Statistical Methods in Medical Research 2008; **00**: 1–26

Efficient sampling approaches to address confounding in database studies

James A Hanley, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada and **Nandini Dendukuri,** Technology Assessment Unit, McGill University Health Centre, and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Administrative and other population-based databases are widely used in pharmacoepidemiology to study the unintended effects of medications. They allow investigators to study large case series, and they document prescription medication exposure without having to contact individuals or medical charts, or rely on human recall. However, such databases often lack information on potentially important confounding variables. This review describes some of the sampling approaches and accompanying data-analysis methods that can be used to assess, and deal efficiently with, such confounding.

1 Introduction

Administrative and other population-based databases are widely used in pharmacoepidemiology to study, nonexperimentally, the unintended – and sometimes too the intended – effects of medications. They allow investigators to assemble large case series, and to document medication exposure (using prescriptions issued or filled). This can be done without having to contact individuals, or their clinical records, or to rely on after-the-adverse-event recall by those in the case series and after-some-time recall by those in the study base, or in a denominator (‘control’) series formed from it.

In the study of intended effects, ‘confounding by indication’ can make the nonexperimental approach infeasible.[1] In the study of unintended effects, contraindications pose a lesser threat, since they are less common, and possibly previously unrecognised.[1] Nevertheless, patients who take a specific medication may have other medical conditions, and take other medications for these, or have behaviours that lead to or protect against the event of interest. Thus, it may be difficult to disentangle unintended effects of the medication of interest from those caused by these other conditions, medications and behaviours. This challenge is all the greater in database research, where information on these ‘confounding’ factors is either not available in the databases themselves, or recorded with considerable imprecision. The large numbers of instances of the event of interest that can be studied via databases may

Address for correspondence: James A Hanley, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada.
E-mail: james.hanley@mcgill.ca

10.1177/0962280208096046
Los Angeles, London, New Delhi and Singapore

well lead to interval estimates of effect that are more precise (narrower) than those available from traditional etiologic studies, but the failure to control for unmeasured or poorly measured confounders may well mean that these precise estimates are 'precisely wrong.'

This review describes some of the approaches that may be used to deal with this problem when access to the information on these confounding factors is limited. Section 2 provides some general orientational remarks on confounding and imprecision, and Section 3 discusses how much it 'costs' – in sample size and variance terms – to deal with confounding. Section 4 briefly reviews some strategies which use external parameter estimates or supplementary individual-level data to quantify, and 'correct' for the potential impact of, suspected confounding. Section 5 deals with situations where supplementary data can be obtained on some of the individuals in the database study. The presentation is not intended to be a comprehensive and highly technical review of the considerable recent work in this area; rather it is aimed primarily at statisticians and epidemiologists who have little familiarity with this topic.

2 Confounding and imprecision – general considerations

2.1 Confounding

The term 'confounding' will be used to denote 'one particular form of the confusion of two effects: the confusion due to extraneous causes, i.e. other factors that really do influence disease incidence, e.g. age, sex, habits or living circumstances.'[2] As a point of departure, we consider first the simplest situation: a large study base of population time, involving a binary exposure E, a binary event indicator Y and a single binary confounder C. What is the impact of ignoring, or being unable to control for, C? To quantify the distortion, suppose that the theoretical *incidence density* (ID) of $Y = 1$ (i.e. the *event rate* one would observe in an infinite amount of experience) follows the multiplicative relationship

$$\mathrm{ID}(Y = 1|C = c, E = e) = \mathrm{ID}_{00} \times \{\psi_C\}^c \times \{\psi_E\}^e. \tag{1}$$

where ID_{00} is the ID in the $(C = 0, E = 0)$ category, ψ_C is the (common) ID ratio contrasting the ID in the $(C = 1, E = 0)$ versus $(C = 0, E = 0)$ category, and in the$(C = 1, E = 1)$ versus $(C = 0, E = 1)$ category, and ψ_E is the (common) ID ratio contrasting the ID in the $(E = 1)$ versus $(E = 0)$ category, within each level of C. Let $P_{C=1|E=1}$ denote the proportion of the exposed (i.e. $E = 1$) population time for which $C = 1$, and $P_{C=1|E=0}$ the corresponding proportion within the unexposed population time. The parameter of interest is

$$\psi_E = \frac{\mathrm{ID}(Y = 1|E = 1, C)}{\mathrm{ID}(Y = 1|E = 0, C)}, \tag{2}$$

which, as is implied by (1), is assumed to be homogeneous across both levels of C. In an infinite amount of experience, this is the (theoretical) value we would observe if we restricted our attention to a domain where there was no variation in C, or if C did not independently affect the ID, and we aggregated the experience across the two levels of C. But what if neither of these two conditions applied?

Suppose one ignores – or is unable to measure – C. In the same infinite amount of experience, how close would the 'crude' ID ratio, $\psi_{E-\text{crude}}$, be to the quantity of interest, ψ_E? The relationship, given in equation 3.1 in Breslow and Day,[3] can be re-written as

$$\psi_{E-\text{crude}} = \psi_E \times \frac{1 + (\psi_C - 1)P_{C=1|E=1}}{1 + (\psi_C - 1)P_{C=1|E=0}}. \tag{3}$$

Thus, for example, in what some authors call 'positive' confounding, if $\psi_C > 1$ and $P_{C=1|E=1} > P_{C=1|E=0}$, then $\psi_{E-\text{crude}} > \psi_E$. The second term on the right hand expression is often referred to as the *confounding ratio*.[4] In addition to establishing *both* $\psi_C \neq 1$ *and* $P_{C=1|E=1} \neq P_{C=1|E=0}$ as the mathematical conditions for confounding, the ratio shows how the magnitude of the distortion of ψ_E is determined jointly by the degree to which ψ_C deviates from 1 and by which $P_{C=1|E=1}$ differs from $P_{C=1|E=0}$. Breslow and Day[3] tabulated values of the confounding ratio for ψ_C = 2, 5 and 10, and a broad range of values of $P_{C=1|E=1}$ and $P_{C=1|E=0}$. Ratios for less extreme situations, such as might be expected in database studies, are tabulated here.

ψ_C	2 →								4 →							
$P_c = 1 \vert E = 0$	0.1		0.2		0.3		0.4		0.1		0.2		0.3		0.4	
$P_c = 1 \vert E = 1$	0.2	0.4	0.3	0.6	0.4	0.6	0.5	0.8	0.2	0.4	0.3	0.6	0.4	0.6	0.5	0.8
Ratio	1.1	1.3	1.1	1.3	1.1	1.2	1.1	1.3	1.2	**1.7**	1.2	**1.8**	1.2	**1.5**	1.1	**1.5**

One notices that a single binary confounder needs to be strong, present in a sizable fraction of the base, and highly correlated with the exposure, in order for its omission to distort the true ID ratio, ψ_C, by 50% or more.

If, in relationship (3), we consider a '*null*' situation where $\psi_E = 1$ but $\psi_C > 1$, and if we divide the numerator and the denominator of the confounding ratio by $(\psi_C - 1)$, we obtain the inequality

$$\frac{P_{C=1|E=1}}{P_{C=0|E=1}} > \frac{1/(\psi_C - 1) + P_{C=1|E=1}}{1/(\psi_C - 1) + P_{C=1|E=0}} = \psi_{E-\text{crude}}.$$

This inequality, with E in place of the A in his 1959 article,[5] C in place of B, and $\psi_{E-\text{crude}}$ in place of r, is a re-statement of Cornfield's fundamental, but often overlooked, result: 'If an agent E, with *no causal effect upon the risk of a disease*, nevertheless, because of a positive correlation with some *causal agent*, C, shows *an apparent risk*, $\psi_{E-\text{crude}}$, for those exposed to E, relative to those not so exposed, then the prevalence of C, among those exposed to E, relative to the prevalence among those not so exposed, must be greater than $\psi_{E-\text{crude}}$. [*italics ours*].' This inequality can also be used to put bounds on the distortion in non-null situations, e.g. when $\psi_E > 1$. When C is multi-valued, or multi-dimensional, the magnitude of the distortions created by not being able to account for it is more difficult to quantify.

Although it may seem like a digression from our topic, Section 2.2 is included to emphasise that there is another component in the $\log\{\widehat{\psi}\}$ derived from most non-experimental studies:

$$\log\{\widehat{\psi}\} = \log\{\psi\} + \log[\text{Confounding Ratio}] + \text{Random Error}.$$

Unless this second component is small, the effects of confounding cannot be isolated.

2.2 Imprecision of statistical estimates

Even when we have access to information on C, and can – via the study design, and in the data-analysis by stratification or statistical modelling – correct for the distortion it would otherwise produce, the resulting *empirical* ID ratio, $\widehat{\mathrm{IDR}}$, or $\widehat{\psi_E}$, derived from any one study will differ from the theoretical ('true') ratio, ψ_E. Some of the difference between $\widehat{\psi_E}$ and ψ_E has to do with factors such as the imperfections in the statistical model used to approximate the true biological situation, and imperfections in the recorded data; its magnitude tends to be unrelated to the size of the case series, and is quite difficult to quantify. What is more readily quantifiable is the expected amplitude of $\widehat{\psi_E} - \psi_E$, or of $\log\{\widehat{\psi_E/\psi_E}\}$: it is governed by the statistical laws that generate the observed *numbers* of exposed (subscript$_1$) and unexposed (subscript$_0$) *cases*, c_1 and c_0, that occur within the two segments of population-time (PT_1 and PT_0) comprising the study base. The c_1 and c_0 comprise the *case series*.

Consider the first simplest situation, where we are dealing with, or can restrict attention to, an otherwise homogeneous study base (i.e. one in which there are no confounding factors), where we know the relative sizes of PT_1 and PT_0, and can assume that c_1 and c_0 are realisations of two independent Poisson random variables. Thus, $\widehat{\psi_E} = (c_1/\mathrm{PT}_1) \div (c_0/\mathrm{PT}_0)$. For such studies, the (large-sample) variance of $\log\{\widehat{\psi_E}\}$ about $\log\{\psi_E\}$ is given by

$$\mathrm{Var}[\log\{\widehat{\psi_E}\}] = 1/\mu_{c_1} + 1/\mu_{c_0}, \tag{4}$$

where μ_{c_1} and μ_{c_0} are the expected numbers of exposed and unexposed cases, respectively.

In many investigations, the absolute or relative sizes of PT_1 and PT_0 that generated these cases are not known. In such situations, a *denominator* series (traditionally referred to as a *control* series), formed from a representative sample, of size d say, of the person-moments in the base, is used to *estimate* their relative sizes. This $\mathrm{PT}_1 : \mathrm{PT}_0$ ratio is then estimated as $d_1 : d_0$, where d_1 and d_0 are the observed numbers of exposed and unexposed (person) moments in the denominator series. From these *quasi-denominators*, and from the numerators c_1 and c_0, the ID ratio ψ_E is estimated as

$$\widehat{\psi_E} = (c_1/\widehat{\mathrm{PT}_1}) \div (c_0/\widehat{\mathrm{PT}_0}) = (c_1/d_1) \div (c_0/d_0) = (c_1/c_1) \div (d_1/d_0).$$

The probability distribution of the $d_1 : d_0$ ratio is governed by the binomial law, and so the large-sample variance of $\log\{d_1/d_0\}$ is $1/\mu_{d_1} + 1/\mu_{d_0}$, where μ_{d_1} and μ_{d_0} are the expected numbers of exposed and unexposed persons in the control series. Thus the variance of $\log\{\widehat{\psi_E}\}$ about $\log\{\psi_E\}$ is given by

$$\text{Var}[\log\{\widehat{\psi_E}\}] = 1/\mu_{c_1} + 1/\mu_{c_0} + 1/\mu_{d_1} + 1/\mu_{d_0}. \tag{5}$$

Epidemiological textbooks usually give its empirical version, i.e. the *estimated* variance, obtained by substituting the observed frequencies c_1, c_0, d_1, and d_0, (or a, b, c and d, as the entries in the 2×2 table are usually called) for the expected quantities. Variations on this fundamental formula provide the basis for measuring many of the statistical 'economies' described in this review. The original variance formula for a single stratum goes back at least as far as Yule, in connection with the cross-product measure of association; today it is usually associated with Woolf,[6] who used it – without derivation – in his 1955 article: for illustration, he combined results from three cities (strata) on the association between blood group and peptic ulcer. Incidentally, his classic article[7] took a remarkably modern approach to rate ratio estimation, using *unrelated* numerator (case) and denominator (control) series. He used h and k (our c_1 and c_0) for the numerators for (i.e. the numbers of cases in) the index and reference blood groups, and H and K (our d_1 and d_0) for the corresponding values from the denominator series. Sadly, it is still common for textbooks to teach that case-control studies *compare cases and controls* with respect to 'exposure odds', i.e. to compare the $h : k$ and $H : K$ ratios. Woolf compared the index and reference categories of the determinant with respect to incidence rates, i.e. he compared h/H and k/K via their ratios. In so doing, he recognised the essence of *the* etiologic study,[8–10] where the main conceptual difference between so-called 'cohort' and 'case-control' approaches is that the former uses *known* population-time denominators, whereas the latter uses an *estimated* ratio of these denominators – and pays an extra price, in terms of statistical variance, for doing so.

Several important design considerations flow from 'Woolf's formula.' If the effort and study budget is proportional to, or otherwise constrained by the size of the case series, $c = c_1 + c_0$, then the more symmetric the $c_1 : c_0$ split, the smaller the variance component $1/c_1 + 1/c_0$. In some situations, the reduction in variance can be achieved by manipulating the magnitudes of PT_1 and PT_0. If the study is constrained by the available amount of exposed experience, PT_1 (and thus by the c_1 it generates), but not seriously by the amount PT_0, then continuing to increase PT_0 so that $c_0 >> c_1$ will continue to reduce the variance component, but according to a 'law of diminishing reductions,' as is evident by evaluating the series $1/c_1 + 1/\{1, 2, 3, \dots\} \times c_0$.

A similar law of diminishing returns prevails when one needs to estimate the $\text{PT}_1 : \text{PT}_0$ ratio using a denominator series of size d. Again, if – as is commonly the situation – the study is constrained by the size of the case series, c, then continuing to increase d so that $d >> c$ will continue to reduce the variance component $1/d_1 + 1/d_0$, but according to the law of diminishing reductions given by the series $(1/c_1 + 1/c_0) + (1/d_0 + 1/d_0)/\{1, 2, 3, \dots\}$. In Volume I,[3] and again in the Design Considerations chapter in Volume II,[11] Breslow and Day study the impact of the $d : c$ ratio in various circumstances.

When the exposure of concern is uncommon, so that $\mathrm{PT}_1 << \mathrm{PT}_0$ and $c_1 << c_0$, and the $\mathrm{PT}_1 : \mathrm{PT}_0$ ratio is known, or can be estimated with such a large denominator series that $(1/d_1 + 1/d_0)$ adds a negligible variance component, then $\mathrm{Var}[\log\{\widehat{\psi_E}\}]$ is dominated by $1/c_1$. We can use this dominant component to derive the *minimum* margin of error associated with $\widehat{\psi_E}$, as a function of the expected number of exposed cases, μ_{c_1}, and thus – by back-calculation – of the expected *total* number of cases.

Expected number of exposed cases, μ_{c_1} :	50	100	200	300	400	500
Min. 95% Margin of Error* for $\widehat{\psi_E}(\times/\div)$:	1.32	1.22	1.15	1.12	1.10	1.09
Total no. of cases, c, if $\psi_E \times \mathrm{Prev}(E) = 0.1$:	500	1000	2000	3000	4000	5000

*Minimum Margin of Error $= \exp\{1.96 \times (\mathrm{Var}[\log\{\widehat{\psi_E}\}])^{1/2}\} = \exp\{1.96 \times (1/\mu_{c_1})^{1/2}\}$

Thus, if a medication is used by a proportion $P(E) = 0.1 = 10\%$ of a population, but has a null effect on the rate of adverse events, a population base that generated 1000 events could nevertheless – with a 5% probability – lead to apparent ID ratio estimates as low as $\widehat{\psi_E} = 1 \div 1.22 = 0.82$ and as high as $\widehat{\psi_E} = 1 \times 1.22 = 1.22$. Likewise, if a medication is used by 5% of a population, but doubles the rate of adverse events, a population base that generated this same number of events could lead to apparent ID ratio estimates as low as $\widehat{\psi_E} = 2 \div 1.22 = 1.64$ and as high as $\widehat{\psi_E} = 2 \times 1.22 = 2.44$. Larger numbers of cases (events) than those tabulated above are required for smaller $P(E) \times \psi_E$ products, for narrower margins of error, for studies where E is more common (the variance component $1/\mu_{c_0}$ was omitted in the above calculations), or if the $\mathrm{PT}_1 : \mathrm{PT}_0$ ratio has to be estimated using a denominator series (since the variance component $1/\mu_{d_1} + 1/\mu_{d_0}$ was also omitted in the above table). The larger c and d numbers can be calculated by replacing the $1/\mu_{c_1}$ in the minimum margin of error formula by $(1/\mu_{c_1} + 1/\mu_{c_0})$, or by $(1/\mu_{c_1} + 1/\mu_{c_0} + 1/\mu_{d_1} + 1/\mu_{d_0})$, if these additional components are non-negligible.

3 Traditional control of confounding: and how much does it cost?

Before multiple logistic regression and Cox regression methods became readily available, the classical methods for the control of confounding were restriction and stratification. The classical methods for the analysis of stratified data from 'cohort' and 'case-control' studies (i.e. from etiologic studies with known and estimated PT denominators, respectively) are well described in the early chapters in the two Breslow and Day volumes.[3,11] When the data for each stratum are plentiful, the observed stratum-specific ID ratio (or difference) estimates can be combined to form an estimate of the (presumed common) summary measure, using the inverses of the estimated variances of the stratum-specific estimates as weights. In the worked example in Woolf's article,[6] the data were stratified by city. The stratum-specific numerators (c: numbers of *c*ases of peptic ulcer) and estimated relative sizes of the underlying population-time 'denominators' (d: numbers of persons in the sample of the base) in the index (Group O, $_{\text{subscript1}}$) and reference (Group A, $_{\text{subscript0}}$) blood groups were in the hundreds and thousands, respectively:

City:	London				Manchester				Newcastle			
c's & d's:	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0
number:	911	4578	579	4219	361	4532	246	3775	396	6598	219	5261

Thus the calculation of a weighted average of the logs of the stratum-specific $\widehat{\psi}$'s, i.e. of [log({911/4578} ÷ {579/4219}) =]0.3716, 0.2008 and 0.3659, all three of which are statistically stable, and using stratum-specific information weights $I_{\text{stratum}} = 1/\hat{\text{V}}\text{ar}(\log\{\widehat{\psi_{E-\text{stratum}}}\}) = [1/(1/911 + 1/4578 + 1/579 + 1/4219 =]304.9$, 136.6 and 134.5, that are also stable, poses no particular problems.

When, however, the data for each stratum are sparse, then – even if there are many such strata – a more careful approach is required in order to avoid unstable effect estimates. The ratio estimator devised by Mantel and Haenszel[12] does not calculate a separate $\hat{\psi}_E$ for each stratum. Instead it calculates a single ratio: the numerator, num_{MH}, of this ratio is the sum, $\sum$, over the strata, of *carefully scaled*, i.e. *downweighted* stratum-specific products, i.e. $\sum(c_1 \times d_0) \times w$, while the denominator den_{MH}, is the corresponding sum $\sum(c_0 \times d_1) \times w$. In the worked example in their classic article, the 12 strata consisted of the combinations of 3 occupations and 4 age groups of women. The 12 stratum-specific numerators (numbers of *c*ases of epidermoid and undifferentiated pulmonary carcinoma) for the index category (+1 pack a day smokers) – denoted by A's by Mantel and Haenszel – ranged from 0 to 4, and amounted to 18 in total. The 12 stratum-specific denominators (numbers in the denominator sample of the base) for the index category – denoted by C's by Mantel and Haenszel – ranged from 0 to 3, and amounted to just 13 in total. The size of the entire case series was $\sum c = 31$, while the entire denominator ('control') series was nine times larger, i.e. $\sum d = 282$. Clearly, Woolf's approach is not suitable for such sparse data: in 6 of the 12 strata, $\log\{\hat{\psi}_E\} = \pm\infty$, its estimated variance is infinite, and thus the associated weight in the weighted average is zero. The Mantel-Haensel summary measure, $\hat{\psi}_E = \text{num}_{\text{MH}} \div \text{den}_{\text{MH}} = 12.825 \div 1.201 = 10.68$ is more stable, and uses information from the 11 informative strata. Estimators for the variance of the summary measure $\log\{\hat{\psi}_E\}$ for the situations where the stratum-specific population-time denominators PT_1 and PT_0 (i) are known (ii) have to be estimated by stratum-specific 'denominators' (control series) have been developed by Breslow,[13] and Robins, Greenland and Breslow,[14] respectively.

A more detailed analysis of the estimator of the variance of Woolf's weighted average of log ratios provides considerable insight into the price, in terms of variance, of adjusting for confounding by combining stratum-specific estimates. Even though in practice, with finite experience, one would not know the true value of ψ_C, we can also examine the price of taking a weighted average when in fact its true value was unity, so that there was no need to control for C. Denote the variance of the stratum-specific log ratio, i.e. the sum of the reciprocals of the four expected frequencies, by V_{stratum}, and the corresponding inverse variance, the Information, by $I_{\text{stratum}} = 1/V_{\text{stratum}}$. The summary estimate is the information-weighed average $\sum(I_{\text{stratum}} \times \log\{\hat{\psi}_{E\text{stratum}}\})/\sum I_{\text{stratum}}$, and its variance is $1/\sum I_{\text{stratum}}$. As a comparison, one can consider the situation where, because of restriction, or otherwise, there was no confounding to be concerned with, and an unstratified analysis was performed, but where the number of cases was the same as the overall number of cases in the stratified study (if denominators are known) or both the numbers of cases and of 'controls' were the same (if the population-time denominators need to be estimated). Generally, but not invariably, the variance for the summary (weighted average) estimator is higher than the variance in the unstratified study of the same size. The additional variance cost, or conversely, the additional sample size needed to maintain the same variance, typically increases with the degree of confounding, i.e. with how far ψ_C departs from 1, and how far apart $P_{C=1|E=1}$ and $P_{C=0|E=1}$ are. One can get a sense of this from a classic confounding example,[15] where the strata are based on the women's ages at the time of the initial survey, the subscripts $_1$ and $_0$ denote smokers and non-smokers, and c's and d's are the respective numbers who died in, and survived, the ensuing 20 years.

Age:	18–44				45–64				65+				Unstratified			
	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0
no.:	19	269	13	327	78	167	52	147	42	7	165	28	139	443	230	502

Because smokers tended to be from the younger generations, the crude odds ratio (OR) contrasting the odds of death in smokers and non-smokers is $(139 \times 502) \div (230 \times 443) = 0.68$, suggesting that death rates are *lower* in smokers than non-smokers. The variance of the crude log OR is $(1/139 + 1/443 + 1/230 + 1/502) = 0.016$. In contrast, the Mantel–Haenszel summary OR of 1.36 suggests the opposite effect, an example of 'Simpson's paradox.' The variance of this log OR is $1/\{(1/19 + 1/269 + 1/13 + 1/327)^{-1} + (1/78 + 1/167 + 1/52 + 1/147)^{-1} + (1/42 + 1/7 + 1/165 + 1/28)^{-1}\} = 0.029$, more than 80% higher than the 0.016 for the log of the crude ratio. The factors that increase the variance have been examined in detail;[16,17] we will return to them in Section 5.

Nowadays, multiple logistic regression and Cox regression methods are often used instead of the classical stratified-data methods for the control of confounding. These regression methods can also be used for 'cohort' and 'case-control' studies that match at the design stage on some of the dimensions of C. Again, these methods for data from such studies are well described in the later chapters in the two Breslow and Day volumes. When the data for each stratum are plentiful, an ID ratio estimate can be obtained by unconditional regression models – Poisson if PT values are known, and logistic if they have been estimated – by representing the information in C as a set of regressor variates. When the stratum-specific data are sparse, and arise from a case-control design, Breslow and Day[3] (p. 250) show that using a separate 'intercept' for each stratum in an unconditional logistic regression model leads to biased estimates of ψ_E, and one must instead resort to conditional logistic regression, where by using stratum-specific likelihood contributions, the effect of the matching factor is not estimated.

The 'Woolf-like' structure of the variance estimator for the $\widehat{\psi_E}$ derived from an unconditional regression model is not widely appreciated. Since it is relevant to the methods described in Section 5, a brief example is given here, using as an illustration the data used by Woolf.[6]

	Stratum-specific, & overall, Information (I)					Var[log $\widehat{\psi_E}$]
	London	Manchester	Newcastle	I_{Overall}	$\widehat{\psi_E}$	($1/I_{\text{Overall}}$)
Crude:	–	–	–	589.4610	1.3482	0.001696465
Woolf:	304.8530	136.5994	134.5333	575.9857	1.3906	0.001736154
Logistic:	306.8512	133.2757	135.7605	575.8874	1.3913	0.001736450

In each instance, the information (I) was calculated as the sum of the reciprocals of four frequencies – the *overall* ones in the crude analysis, the *observed* stratum-specific ones in the Woolf approach, and the *fitted* stratum-specific ones obtained from the four parameter logistic regression model that contained an intercept for the reference city, two indicator variables for the other two cities, and an indicator for the index blood group. In this example, there was minimal confounding, and thus the variances for the adjusted ID ratios are only slightly higher than the one accompanying the crude estimate. Unless there is considerable confounding, the expected overall frequencies can often be used as a rough guide to project, at the design stage, the expected variance for the logistic-regression adjusted estimates.

When there is extreme confounding, such as the Simpson's paradox example,[15] the variance of the adjusted log OR can be substantially higher than that for the crude one (recall that the the variance of the crude log OR was 0.016). A four parameter logistic regression model that contained an intercept for the reference age-group, two indicator variables for the other age-groups, and an indicator for smoking, yielded an OR of 1.36; the variance of its log was 0.029, identical to three decimal places to that obtained for Woolf's weighted average of log OR. Thus, the cost of adjustment in this extreme example amounts to a variance – or sample size – inflation of $\sim$ 80%.

3.1 Confounding, matched sets, and unused information

Some studies use matching, rather than regression, as the primary control for key confounders. In such studies, it is important that investigators not spend resources on the acquisition, cleaning and computerisation of data that ultimately make no contribution to the adjusted – or unadjusted – estimate of the ID ratio. This issue arises in both 'matched cohort', and 'case-crossover' studies. Two examples will illustrate. The first of these[18] is a matched cohort study which included 48,857 persons with food-borne infections, each one matched with 10 non-infected persons, matched for age, sex, and county of residence. The authors compared the mortality of the infected and non-infected, 'using conditional proportional hazard regression,' effectively conditional logistic regression. A comorbidity index was included as an important, but unmatched, confounding variable. The authors reported their comparisons using relative mortality ratios (effectively hazard, or ID ratios), over the entire 12 month follow-up window, and – because of the sharp decline in this ratio over the follow-up period – in several sub-windows. 'Elevens' (i.e. matched sets) in which there was no event (death) during the time window do not contribute to the (partial) likelihood, and so could have been omitted from the analysis. If each of the deaths (4707 in all) occurred in a different matched set (and probably most did), then $48,857 - 4707 = 44,150$ (i.e. $>90\%$) of the matched sets did not contribute to the fitted mortality ratio. Since there was no specific mention of it in their report, it appears that the authors did not take advantage of this considerable potential economy. Even though they were largely obtained from administrative databases, the marginal cost of obtaining and processing these uninformative 441,500 records was hardly minimal. As Walker[19] emphasises, if obtaining important exposure or covariate information involves substantial unit costs, these should be expended on the informative, i.e. event-containing, matched sets. The accompanying commentary on the matched cohort study pointed out that 'cohort studies usually have to be very large to obtain a sufficient number of outcome events.' To this, one might add 'once the large number of events has been generated, we should use both the exposure and confounder data in the most cost-efficient and statistically-efficient way.'

This economy *was* exploited in a study of the risk of percutaneous injuries among more than 2000 medical interns during standard and extended work shifts.[20] To assess the relationships between injury risk and either time of day or duration of work, the authors 'used a within-person [...] design in which each participant acted as a separate stratum, and a combined OR was generated using a Mantel–Haenszel [summary ratio]. Because each participant acts as his or her own control, the [...] study design eliminated the need to account for potential between-subject confounders such as differences in age, sex or medical specialty.' Thus, the data-preparation could be limited to the within-participant information just for the approximately 200 participants who suffered a percutaneous injury.

One variation on this design has become known as the *case-crossover* design, a term generally ascribed to Maclure,[21] although variations on this design have been used in epidemiology for quite some time (e.g. Haddon[22]), and by individuals since time immemorial to investigate the origins of rashes, headaches, computer crashes and other untoward personal events. The statistical analyses are just as in the matched case-control examples discussed above. We wonder why the design – or at least the analysis – was not given the more informative 'self-paired case control' label.

In these two examples, the efficiency stemmed from ignoring the sets (elevens in the infections study, interns in the injuries study) in whom there were no events (deaths, injuries). The case-crossover study also ignores cases in whom there is no variation in exposure, unless there are 'within-set' confounding variables – in which situation the exposure-concordant sets contribute to estimating their effects, and thus ultimately to the control of confounding.

4 `External' control of, or allowance for, confounding

Section 3 described instances where the information on potential confounding variables *was readily available*, but either *did not matter*, or – even if the variable were important – the data for the majority of the matched sets were *uninformative*. But what of the opposite situation, where the data on potential confounders *might well matter*, but it would be prohibitively costly, or maybe even impossible, to obtain them on all subjects?

Restricting the study population, thereby increasing homogeneity, is often the first line of defense against the effects of potential confounding factors. Schneeweiss *et al.*[23] recently illustrated the benefits of this approach in pharmacoepidemiologic database studies: they compared the results of using increasing levels of restriction with those from randomised trial results. They found that 'restricting to incident drug users, similar comparison groups, patients without contraindication, and to adherent patients was a practical strategy, which limited the effect of confounding, as these approaches yield results closer to those seen in trial results.'

Another option is to conduct formal sensitivity i.e. 'what if' , analyses, using external information, or external expert opinion. In the simplest situation, where both E and C are unidimensional binary variables, outside estimates – subjective or objective – of the confounding ratio in expression (3) can help to adjust the point and interval estimates obtained. A Bayesian analysis can be used to combine the sampling variability in the statistical estimate of $\widehat{\psi}_{E-\text{crude}}$ with the uncertainty in the elicited, or literature-based, estimates of the ID ratio parameter ψ_C and of the prevalence proportions $P_{C=1|E=0}$ and $P_{C=1|E=0}$. If the latter proportions are difficult to document/estimate directly, it may be more feasible to substitute estimates of their components into $P_{C|E} = (P_{E|C} \times P_C)/P_E$. In many situations, there will be more than one potential confounding variable, and so expression (3) would need to be extended accordingly. However Greenland[24] provides arguments why, under certain causal models, unmeasured confounding be modelled via a single latent variable. Section 3.7 of his article provides a worked example of how to implement this. The article also describes how to deal with other sources of uncertainty, such as possible uncontrolled associations of exposure and disease with selection and participation (sampling and response biases) and – especially important for database studies in pharmaco-epidemiology – the effects of measurement errors.

Schneeweiss[25] describes four approaches to sensitivity analyses to investigate the impact of residual confounding in pharmacoepidemiologic studies that rely on health care utilisation databases (1) sensitivity analyses based on an array of informed assumptions; (2) analyses to identify the strength of residual confounding that would be necessary to explain an observed drug-outcome association; (3) external adjustment of a drug-outcome association given additional information on single binary confounders from survey data using algebraic solutions; (4) external adjustment considering the joint distribution of multiple confounders of any distribution from external sources of information using propensity score calibration. Given the availability of easy-to-apply techniques, he like Greenland, advocates greater use of formal sensitivity analyses, rather than the current culture of qualitative discussions of residual confounding. The articles by Strümer *et al.*[26] and by Schneeweiss *et al.*[27] provide good examples of some of these approaches.

Methods that use individual validation data – internal or external – to correct the parameter estimate of interest for measurement errors in the regressor variables have been available for some time.[28] Although, they have not been widely used elsewhere, they can be particularly helpful for database studies, where the limited information on each person in the database can be thought of as an imperfect version of what would have been provided by the full set of relevant covariates. The essence of this correction (often called regression calibration) can be illustrated using the simplified case-control example that will be used extensively in Section 5. If the investigator only had access to the dichotomised version (C^*) of the quantitative confounding variable (C) – the subject's background risk – on each of the 44,199 study subjects, the data would yield $\log\{\widehat{\psi_E}\} = 0.545$, and $\log\{\widehat{\psi_{C^*}}\} = 1.12$. Suppose that we estimated, from an auxiliary sample, that the correlation between C^* and C was 0.80, and that $\hat{C} = 0.35 + 0.096 \times E + 1.08 \times C^*$. Then, the regression calibration would yield $\log\{\widehat{\psi_E}\} = 0.545 - 0.096 \times 1.12 \div 1.08 = 0.445$. This lower value reflects the greater control of residual confounding by using a more refined version of the confounding variable.

If there are several confounding variables, they can be converted to a scalar quantity for each study subject via a propensity score. The score is the probability, estimated using a logistic regression model based on these variables, that a subject with this covariate pattern was exposed. If a suitable auxiliary database or clinical subsample exists, the relationship between the more refined, but unavailable, propensity score and the less refined version, available from the study database, can be estimated from the exposure and covariate information in this auxiliary source. The estimated parameters of this relationship are then used to correct the parameter estimate of interest obtained from the database-only regression model. Stürmer et al.[29] describe a striking pharmaco-epidemiologic example of this combination of propensity scores and regression calibration, and the software implementation. The main study population consisted of just over 100,000 community-dwelling New Jersey residents aged 65 years or older who filled prescriptions within the Medicaid program or the Pharmaceutical Assistance to the Aged and Disabled program and who were hospitalised at any time between January 1, 1995, and December 31, 1997.The auxiliary validation sample of just over 5000 was drawn from the the Medicare Current Beneficiary Survey (MCBS) – a sample of beneficiaries selected each year to be representative of the current Medicare population. Data, including data on medication use over the past 4 months (verified by inspection of medication containers), are obtained from face-to-face interviews and linked to Medicare claims data. The authors assessed the relation between non-steroidal anti-inflammatory drugs (NSAIDs) and 1-year mortality (22,000 deaths). Adjustment based only on the variables available for the 100,000 resulted in a hazard ratio for NSAID users of 0.80 as compared with an unadjusted hazard ratio of 0.68. Propensity score calibration resulted in a 'more plausible' hazard ratio of 1.06.

5 Using confounding data on sub-samples of cases and controls

5.1 Motivation and data for illustration

How much can be accomplished if we obtain 'internal' confounder data on a judiciously selected subset of those in the database study? To illustrate, we use information from a study that examined whether, and if so by how much, selective serotonin reuptake inhibitor antidepressants (SSRIs) increase the risk of upper gastrointestinal haemorrhage.[30] This population-based case control study was conducted using the General Practice Research Database. In addition to documenting concomitant medications considered to increase/decrease the risk of GI haemorrhage, this database also contains information on items such as the patient's smoking history, body mass index, and history of heavy alcohol use, items not available in administrative databases. Some 4028 persons who suffered a first episode of upper GI haemorrhage were matched with up to 10 persons, sampled from those at risk, i.e. in the *riskset*, on the date of the haemorrhage (the 'index date'). Those who had been issued an SSRI prescription in the 90 days before the index date were considered to have been 'exposed'. Since omitting the matching variables age, practice and index-date from the regression analyses had little effect, we ignore them here. In order to simplify the illustration, and be able to show the detailed calculations, we combined several other variables into a single confounder score, and then dichotomised it: individuals who are 'positive' on this 0/1 scale are – independently of SSRI exposure – at higher risk of bleeding, and would be suspected to be more common among those who have recently received SSRIs. We refer to this potential 'confounding' variable by the generic letter 'C'.

5.2 Additional data available only for cases

In some studies, it may be possible to access the medical records of those in the *case*-series but not for those in the *denominator* (control) series. Since the information on this '*background-risk*' was in fact available for each person in the GPRD database, we first show what the researchers *would* have observed if they had access to it for *each* of the $c = 4028$ in the case series but for *none* of the $d = 40,171$ persons in the denominator series.

Stratum:	Lower risk ($C = 0$)				Higher risk ($C = 1$)				Unstratified			
c's & d's:	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0
number:	135	–	1768	–	200	–	1925	–	335	1780	3693	38391

To understand what one can and cannot learn from these supplementary *case-only* data, let the four proportions $P_{E=0,C=0}$ to $P_{E=1,C=1}$ denote the frequency distribution of the total population-time in the study base. If the 'no effect modification' (no interaction) model (1) holds, then the numbers of cases ($135, \ldots, 1925$ in our example) in these four stratum-specific 'cells' should – apart from sampling variability – be proportional to

$$P_{E=0,C=0}, \quad P_{E=1,C=0} \times \psi_E, \quad P_{E=0,C=1} \times \psi_C, \quad P_{E=1,C=1} \times \psi_E \times \psi_C.$$

Ray and Griffin[31] proposed that this mathematical relationship could be used to infer that, *if the 'no effect modification' model holds*, and *if the cross-product ratio of these four frequencies is close to 1*, then the distributions of E and C in the study base are independent of each other, and thus one of the two necessary conditions for confounding is absent. They used this approach in a population-based case-control study[32] of the role of cyclic antidepressants in the risk of hip fracture, using administrative data. Medical record review for a sample of 164 of their 4501 cases suggested the observed ID ratio of 1.6 was not due to confounding by body mass, impaired ambulation, functional status or dementia.

If however, the cross-product ratio is substantially different from 1, so that confounding is definitely possible, one cannot – without further external information on the plausible values for ψ_C, and on the prevalence of C in the various exposure categories in the base – adjust for the confounding. Moreover, the limitation of this use of 'case-only' confounder data is that a null cross-product ratio only rules out confounding if the uncheckable 'no effect modification' (no interaction) model (1) holds. Suissa and Edwardes[33] proposed 'simple conditions under which an adjusted estimate of the relative risk can be obtained' when data on a confounder are available only for the cases, and derived formulae for the estimator and its confidence limits. The method relies on an external estimate of the confounder prevalence or, additionally, of the confounder-exposure OR. Unfortunately, one of these conditions is the correctness of the no-effect-modification model (1).

In our SSRI scenario, the cross-product ratio of $(135 \times 1925) \div (1768 \times 200) = 0.73$ would have left the researchers in a quandary, and might have prompted them to either use additional external information, and some additional assumptions, or else to pursue background-risk data on at least some of those in the denominator series.

5.3 Additional data available for samples of cases and controls

We will suppose that the unit costs of obtaining the additional documentation on this confounding factor were the same for 'cases' and for 'controls'. If one had a fixed budget, how should the 'phase-2' sub-sample be selected? How should the resulting phase-1 and phase-2 data be included in the analysis? How much wider would the confidence intervals for the estimate of $\mathrm{IDR}_{\mathrm{SSRI}}$ be? In a 1996, the authors of a database study were faced with a similar question when planning a study to investigate the role of non-steroidal anti inflammatory drugs in the prevention of breast cancer, and the role of antidepressant medications in the etiology of breast cancer. A fuller account of how we proceeded can be found elsewhere.[34–36] From the Saskatchewan databases, we had detailed prescription drug information on the 1440 cases diagnosed in the years 1991 to 1995. We could, if we wished, have obtained the corresponding medication histories for *all* of these women's peers (i.e. the entire dynamic population). Instead, we obtained this information for four 'controls' for each woman with breast cancer, with the controls for each case chosen at random from the 'risk set' of women of the same age who had not been diagnosed with breast cancer by the date of diagnosis of the case. Despite this detailed medication history on almost 7500 subjects, we still lacked critical information on their individual breast cancer risk profiles, i.e. family history of breast cancer, age at menarche, reproductive history, use of alcohol and over-the-counter (non-prescription) medications, etc. These items are known risk factors for breast cancer, and that have different distributions in those exposed and not exposed to the medication of interest. Information on these potential confounding factors could only be obtained from the women themselves, or if they were no longer alive, from family members.

Locating and interviewing this number of women or their families was not economically feasible, and so we were forced to consider a sampling approach. We realised that even if we were able to interview as many as a 10% random sub-sample, these 750 would still be a relatively uninformative subsample: the relative rarity of long-term exposure to the medications of interest would mean that some of the cells frequencies in the 2 × 2 tables (and in the logistic regression which adjusted for the confounding variables measured) would be single- or barely double-digit numbers. We decided to deliberately over-sample some of the cells, and to – 'somehow' – account for the sampling at the analysis stage. We were able, by algebra and by simulations, to see that this selection bias, deliberately introduced, could be removed in the analysis by using the known sampling fractions. But would the rest of the epidemiology world believe us? And would the funding agency be convinced by our sample size considerations? Two-phase designs, in which in which basic and easily obtained data are determined for a large first-stage sample, but additional more expensive data are measured on a second-stage sub-sample, have a long history in survey sampling, but we had not seen them used in epidemiology. After an intense search, we found the articles by White[37] and by Walker,[38] who introduced this design into etiologic studies in epidemiology, and the regression-based extensions introduced by Breslow and Cain[39–40] (some of the other work since then will be described in Section 5.3).

To motivate and illustrate the efficiency implications of various sampling strategies, and how in fact the two stage data are combined, we again use the case-control study of SSRIs, with the binary background-risk data item as the single potential confounder. This allows us to simulate the phase-2 sampling variation and, by hand-calculating the estimated variance of the parameter of interest, to see what influences it. To make the exercise realistic, we simulate a situation where the budget for obtaining the information on the background-risk factor limits us to a phase-2 sample of 1000. We compute $\widehat{\psi}_{\text{adj}}$ and SE[log{$\widehat{\psi}_{\text{adj}}$}] using the `two-phase` and `svyglm` functions in the `survey`[41] package in R (we will describe this software, and the analysis methods later). In order to give some sense of the sampling variation, we show the data and the results in five of the infinitely many possible samples.

	Lower risk				Higher risk					
Possible sample	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	$\widehat{\psi}_{\text{adj}}$	SE[log{$\widehat{\psi}_{\text{adj}}$}]
1	2	30	40	645	4	14	50	215	1.17	0.458
2	0	23	43	645	2	19	40	228	0.40	0.724
3	3	27	42	612	5	11	53	247	1.87	0.410
4	8	19	42	633	3	17	43	235	2.84	0.392
5	3	26	41	631	4	18	51	226	1.25	0.440

Clearly a two-stage sample of 1000, even when coupled with the SSRI data available on all of the the 44,199 subjects in phase-1, could lead to a very imprecise estimate of ψ_{SSRI}. Even though in reality a research team does not have the luxury that we have here to judge how far off the target their point estimate may have been, the large SE[log{$\widehat{\psi}_{\mathrm{adj}}$}] indicates that the point estimate may be very wide of the mark.

Epidemiologists are taught early in their training that it is permissible to sample 'by *Y* value without regard to the *E* value,' or 'by *E* value without regard to the *Y* value,' but not by *both*. Thus, they would probably have tried to obtain a more informative, i.e. balanced second-stage sample of 1000 by selecting 500 from the case series and 500 from the denominator series. This strategy might have led to one of the following data sets:

	Lower risk				Higher risk					
Possible sample	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	$\widehat{\psi}_{\mathrm{adj}}$	SE[log{$\widehat{\psi}_{\mathrm{adj}}$}]
1	14	10	233	358	22	6	231	126	2.06	0.327
2	14	13	205	370	30	5	251	112	2.33	0.316
3	22	21	215	349	25	9	238	121	1.55	0.263
4	20	23	213	333	23	7	244	137	1.57	0.260
5	14	18	219	348	27	11	240	123	1.25	0.270

This approach leads to somewhat less variable estimates, since instead of having two single-digit frequencies (both c_1's) as 'weak links' (recall the heuristics from the Woolf variance formula), we now have only one – the d_1 in the higher background-risk stratum. However, since the marginal frequencies in the 2 × 2 stage-1 data are more imbalanced with respect to SSRI exposure than with respect to case:control status, some epidemiologists would probably have selected 500 from those exposed to SSRIs and 500 from those not. Possible results of this strategy are:

	Lower risk				Higher risk					
Possible sample	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	$\widehat{\psi}_{\mathrm{adj}}$	SE[log{$\widehat{\psi}_{\mathrm{adj}}$}]
1	33	275	14	344	38	154	24	118	1.76	0.223
2	30	280	16	346	46	144	26	112	1.69	0.214
3	29	268	28	343	52	151	26	103	1.35	0.196
4	29	257	21	324	55	159	21	134	2.00	0.204
5	40	266	19	353	63	131	14	114	3.33	0.214

This approach produces a further slight improvement, but continues to be hampered, albeit to a lesser extent than before, by the low c_0 frequencies. This leads to the obvious strategy: select 250 persons from each of the 2 (SSRI+, SSRI−) × 2 ('case', 'control') = 4 cells.

This approach produces a large reduction in variance. Furthermore, it raises the natural question as to how far more we can reduce it, and what uncertainty would remain if the researchers had a budget that allowed them to obtain this information

	Lower risk				Higher risk					
Possible sample	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	$\widehat{\psi}_{\text{adj}}$	SE[log{$\widehat{\psi}_{\text{adj}}$}]
1	103	154	116	177	147	96	134	73	1.78	0.079
2	100	149	125	187	150	101	125	63	1.66	0.087
3	105	160	118	193	145	90	132	57	1.64	0.092
4	95	160	130	190	155	90	120	60	1.69	0.085
5	103	162	125	186	147	88	125	64	1.75	0.082

on the background-risk status of *each* of the $c = 4028$ persons in the case-series *and on each* of the $d = 40, 171$ persons in the denominator series? The entire data, and the resulting adjusted IDR estimates they would have derived from them – under the assumption of no-effect modification – are as follows (MH: Mantel–Haenszel; LR: logistic regression).

Stratum:	Lower risk				Higher risk				Unstratified			
	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0	c_1	d_1	c_0	d_0
no:	135	1126	1768	28,415	200	654	1925	9976	335	1780	3693	38,391
$\hat{\psi}$	1.93				1.58				1.96			
$\hat{\psi}_{\text{adj.}}$	MH → 1.74; 1.72 ← LR											
SE[log]	MH → 0.063; 0.063 ← LR								0.062			

As it turns out, the *full* data show little evidence of effect modification: in a formal test of homogeneity, the p-value was 0.12. The full denominator series allows us to quantify the prevalence of the higher-risk profile in the study base: it was present in an estimated $654/(654 + 1126) = 37\%$ of those who were exposed to SSRIs, and in $9976/(9976 + 28415) = 26\%$ of those who were not. The combined data from the denominator and case series also allow us to use a logistic regression model containing indicator terms for both SSRI exposure and the higher-background-risk to estimate that, independently of SSRI exposure, the ID of GI bleeding is just over 3 times higher in those with the higher- than the lower-risk profile. This same logistic regression model allows us to estimate that after adjustment for this confounding, the ID of GI bleeding is 1.72 times higher in those exposed to SSRIs than in those who were not. Of particular note is the standard error obtained using the background-risk status of all 44,000 subjects – not that much smaller than the one achieved when we had this information on just 1000 of them. The slight exaggeration in the crude ID ratio (1.96) relative the the adjusted one (1.72) is reflected in the confounding ratio: $\{1 + (3.05 - 1) \times 0.37\}/\{1 + (3.05 - 1) \times 0.26\} = 1.15.$ As expected, in this simple situation, the Mantel–Haenszel summary ratio, 1.74, is very close to the one derived by logistic regression.

We now explain how the data from the two stages are combined, how the selection bias[42–p96] that was deliberately introduced at stage 2 is removed, how the variance of the parameter estimate of interest is calculated, and how it can be decomposed in a way that helps us plan the size of a 2-phase case-control study. To illustrate, we use the data from the first 'possible sample' of 250 from each of the four stage-1 cells, i.e. the one that yielded $\widehat{\psi_{\text{adj}}} = 1.78$ and SE[log{$\widehat{\psi_{\text{adj}}}$}] = 0.079. White's approach[37] was to multiply the observed stratum-specific two-stage frequencies by the inverses of the 4 sampling factions – in this example by 1.34 to 154 – in order to project what the full (but unobserved) first-stage frequencies must have been.

Phase				
	c'_1	d'_1	c'_0	d'_0
1:	335	1780	3693	38391
	c''_1	d''_1	c''_0	d''_0
2:	250	250	250	250
	c'_1/c''_1	d'_1/d''_1	d'_0/c''_0	d'_0/d''_0
Ratio:	1.34	7.12	14.8	154

	Lower risk				Higher risk				
	c''_1	d''_1	c''_0	d''_0	c''_1	d''_1	c''_0	d''_0	
2:	103	154	116	177	147	96	134	73	← Observed
1:	138	1096	1714	27181	197	684	1979	11210	← Projected, based on Ratio
		$\widehat{\psi} = 2.00$				$\widehat{\psi} = 1.63$			

As one might expect, the cross-product ratios involving the projected phase-1 frequencies are used as the two stratum-specific point estimates of ψ. White[37] showed that in order to obtain the variance of the log of each stratum-specific point estimate, one needs to subtract the quantity $K = (1/250 + \cdots + 1/250) - (1/355 + \cdots + 1/38391) = 0.0122$ from the 'Woolf' variance based on the phase-2 frequencies. Thus, for the lower-risk stratum, for example, the variance is $(1/103 + \cdots + 1/177) - K = 0.0183$. For the higher-risk stratum, the variance is 0.0262. The common set of ratios used in the projected frequencies means that the logarithms of the two point estimates have a (negative) covariance, of magnitude $-K$.

One can, as White did, combine these two negatively correlated point estimates using a weighted average. Or, one can estimate the common ID ratio using regression methods. The more recent of these, implemented in the `svyglm` function in the `survey`[41] package in `R`, fits a generalised linear model to data from complex survey designs, with inverse-probability weighting and design-based standard errors. It produced the $\widehat{\psi_{\text{adj}}} = 1.78$ and SE[log{$\widehat{\psi_{\text{adj}}}$}] = 0.079, shown above. In the original logistic-regression-based approach devised by Breslow and Cain,[39] one fits a logistic regression to the stage-2 data, with $\log\{(c''_1 \times d'_1)/(d''_1 \times c'_1)\}$ and $\log\{(c''_0 \times d'_0)/(d''_0 \times c'_0)\}$ as offsets, and modifies the covariance matrix that is output by the software. The R statements

```
E    = c(  1,  0,   1,  0);
hi.R = c(  0,  0,   1,  1);
c    = c(103, 116, 147, 134);
d    = c(154, 177,  96,  73);
o    = log( c( 250*1780/(250*335), 250*38391/(250*3693) ) ); o=c(o,o);

glm( cbind(c,d) ~ hi.R + E, family=binomial,offset=o )
```

yield $\widehat{\psi} = \exp(0.6021) = 1.83$. Since in this example both the stratum and exposure variables are binary, the variance obtained from the logistic regression program (0.0170) simply needs to be reduced by the same correction factor $K = 0.0122$ derived by White, to give $\text{SE}[\log\{\widehat{\psi}\}] = (0.0170 - 0.0122)^{1/2} = 0.069$. See Breslow and Cain[39] for the (matrix) variance calculations when the regressor variables are more complex.

5.4 Sample size considerations

Many of the early methodological papers focused on the *relative* efficiencies of various designs and methods of data analysis, using simulated and already assembled data sets. A number of planning tools have been developed to help end users calculate statistical precision/power in *absolute* terms. The earliest[43,44] only accommodates case-control studies with a binary exposure and a single binary confounder. We have recently extended this[45] to accommodate multiple confounding variables and/or covariates and either a binary or a categorical exposure; we also indicated how to proceed when exposure is represented as a quantitative variate. Schill and Wild[46] developed a strategy to obtain optimised sampling fractions to estimate a parameter vector. They note that no global optimal design exists and that local optimal designs depend on scenarios comprising the true disease model and the association between the phase-1 and phase-2 information. Thus they develop 'an admissibility test that rejects scenarios inconsistent with the phase-1 data and, for the selected scenarios, determine a minmax D- or A-optimal design that protects against worst-case scenarios.' More recently, Schill *et al.*[47] developed a software tool for planning two-phase case-control studies assuming categorical covariates. It offers a graphical user interface to organise and input the relevant anticipated entities and calculates a normed, expected two-phase case-control study. The planning tool helps to select a stratification.

The basis of our approach[45] to sample size considerations can be illustrated heuristically using the worked example used above, involving one binary exposure, and one binary confounder, and the structure of the variance formula for $\log\{\widehat{\psi}\}$ developed by Breslow and Cain,[39] namely

$$\text{Var}[\log\{\widehat{\psi}\}] = V_{\text{LR}_2} - K.$$

V_{LR_2} (0.0170 in the worked example) is the variance obtained from the logistic regression, with offsets, fitted to the stage-2 data. K (0.0122 in the example) is the difference between the 'Woolf' variance calculated using the four stage-2 frequencies, i.e. the $V_{W_2} = 1/250 + \cdots + 1/250 = 0.0160$ in the example, and the corresponding one calculated using the the four stage-1 frequencies, i.e. $V_{W_1} = 1/335 + \cdots + 1/38391 = 0.0038$. In this simple case, the Breslow-Cain variance (0.0048 in the example) can be re-written as

$$\begin{aligned}\text{Var}[\log\{\widehat{\psi}\}] &= V_{W_1} + (V_{\text{LR}_2} - V_{W_2})\\ &= 0.0038 + (0.0170 - 0.0160) = 0.0038 + 0.0010 = 0.0048.\end{aligned}$$

Re-arranging it this way shows the two separate variance components associated with the sample sizes used in each stage. The first component, V_{W_1}, is determined by the four stage-1 frequencies. Thus, its magnitude, for case and denominator series of planned sizes c and d, can be projected from the expected frequencies of exposed and unexposed, c_1, c_0, d_1 and d_0. The second, $V_{\text{LR}_2} - V_{W_2}$, is the difference between the variances from two logistic regression models applied to the stage 2 data: one that includes and one that excludes the (binary) confounding variable. The difference can be re-written as $V_{\text{LR}_2} - V_{W_2} = \text{VIF} \times V_{W_2}$, where VIF is the variance inflation factor associated with including the additional variable in the logistic regression. This VIF been extensively studied,[16,17,48] and so allows us to give a rough guide for the overall variance:

$$\text{Var}[\log\{\widehat{\psi}\}] = V_{W_1} + (\text{VIF} - 1) \times V_{W_2}.$$

This representation can then be used to plan the sizes of the phase-1 and phase-2 samples. See Hanley *et al.*[45] for a more detailed investigation of the (VIF − 1) quantity.

5.5 Additional literature and applications

For a helpful orientation to the literature on this, and allied designs, the reader is first referred to the encyclopedia entry by Breslow[49] and the references therein. Much of the theoretical work on these designs has focused on more efficient data-analysis models, and on the tradeoffs between efficiency (when using the correct model) and bias (when not). The main approaches include Horvitz-Thompson estimating equations, non-parametric maximum likelihood, and pseudo-likelihood, estimators for logistic regression coefficients, and a pseudo-score method, in which the scores for subjects not included in phase-2 are estimated from the regression model. The semi-parametric methods incorporate data from phase-I subjects when the covariate information can be summarised into a finite number of strata. Chatterjee and Chen[50] have recently extended their pseudo-score approach to incorporate information on continuous phase-1 covariates. The special March 2007 issue on Statistical Analysis of Complex Event History Data in the Scandinavian Journal of Statistics and the December 2007 issue of Lifetime Data Analysis have further theoretical articles on some of these issues.

Whereas, the previous examples in this section were mainly case-control studies, the same principles apply to cohort studies, and to 'case-cohort' studies that collect data on additional risk factors from a sub-sample of the cohort. Eng *et al.*[51] describe an application of the case-cohort design to assess residual confounding by risk factors not captured in the comparison of the event rates in two (propensity score-matched) cohorts of $n_1 = 22,429$ and $n_0 = 44,858$ derived from medical claims for members of a large health plan. Supplementary data on risk factors not measured in the two cohorts were collected from medical records for 701 of the 67,287. The authors estimated the ID ratio of interest adjusted for the supplementary variables . The ID ratio of interest adjusted for the supplementary variables using Cox regression modified for a case-cohort design was 0.90 (95% CI): 0.49 to 1.68). This was similar to the ID ratio from the cohort study itself (0.92; 95%CI: 0.50, 1.63). The authors took this to indicate that there was negligible confounding by the supplementary variables in the cohort study. The most recent two-stage methods described by Breslow[52] include several ways to include *all* of the available information in a single analysis.

The methods described above have focused on the efficiency achievable by the biased selection of subjects for further data collection were for *unmatched* case control studies. A stratified version of nested case-control sampling was introduced by Langholz and Clayton.[53] This design, which they call 'countermatching', uses data available for all cohort members to select a case-control sub-sample for whom additional information will be collected. Rather than *match* controls to make them as similar as possible to cases, they make them as different as possible on exposure, to ensure the maximum contrast. Relative to a simple random sample of controls, countermatching can yield a substantial efficiency gain when a surrogate measure of exposure is available for the full cohort, but accurate exposure data is to be collected only in a nested case-control study, and when exposure data are available for the whole cohort but data concerning important confounders are not. Further theory can be found in Langholtz and Borgan.[54] Countermatching can also be an efficient way to study interaction, when selecting controls in nested case-control studies of the joint effects of multiple risk factors when one is previously measured in the full cohort.[55]

6 Concluding remarks

Since epidemiologic studies of the unintended of medications usually rely on non-experimental comparisons, confounding by ummeasured factors must be considered when evaluating the resulting estimates of the comparative parameter of interest. This review has considered statistical strategies for situations where information on these factors is not available in the large administrative databases that are used in pharmacoepidemiology research. Although, it may have seemed like a digression, Section 2 (and part of Section 4) were included to emphasise that there are at least two additional components in the $\log\{\widehat{\psi}\}$ derived from most non-experimental studies:

$$\log\{\widehat{\psi}\} = \log\{\psi\} + \log[\text{Confounding Ratio}] + \log[\text{Attenuation Ratio}] \\ + \text{Random Error}.$$

Thus, while the emphasis in this review has been on efficient study designs and statistical strategies to reduce the effects of confounding, and of attenuation due to measurement errors, one also needs to have a large case series in order to reduce the imprecision (random error) in the estimate.

In some situations, the unmeasured risk factors are unlikely to be uncorrelated with the choice of medication, and thus to distort the comparison. For example, compared medications in the same class are likely to be prescribed without considering the subtle details in their chemical composition (e.g. a carbon vs. a nitrogen atom at a specific position in the central ring[35]) and without knowledge of possible differences in unintended effects (e.g. genotoxicity[35]).

In other situations, it may be possible to use external population-level data to quantify the maximum bias that would ensue from not being able to adjust for an unmeasured factor. In a different context, Wacholder *et al.*[56] used the confounding ratio and population-based estimates of the components of this ratio to measure the potential bias from a particular form of confounding in genetics studies. In 1972, Miettinen[4] suggested that routine assessment of the amount of confounding in the crude ID ratio 'would help accumulate valuable experience for use in the planning and evaluation of other studies.'

As described in Section 4, sophisticated statistical tools and accessible software are becoming available to incorporate quantitative sensitivity analyses into pharmacoepidemiology studies based on databases. These corrections can be based on expert opinion, external population-level data, or where available, on individual level data on auxiliary samples such as those used by Stürmer *et al.*[29]

As described in Section 5, two-stage sampling to provide additional data on the subjects in the main data-based study can offer large economies; the statistical properties of the various estimators are beginning to be better understood, and easy-to-use software is becoming more readily available. But, despite its 25-year existence, two-stage sampling continues to be under-used. We identified the approximately 200 citations of the initial articles by White, Walker, Breslow and Cain. The vast majority of citations were in methodological articles; only about a dozen were in reports of actual epidemiologic studies. We suspect that some of the problems may stem from the nature of the administrative databases – many of them government operated – used for the studies, and on the privacy issues and other difficulties involved in contacting patients. We contracted with the Government of Saskatchewan to send questionnaires on our behalf to a subsample of those in our case control study of NSAIDs and breast cancer,[36] $\sim$ 50% of the cases and 4% of the controls responded – far fewer than is customary with traditional case-control studies where the investigators themselves have direct access to subjects. In a study which used birth-certificates to study the relationship between ambient air pollution and preterm birth, the response rate in the second-stage sample was only 40%.[57] In some contexts, investigators have been able to obtain the relevant additional information directly from hospital or medical charts[58] or from health maintenance organisations.[59]

Given these access difficulties, the techniques in Section 4, particularly propensity score calibration, may offer a valuable alternative. These difficulties also reinforce the the plea[4] to researchers who do have access to more comprehensive data, to 'routine[ly] assess […] the amount of confounding in the crude RR' and to share this information to 'help accumulate valuable experience for use in the planning and evaluation of other studies.'

Acknowlegment

The authors thank Jean-François Boivin, Chris Delaney and Ernest Lo for their input.

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and Le Fonds québécois de la recherche sur la nature et les technologies.

References

1 Miettinen OS. The need for randomization in the study of intended effects. *Statistics in Medicine* 1983; **2**: 267–71.

2 Vandenbroucke JP. The history of confounding. *Sozial- und Präventivmedizin/Social and Preventive Medicine* 2004; **47**: 216–24.

3 Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. IARC Scientific Publication number 32. 1980.

4 Miettinen OS. Components of the crude risk ratio. *American Journal of Epidemiology* 1972; **96**: 168–72.

5 Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, Wynder E. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 1959; **22**: 173–203.

6 Woolf B. On estimating the relation between blood group and disease. *Annals of Human Genetics* 1955; **11**: 251–3.

7 Greenland S (Editor). Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods. Epidemiology Resources Inc. Ch. 4. 1987.

8 Miettinen OS (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology* **103**: 226–35.

9 Miettinen OS. Epidemiology: quo vadis? *European Journal of Epidemiology* 2004; **19**: 713–8.

10 Miettinen OS. Important concepts in epidemiology. Chapter 2 in *Teaching Epidemiology: A Guide for Teachers in Epidemiology, Public Health and Clinical Medicine*, 3rd edn, Olsen, Saracci, Trichopoulos (eds.) 2008 to appear.

11 Breslow NE, Day NE. Statistical Methods in Cancer Research: Volume II: The Design and Analysis of Cohort Studies, IARC Scientific Publication, No 82. 1987.

12 Mantel N and Haenszel W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**: 719–48.

13 Breslow NE. Elementary methods of cohort analysis. *Int J Epidemiol* 1984; **13**:112–5.

14 Robins J, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**: 311–23.

15 Rothman KJ. *Epidemiology: an introduction* Oxford University Press, New York, 2002.

16 Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology* 1984; **13**: 356–65.

17 Smith PG, Day NE: Matching and confounding in the design and analysis of epidemiological case-control studies. Perspectives in Medical Statistics, J.F. Bithell, R. Coppi, eds. Academic Press, London, 1987: 39–64.

18 Helms M, Vastrup P, Gerner-Smidt P, Mlbak K. Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study. *BMJ* 2003; 15; **326**(7385): 357.

19 Walker AM. Efficient assessment of confounder effects in matched follow-up studies. *Applied Statistics* 1982; **31**: 293–7.

20 Ayas NT, Barger LK, Cade BE, Hashimoto DM, Rosner B, Cronin JW, Speizer FE, Czeisler CA. Extended work duration and the risk of self-reported percutaneous injuries in interns. *JAMA* 2006; **296**: 1055–62.

21 Maclure M. The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology* 1991; **133**: 144–53.

22 Haddon W, Valien P, McCarroll JR, Umberger CJ. A controlled investigation of the characteristics of adult pedestrians fatally injured by motor vehicles in Manhattan.

Journal of Chronic Diseases 1961; **14**: 655–78.

23 Schneeweiss S, Patrick AR, Stürmer T, Brookhart MA, Avorn J, Maclure M, Rothman KJ, Glynn RJ. Increasing levels of restriction in pharma-coepidemiologic database studies of elderly and comparison with randomized trial results. *Medical Care* 2007; **45**: S131–42.

24 Greenland S. Multiple-bias modeling for analysis of observational data. *Journal of the Royal Statistical Society A* 2005; **168**: 267–306.

25 Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety* 2006; **15**: 291–303.

26 Stürmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Medical Care* 2007; **45**: S158–65.

27 Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology* 2005; **16**: 17–24.

28 Thürigen D, Spiegelman D, Blettner M, Heuer C, Brenner H. Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research* 2000; **9**: 447–74.

29 Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology* 2005; **162**: 279–89.

30 Opatrny L, Delaney JAC, Suissa S. Gastrointestinal haemorrhage risks of selective serotonin receptor antagonist therapy: a new look. *British Journal of Clinical Pharmacology* DOI:10.1111/j.1365–2125.2008.03154.x

31 Ray, W.A. and Griffin, M.R. Use of Medicaid data for pharmacoepidemiology. *American Journal of Epidemiology* 1989; **129**: 837–49.

32 Ray WA, Griffin MR, Malcolm E. Cyclic antidepressants and the risk of hip fracture. *Archives of Intern Medicine* 1991; **151**: 754–56.

33 Suissa S, Edwardes MD. Adjusted odds ratios for case-control studies with missing confounder data in controls. *Epidemiology* 1997; **8**: 275–80.

34 Hanley JA. Two-phase case-control studies: a personal account. *Australasian Epidemiologist* 1998; **5**: 12–3.

35 Sharpe CR, Collet JP, Belzile E, Hanley JA, Boivin JF. The effects of tricyclic antidepressants on breast cancer risk. *Br J Cancer* 2002; **86**: 92–7.

36 Sharpe CR, Collet JP, McNutt M, Belzile E, Boivin JF, Hanley JA. Nested case-control study of the effects of non-steroidal anti-inflammatory drugs on breast cancer risk and stage. *Br J Cancer* 2000; **83**: 112–120.

37 White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* 1982; **115**: 119–28.

38 Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* 1982; **38**: 1025–32.

39 Breslow NE, Cain KC. Logistic regression for two stage case-control data. *Biometrika* 1988; **75**: 11–20.

40 Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology* 1988; **128**: 1198–206.

41 Lumley T. Analysis of Complex Survey Samples. *Journal of Statistical Software* 2004; **9**: 1–19.

42 Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research : principles and quantitative methods.Van Nostrand Reinhold, New York, 1982.

43 Schaubel D, Hanley J, Collet JP, Bolvin JF, Sharpe C, Morrison HI, Mao Y. Two-stage sampling for etiologic studies: sample size and power. *Am J Epidemiol* 1997; **146**: 450–8.

44 Collet JP, Schaubel D, Hanley J, Sharpe C, Boivin JF. Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. *Epidemiology* 1998; **9**(3): 309–15.

45 Hanley JA, Csizmadi I, Collet JP. Two-stage case-control studies: precision of parameter estimates and considerations in selecting sample size. *American Journal of Epidemiology* 2005; **162**: 1225–34.

46 Schill, W and Wild, P. Minmax designs for planning the second phase in a two-phase case-control study. *Statistics in Medicine* 2006; **25**: 1646–59.

47 Schill W. Wild P. Pigeot I. A planning tool for two-phase case-control studies *Computer Methods and Programs in Biomedicine* 2007; **88**: 175–81.

48 Edwardes MD. Sample size requirements for case-control study designs. *BMC Medical Research Methodology* 2001; 1:11doi:10.1186/1471-2288-1-11

49 Breslow NE. Case-control study, two-phase. In Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*, 2nd Edition. John Wiley & Sons, Chichester, West Sussex, England; Hoboken, NJ, 2005.

50 Chatterjee N, Chen YH. A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-I covariates. *Lifetime Data Analysis* 2007; **13**: 607–622. Epub 2007 Nov 14.

51 Eng PM, Seeger JD, Loughlin J, Clifford CR, Mentor S, Walker AM. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiology and Drug Safety* 2008 ; **17**: 297–305.

52 Breslow NE. Using the whole cohort in the analysis of case-control and case-cohort data. Under review.

53 Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environmental Health Perspectives* 1994; **102(Suppl 8)**: 47–51.

54 Langholtz B and Borgan O. Counter-matching: A stratified nested case-control sampling method. *Biometrika* 1995; **82** 69–79.

55 Cologne JB, Sharp GB, Neriishi K, Verkasalo PK, Land CE, and Nakachi K. Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. *Int. J. Epidemiology* 2004; **33**: 485–92.

56 Wacholder S, Rothman N, Caporaso N. Population Stratification in Epidemiologic Studies of Common Genetic Variants and Cancer: Quantification of Bias *Journal of the National Cancer Institute* 2000; **92**: 1151–8.

57 Ritz B, Wilhelm M, Hoggatt KJ, Ghosh GKC. Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles *American Journal of Epidemiology* 2007; **166**: 1045–52.

58 Blais L, Beauchesne MF, Rey E, Malo JL, Forget A. Use of inhaled corticosteroids during the first trimester of pregnancy and the risk of congenital malformations among women with asthma. *Thorax* 2007; **62**: 320–28.

59 Haselkorn T, Whittemore AS, Udaltsova N, Friedman GD. Short-term chloral hydrate administration and cancer in humans. *Drug Safety* 2006; **29**: 67–77.