European Journal of Epidemiology (2018) 33:897–907
https://doi.org/10.1007/s10654-018-0434-4

ESSAY

Individually-matched etiologic studies: classical estimators made new again

James A. Hanley[1]

Received: 12 April 2018 / Accepted: 18 August 2018 / Published online: 24 August 2018

Abstract

With greater access to regression-based methods for confounder control, the etiologic study with individual matching, analyzed by classical (calculator) methods, lost favor in recent decades. This design was costly, and the data sometimes mis-analyzed. Now, with Big Data, individual matching becomes an economical option. To many, however, conditional logistic regression, commonly used to estimate the incidence density ratio parameter, is somewhat of a black box whose output is not easily checked. An epidemiologist-statistician pair recently proposed a new estimator that is easily applied to data from individually-matched series with a 2:1 ratio (and no other confounding variables) using just a hand calculator or spreadsheet. Surprisingly—or possibly not—they overlooked classical estimators developed in earlier decades. This prompts me to re-introduce some of these, to highlight their considerable flexibility and ease of use, and to update them. Nowadays, for any matching ratio (M:1), the Maximum Likelihood result can be easily computed from data gathered under the matched design in two different ways, each using just the summary data. One is via any binomial regression program that allows offsets, applied to just M 'rows' of data. The other is by hand! The aim of this note is not to save on computation; instead, it is to make connections between classical and regression-based methods, to promote terminology that reflects the concepts and structure of the etiologic study, and to focus attention on what parameter is being estimated.

Keywords Woolf · Mantel and Haenszel · Miettinen · Breslow and Day · Clayton · Nelder and Wedderburn

Abbreviations

ML	Maximum likelihood
MLE	Maximum likelihood estimator or estimate
IDR	Incidence density ratio
BD	Breslow and Day
MH	Mantel–Haenszel
SE	Standard error

Introduction

The essence of an etiologic study was originally taken to be that of a 'case–control' study, which involves a group of cases of the illness in question and a comparable control group without the illness; and these groups are compared with respect to the histories of the etiologic factor under study.

That conception of the etiologic study, while still common, is at variance with the principle that "any empirical study has, by definition, some particularistic (spatio-temporally specific) experience as its base, and the results of the study apply in a direct, technical sense to that particular experience. The base thus is the direct referent of the empirical information" [1]. So-called 'case–control' studies do not have an explicit study base.

In the newer conception of it [2], an etiologic study is to be constructed on a defined aggregate of *study population-time*, constituting the base of the study and, hence, the referent of the study result. Its elements, in reference to its *study base*, are: (1) the suitably documented case series, constituted by the entirety of the cases (as defined) occurring in the study base; (2) the similarly documented base series, derived as a fair sample of the study base; and (3) the data on these two series (of person-moments) translated into the corresponding value for the confounder-conditional rate-ratio of the occurrence of the illness in the study base, and into its associated inferential statistic(s). In

✉ James A. Hanley
james.hanley@mcgill.ca
http://www.biostat.mcgill.ca/hanley

[1] Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC H3A 1A2, Canada

this, the study base is "the aggregate of population-time for which the outcome's rate of occurrence is documented." The result is an incidence-density ratio, free of any 'rare-disease assumption' [3].

With increasing access to regression-based adjustment methods, the use of individual matching in etiologic studies—and classical (calculator-based) data-reduction—lost favor in the later decades of the 20th century. This design was common earlier, but costly to implement. But with the increasing availability of Big Data, tighter matching in the selection of the base ('denominator') series is an attractive and economical possibility. Conditional logistic regression allows individual sets to be tightly matched on the most important confounders, and the less critical unmatched ones used as regression variates. This 'match if possible, model if you must' approach provided by conditional logistic regression makes the results more explainable, and less reliant on the form of the adjustment model.

This important tool was developed in parallel in the social sciences (one of its developers, Daniel McFadden received the Nobel Prize in Economics in the year 2000) and in biostatistics. These 'two solitudes' were the subject of Norman Breslow's presidential address [4] to the International Biometric Society in 2002. Unfortunately, similar solitudes also now divide 'classical' and 'clinical' epidemiology: in the 2017 Journal of Clinical Epidemiology article [5] that prompted the present piece, the journal's referees and editors did not notice that this eminent developer of statistical methods for epidemiology was referred to as 'Bresolow.'

Seemingly unaware of classical methods for individually-matched etiologic studies developed over the last six decades, and concerned that conditional logistic regression is not as transparent as epidemiologists would wish, Redelmeier and Tibshirani (RT) [5] recently proposed a new incidence density ratio estimator that uses data from a 2:1 design and is easily implemented on a hand calculator. I re-analyze their illustrative data by historical classical methods, and find them as easy to use as, and more flexible than, the proposed one. Thus, in this piece, aimed especially at the newer generation, I welcome the chance to dust off some of these classics, highlight some extensions, and make new connections. I show how the Maximum Likelihood (ML) estimate using data from the M:1 design can be derived in two different ways, both using just the summary data. One is via any binomial regression program that allows offsets, applied to just M 'rows' of data. The other is by hand or by spreadsheet!

The emphasis will not be so much on showing how to save on computation, but on heuristics; on Mantel's inspired (and nearly ML) choice of estimator and of weights for the cross-products; on making connections between classical and regression-based methods; and on truly understanding our statistical tools, and the target parameters we aim at. Intended readers of this piece include the investigator/supervisor whose assistant/graduate student brings a point estimate and confidence interval (CI) based on a conditional logistic regression, or the reviewer/reader who is digesting the results in an article. Just from the sufficient statistics, and with a hand calculator or smart phone, is that person able to check that the point estimate makes sense, and that the CI is not an order of magnitude too wide/narrow? By using simple 'tabular' methods to reality-check the numbers from black boxes, both the producers and consumers of the results will also get a better sense of what precision to expect with various amounts of data. In some instances, the study results and inferential statistics obtained from the 'tabular' data will be sufficient, and in any event such results are more likely to be understood by readers.

Before re-applying these older methods—and applying the updated ones—to a 'ragged' dataset long used in teaching, I will first use the simpler matched data that RT used to illustrate their method, namely weather information from 6962 place-, day-of week- and time-of-day- matched triplets; the two selected place-moments in each triplet were 7 days before and after the crash moment. RT asked whether, and if so, by how much, the rate of traffic crashes is lower in overcast weather compared with other types of weather. The hypothesized causal factor is "the cautiousness induced by gloomy circumstances." Curiously, their justification for using the 'odds ratio' was that "the baseline risk of a crash is low (< 1%) during an average day, thereby making an odds ratio a good estimate of relative risk."

I will adopt the 'case-referent' framework [1–3], where the study base is the referent of both the case and the base/referent series and, hence, of the study result; thus the 'odds ratio' actually is incidence-density ratio, with no rare-disease proviso. Throughout, I will refer to a '*case series*' or '*numerator series*', and a selected '*base series'* or '*denominator series'* that *probes* into, and thus represents or 'refers to' the base (the aggregate of population-time in the defined study population's movement over a defined span of time) from which the cases emerged (see Fig. 1). Viewed from this perspective, one that was already alluded to in 1955, these two series allow a natural and direct comparison of incidence rates (incidence densities) by estimating the relative amounts of population time in the two categories of the etiologic factor at issue. The object is not to compare 'cases' with 'controls,' but to compare incidence densities between the compared segments of the study base.

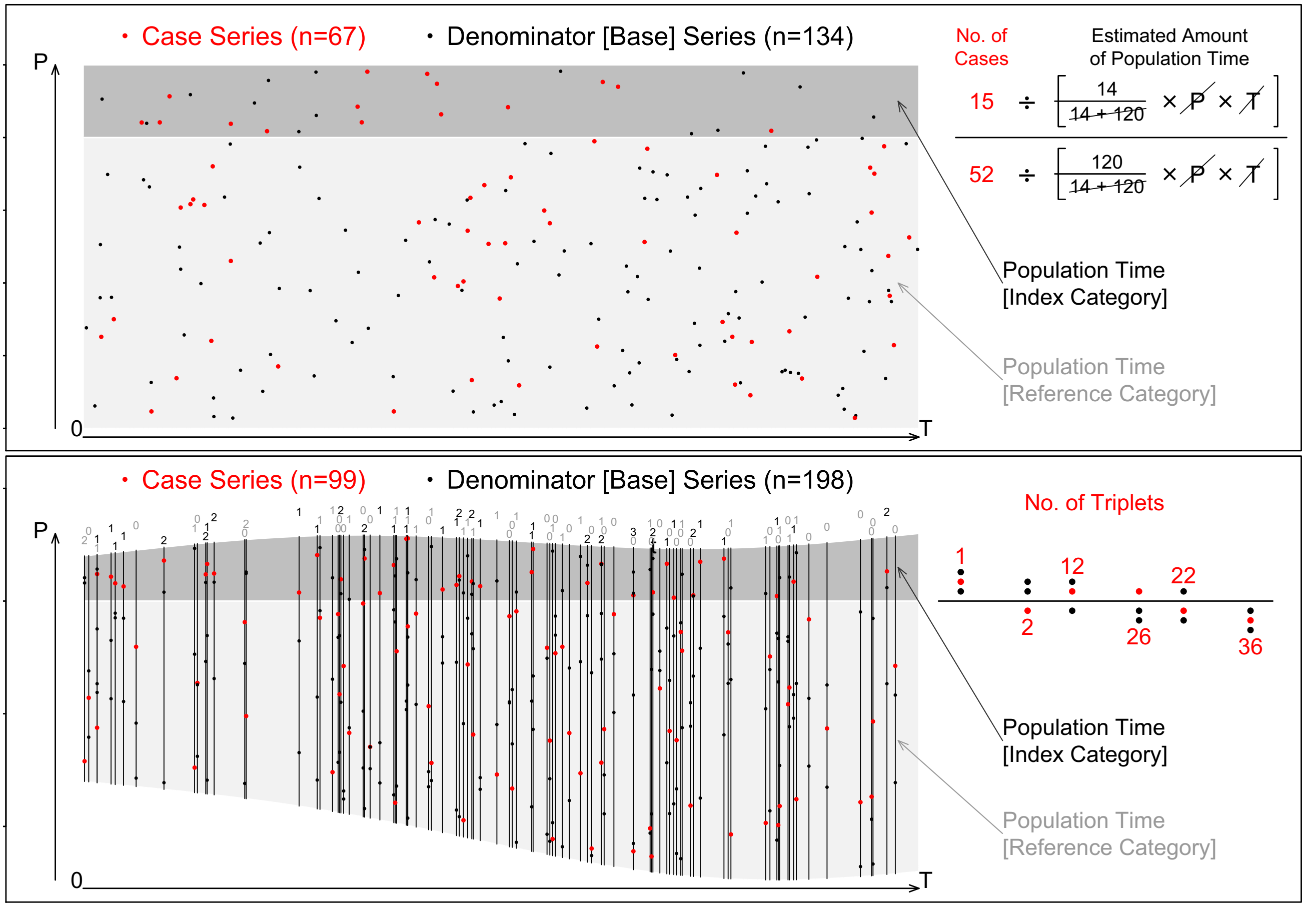

• Case Series (n=67)
• Denominator [Base] Series (n=134)
P
0
T
No. of Cases
Estimated Amount of Population Time
15 ÷ $\left[\frac{14}{14+120} \times \cancel{P} \times \cancel{T}\right]$
52 ÷ $\left[\frac{120}{14+120} \times \cancel{P} \times \cancel{T}\right]$
Population Time [Index Category]
Population Time [Reference Category]
• Case Series (n=99)
• Denominator [Base] Series (n=198)
No. of Triplets
1
2
12
26
22
36
Population Time [Index Category]
Population Time [Reference Category]

Fig. 1 Top panel: events (red dots) within a fictional dynamic population of fixed (but possibly unknown) size P over a timespan of length T, during which the distribution (unknown) of the factor at issue, or the event rate in the population time in the reference category, does not change. Estimation of an incidence density ratio begins by classifying the events in the case series into those that occurred in the index and reference categories and tallying them as 'numerators.' The (unknown) relative amounts of population time in these contrasted categories are then estimated by classifying the indiscriminately sampled person-moments (denominator series or base series, black dots, that 'probe' the base) into the two categories. The empirical incidence density ratio is computed from the estimated amounts of population time just as if the amounts were known, but the estimation of the underlying denominators involves an additional variance component. Bottom panel: events (red dots) arising in a fictional, dynamic population in which the distribution of the factor at issue, and event rates in the population time in the reference category, both vary over the timespan. Ignoring this double time-dependency would yield a confounded incidence ratio. In the Redelmeier and Tibshirani (RT) study, each set of 3 moments is matched on place, time-of-day, and day of week. For each individually-matched triplet, how many of the 3 moments were classified as being in the index category of the factor at issue is shown above the triplet. Shown at the right are the frequencies with which the triplets are distributed over the 6 configurations. Estimators based on these frequencies are addressed in Figs. 2 and 3

Classical estimators revisited—and updated

Breslow and Day [6] (BD) divide their four key data-analysis chapters into classical methods for grouped and matched data, and their respective counterparts unconditional and conditional logistic regression for large strata and matched sets. The last of these chapters shows that an *unconditional* logistic regression of individually matched sets (such as the 99 in the lower schematic in Fig. 1), with a separate nuisance parameter for each set, results in an incidence density rate ratio (IDR) estimate which is *further from* the null than the correct, conditional one, and that the estimate obtained by *ignoring the matching* tends on average to be *too close* to the null.

The following sections address, and attempt to connect, the classical and conditional logistic regression estimators of the IDR using data from individually matched sets, and using terminology that is more appropriate to what is involved in *the* etiologic study.

A 'nearly maximum likelihood' incidence density ratio estimator (1959)

A natural point of departure for classical methods for matched case-base data is the 1959 citation classic [7] by

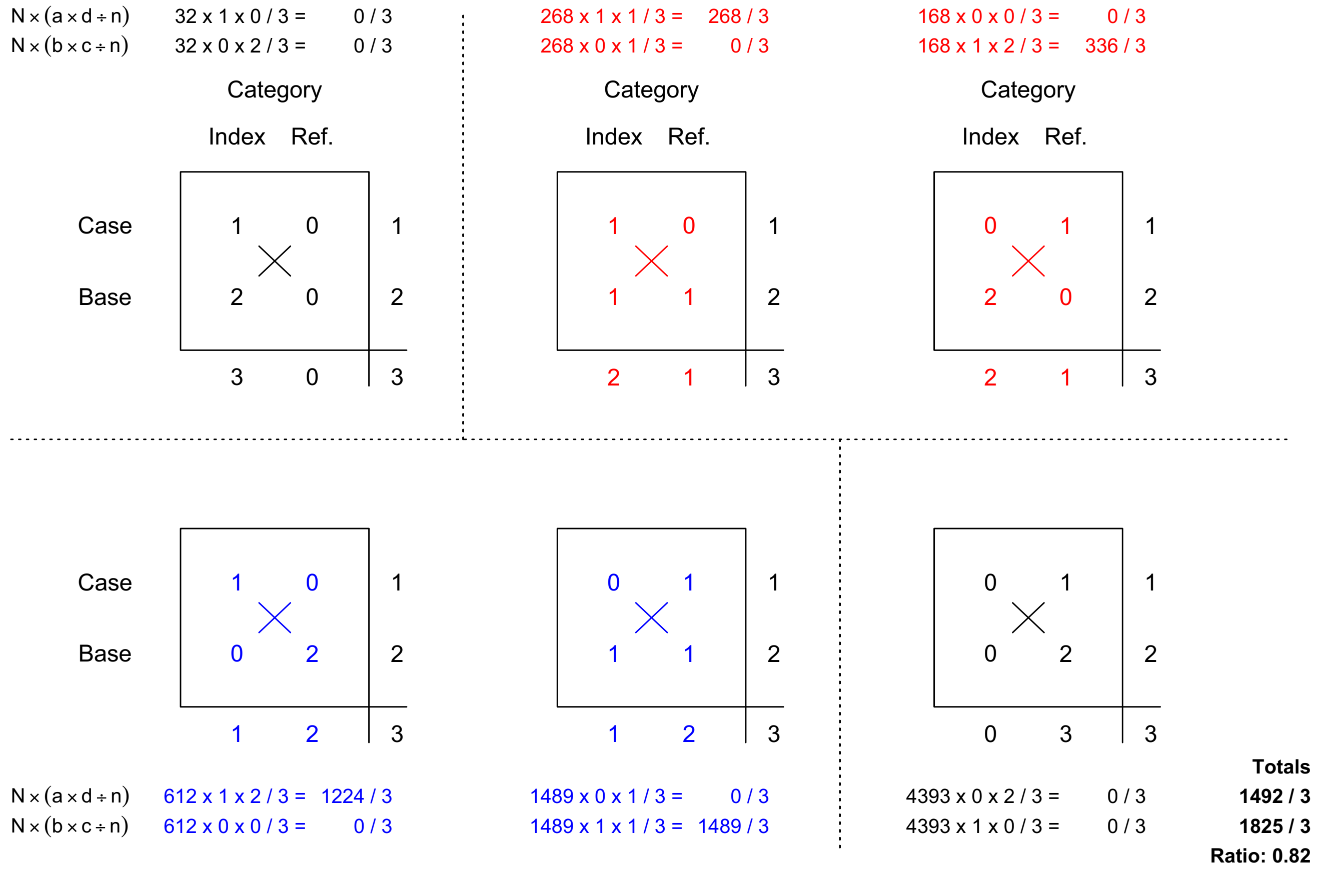


Fig. 2 The Mantel–Haenszel estimate of incidence density ratio, computed from the data on the 6962 matched triplets analyzed by Redelmeier and Tibshirani (RT). Each cross-product, reduced by 3, is multiplied by the number of triplets (N) with the indicated data configuration. Four of the six configurations (involving 6962 − (32 + 4393) = 2537 triplets) contribute to the IDR estimate, while the two 'concordant' ones do not

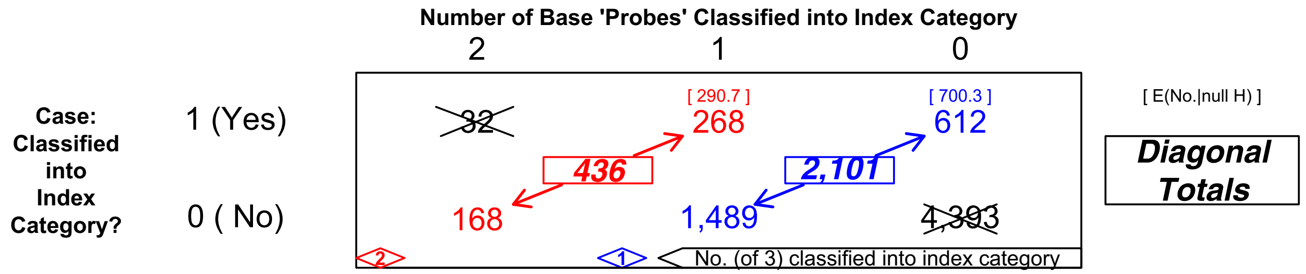
Mantel and Haenszel [8] (MH). More important than the statistic they developed to *test* the null hypothesis that the incidence density ratio equals unity is their *estimator of the ratio*. And, unlike the 'large-sets' one developed by Woolf in 1955 [9], the estimator works well with individually matched sets as small as two (matched pairs). Indeed, in this limiting case, MH noted that the estimator has a particularly simple form, namely the ratio of the numbers of discordant pairs of each type. Presumably because of its focus on these two cells, RT write of it as if it were the 'McNemar' estimator. However, McNemar [10] focused on the statistical variability of the difference of two correlated proportions to be used in hypothesis testing, and was not concerned with estimating an IDR. But his *table layout* becomes important below.

Figure 2 shows the contributions to the IDR estimate obtained via the MH estimator from the data on the 6962 matched triplets. It reduces to

$$IDR_{MH} = \frac{268 \times 1 \times 1 \div 3 + 612 \times 1 \times 2 \div 3}{168 \times 1 \times 2 \div 3 + 1489 \times 1 \times 1 \div 3} = \frac{1492/3}{1825/3}$$
$$= 0.82.$$

Even though others since then have recast the Sum(*ad*/*n*)/Sum(*bc*/*n*) expression as a weighted (by '*bc*/*n*') sum of the ('*ad*/*bc*') IDR estimates, Mantel would not have endorsed this formulation. To him, and to those who have developed ratio estimators elsewhere in statistics, it is the '*single* ratio of two sums' structure that gives this classic estimator its statistical stability. In his teaching, Miettinen used an even simpler example to make this point. Imagine you were asked to measure the sex ratio in small (in-home) day-care centers, using a sample of such facilities. Would you take a mean or median of the individual facility-specific sex-ratios, or would you take the single ratio of the sum of all boys sampled divided by the corresponding sum for girls?

As will be seen, the MH estimate is quite close to the ML one. Even though Mantel developed the estimator more from intuition than from any formal statistical model, the table-specific variances in the *test statistic* are based on


Number of Base 'Probes' Classified into Index Category
2
1
0
Case: Classified into Index Category?
1 (Yes)
0 (No)
~~32~~
[290.7]
268
[700.3]
612
436
2,101
168
1,489
~~4,393~~
2
1
No. (of 3) classified into index category
[E(No.|null H)]
Diagonal Totals

SE's for log(MH IDR estimate)

```
                 No. [m] of triplet members
               classified into index category
                         2           1
MH numerators & denominators                [SUM]
  r = ad/3      (268) 1/3     (612) 2/3      R=1492/3
  s = bc/3      (168) 2/3    (1489) 1/3      S=1825/3
                  MH point estimate = R/S = 0.8175

No. triplets, case classified into index category
Observed(O)          268         612
Expected(Null,E)     290.7       700.3
O-E:                 -22.7       -88.3        -111

Null Variances  436(2/3)(1/3)  2101(1/3)(2/3)
                         97          467          V=566

                  CHI.SQ = (-111)^2 / 566 = 21.84

* TEST-BASED SE (Miettinen, 1976)
   SE = abs[log(0.8175)] / sqrt(21.84)    = 0.0428

* GENERAL SE (Robins,Greenland,Breslow, RGB 1986)
  r*p [Notes]    (268) 2/9    (612) 6/9       4208/9
  r*q            (268) 1/9    (612) 0/9        268/9
  s*p            (168) 0/9   (1489) 1/9       1489/9
  s*q            (168) 6/9   (1489) 2/9       3986/9

         4209         268 + 1489          3986
 Var =  ---------- + --------------- + ----------
        2 x 1492^2   2 x 1492 x 1825   2 x 1825^2

  SE = sqrt(V)  = sqrt(0.001866186)         = 0.0431

* NOT WIDELY KNOWN SE (Clayton & Hills, CH, 1993)
   SE = sqrt[566/([1492/3] x [1825/3])]    = 0.0431
```

DIRECT, AND REGRESSION-BASED, MLE of IDR

```
* DIRECT: Miettinen, 1970                                  .
  (closed form when 2:1 individual matching)               .
                                     436+2,101     2,537
                                       268+612       880
                               5x880-2,101-4x436     555
                                    4x(2,537-880)  6,628
                                       555/6,628  0.0837
                               880/(2,537 - 880)  0.5311
           MLE = 0.0837+sqrt(0.0837^2+0.5311)     0.8173

* REGRESSION-BASED, 2018                                   .
  (Generalized Linear Model, R code shown)                 .

                                 Pos =    c(268, 612)      .
                                 Neg =    c(168,1489)      .
                                 O   = log(c(2,1/2))       .

   ( fit=summary(glm(cbind(Pos,Neg) ~ 1+ offset(O),        .
                  family=binomial)) )                       .
   >          Estimate Std. Error z value Pr(>|z|)          .
   >Intercept -0.20177    0.04322   -4.669 3.03e-06         .

   [MLE = exp(-0.20177);  0.04322 = SE of log MLE          ]

   round(        exp(fit$coefficients[1]+                  .
     c(-1.96,0,1.96)*fit$coefficients[2]),4)               .
   >  0.7509 0.8173 0.8895                                 .
```

[NOTES]

```
MH: Mantel-Haenszel;
[SUM] is over INFORMATIVE triplets (m=2,1 in Breslow&Day notation.)
[E, null] = m*1/3 ; null variance = m*(3-m)*1*2/[(3^2)*(3-1)].
Null variance is under conditional (central hypergeometric) model;
if individual matching, the 'a' frequency is a random variable with
a Bernoulli distribution, so null variance = (m*1/3)*(1 - m*1/3).
RGB: p = (a+d)/3 &  q = (b+c)/3 ; see text, and [11], for details.
```

Fig. 3 Layout (after Miettinen, 1970) of data frequencies from Redelmeier and Tibshirani) (RT)'s 2:1 individually matched case-base study, showing 2 non-informative and 4 informative triplet configurations. Shown under it on left are 3 possible SEs for the log estimate, from which to calculate a confidence interval to accompany the Mantel–Haenszel (MH) estimate of the incidence density ratio (IDR). Shown under it on right is manual calculation of maximum likelihood estimate (MLE) of the IDR described in the 1970 article. Shown below this is the generalized linear model code that allows the fitting of the single parameter (IDR) Bernoulli model whose fitted frequencies are a function of that single parameter, and the number, m, of triplet members classified into the index category of the 'exposure' (see section on a 'familiar logistic regression')

the hypergeometric model obtained by *conditioning* on all four marginal totals of each table, rather than the unconditional (two independent binomials) model used by Cochran.

Mantel and Haenszel did not provide an expression for the precision of the MH estimate. Some 17 years later, by reverse-engineering a (null) standard error (SE) from this MH test statistic, Miettinen [3] provided a simple test-based confidence interval for it. Several standard errors specific to the MH-estimator were proposed over the following decade; the most general is the one developed by Robins, Greenland and Breslow ('RGB') [11]. I return to various SE versions in another later subsection.

Maximum likelihood IDR estimator, with closed form in 2:1 design (1970)

Section 3 of the 'Classical Methods of Analysis of Matched Data' chapter 5 in Breslow and Day's textbook [Ref. 11 in RT] is devoted to "1:M matching: dichotomous exposures". There they explain that the conditional Maximum Likelihood Estimate (MLE) in general requires iterative numerical calculations, but that in 1970 Miettinen [12] provided a closed form expression for the case M = 2. Since this is the very design that RT address, it is instructive to revisit this 1970 publication. The purpose here is not to promote the use of his closed form estimator, but to introduce readers to his extension of McNemar's table layout, and his use of the Maximum Likelihood (ML) fitting criterion. This 2 × 3 table, with its two familiar uninformative corner frequencies—just as in the McNemar layout—but with *two* diagonals, is shown in the top half of Fig. 3. Because it treats the observations from the two probes of the base as exchangeable, this layout is more economical and extensible than RT's Venn diagram [5]. The ML point and interval estimates are based on a separate (binomial) statistical model for the frequencies in each diagonal (each binomial is induced by conditioning on the sum), and on the sufficient statistics, whereas, as Miettinen [12] pointed out, the MH estimator does not reduce to a function of these.

Standard errors for the log of the Mantel–Haenszel estimator (1976–1993)

Also shown in the bottom left of Fig. 3 are three versions of a standard error from which, beginning on the log IDR scale, one can construct a confidence interval to accompany the MH estimate. The first of these is Miettinen's test-based version [3]. It uses the square root of the Null Chi Square statistic (here sqrt[21.84] = 4.7) to measure that, in this example, the observed number of exposed cases (880) deviates from its null expectation (991) by 4.7 standard errors (SE's). He takes this as equivalent to observing that the log of the observed MH estimate (log[0.8175]) deviates from its null expectation (0) by 4.7 standard errors (SE's), implying a standard error for the log[0.8175] of (log[0.8175] − 0)/4.7 = 0.0428. This inferred null SE of the log of the IDR estimate is then multiplied by 1.96 to compute a 95% CI for the log of the IDR estimate. Its exponentiated form provides a CI for the IDR itself.

The second of these, the Robins-Greenland-Breslow expression [11], yields a standard error of 0.0431. This expression, obtained by summing six quantities from each stratum/triplet (see Fig. 3), was developed in 1986 to replace seven previously proposed estimators, to be easily computed, and to be used in the analysis of data from individually matched, grouped, and unstratified case-base series.

Writing in 1993, Clayton and Hills [13] noted that most of the several standard error expressions for the log IDR estimate (Ref. [11] considered eight of these) that had been developed over the previous decades were "rather awkward to calculate," and suggested that "for most practical purposes, a good estimate is provided by the ("not widely known") expression" sqrt[V/(QR)], where V is the sum (here 564) of the (null) score variances

$$\left(436 \times \frac{1 \times 2 \times 1 \times 2}{3 \times 3 \times 2} = 97 \text{ and } 2101 \times \frac{1 \times 2 \times 2 \times 1}{3 \times 3 \times 2} = 467\right)$$

that forms the denominator of the MH test statistic, and Q (= 1492/3) and R (= 1825/3) are the numerator and denominator used to calculate the MH estimate itself (0.82). The much shorter expression also yields a standard error of 0.0431. Thus, in this situation *where the IDR estimate is close to the null*, and the 3 computed SEs are likely to be very close to each other, the choice of the approximate one(s) to calculate when one's laptop is out of power—or one did not bring it on the trip, or the internet is out of reach, or one doesn't have the full data file to run the conditional logistic regression itself—can be based on which of the three formulae one can confidently remember, and has the fewest steps. But, as we will see with a later example, this near-equivalence of the three does not extend over the full IDR range.

Miettinen [12] also considered designs with a fixed matching ratio, M, for M > 2. In such designs, the MLE involves iterative calculations. The advent of programmable calculators in the early 1970s, and spreadsheets in the early 1980s, made calculation of these MLEs less tedious.

Classical methods of analysis of (individually) matched case-base data (1980)

Section 3 of the BD chapter provides a comprehensive treatment. It addresses both the MH and ML estimators for each of the 1:1 (matched pairs), M:1 (M fixed), and 'variable M' designs. Throughout, it uses the (mirror image of the) Miettinen layout, as does chapter 19 of the 1993 textbook by Clayton and Hills [13].

Unlike Miettinen [12], who used the Newton–Raphson approach, BD obtain the MLE as the root of their estimating equation (5.17), which in this example is

$$268 + 613 = 436 \times \frac{2 \times IDR_{ML}}{2 \times IDR_{ML} + 1} + 2101 \times \frac{1 \times RR_{ML}}{RR_{ML} + 2}.$$

Following Clayton's re-expression of the ML estimator when denominators are known [14, 15], this ML estimating equation can be re-arranged in a more familiar 'MH-like' ratio-estimator form:

$$IDR_{ML} = \frac{268 \times 1 \times 1 \div (2 \times IDR_{ML} + 1) + 612 \times 1 \times 2 \div (IDR_{ML} + 2)}{168 \times 1 \times 2 \div (2 \times IDR_{ML} + 1) + 1489 \times 1 \times 1 \div (IDR_{ML} + 2)} \quad [iMH]^{\#}$$

Since the IDR_{ML} appears on both sides, entering the null value IDR = 1 into the expression on the right side yields an initial estimate, IDR = 0.8175. This is the MH estimate. Entering this estimate then gives IDR = 0.8173; the 4 decimal places remain unchanged when one iterates further. So, even though MH did not use a specific criterion to derive it, their estimator turns out to be the first iteration on the way to the ML one obtained by applying conditional logistic regression to the individually matched sets. And in most applications, unless the IDR departs considerably from unity, it provides more than adequate accuracy. The iterated Mantel–Haenszel estimator (iMH) convergences very rapidly; as Clayton notes elsewhere [14], “a single refinement stage is usually all that will be required.”

Miettinen [personal communication] once asked Mantel why he favored his particular estimator among the five estimators he and Haenszel considered in their 1959 paper. He also asked why Mantel chose to sum the *ad*/*n* and *bc*/*n* products, rather than just the *ad* and *bc* products themselves. For example, why not a sixth estimator, Sum(*ad*)/Sum(*bc*), which too would be to the left of the target in 50% of shots, and to the right of it in 50%? Mantel answered that the amount of information concerning the parameter of interest contained in the *ad* and *bc* products is proportional to the square root of the product rather than the product itself, and so he used the divisor of *n* = *a*+*b* + *c*+*d* to ‘slap down’ each product so that it made a more appropriate (and less noisy) contribution. [minute 17 of 2011 interview [16]]. Today, we see how insightful and inspired this was.

In the present application, at the MLE reached by the *iMH* procedure, the final weights are 1/(2 × 0.8173 + 1) = 1/2.6346 and 1/(0.8173 + 2) = 1/2.8173, rather than the common 1/3 for each that MH proposed. Thus the two diagonals contribute to the MLE in the ratio of 436/2.6346 to 2101/2.8173, i.e., approximately 18:82; RT’s estimate uses 17:83.

Proposed classical approximate MLE (2017)

RT [5] also take a binomial-model-based approach, but combined the two independent $\mathrm{IDR}_{\mathrm{ML}}$ estimates, 2 × [612/1489] = 0.8228 and [1/2] × [268/168] = 0.7976, manually using a novel, but somewhat mixed, approach. *Directly in the ratio scale*, they took a weighted average of the two estimates, using weights 0.8281 and 0.1719 proportional to the numbers of informative triplets of each of the two types (2101 and 436), arriving at a summary estimate of 0.8178. They calculated the variance of the log of the summary estimate as if the weighted average itself was computed in the log scale, and used it to calculate a multiplicative margin of error for the summary estimate.

In the situations we have investigated, the CI thus obtained is close to that obtained by the more common approach to combining separate IDR estimates: combining *on the log scale*, using *inverse-variance* (i.e., *information*) weights. Often referred to as ‘Woolf’s method’, this approach implicitly assumes that the estimates being combined are estimates of the same parameter; the combination respects both the minimum-variance and ML criteria; the RT one does not.

MLE and SE of its log: by a familiar logistic regression, or by hand calculator (2018)

RT’s motivations for their approach were that “conditional logistic regression requires accessing the original individual-level database, is vulnerable to programming errors, and provides readers no easy way to verify results.”

If there are relevant within-triplet variables that the investigator wishes to control for by including them in a conditional logistic regression, then at each iteration, the log-likelihood contribution from each of the large number (here 6962 × 3) of rows of data must indeed be computed anew. But *if*, as here, *there are no such variables*, then the conditional logistic regression can be fitted using just M (here just 2) rows of data, using regular (*unconditional*) logistic regression software. As is seen in the R code in Fig. 3, one merely sets up two vectors/columns of length M, one for the numbers (268 and 612) of triplets where the case arose from the index category of the factor at issue (‘Pos’) and one for the numbers (168 and 1489) where it did not (‘Neg’), along with a third one (‘O’) containing the logs of the multiples by which each Pos/Neg ratio would, if the factor were ignored, distort the IDR estimate (here 2 and ½). The logs of these multiples are referred to in generalized models [17] as ‘offsets’, and can be thought of as variates whose regression coefficients are forced to be 1, rather than estimated. They are used to compensate for the fact that the ratio 268/168 is an estimate of 2 × IDR, and that 612/1489 is an estimate of (1/2) × IDR. Including their logs in this simple logit model

$$\log(\mathrm{E[Pos]/[E(neg)]} = \log(\mathrm{IDR}) + 1 \times \log(\mathrm{multiple})$$

allows the antilog of the intercept (exp[-0.20177] = 0.8173) to be directly used as an estimate of the target parameter IDR, and its standard error used to compute a multiplicative margin of error. Moreover, the fitted proportions, 0.6204 and 0.2901, can be applied to the two diagonal totals (436 and 2101) to arrive at the fitted frequencies 270.5 and 609.5, thereby providing a check on the coding, and a test of fit.

Thanks to the iterated MH (iMH) procedure proposed above, epidemiologists who prefer a handheld calculator or spreadsheet and being close to their data now have an easy way to arrive at the 'deluxe' ML estimate and its SE without having to run *any* logistic regression. Here, again using the RT data, are the steps.

1. Use the 'iMH' formula to calculate $\mathrm{IDR}_{\mathrm{ML}} = 0.8173$.
2. Compute the fitted values:
3. $436 \times \frac{2\times 0.8173}{2\times 0.8173+1} = 270.5$ and $2,101 \times \frac{0.8173}{0.8173+2} = 609.5$
4. and their complements 436-270.6 = 165.5 and 2101-609.5 = 1491.5.
5. Compute and sum the amounts of Information about $\log(\mathrm{IDR}_{\mathrm{ML}})$.
6. $I = \frac{270.5\times 165.5}{436} + \frac{609.5\times 1491.5}{2101} = 535.3572.$
7. Calculate the SE of $\log(\mathrm{IDR}_{\mathrm{ML}})$ as 1/sqrt(535.3572) = 0.04322.

The two amounts of information are the reciprocals of the two variance expressions used by RT, but using the *fitted* frequencies rather than the *observed* ones. The form provides a helpful connection between logistic regression and classical methods: the variance for the log(IDR) estimate from a logistic regression fitted to several 2 × 2 tables is merely a 'smoothed' version of the one for Woolf estimate, i.e., with the observed frequencies replaced by their fitted frequencies.

More generally

RT's study [5] addressed an agent (weather) one cannot modify, and was limited to a fixed M = 2 matching ratio. Lest the methods above be seen to be limited to sets that are all of the same size, and in order to show the estimators side by side, I now apply them to a 'ragged' data table from a classic teaching dataset from Breslow and Day's textbook [6], also used in reference 11. It is derived from the Los Angeles Retirement Community Study of the effect of exogenous estrogens on the risk of endometrial cancer. The *study base* consisted of the experience from 1971 to 1975. Although the investigators did not attempt to measure it, it helps to denote its (unknown) size by *B* woman-years (w-y). The *case series* consisted of 63 instances of endometrial cancer. For each case, four woman-moments (base-probes) were selected from the base; this *individually matched base-series* was contributed by women who were alive and living in the community at the time the case was diagnosed, were born within 1 year of the case, had the same marital status, had entered the community at approximately the same time, and had not had a hysterectomy prior to the time the case was diagnosed. Information on the history of use of several specific types of medicines, including conjugated estrogens, was abstracted from the medical record of each of the 63 × 5 women. Since the histories of some (including 4 in the case series) had missing values, the *data-synthesis* was applied to the information from 55 quintuplets and 4 quadruplets. The resulting distribution of the 55 × 5 + 4 × 4 = 291 medication histories, each reduced to a binary—or +, is shown at the left of Fig. 4.

If one ignores the matching, and contrasts the incidence density of 47 cancers among an estimated exposed segment of [(96/232) × B] w-y with the 12 among the estimated unexposed segment of [(136/232) × B] w-y, the IDR result is (47/96)/(12/136) = 5.55.

The calculations involved in, and results obtained by the various estimators applied to the data from the individually matched sets are shown in the remaining columns of Fig. 4.

The Mantel–Haenszel estimate is 5.75. The standard error for its log, obtained by the lengthier Robins-Greenland-Breslow (RGB) expression [11], is 0.38. Not surprisingly, given how far the 5.75 is from 1, the standard error of 0.33 produced by the test-based approach is narrower. Of considerable interest, particularly to those who would like a shorter alternative to the RGB standard error, is the 0.38 produced by the Clayton-Hills expression [13], using nothing more than the (null) variance calculation used in the denominator of the null Chi square test-statistic, along with the numerator and denominator inputs to the MH estimate itself.

The ML result is obtained by finding the IDR value that equates the sums of the observed numbers of sets where the history was positive (45) with the sum of their 'fitted' frequencies. It is both easy and instructive to use trial and error on a spreadsheet to solve the balancing ("estimating") equation. One can then use the fitted proportions to compute (and sum) the amounts of (Fisher) information concerning the log of the IDR provided by each of the six pairs of frequencies. The square root of the reciprocal of the sum provided a standard error (SE) for a multiplicative margin of error.

Also of note is how well the sixty-year-old MH estimator performs, and how it quickly it leads on to the ML result.

The setup in the "logistic regression" column leads directly to a rapid deluxe ML answer, while avoiding trial and error, iterations and conditional logistic regression. This familiar unconditional logistic regression also provides fitted frequencies.

Why is it that these ML results obtained from *conditional* logistic regression, can also be obtained directly from this *unconditional* logistic regression? The answer lies in the *individual* matching, where conditioning on *both* margins of the 2 × 2 table derived from a matched set

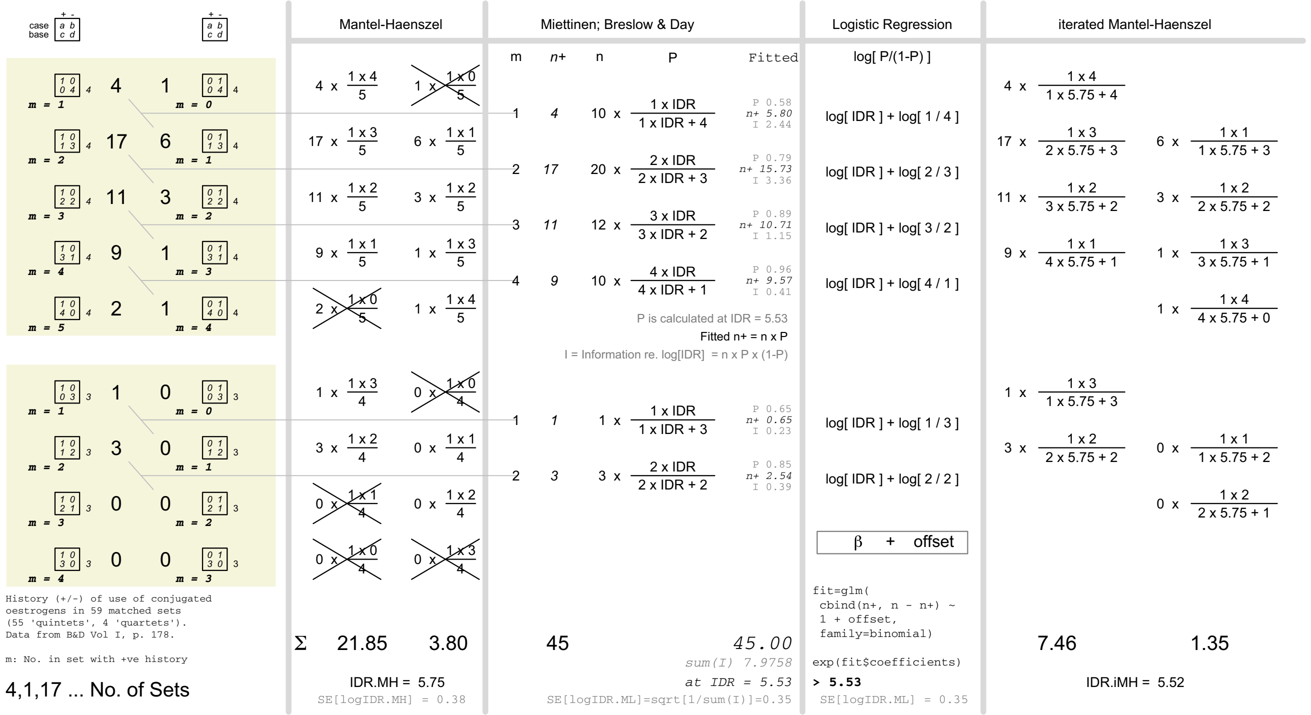


Fig. 4 Data frequencies from an individually matched case-base study widely used in teaching [6], together with IDR incidence density ratio (IDR) results obtained by tabular and regression-based data-synthesis approaches (see text). In the leftmost column, the information from each configuration of M = 5 or M = 4 is shown as a 2 (case or base) × 2 (+ve or −ve history) frequency table; the multiplicities sum to 59. The IDR result (5.53) obtained using the ML criterion was obtained either by solving an estimating equation; or using binomial regression with a logit link, and an offset that is a function of m (the number of positive histories in the set) and M. The weights in the 'iterated MH' estimator further reduce ("slap down") the 'ad' and 'bc' products by more than the divisors of 5 and 4 do in the original MH round. In one further iteration (not shown), the iMH result is the same as the ML result to 3 digits

leads to a *Bernoulli* random variable whose expected value P has a closed form even when the IDR is non-null. A similar simplification is achieved when, with the person-time denominators PT_1 and PT_0 *known*, the distributions of the two independent Poisson random variables C_1 and C_0 can, by conditioning on their sum, be converted to a single Binomial distribution with 'n' = $C_1 + C_0$ and $P = (IDR \times PT_1)/(IDR \times PT_1 + PT_0)$. No such closed form exists for the more general (non-central hypergeometric) random variable that arises from conditioning on both margins of stratified frequency tables where the smallest of the four marginal frequencies exceeds 1 (such data are referred to as 'grouped data' (rather than 'matched' data) in chapter 4 of Breslow and Day's textbook [6]).

Readers will no doubt have noted that in *this* individually matched case-base series, the result from the matched synthesis is (just slightly) closer to the null than the one that ignores it, contrary to repeated warnings about the dangers of ignoring the matching. One can surmise that this is one of those situations where the matching was on variables that do not matter. Although Breslow and Day devoted section 7.3 of their text to the validity and efficiency issues involved in such matched studies, it has taken quite a long time for the principles concerning matching [6, 18–24] to be fully understood, and for the warnings derived from extreme theoretical scenarios to be tempered by what is observed in real life examples.

Discussion

Investigators are increasingly relying on not just big, but also clinically rich, data bases, and thus at little or no additional cost can include 10–20 base-probes per case [25]—enough that virtually all matched sets contribute to the synthesis [cf. Section 7.7 of Volume II of Breslow and Day [26] for formal efficiency considerations]. In such applications, the cost constraints that led Miettinen 'up from matching' [27] are no longer a deterrent.

Given the expanding opportunities for conducting case-base studies with individually matched base series, RT's pursuit of simplicity, transparency and ways to check the analysis is welcome, even if they limited their attention to methods for designs with a 2:1 matching ratio [5]. But it is disappointing that the referees and editors of the Journal of

Clinical Epidemiology [5] were unable to help when RT lamented that “Methods for extending McNemar’s test for analyzing one-to-two matched control studies have been sought but have not been previously developed.” In this example, the near sightedness of modern-day ‘epidemiologic academia’ [28] is all the more remarkable: the overlooked sources were to be found in the very 1980 textbook [6] that RT cited for their ‘simple Mantel–Haenszel analysis of the two McNemar estimates’ whose ‘estimated variance is slightly too narrow because of the double count of some days.’ The present piece is a call to be more far-sighted, and to not forget our still-relevant ‘tabular’ foundations.

If we are to learn from classic articles (RT [5] cite McNemar’s 1947 article [10]), then a relevant one in this IDR context is the 1955 one by Barnet Woolf [9]. It—like three of Miettinen’s—is reprinted in Greenland’s collection [29]. In it, without explicitly mentioning the amount of time during which the case series arose, but with remarkable clarity, Woolf made the case for directly “work[ing] with (i.e. contrasting) incidence rates”. “The data usually do not permit calculation of absolute rates, nor are they needed. What is wanted and readily obtained is an estimate of the ratio of one rate to another.” He denoted by *h* and *k* the numbers of the diseased series in the index and reference categories (blood groups in his example), and by *H* and *K* the corresponding numbers in the control (denominator) series. Then,

> the incidence in the [population time in the index category] will be h/H × some constant, and that in the reference category will be k/K × the same constant. An estimate of the ratio will be hK/Hk, and it may readily be shown that this is the maximum-likelihood estimate.

Today’s epidemiologists might note that *neither Woolf* [9], *not Haldane* [30] *in his refinement a year later, used the term ‘odds ratio*.’ They might even adopt as a principle, stated in the same words used by Woolf: “*nor is it needed*.”

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

1. Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. New York: Wiley; 1985. p. 46.
2. Miettinen OS. Epidemiological research: terms and concepts. Dordrecht: Springer; 2011. p. 115–31.
3. Miettinen O. Estimability and estimation in case-referent studies. Am J Epidemiol. 1976;100:226–35.
4. Breslow NE. Are statistical contributions to medicine undervalued? Biometrics. 2003;59:1–8.
5. Redelmeier DA, Tibshirani RJ. A simple method for analyzing matched designs with double controls: McNemar’s test can be extended. J Clin Epidemiol. 2017;81:51–55.e2. https://doi.org/10.1016/j.jclinepi.2016.08.006. Epub 2016 Aug 24.
6. Breslow NE, Day NE. Statistical methods in cancer research. Volume 1—the analysis of case–control studies. International Agency for research on Cancer. Scientific Publication No. 32. Lyon 1980.
7. This Week’s Citation Classic. No. 26 June 29, 1981. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719–748.
8. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22:719–48.
9. Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet. 1955;19:251–3.
10. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12:153–7.
11. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. Am J Epidemiol. 1986;124:719–23.
12. Miettinen OS. Estimation of relative risk from individually matched series. Biometrics. 1970;26:75–86.
13. Clayton D, Hills M. Statistical models in epidemiology. New York: Oxford University Press; 1993. p. 178.
14. Clayton DG. The analysis of prospective studies of disease aetiology. Commun Stat Theory Methods. 1982;11:2129–55.
15. Clayton D, Hills M. Statistical models in epidemiology. New York: Oxford University Press; 1993. p. 144.
16. Miettinen O. A conversation with Olli Miettinen. Interview by James A. Hanley. Epidemiology. 2012;23:495–9. https://doi.org/10.1097/EDE.0b013e31824968b9.
17. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A (Gen). 1972;135(3):370–8.
18. Cole P, Mack T, Rothman K, Henderson B, Newell G. Tonsillectomy and Hodgkin’s disease (Letter). N Engl J Med. 1973;288:634.
19. Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case–control studies estimate? Survey of methods and assumptions in published case–control research. Am J Epidemiol. 2008;168. https://doi.org/10.1093/aje/kwn217. Advance Access publication September 15, 2008.
20. Vandenbroucke JP, Pearce N. Case–control studies: basic concepts. Int J Epidemiol. 2012;41:1480–9. https://doi.org/10.1093/ije/dys147.
21. Vandenbroucke JP, Pearce N. Incidence rates in dynamic populations. Int J Epidemiol. 2012;41:1472–9. https://doi.org/10.1093/ije/dys142.
22. Pearse N. Analysis of matched case–control studies. BMJ. 2016;352:i969. https://doi.org/10.1136/bmj.i969.
23. Mansournia MA, Jewell NP, Greenland S. Case–control matching: effects, misconceptions, and recommendations. Eur J Epidemiol. 2018;33:5–15.
24. Pearse N. Bias in matched case–control studies: DAGs are not enough. Eur J Epidemiol. 2018;33:1–4. https://doi.org/10.1007/s10654-018-0362-3(0123456789(),-volV)(0123456789().,-volV).
25. Filion KB, et al. A multicenter observational study of incretin-based drugs and heart failure. N Engl J Med. 2016;374(12):1145–54.

26. Breslow NE, Day NE. Statistical methods in cancer research. Volume 2—the design and analysis of cohort studies. International Agency for research on Cancer. Scientific Publication No. 82. Lyon 1987.
27. Miettinen OS. Theoretical developments. In: Holland WW, Olsen K, Florey C, editors. The development of modern epidemiology: personal reports from those who were there. Print publication date: 2007. Print ISBN-13: 9780198569541. Published to Oxford Scholarship Online: September 2009.
28. Miettinen OS. On progress in epidemiologic academia. Eur J Epidemiol. 2017;32(3):173–9. https://doi.org/10.1007/s10654-017-0227-1 **(Epub 2017 Mar 8)**.
29. Greenland S. Evolution of epidemiologic ideas: annotated readings on concepts and methods. Epidemiology Resources. Chestnut Hill, MA. 1987.
30. Haldane JBS. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet. 1956;20:309–11.