved, for example by normalization of one set of parameters). See Holford (1980) for the relation of this algorithm to iterative proportional fitting of log-linear models in contingency tables. Neither Mantel and Stark, Osborn, nor Breslow and Day cited Kilpatrick or Kermack, McKendrick and McKinlay.

Logistic models for tables of proportions

We have seen that Cochran (1954) had suggested that analysis of the comparison of two groups with respect to a binary response in the presence of a confounding factor (an analysis of a $2 \times 2 \times K$ contingency table), could be approached by fitting formal models to the $2 \times K$ table of proportions, using a transformation such as the logit or probit transformation. But such analyses, given computational resources available at that time, were extremely laborious. Cochran cited the pioneering work of Dyke and Patterson (1952) who developed a method for fitting the logit regression model to fitted probabilities of response, $\pi_{ijk\dots}$, in a table :

$$\log \frac{\pi_{ijk\dots}}{1 - \pi_{ijk\dots}} = \mu + \alpha_i + \beta_j + \gamma_k + \dots$$

by maximum likelihood, illustrating this technique with an analysis estimating the independent contributions of newspapers, radio, “solid” reading, and lectures upon knowledge of cancer. Initially they applied an empirical logit transformation to the observed proportions, $p_{ijk\dots}$, and fitted a linear model by weighted least squares. They then developed an algorithm to refine this solution to the true maximum likelihood, an algorithm which was later generalized by Nelder and Wedderburn (1972) to the wider class of generalized linear models — the now familiar iteratively reweighted least squares (IRLS) algorithm. Since, in their example, the initial fit to the empirical data provided a good approximation to the maximum likelihood solution, only one or two steps of the IRLS algorithm were necessary — perhaps fortunate since the calculations were performed without recourse to a computer.

Although, in its title, Dyke and Patterson referred to their paper as concerning “factorial arrangements”, they explicitly drew attention to its uses in dealing with confounding in observational studies:

It is important to realise that with this type of data there are likely to be a number of factors which may influence our estimate of the effect of say, solid reading but which have not been taken into account. The point does not arise in the case of well conducted experiments but is common in survey work.

Log-linear models and the National Halothane Study

Systematic theoretical studies of multiple cross-classifications of discrete data date back at least to Yule (1900), quoted above. For three-way tables, Bartlett (1935) discussed estimation and hypothesis testing regarding the second-order interaction, forcefully followed up by Birch (1963) in his study of maximum likelihood estimation in the three-way table.

However, as will be exemplified below in the context of The National Halothane Study, the real practical development in the analysis of large contingency tables needed large computers for the necessary calculations. This development largely happened around 1970 (with many contributions from L.A. Goodman in addition to those already mentioned), and the dominating method was straightforward maximum likelihood. Particularly influential were the dissertation by Haberman (1974), which also included important software, and the authoritative monograph by Bishop et al. (1975).

The National Halothane Study Halothane is an anaesthetic which around 1960 was suspected in the U.S. for causing increased rates of hepatic necrosis, sometimes fatal. A subcommittee under the U.S. National Academy of Sciences recommended that a large cooperative study be performed, and this was started in July 1963. We shall here focus on the study of “surgical deaths”, i.e. deaths during the first 6 weeks after surgery. The study was based on retrospective information from 34 participating medical centres, who reported all surgical deaths during the four years 1959-62 as well as provided information on a random sample of about 38,000 from the total of about 856,000 operations at these centres during the four years. The study was designed and analyzed in a collaborative effort between leading biostatisticians at Stanford University, Harvard University and Princeton University/Bell Labs and the report (Bunker et al., 1969) is unusually rich in explicit discussions about how to handle the adjustment problem with the many variables registered for the patients and the corresponding “thin” cross-classifications. For a very detailed and informative review, see Stone (1970). The main problem in the statistical analysis was whether the different anesthetics were associated with different death rates, after adjusting for a range of possible confounders, as we would say today. In a still very readable introduction by B.W. Brown et al. it was emphasized (p. 185) that

the analysis of rates and counts associated with many background variables is a recurring and very awkward problem. . . It is appropriate to create new methods for handling this nearly universal problem at just this time. High-speed computers and experience with them have now developed to such a stage that we can afford to execute extensive manipulations repeatedly on large bodies of data with many control variables, whereas previously such heavy arithmetic work was impossible. The presence of the large sample from the National Halothane Study has encouraged the investigation and development of flexible methods of adjusting for several background variables. Although this adjustment problem is not totally solved by the work in this Study, substantial advances have been made and directions for further profitable research are clearly marked.

The authors here go on to emphasize that the need for adjustment is not restricted to “nonrandomized” studies.

Pure or complete randomization does not produce either equal or conveniently proportional numbers of patients in each class; attempts at deep post-stratification are doomed to failure because for several variables the number of possible strata quickly climbs beyond the thousands. . . . Insofar as we want rates for special groups, we need some method of estimation that borrows strength from the general pattern of the variables. Such a method is likely to be similar, at least in spirit, to some of those that were developed and applied in this Study. At some stage in nearly every large-scale, randomized field study (a large, randomized prospective study of postoperative deaths would be no exception), the question arises whether the randomization has been executed according to plan. Inevitably, adjustments are required to see what the effects of the possible failure of the randomization might be. Again, the desired adjustments would ordinarily be among the sorts that we discuss.

The National Halothane Study has perhaps become particularly famous among statisticians for the early multi-way contingency table analyses done by Yvonne M.M. Bishop supervised by F. Mosteller. This approach is here termed “smoothed contingency-table analysis”, reflecting the above mentioned recognized need for the analysis to “borrow strength from the general pattern”. Bishop did her Ph.D. thesis in this area cf. the journal publications (Bishop, 1969, 1971) and combined efforts with S.E. Fienberg and P. Holland in their very influential monograph on “Discrete Multivariate Analysis” (Bishop et al., 1975). But the various versions of data-analytic (i.e. model-free) generalizations of standardization are also of interest, at least as showing how broadly these statisticians struggled with their task: to adjust discrete data for many covariates in the computer age.

The analysis began with classical standardization techniques (L. Moses), which were soon overwhelmed by the difficulty in adjusting for more than one variable at a time. Most of the subsequent approaches use a rather special form of stratification where the huge, sparse multidimensional contingency table generated by cross-classification of covariates other than the primary exposure variables (the anesthetic agents) are aggregated to yield “strata” with homogeneous death rates, the agents subsequently compared by standardizing across these strata. Several detailed techniques were developed for this purpose by J.W. Tukey and colleagues, elaborately documented in the report and briefly quoted by Tukey (1979, 1991); however, criticisms were also raised (Stone, 1970; Scott, 1978) and the ideas do not seem to have caught on.

Clogg’s “purging” of contingency tables

Clifford Clogg was a Ph.D. student of Hauser, Goodman and Kitagawa at the University of Chicago, writing his dissertation in 1977 on Hauser’s theme of using a broader measure of *underemployment* (as opposed to just *unemployment*) as social indicator, in the climate of Goodman’s massive recent efforts on loglinear modelling and Kitagawa’s strong tradition in standardization. We shall briefly present Clogg’s attempts at combining the latter two worlds in the *purging* techniques (Clogg (1978), Clogg and Eliason (1988) and many other articles). A useful concise summary was provided by Sobel (1996, pp.11–14) in his tribute to Clogg after Clogg’s early death, and a recent important discussion and generalization was given by Yamaguchi (2011).

Clogg considered a *composition* variable C with categories $i = 1, \dots, I$, a *group* variable G with categories $j = 1, \dots, J$, and a *dependent* variable D with categories $k = 1, \dots, K$. The composition variable may itself have been generated by cross-classification of several composition variables. The object is to assess

the possible association of D with G adjusted for differences in the compositions across the groups. Clogg assumed that the three-way $C \times G \times D$ classification has already been modelled by a loglinear model, and the purging technique was primarily promoted as a tool for increased accessibility of the results of that analysis. Most of the time the saturated model is assumed, although in our view the purging idea is much easier to assimilate when there is no three-factor interaction.

A brief version of Clogg's explanation is as follows. The $I \times J \times K$ table is modelled by the saturated log-linear model

$$\pi_{ijk} = \eta \tau_i^C \tau_j^G \tau_k^D \tau_{ij}^{CG} \tau_{ik}^{CD} \tau_{jk}^{GD} \tau_{ijk}^{CGD}$$

where the disturbing interaction is τ_{ij}^{CG} ; the composition-specific rate

$$r_{ij(k)} = \pi_{ijk} / \sum_k \pi_{ijk} = \pi_{ijk} / \pi_{ij}$$

is independent of τ_{ij}^{CG} , but the overall rate of occurrence

$$r_{\cdot j(k)} = \sum_i \pi_{ijk} / \sum_{i,k} \pi_{ijk} = \pi_{\cdot jk} / \pi_{\cdot j}$$

does depend on τ_{ij}^{CG} .

Now *purge* π_{ijk} of the cumbersome interaction by defining purged proportions proportional to

$$\pi_{ijk}^{**} = \pi_{ijk} / \tau_{ij}^{CG} \quad (\text{i.e. } \pi_{ijk}^* = \pi_{ijk}^{**} / \pi_{ij}^{**}).$$

Actually

$$\pi_{ijk}^* = \eta^* \tau_i^C \tau_j^G \tau_k^D \tau_{ik}^{CD} \tau_{jk}^{GD} \tau_{ijk}^{CGD}, \quad \eta^* = \eta / \pi_{ij}^{**}$$

i.e. the π_{ijk}^* specify a model with all the same parameters as before except that τ_{ij}^{CG} has been replaced by 1.

Rates calculated from these adjusted proportions are now purged of the $C \times G$ interaction but all other parameters are as before. Clogg noted the fact that this procedure is not the same as direct standardization and defined a variant, *marginal CG-purging*, which is equivalent to direct standardization.

Purging was combined with further developments of additive decomposition methods by Xie (1989) and Liao (1989) and was still mentioned in the textbook by Powers and Xie (2008, Section 4.6), but seems otherwise to have played a modest part in recent decades. A very interesting recent application is by Yamaguchi (2011), who used purging in counterfactual modeling of the mediation of the salary gap between Japanese males and females by factors such as differential educational attainment, use of part-time jobs, and occupational segregation.

Multiple regression in epidemiology

By the early 1960's epidemiologists, in particular cardiovascular epidemiologists, were wrestling with the problem of *multiple causes*. It was clear that methods based solely on cross-classification would have limited usefulness. As put by Truett et al. (1967):

Thus, if 10 variables are under consideration, and each variable is to be studied at only three levels, ... there would be 59,049 cells in the multiple cross-classification.

Cornfield (1962) suggested the use of Fisher’s discriminant analysis to deal with such problems. Although initially he considered only two variables, he set out the idea more generally. This model assumes that the vector of risk factor values is distributed, in (incident) cases of a disease and in subjects who remain disease free, as multivariate normal variates with different means but equal variance–covariance matrices. Reversing the conditioning by Bayes theorem shows that the probability of disease given risk factors is then given by the multiple logistic function. The idea was investigated in more detail and for more risk factors by Truett et al. (1967) using data from the 12–year follow-up of subjects in the Framingham study. A clear concern was that the multivariate normal assumption was clearly wrong in the situations they were considering, which involved a mixture of continuous and discrete risk factors. Despite this they demonstrated that there was good correspondence between observed and expected risks when subjects were classified according to deciles of the discriminant function.

Truett et al. discussed the interpretation of the regression coefficients, at some length, but did not remark on the connection with multiplicative models and odds ratios, although Cornfield had, 15 years previously, established the approximate equivalence between the odds ratio and a ratio of rates (see Section 5). They did note that the model is *not* additive:

The relation between logit of risk and risk is illustrated in Fig. 1 . . . a constant increase in the logit of risk does not imply a constant increase in risk.

and preferred to present the coefficients of the multiple logistic function as multiples of the standard deviation of the corresponding variable. They did, however, make it clear that these coefficients represented an estimate of the effect of each risk factor *after holding all others constant*. They singled out the effect of weight in this discussion:

The relative unimportance of weight as a risk factor . . . when all other risk factors are simultaneously considered is noteworthy. This is not inconsistent with the possibility that a reduction in weight would, by virtue of its effect on other risk factors, e.g. cholesterol, have important effects on the risk of CHD.

Finally they noted that the model assumes the effect of each risk factor to be independent of the levels of other risk factors, and noted that first order interactions could be studied by relaxing the assumption of equality of the variance-covariance matrices.

The avoidance of the assumption of multivariate normality in the logistic model was achieved by use of the method of maximum likelihood. In the epidemiological literature, this is usually credited to Walker and Duncan (1967) who used a likelihood based on conditioning on the values, x , of the risk factors, and computing maximum likelihood estimates using the same iteratively reweighted least squares algorithm proposed by Dyke and Patterson (1952). However, use of maximum likelihood in such models had also been anticipated by Cox (1958) although he had advocated conditioning both on the observed set of risk factors, x , and on the observed values of the disease status indicators, y . This is the method, now known as “conditional” logistic regression, which is important in the analysis of closely matched case–control studies. Like Truett et al., Walker and Duncan gave little attention to interpretation of the regression coefficients, save for advocating standardization to SD units in an attempt to demonstrate the relative importance of different factors. The main focus seems to have been in risk prediction given multiple risk factors. Cox (1958), however, discussed the interpretation of the regression coefficient of a dichotomous variable as a log odds ratio, even applying

this to an example, cited by Cornfield (1956), concerning smoking and lung cancer in a survey of physicians.

A limitation of logistic regression for the analysis of follow-up studies is the necessity to consider, as did Truett et al. (1967), a fixed period of follow-up. A further rationalization of analytical methods in epidemiology followed from the realization that such studies generate right-censored, and left-truncated, *survival data*. Mantel (1966) pioneered the modern approach to such problems, noting that such data can be treated as if each subject undergoes a series of Bernoulli trials (of very short duration). He suggested, therefore, that the comparison of survival between two groups could be treated as an analysis of a $2 \times 2 \times K$ table in which the K “trials” are defined by the time points at which deaths occurred in the study (other time points being uninformative). In his famous paper, Cox (1972), described a regression generalization of this idea, in which the instantaneous risk, or “hazard”, is predicted by a log-linear regression model so that effects of each risk factor may be expressed as hazard ratios. Over subsequent decades this theory was further extended to encompass many types of event history data. See Andersen et al. (1993) for a comprehensive review.

Confounder scores and propensity scores

Miettinen (1976b) put forward an alternative proposal for dealing with multiple confounders. It was motivated by three shortcomings he identified in the multivariate methods then available:

1. they (discriminant analysis in particular) relied on very dubious assumptions,
2. they (logistic regression) were computationally demanding by the standards then applying, and
3. they were poorly understood by substantive scientists.

His proposal was to carry out a preliminary, perhaps crude, multivariate analysis from which could be computed a “confounder score”. This score could then be treated as a single confounder and dealt with by conventional stratification methods. He suggested two ways of computing the confounder score. An *outcome function* was computed by an initial regression (or discriminant function) analysis of the disease outcome variable on all of the confounders plus the exposure variable of interest, then calculating the score for a fixed value of exposure so that it depended solely on confounders. Alternatively, an *exposure function* could be computed by interchanging the roles of outcome and exposure variables, regressing exposure on confounders plus outcome.

Rosenbaum and Rubin (1983) later put forward a superficially similar proposal to the use of Miettinen’s exposure function. By analogy with randomized experiments they defined a *balancing score* as a function of potential confounders such that exposure and confounders are conditionally independent given the balancing score. Stratification by such a score would then simulate a randomized experiment within each stratum. They further demonstrated, for a binary exposure, that the coarsest possible balancing score is the *propensity score*, the probability of exposure conditional upon confounders, which can be estimated by logistic regression. Note that, unlike, Miettinen’s exposure score, the outcome variable is not included in this regression. The impact of estimation of the nuisance parameters of the propensity score upon the test of exposure effect was later explored

by Rosenbaum (1984). Hansen (2008) later showed that a balancing score is also provided by the “prognostic analogue” to the propensity score which is to Miettinen’s outcome function as the propensity score is to his exposure function i.e. the exposure variable is omitted when calculating the prognostic score.

Given this later work on balancing scores, it is interesting to note that Miettinen discussed at some length why he believed it necessary to include the “conditioning variable” (either the exposure of interest or the outcome variable) when computing the coefficients of the confounder score, noting that the need for this was “puzzling to some epidemiologists”. His argument comes down to the requirement to obtain an (approximately) unbiased estimate of the conditional odds ratio for exposure versus outcome; omission of the conditioning variable means that the confounder score potentially contains a component related to only one of the two variables of interest and, owing to non-collapsibility of the odds ratio, this leads to a biased estimate of the conditional effect. Unfortunately, as demonstrated by Pike et al. (1979), Miettinen’s proposal for correcting this bias comes at the cost of inflation of the type 1 error rate for the hypothesis test for an exposure effect. To demonstrate this, consider a logistic regression of an outcome, y , on an exposure of interest, x , and multiple confounders, z . Miettinen proposed to first compute a confounder score $s = \hat{\gamma}^T z$, where $\hat{\gamma}$ are the coefficients of z in the logistic regression of x on y and z , and then to fit the logistic regression of y on x and s . While this regression yields an identical coefficient for x as the full logistic regression of y on x and z , and has the same maximized likelihood, in the test for exposure effect this likelihood is compared with the likelihood for the regression of y on s which, in general, will be rather less than that for y on z — the correct comparison point. Thus, the likelihood ratio test in Miettinen’s procedure will be inflated.

Rosenbaum and Rubin circumvented the estimation problem posed by omission of the conditioning variable when calculating balancing scores, by estimating a *marginal* causal effect using direct standardization with appropriate population weights. Equivalently, inverse probability weights based on the propensity score can be used.

Owing to the focus on conditional measures of effect, the propensity score approach was little used in epidemiology during the latter part of the 20th century. However, the method has gained considerably in popularity over the last decade. For a recent case study of treatment effect estimation using propensity score and regression methods, see Kurth et al. (2006). They emphasised that, as in classical direct standardization, precise identification of the target population is important when treatment effects are non-uniform.

Time-dependent confounding

Cox’s life table regression model provided an exceedingly general approach to modelling the probability of a failure event conditional upon exposure or treatment variables and upon extraneous covariates or confounders, the mathematical development extending quite naturally to allow for variation of such variables over time. However, shortly after its publication, Kalbfleisch and Prentice (1980, p. 124–126) pointed out a serious difficulty in dealing with “internal” (endogenous) time-dependent covariates. Referring to the role of variables such as the general condition of patients in therapeutic trials which may lie on the causal path between earlier treatment and later outcome and, therefore, carry part of the causal treatment effect, they wrote

A censoring scheme that depends on the level of a time dependent covariate $z(t)$ (e.g. general condition) is . . . not independent if $z(t)$ is not included in the model. One way to circumvent this is to include $z(t)$ in the model, but this may mask treatment differences of interest.

Put another way, to ignore such a variable in the analysis is to disregard its confounding effect, but its inclusion in the conditional probability model could obscure some of the true causal effect of treatment.

While Kalbfleisch and Prentice had identified a fundamental problem with the conditional approach to confounder adjustment, they offered no convincing remedy. This was left to Robins (1986). In this and later papers Robins addressed the “healthy worker” effect in epidemiology — essentially the same problem identified by Kalbfleisch and Prentice. Robins proposed two lines of attack which we may classify as ‘marginal’ and ‘conditional’ in keeping with a distinction that have come up throughout our exposition. The original approach was ‘g-computation’ which may be loosely conceived as sequential prediction of ‘what would have happened’ under various specified externally imposed ‘treatments’ and thus generalizes (direct) standardization, basically a marginal approach. A further development in this direction was inverse probability weighting of marginal structural models, i.e. models for the counterfactual outcomes (Robins et al., 2000). Here, the essential idea is to estimate a marginal treatment effect in a population in which the association between treatment (exposure) and the time-dependent covariate is removed. Sato and Matsuyama (2003) and Vansteelandt and Keiding (2011) gave brief discussions of the relationship between g-computation, inverse probability weighting and classical standardization in the simplest (non-longitudinal) situation. On the other hand the approach via structural nested models (e.g. Robins and Tsiatis (1992), Robins et al. (1992)) focuses on the effect of a ‘blip’ of exposure at time t conditional on treatments and covariate values before t . In the latter models, time-varying effect modification may be studied. See the recent tutorial surveys by Robins and Hernán (2009) or Daniel et al. (2013) for details.

7. PREDICTION AND TRANSPORTABILITY

We saw that in the National Halothane Study standardization methods were used analytically, in order to control for confounders strictly within the frame of the concrete study. The general verdict in the emerging computer age regarding this use of standardization was negative, as formulated by Fienberg (1975), in a discussion of a careful and detailed survey on observational studies by McKinlay (1975):

The reader should be aware that standardization is basically a descriptive technique that has been made obsolete, for most of the purposes to which it has traditionally been put, by the ready availability of computer programs for loglinear model analysis of multidimensional contingency tables.

However, the original use of standardization not only had this analytical ambition, it also aimed at obtaining meaningful generalizations to other populations — or other circumstances in the original population. Before we sketch the recovery since the early 1980s of this aspect of standardization, it is useful to record the attitude to generalization by influential epidemiologists back then. Miettinen (1985, p. 47), in his long-awaited text-book, wrote

In science the generalization from the actual study experience is not made to a population of which the study experience is a sample in a technical sense of probability sampling . . . In science the generalization is from the actual study experience to the abstract, with no referent in place or time

and thus did not focus on specific recommendations as to how to predict precisely what might happen under different concrete circumstances. A similar attitude was voiced by Rothman (1986, p. 95), in the first edition of *Modern Epidemiology* and essentially repeated in the following editions of this central reference work (Rothman and Greenland (1998, pp. 133–134), Rothman et al. (2008, pp. 146–147)). The immediate consequence of this attitude would be that all that we need are the parameters in the fitted statistical model and assurance that no bias is present in the genesis of the concretely analyzed data.

However, as we have seen, Clogg (1978) (and later) had felt a need for interpreting the log-linear models in terms of their consequences for summary tables. Freeman and Holford (1980) wrote a useful guide to the new situation for survey analysis where the collected data had been analyzed using the new statistical models. They concluded that much favoured keeping the reporting to the model parameters: these would then be available to other analysts for comparative purposes, the model fit was necessary to check for interactions (including possibly identifying a model where there is no interaction). But

in many settings these advantages are overshadowed by the dual requirements for simplicity of presentation and immediacy of interpretation

and Freeman and Holford (1980) therefore gave specific instructions on how to calculate “summary rates” for the total population or other populations. The main requirement for validity of such calculations is that there is no interaction between population and composition.

Interestingly, an influential contribution in 1982 came from a rather different research environment: the well-established agricultural statisticians P.W. Lane and J.A. Nelder (Lane and Nelder, 1982). In a special issue of *Biometrics* on the theme “the analysis of covariance”, they wrote a short note containing several germs of the later so important potential outcome view underlying modern causal inference, and placed the good old (direct) standardization technique right in the middle of it.

Their view was that the purpose of a statistical analysis such as analysis of covariance is not only to estimate parameters, but also to make what they called *predictions*.

An essential feature is the division into effects of interest and effects for which adjustment is required. . . . For example, a typical prediction from a variety trial is the yield that would have been obtained from a particular variety if it had been grown over the whole experimental area. When a covariate exists the adjusted treatment mean can be thought of as the prediction of the yield of that variety grown over the whole experimental area with the covariate fixed at its mean value. . . . The predictions here are not of future events but rather of what would have happened in the experiment if other conditions had prevailed. In fact no variety would have been grown over the whole experimental area nor would the covariate have been constant.

Lane and Nelder proposed to use the term *predictive margin* for such means, avoiding the term “population treatment mean” suggested by Searle et al. (1980) to replace the cryptic SAS-output term “least square means”. Lane and Nelder emphasized that these means might either be

conditional on the value we take as standard for the covariate

or

marginal to the observed distribution of covariate values

and Lane and Nelder went on to explain to this new audience (including agricultural statisticians) that there exist many other possibilities for choice of standard.

We find it interesting that Lane and Nelder used the occasion of the special issue of *Biometrics* on analysis of covariance to point out the *similarities* to standardization, and to phrase their “prediction” in much similar terms as the later causal analysis would do. Of course it should be remembered that Lane and Nelder manoeuvred within the comfortable framework of randomized field trials. Rothman et al. (2008, pp. 386 ff.), described how these ideas have developed into what is now termed *regression standardization*.

An example: cancer trends

A severe practical limitation of the modelling approach is that the model must encompass all the data to be compared. However, many official statistics are published explicitly to allow comparisons with other published series. Even within a single publication it may be inappropriate to fit a single large and complex model across the entire dataset.

An example of the latter situation is the I.A.R.C. monograph on Trends in Cancer Incidence and Mortality (Coleman et al., 1993). The primary aim of this monograph was to estimate cancer trends across the population-based cancer registries throughout the world and this was addressed by fitting age-period-cohort models to the data from each registry. But comparisons of rates between registries at specific time points were also required and, since the age structures of different registries differed markedly, direct standardization to the world population, ages 30–74 was used. However, some of the cancers considered were rare and this exposes a problem with the method of direct standardization — that it can be very inefficient when the standard population differs markedly from that of the test group. The authors therefore chose to apply direct standardization to the *fitted* rates from the age-period-cohort models.

Transportability across studies

Pearl and Barenboim (2012) noted that precise conditions for applying concrete results obtained in a study environment to another “target” environment

remarkably... have not received systematic formal treatment... The standard literature on this topic... consists primarily of “threats”, namely verbal narratives of what can go wrong when we try to transport results from one study to another... Rarely do we find an analysis of “licensing assumptions”, namely, formal and transparent conditions under which the transport of results across differing environments or populations is licensed from first principles.

After further outlining the strong odds against anyone who dares formulate such conditions, Pearl and Barenboim then set out to propose one such formalism, based on the causal diagrams developed by Pearl and colleagues over the last decades, cf. Pearl (2009).

In the terminology of Pearl and Barenboim, the method of direct standardization, together with the “predictions” of Lane and Nelder, is a *transport formula* and, as they state,

the choice of the proper transport formula depends on the causal context in which population differences are embedded.

Although a formal treatment of these issues is overdue, it has been recognized in epidemiology for many years that the concept of confounding cannot be defined solely in terms of a third variable being related to both outcome and exposure of interest. A landmark paper was that of Simpson (1951) which dealt with the

problem of interpreting associations in three-way contingency tables. As we saw in Section 3, although “Simpson’s paradox” is widely regarded as synonymous with Yule’s paradox, Simpson’s primary concern was the role of the causal context in deciding whether the conditional or marginal association between two of the three factors in a table is of primary interest. The point has been understood by many (if not all) epidemiologists writing in the second half of the 20th century as, for example, is demonstrated by the remark of Truett et al. (1967), cited in Section 6, concerning interpretation of the coefficient of body weight in their regression equation for coronary disease incidence. However, as far as we can tell the issue does not seem to have concerned 19th century writers; for example, no consideration seems to have been given to the possibility that age differences between populations could, in part, be a consequence of differences in “the force of mortality” and, if so, the implication for age standardization.

8. CONCLUSION

In the fields of scientific enquiry with which we are concerned here, the causal effect of a treatment, or exposure, cannot be observed at the individual level. Instead, the effect measures we use contrast the distributions of responses in populations with differing exposures, but in which the distributions of other factors do not differ. In randomized studies, this equality of distribution of extraneous factors is guaranteed by randomization and causal effects are simply measured. In observational studies, however, differences between the distributions of relevant extraneous factors between exposure groups (what epidemiologists call ‘confounding’) is ubiquitous and we must rely on the assumption of ‘no unmeasured confounders’ to allow us to estimate the causal effect.

In much recent work, the problem is approached by postulating that each individual has a number of potential responses, one for each possible exposure; only one of these is observed, the other counterfactual responses being assumed to be ‘missing at random’ given measured confounders. Alternatively we can restrict ourselves to dealing with observed outcomes, assuming that the mechanism by which exposure was allocated in the experiment of nature we have observed did not depend on unmeasured confounders. The choice between these positions is philosophical and, to the applied statistician, largely a matter of convenience. The more serious concern, with which we have been largely concerned in this review, is the choice of effect measure; we can choose to contrast the marginal distribution of responses under equality of distribution of extraneous factors, or to contrast response distributions which condition on the values of these factors.

Standardization grew up in response to obvious problems of age-confounding in actuarial (18th century) and demographic (19th century) comparative studies of mortality. The simple intuitive calculations considered scenarios in which either the age distributions did not differ (‘direct’ standardization) or age-specific rates did not differ (‘indirect’ standardization) between study groups. However, formal consideration of such indices as effect measures came later, the contribution of Yule (1934) being noteworthy.

There would probably be widespread agreement that describing causal effects in relation to *all* potential causes, as in the conditional approach, must represent the most complete analysis of a dataset and we have described how this approach developed throughout the 20th century, starting with the influential paper of Yule

(1900). This impressive paper clearly described the proliferation of ‘partial’ association measures introduced by the conditional approach, and drew attention to the consequent need to use measures which remain relatively stable across subgroups. In later work Yule (1934) revisited classical standardization in terms of an average of conditional (i.e. stratum-specific) measures. Such early work sowed the seeds of the ‘statistical’ approach based on formal probability models, leading eventually to the widespread use of logistic regression and proportional hazard (multiplicative intensity) models, the contributions of Cochran (1954) and Mantel and Haenszel (1959) providing important staging posts along the way (even though the latter authors explicitly denied any reliance on a model). By the end of the century such approaches dominated epidemiology and biostatistics.

Towards the end of the 20th century, the use of marginal measures of causal effects emerged from the counterfactual approach to causal analysis in the social sciences, the idea of propensity scores (Rosenbaum and Rubin, 1983) being particularly influential. However, these methods only found their way into mainstream biostatistics when applications arose for which conventional conditional probability models were not well-suited, the foremost of which being the problem of time-dependent confounding. Whereas, for simple problems, the parameters of logistic and multiplicative intensity models have an interpretation as measures of (conditional) causal effects, Kalbfleisch and Prentice (1980) noted that this ceased to be the case in the presence of time-dependent confounding. While the statistical modelling approach can be extended by joint modelling of the event history and confounder trajectories, such models are complex and, again, causal effects do not correspond with model parameters and would need to be estimated by simulations based on the fitted model. Such models continue to be studied (see, for example, Henderson et al. (2000)) but, although it could be argued that these offer the opportunity for a more detailed understanding of the nature of causal effects in this setting, the simpler marginal approaches pioneered by Robins (1986) are more attractive in most applications. The success of this latter approach in offering a solution to a previously intractable problem encouraged biostatisticians to further explore methods for estimation of marginal causal effects; for example propensity scores are now widely used in this literature.

So where are we today? Both approaches have strengths and weaknesses. The conditional modelling approach relies on the assumption of homogeneity of effect across subgroups or, when this fails to hold, to a multiplicity of effect measures. The marginal approach, while seemingly less reliant on such assumptions, encounters the same issue when considering the transportability of effect measures to different populations.

REFERENCES

- AALEN, O. (1978). Non-parametric inference for a family of counting processes. *Ann. Stat.*, **6** 701–726.
- ALDRICH, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Stat. Sci.*, **10** 364–376.
- ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- BARTLETT, M. S. (1935). Contingency table interactions. *Suppl. J. Roy. Stat. Soc.*, **2** 248–252.
- BELLHOUSE, D. (2008). Review of “Disciplining Statistics: Demography and Vital Statistics in France and England, 1830-1885” by L. Schweber. *Hist. Math.*, **35** 249–252.

- BELSON, W. A. (1956). A technique for studying the effects of a television broadcast. *Appl. Stat.*, **5** 195–202.
- BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Stat. Soc. Ser. B*, **25** 220–233.
- BIRCH, M. W. (1964). The detection of partial association, I: the 2×2 case. *J. Roy. Stat. Soc. Ser. B*, **26** 313–324.
- BISHOP, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, **25** 383–399.
- BISHOP, Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics*, **27** 545–562.
- BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.
- BORGAN, O. (1984). Maximum-likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Stat.*, **11** 1–16.
- BRESLOW, N. E. and DAY, N. E. (1975). Indirect standardization and multiplicative models for rates, with reference to age adjustment of cancer incidence and relative frequency data. *J. Chron. Dis.*, **28** 289–303.
- BRESLOW, N. E., LUBIN, J. H., MAREK, P. and LANGHOLZ, B. (1983). Multiplicative models and cohort analysis. *J. Am. Stat. Ass.*, **78** 1–12.
- BUNKER, J., FORREST JR., W., MOSTELLER, F. and VANDAM, L. (eds.) (1969). *The National Halothane Study*. National Institute of General Medical Sciences.
- CHADWICK, E. (1844). On the best modes of representing accurately, by statistical returns, the duration of life, and the pressure and progress of the causes of mortality amongst different classes of the community, and amongst the populations of different districts and countries. *J. Stat. Soc. London*, **7** 1–40.
- CHEVAN, A. and SUTHERLAND, M. (2009). Revisiting Das Gupta: Refinement and extension of standardization and decomposition. *Demography*, **46** 429–449.
- CLAYTON, D. and HILLS, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.
- CLOGG, C. C. (1978). Adjustment of rates using multiplicative models. *Demography*, **15** 523–539.
- CLOGG, C. C. and ELIASON, S. R. (1988). A flexible procedure for adjusting rates and proportions, including statistical-methods for group comparisons. *Am. Sociol. Rev.*, **53** 267–283.
- COCHRAN, W. (1954). Some methods of strengthening the common χ^2 test. *Biometrics*, **10** 417–451.
- COCHRAN, W. G. (1969). Use of covariance in observational studies. *Appl. Stat.*, **18** 270–275.
- COLEMAN, M. P., ESTÈVE, J., DAMIECKI, P., ARSLAN, A. and RENARD, H. (1993). *Trends in Cancer Incidence and Mortality*. No. 121 in IARC Scientific Publications, International Agency for Research on Cancer, Lyon.
- CORNFIELD, J. (1951). A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix. *J. Natl. Cancer. Inst.*, **11** 1269–1275.
- CORNFIELD, J. (1956). A statistical problem arising from retrospective studies. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health* (J. Neyman, ed.). University of California Press, Berkeley, 135–148.
- CORNFIELD, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fed. Proc.*, **21** 58–61.
- COX, D. R. (1958). The regression-analysis of binary sequences. *J. Roy. Stat. Soc. Ser. B*, **20** 215–242.
- COX, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Stat. Soc. Ser. B*, **34** 187–220.
- DANIEL, R., COUSENS, S., DE STAVOLA, B., KENWARD, M. and STERNE, J. (2013). Methods for dealing with time-dependent confounding. *Stat. Med.*, **32** 1584–1618.
- DAVIDSON, A. and REID, A. (1926–1927). On the calculation of rates of mortality. *T. Fac. Actuaries*, **11** 183–232.
- DAY, N. E. (1976). A new measure of age standardized incidence, the cumulative rate. In *Cancer Incidence in Five Continents* (J. Waterhouse, C. Muir, P. Correa and J. Powell, eds.), vol. 3 of *IARC Scientific Publications*, chap. 8. International Agency for Research on Cancer, Lyon, 443–445.

- DERRICK, V. P. A. (1927). Observations on (1) errors of age in the population statistics of England and Wales, and (2) the changes in mortality indicated by the national records. *J. Inst. Actuaries*, **58** 117–159.
- DYKE, G. and PATTERSON, H. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8** 1–12.
- EDGEWORTH, F. Y. (1917). Review of “Scope and Method of Statistics” by Harald Westergaard. *J. Roy. Stat. Soc.*, **80** 546–551.
- ELANDT-JOHNSON, R. C. (1975). Definition of rates: some remarks on their use and misuse. *Am. J. Epidemiol.*, **102** 267–271.
- FARR, W. (1859). *Letter to the Registrar General*. General Registrar Office, London.
- FIBIGER, J. (1898). Om serumbehandling af difteri. *Hospitalstidende*, **6** 337–350.
- FIENBERG, S. E. (1975). Design and analysis of observational study — Comment. *J. Am. Stat. Ass.*, **70** 521–523.
- FISHER, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P . *J. Roy. Stat. Soc.*, **85** 87–94.
- FREEMAN, D. H. and HOLFORD, T. R. (1980). Summary rates. *Biometrics*, **36** 195–205.
- GRAUBARD, B. I., RAO, R. S. and GASTWIRTH, J. L. (2005). Using the Peters-Belson method to measure health care disparities from complex survey data. *Stat. Med.*, **24** 2659–2668.
- GREENLAND, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Stat. Med.*, **1** 217–227.
- HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.
- HAENSZEL, W., SHIMKIN, M. and MANTEL, N. (1958). A retrospective study of lung cancer in women. *J. Natl. Cancer Inst.*, **21** 825–842.
- HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, **95** 481–488.
- HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1** 465–480.
- HERNÁN, M. A., CLAYTON, D. and KEIDING, N. (2011). The Simpson’s paradox unraveled. *Int. J. Epidemiol.*, **40** 780–785.
- HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health*, **60**.
- HOBSCRAFT, J., MENKEN, J. and PRESTON, S. (1982). Age, period, and cohort effects in demography — a review. *Popul. Index*, **48** 4–43.
- HOEM, J. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates — a review. *Int. Stat. Rev.*, **55** 119–152.
- HOLFORD, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, **36** 299–305.
- HROBJARTSSON, A., GOTZSCHE, P. C. and GLUUD, C. (1998). The controlled clinical trial turns 100 years: Fibiger’s trial of serum treatment of diphtheria. *Br. Med. J.*, **317** 1243–1245.
- KALBFLEISCH, J. and PRENTICE, R. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- KALTON, G. (1968). Standardization — a technique to control for extraneous variables. *Appl. Stat.*, **17** 118–136.
- KEIDING, N. (1987). The method of expected number of deaths, 1786–1886–1986. *Int. Stat. Rev.*, **55** 1–20.
- KEIDING, N. (2011). Age-period-cohort analysis in the 1870s: Diagrams, stereograms, and the basic differential equation. *Can. J. Stat.*, **39** 405–420.
- KENDALL, P. L. and LAZARSELD, P. F. (1950). Problems of survey analysis. In *Continuities in Social Research: Studies in the Scope and Method of “The American Soldier”* (R. Merton and P. Lazarsfeld, eds.). The Free Press, Glencoe, Ill., 133–196.
- KERMACK, W. O., MCKENDRICK, A. G. and MCKINLAY, P. L. (1934a). Death-rates in Great Britain and Sweden — some general regularities and their significance. *Lancet*, **1** 698–703.
- KERMACK, W. O., MCKENDRICK, A. G. and MCKINLAY, P. L. (1934b). Death-rates in Great Britain and Sweden: Expression of specific mortality rates as products of two factors, and some consequences thereof. *J. Hyg.*, **34** 433–457.
- KILPATRICK, S. J. (1962). Occupational mortality indexes. *Popul. Stud.*, **16** 175–187.
- KITAGAWA, E. M. (1955). Components of a difference between 2 rates. *J. Am. Stat. Ass.*, **50** 1168–1194.
- KITAGAWA, E. M. (1964). Standardized comparisons in population-research. *Demography*, **1**

- 296–315.
- KITAGAWA, E. M. (1966). Theoretical considerations in selection of a mortality index and some empirical comparisons. *Human Biology*, **38** 293–308.
- KÖRÖSI, J. (1892–1893). Report of an international mortality standard, or mortality index. *Pub. Am. Stat. Ass.*, **3** 450–462.
- KURTH, T., WALKER, A. M., GLYNN, R. J., CHAN, K. A., GAZIANO, J. M., BERGER, K. and ROBINS, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.*, **163** 262–270.
- LANE, P. W. and NELDER, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38** 613–621.
- LEXIS, W. (1876). Das Geschlechtsverhältnis der Geborenen und die Wahrscheinlichkeitsrechnung. *Jahrbücher für Nationalökonomie und Statistik*, **27** 209–245.
- LIAO, T. (1989). A flexible approach for the decomposition of rate differences. *Demography*, **26** 717–726.
- MACKENZIE, D. (1978). Statistical theory and social interests — case study. *Soc. Stud. Sci.*, **8** 35–83.
- MACKENZIE, D. A. (1981). *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh University Press, Edinburgh.
- MACKINNON, D. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum ass, New York.
- MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50** 163–170.
- MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer. Inst.*, **22** 719–748.
- MANTEL, N. and STARK, C. R. (1968). Computation of indirect-adjusted rates in presence of confounding. *Biometrics*, **24** 997–1005.
- McKENDRICK, A. G. (1925–26). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.*, **43–44** 98–130.
- McKINLAY, S. M. (1975). Design and analysis of observational study — review. *J. Am. Stat. Ass.*, **70** 503–520.
- MIETTINEN, O. (1972a). Standardization of risk ratios. *Am. J. Epidemiol.*, **96** 383–388.
- MIETTINEN, O. (1976a). Estimability and estimation in case-referent studies. *Am. J. Epidemiol.*, **103** 230–235.
- MIETTINEN, O. S. (1972b). Components of the crude risk ratio. *Am. J. Epidemiol.*, **96** 168–172.
- MIETTINEN, O. S. (1976b). Stratification by a multivariate confounder score. *Am. J. Epidemiol.*, **104** 609–620.
- MIETTINEN, O. S. (1985). *Theoretical Epidemiology*. Wiley, New York.
- NEISON, F. (1844). On a method recently proposed for conducting inquiries into the comparative sanitary condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *J. Stat. Soc. London*, **7** 40–68.
- NEISON, F. G. P. (1851). On the rate of mortality among persons of intemperate habits. *J. Stat. Soc. London*, **14** 200–219.
- NELDER, J. and WEDDERBURN, R. (1972). Generalized linear models. *J. Roy. Stat. Soc. Ser. A*, **135** 370–384.
- NELSON, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14** 945–966.
- NEWSHOLME, A. and STEVENSON, T. H. C. (1906). The decline of human fertility in the United Kingdom and other countries as shown by corrected birth-rates. *J. Roy. Stat. Soc.*, **69** 34–87.
- OGLE, W. (1892). Proposal for the establishment and international use of a standard population, with fixed sex and age distribution, in the calculation and comparison of marriage, birth, and death rates. *Bull. Int. Stat. Inst.*, **6** 83–85.
- OSBORN, J. (1975). Multiplicative model for analysis of vital statistics rates. *Appl. Stat.*, **24** 75–84.
- PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufman, San Francisco, 411–420.
- PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

- PEARL, J. and BARENBOIM, E. (2012). Transportability across studies: A formal approach. Tech. rep., Computer Science Department, University of California, Los Angeles.
- PEARSON, K. (1900). On the correlation of characters not quantitatively measurable. *Phil. T. Roy. Soc. A*, **195** 1–47.
- PEARSON, K. (1910). *The Grammar of Science*. 3rd ed. Black, Edinburgh.
- PEARSON, K., LEE, A. and BRAMLEY-MOORE, L. (1899). Genetic (reproductive) selection: inheritance of fertility in man and of fecundity in thoroughbred racehorses. *Phil. T. Roy. Soc. A*, **192** 534–539.
- PEARSON, K. and TOCHER, J. F. (1915). On criteria for the existence of differential death rates. *Biometrika*, **11** 159–184.
- PETERS, C. C. (1941). A method of matching groups for experiment with no loss of population. *J. Educ. Res.*, **34** 606–612.
- PETERSEN, M., SINISI, S. and VAN DER LAAN, M. (2006). Estimation of direct causal effects. *Epidemiology*, **17** 276–284.
- PIKE, M., ANDERSON, J. and DAY, N. (1979). Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiol. Comm. Health*, **33** 104–106.
- POWERS, D. A. and XIE, Y. (2008). *Statistical Methods for Categorical Data Analysis: 2nd Edition*. Emerald.
- POWERS, D. A. and YUN, M.-S. (2009). Multivariate decomposition for hazard rate models. *Sociol. Methodol.*, **39** 233–263.
- PRESTON, S. H., HEUVELINE, P. and GUILLOT, M. (2001). *Demography*. Blackwell.
- ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Math. Modelling*, **7** 1393–1512.
- ROBINS, J., BLEVINS, D., RITTER, G. and WULFSOHN, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, **14** 79–97.
- ROBINS, J. and HERNÁN, M. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis* (G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs, eds.). Chapman and Hall, New York, 533–599.
- ROBINS, J., HERNÁN, M. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11** 550–560.
- ROBINS, J. and TSIATIS, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, **79** 311–319.
- ROSENBAUM, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Am. Stat. Ass.*, **79** 565–574.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70** 41–55.
- ROSENBERG, M. (1962). Test factor standardization as a method of interpretation. *Soc. Forces*, **41** 53–61.
- ROTHMAN, K. (1986). *Modern Epidemiology*. Little, Brown and Company.
- ROTHMAN, K. and GREENLAND, S. (1998). *Modern Epidemiology*. 2nd ed. Lippincott Williams and Wilkins.
- ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*. Third edition ed. Wolters Kluwer.
- SATO, T. and MATSUYAMA, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology*, **14** 680–686.
- SCHWEBER, L. (2001). Manipulation and population statistics in nineteenth-century France and England. *Soc. Res.*, **68** 547–582.
- SCHWEBER, L. (2006). *Disciplining Statistics: Demography and Vital Statistics in France and England 1830–1885*. Duke University Press, Durham.
- SCOTT, R. C. (1978). Bias problem in smear-and-sweep analysis. *J. Am. Stat. Ass.*, **73** 714–718.
- SEARLE, S. R., SPEED, F. M. and MILLIKEN, G. A. (1980). Population marginal means in the linear model — an alternative to least-squares means. *Am. Stat.*, **34** 216–221.
- SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Stat. Soc. Ser. B*, **13** 238–241.
- SOBEL, M. E. (1996). Clifford Collier Clogg, 1949–1995: A tribute to his life and work. *Sociol. Methodol.*, **26** 1–38.

- STIGLER, S. M. (1986). *The History of Statistics - The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University Press.
- STONE, M. (1970). Review of National Halothane Study — Bunker, J.P., Forrest, W.H., Mosteller, F. and Vandam, L.D.D. *J. Am. Stat. Ass.*, **65** 1392–1396.
- STOUFFER, S. A. and TIBBITTS, C. (1933). Tests of significance in applying Westergaard's method of expected cases to sociological data. *J. Am. Stat. Ass.*, **28** 293–302.
- THIELE, T. (1881). H. Westergaards Mortalitets-Statistik. *Nationaløkonomisk Tidsskrift*, **18** 219–227.
- TRUETT, J., CORNFIELD, J. and KANNEL, W. (1967). A multivariate analysis of coronary heart disease risk in Framingham. *J. Chron. Dis.*, **20** 511–524.
- TUKEY, J. W. (1979). Methodology, and the statisticians responsibility for both accuracy and relevance. *J. Am. Stat. Ass.*, **74** 786–793.
- TUKEY, J. W. (1991). Use of many covariates in clinical trials. *Int. Stat. Rev.*, **59** 123–137.
- TURNER, R. H. (1949). The expected-cases method applied to the nonwhite male labor force. *Am. J. Sociol.*, **55** 146–156.
- VANSTEELENDT, S. and KEIDING, N. (2011). Invited commentary: G-computation – lost in translation? *Int. J. Epidemiol.*, **173** 739–742.
- VON BORTKIEWICZ, L. (1904). Über die methode der “standard population”. *Bull. Int. Stat. Inst.*, **14** 417–437.
- WALKER, S. H. and DUNCAN, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, **54** 167–179.
- WESTERGAARD, H. (1882). *Die Lehre von der Mortalität und Morbilität*. Jena: Fischer.
- WESTERGAARD, H. (1901). *Die Lehre von der Mortalität und Morbilität*. 2nd ed. Auflage. Jena: Fischer.
- WESTERGAARD, H. (1916). Scope and method of statistics. *Pub. Am. Stat. Ass.*, **15** 229–276.
- WESTERGAARD, H. (1918). On the future of statistics. *J. Roy. Stat. Soc.*, **81** 499–520.
- WOODBURY, R. M. (1922). Westergaard's method of expected deaths as applied to the study of infant mortality. *J. Am. Stat. Ass.*, **18** 366–376.
- XIE, Y. (1989). An alternative purging method — controlling the composition-dependent interaction in an analysis of rates. *Demography*, **26** 711–716.
- YAMAGUCHI, K. (2011). Decomposition of inequality among groups by counterfactual modeling: an analysis of the gender wage gap in Japan. *Sociol. Methodol.*, **41** 223–255.
- YULE, G. U. (1900). On the association of attributes in statistics. *Phil. T. Roy. Soc. A*, **194** 257–319.
- YULE, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, **2** 121–134.
- YULE, G. U. (1906). On the changes in the marriage- and birth-rates in England and Wales during the past half century; with an inquiry as to their probable causes. *J. Roy. Stat. Soc.*, **69**.
- YULE, G. U. (1911). *An Introduction to the Theory of Statistics*. Griffin, London.
- YULE, G. U. (1912). On the methods of measuring association between two attributes. *J. Roy. Stat. Soc.*, **75** 579–652.
- YULE, G. U. (1920). *The Fall of the Birth-Rate*. Cambridge University Press, Cambridge. A paper read before the Cambridge University Eugenics Society, 20 May 1920.
- YULE, G. U. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality (with discussion). *J. Roy. Stat. Soc.*, **97** 1–84.