Case-Control Studies, Inference in

Gary King
Harvard University, Cambridge, Massachusetts, U.S.A.

Langche Zeng
George Washington University, Washington, District of Columbia, U.S.A.

INTRODUCTION

Classic (or ''cumulative'') case-control sampling designs do not admit inferences about quantities of interest other than risk ratios and then only by making the rare events assumption. Probabilities, risk differences, number needed to treat, and other quantities cannot be computed without knowledge of the population incidence fraction. Similarly, density (or ''risk set'') case-control sampling designs do not allow inferences about quantities other than the rate ratio. Rates, rate differences, cumulative rates, risks, and other quantities cannot be estimated unless auxiliary information about the underlying cohort such as the number of controls in each full risk set is available. Most scholars who have considered the issue recommend reporting more than just risk and rate ratios, but auxiliary population information needed to do this is not usually available. We address this problem by developing methods that allow valid inferences about all relevant quantities of interest from either type of case-control study when completely ignorant of or only partially knowledgeable about relevant auxiliary population information.

OVERVIEW

Moynihan et al.[1] express the conclusions of nearly all who have written about the standards of statistical reporting for academic and general audiences:

> In general, giving only the absolute or only the relative benefits does not tell the full story; it is more informative if both researchers and the media make data available in both absolute and relative terms. For individual decisions...consumers need information to weigh the probability of benefit and harm; in such cases it [also] seems desirable for media stories to include actual event [probabilities] with and without treatment.

Unfortunately, existing methods make this consensus methodological advice impossible to follow in the most used research design in many areas of medical research: case-control studies. In practice, medical researchers have historically used classic (i.e., ''cumulative'') case-control designs along with the rare events assumption (i.e., that the exposure and nonexposure incidence fractions approach zero) to estimate risk ratios—or they abandon these quantities of interest altogether and merely report odds ratios. Although virtually no one supports the publication of odds ratios alone, this remains the dominant practice in the field.

In recent years, medical researchers have been switching to density sampling, which requires no such assumption for estimating rate ratios. (We discuss the failure-time matched version of density sampling called risk-set sampling by biostatisticians.) Unfortunately, researchers rarely have the population information needed by existing methods to estimate almost any other quantity of interest, such as absolute risks and rates, risk and rate differences, attributable fractions, or numbers needed to treat. We provide a way out of this situation by developing methods of estimating all relevant quantities of interest under classic and density case-control sampling designs when completely ignorant of, or only partially knowledgeable about, the relevant population information.

We begin with theoretical work on cumulative case-control sampling by Manski,[2,3] who shows that informative bounds on the risk ratio and difference are identified for this sampling design even when no auxiliary population information is available. We build on these results and improve them in several ways to make them more useful in practice. First, we provide a substantial simplification of Manski's risk difference bounds, which also makes estimation feasible. Second, we show how to provide meaningful bounds for a variety of quantities of interest in situations of partial ignorance. Third, we provide confidence intervals for all quantities and a ''robust Bayesian'' interpretation of our methods that work even for researchers who are completely ignorant of prior information. Fourth, through the reanalysis of the hypothetical example from Manski's work and a replication and extension of an epidemiological study of

bacterial pneumonia in individuals infected with human immunodeficiency virus (HIV), we demonstrate that adding information in the way we suggest is quite powerful as it can substantially narrow the bounds on the quantities of interest. Fifth, we extend our methods to the density case-control sampling design and provide informative bounds for all quantities of interest when auxiliary information on the population data is not available or only partially available. Finally, we suggest new reporting standards for applied research and offer software in Stata and in Gauss that implements the methods developed in this paper (available at http://GKing.Harvard.edu).

QUANTITIES OF INTEREST

For subject i ($i = 1, \ldots, \Delta n$), define the outcome variable $Y_{i,(t,t+\Delta t)}$ as 1 when one or more ''events'' (such as disease incidence) occur in interval $(t, t+\Delta t)$ (for $\Delta t > 0$) and 0 otherwise. The variable t usually indexes time but can denote any continuous variable. In etiological studies, we shall be interested in $Y_{it} \equiv \lim_{\Delta t \to 0} Y_{i,(t,t+\Delta t)}$. In other studies, such as of perinatal epidemiology, conditions with brief risk periods such as acute intoxication, and some prevalence data, scholars only measure, or only can measure, $Y_{i,(t,t+\Delta t)}$, which we refer to as Y_i, because the observation period in these studies is usually the same for all i.[4] Define a k-vector of covariates and constant term as X_i.[a] In addition, let X_0 and X_ℓ each denote k-vectors of possibly hypothetical values of the explanatory variables (often chosen so that the treatment variable changes and the others remain constant at their means).

Quantities of interest that are generally a function of t include the *rate* (or ''hazard rate'' or ''instantaneous rate''), $\lambda_i(t) = \lim_{\Delta t \to 0} \Pr(Y_{i,(t,t+\Delta t)} = 1 | Y_{is} = 0, \ \forall s < t, X_i)/\Delta t$, and functions of the rate. For example, the *rate ratio*, $\mathrm{rr}_t = \lambda_\ell(t)/\lambda_0(t)$, and the *rate difference*, $\mathrm{rd}_t = \lambda_\ell(t) - \lambda_0(t)$, indicate how rates differ as the explanatory variables change from values X_0 to X_ℓ.

Quantities of interest that are cumulated over an interval of time include the *risk* (also called the ''probability'' or ''conditional probability of disease''),

$$\pi_i = \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)}$$

$$= 1 - \exp\left(-\int_t^{t+\Delta t} \lambda_i(s)\mathrm{d}s\right) \tag{1}$$

[where $H(T_i, X_i)$ is the *cumulative hazard rate* for individual i over time interval $(t, t+\Delta t)$], the *risk ratio*,

$\mathrm{RR} \equiv \Pr(Y = 1 | X_\ell)/\Pr(Y = 1 | X_0)$, and the *risk difference* (or ''first difference,'' as it is called in political science, or the ''attributable risk,'' as economists and some epidemiologists call it), $\mathrm{RD} \equiv \Pr(Y = 1 | X_\ell) - \Pr(Y = 1 | X_0)$. The probability is evaluated at some values of the explanatory variables, such as X_0 or X_ℓ. The quantity RD is the increase in probability, and RR is the factor by which the probability increases [an $(\mathrm{RR} - 1) \times 100$ percent increase] when the explanatory variables change from X_0 to X_ℓ.

The quantities $\lambda(t)$, rr_t, rd_t, π, RR, and RD are normally used to study incidence in etiological analyses, but they are sometimes used to study prevalence by changing the definition of an ''event.'' Functions of these quantities, such as the proportionate increase in the risk or rate difference, the attributable fraction, or the expected number of people needed to treat to prevent one adverse event (or, when the ''treatment'' has an adverse effect, the expected number of people needed to harm to result in one adverse event), can also be computed as a function of these quantities with the methods described below. For example, to compute the number needed to treat (NNT), we merely calculate:

$$\mathrm{NNT} = \frac{1}{|\mathrm{RD}|} \tag{2}$$

$$= \frac{1}{|\Pr(Y = 1 | X_\ell) - \Pr(Y = 1 | X_0)|} \tag{3}$$

That is, we simply use the reciprocal of the size of the risk difference between the ''treatment'' and ''control'' groups. The number needed to harm (NNT) is obtained in the same way.

The other quantity frequently discussed and presented in the epidemiological literature is the *odds ratio*

$$\mathrm{OR} \equiv \frac{\Pr(Y = 1 | X_\ell)/\Pr(Y = 0 | X_\ell)}{\Pr(Y = 1 | X_0)/\Pr(Y = 0 | X_0)}$$

$$= \frac{\Pr(X_\ell | Y = 1)\Pr(X_0 | Y = 0)}{\Pr(X_0 | Y = 1)\Pr(X_\ell | Y = 0)} \tag{4}$$

where the second equality holds by Bayes theorem. The main advantage of OR is that it has been easier to estimate than the other quantities. In logit and other multiplicative intercept models (but not generally), OR also has the attractive feature of being invariant with respect to the values at which control variables are held constant. The disadvantage of OR is understanding what it means, and when OR is not the quantity of interest—which is almost all the time—then its ''advantages'' are not sufficient to recommend its use. Some statisticians seem comfortable with OR as their ultimate quantity of interest, but this is not common. Even more unusual is to find anyone who feels

---

[a]We assume for the purpose of this paper that the X's are not time-dependent.

more comfortable with OR than the other quantities defined above; we have found no author who claims to be more comfortable communicating with the general public using an odds ratio than one of the other quantities.[5] The odds ratio has been used ''largely because it serves as a link between results obtainable from follow-up studies and those obtainable from case-control studies.''[6] We provide this link with other quantities so that OR is no longer the only choice available.

Concluding a controversy on this subject in the *British Medical Journal*, Davies et al.[7] write ''On one thing we are in clear agreement: odds ratios can lead to confusion and alternative measures should be used when these are available.''

We show here how to make all relevant alternative measures *always* available, even in case-control data. With the results in this article, scholars should never again feel forced into presenting odds ratios.[b]

CASE-CONTROL SAMPLING DESIGNS

We now introduce the two key case-control sampling designs. With appropriate modifications, most of the methods we introduce also work with many variants of them.

Classic (or *cumulative*) *case-control* designs involve sampling (usually all) ''cases'' (subjects for which $Y_i=1$) and a random sample of ''controls'' (subjects for which $Y_i=0$) at the end of the study period. This design is most appropriate in studies of closed populations and when Y_i but not Y_{it} is observed. Statistical models for such data specify the risk $\Pr(Y_i=1|X_i)$ as a function of the input X_i. The most commonly used model is the logistic:

$$\Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{-X_i\beta}} \qquad (5)$$

Some examples in medical research using the cumulative case-control design are studies of congenital malformation, analyses of ''chronic conditions with ill-defined onset times and limited effects on mortality, such as obesity and multiple sclerosis, and studies of health services utilization''[8] and research that has as its goal descriptive, rather than causal, inferences (such as ascertaining the types of people that now have lung cancer so we can better prepare health-care facilities in anticipation).

The newer *density case-control* sampling design involves, in the study of cohort data organized by risk sets, sampling ''cases'' (subjects for which $Y_{it}=1$) at their

failure times and a subset of ''controls'' (subjects for which $Y_{it}=0$) from all individuals at risk at the time of each failure, possibly matched on a set of other control variables. So each *sampled* risk set R_j ($j=1,\ldots,M$, where M is the total number of cases in the data) is composed of one case (or more in the case of timing ties in their occurrence) and a random sample of controls at the same time (or other continuous index) t. A subject may appear in multiple risk sets. This administratively convenient data collection strategy can result in a substantial reduction in the resources required for the study and also has the advantage of controlling nonparametrically (i.e., without functional form assumptions) for all omitted variables, and unmeasured heterogeneity, related to t. Statistical models for such data specify the incidence (hazard) rate as a function of t and X. The most commonly used model is the Cox proportional hazard:

$$\lambda_i(t) = \lambda(t)r(X_i, \beta) \qquad (6)$$

where, in most applications, $r(X_i,\beta)=e^{X_i\beta}$ and the baseline hazard $\lambda(t)$ does not vary over subjects.[c]

INFERENCE WITHOUT POPULATION INFORMATION

Classic Case-Control

Without additional population information or assumptions, the literature provides no method for estimating any quantity of interest from classic case-control data. Historically, medical researchers have used the *rare events assumption* to estimate the risk ratio by the odds ratio. Denote the population fraction of incident cases by τ, which is the key piece of auxiliary information about the population not reflected in the sample [although other quantities such as $P(X)$, the possibly multivariate density of X, are also sufficient]. The rare events assumption states that τ is arbitrarily small [while $P(X)$ stays bounded away from zero, or instead that $\Pr(Y_i=1|X_j)\to0$ for $j=0, \ell$]. This assumption is not merely that cases are ''rare,'' but that they occur, at the limit, with zero probability. The advantage of this assumption is that, when correct, OR is a good approximation to the risk ratio: $\lim_{\tau\to0}OR=RR$.

This limit result is attractive because OR is often easy to estimate in case-control designs. For example, if X is a single binary variable, OR could be estimated by replacing elements in the second line of Eq. 4 with their sample analogs [e.g., $\Pr(X_\ell|Y=1)$ can be estimated by the fraction

---

[b]Unfortunately, the term ''relative risk'' no longer seems useful because it has been co-opted to denote such diverse quantities as rr, RR, and OR.

[c]The proportionality assumption holds when X_i is time-invariant. When this is not appropriate, a model that allows time-dependent covariates (such as the so-called Cox regression model) can be used.

of cases for which $X_i = X_\ell$ among all observations where $Y_i = 1$]. For another example, in logistic regression Eq. 5, case-control sampling only biases the intercept term, which drops out in the odds ratio expression, $\text{OR} = e^{(X_\ell - X_0)\beta}$. The coefficients of the control variables also drop out (because the elements of $X_\ell - X_0$ corresponding to those parameters are zero).

As is well known (e.g., Ref. [8]), the rare events assumption is problematic when τ is not nearly zero, and as a result, the odds ratio overestimates RR (when both are above 1; OR underestimates RR otherwise). In addition, this assumption implies, implausibly for most applications, that the risk difference tends to zero ($\lim_{\tau \to 0} \text{RD} = 0$), no matter how strong the real effect. In practice, RD is therefore not estimated, even when it is of more interest than RR.

When τ is very small but not zero, the bias introduced by using OR as if it were RR can be ignored without practical consequence. The rare events assumption is inappropriate in etiology studies with more commonly occurring events, such as in highly infectious diseases. The assumption is also often not plausible when studying diseases with nonabsorbing states or when studying prevalence.

Density Case-Control

In the commonly used proportional hazard model (6), the coefficients β can be consistently estimated without any auxiliary information on the population risk sets. The contribution of each sampled risk set to the likelihood function is the ex ante incidence probability for the individual who will actually acquire the disease, conditional on the total number of cases in the set being one:

$$\Pr(Y_{it_j} = 1 | R_j) = \frac{\Pr(Y_{it_j} = 1)}{\sum_{k \in R_j} \Pr(Y_{kt_j} = 1)} = \frac{e^{X_i\beta}}{\sum_{k \in R_j} e^{X_k\beta}} \quad (7)$$

where i indexes the case in the risk set and the summation in each denominator is over all observations in the risk set, $k = 1, \ldots, n_j$, where n_j is the number of observations, or the size, of R_j. The likelihood function then is the product of terms such as Eq. 7 over all M sampled risk sets (see, e.g., Ref. [9]). When a risk set includes multiple cases, because of timing ties, the conditional probability expression is more complicated, but the approach remains the same. This same estimator was independently proposed in econometrics by Chamberlain[10] to estimate consistently the parameters in a fixed-effect logit model when the number of subjects per group remains fixed even as the total number of subjects grows. Note also that Eq. 7 takes the same form as the probability expressions in the popular multinomial logit model (or McFadden's choice model as some call it), and indeed the same software can

be used on the data organized in risk sets to estimate β. Eq. 7 also takes the same form as the conditional probability expression for matched cohort data (for which Greenland[11] gives an alternative modeling strategy for estimating the risk ratio).

Eq. 7 takes the same form as the likelihood function contribution of a *full* risk set in a Cox regression model for full cohort data, only that now the *sampled* risk sets are used. Fortunately, this does not bias the first-order condition and preserves the consistency and asymptotic normality of the maximum likelihood estimator (see Refs. [12–14]). Moreover, values of n_j of only about 6 can result in nearly full efficiency.[12,15] For a discussion of biased selection of the controls, see Refs. [16,17] and see Ref. [18] for a discussion of alternative sampling designs.

With β consistently estimated by the conditional logit approach, we can compute the rate ratio using the incidence model (6), although the baseline hazard rate $\lambda(t)$ is not estimated:

$$\text{rr} = \frac{\lambda_\ell(t)}{\lambda_0(t)} = \frac{\lambda(t)e^{X_\ell\beta}}{\lambda(t)e^{X_0\beta}} = \frac{e^{X_\ell\beta}}{e^{X_0\beta}} = e^{(X_0 - X_\ell)\beta} \quad (8)$$

Thus a key advantage of density designs over classic case-control designs is that we can estimate rr without any additional (rare events or other) assumptions or any population information. However, without auxiliary information on the population, the literature provides no method for estimating any other quantity of interest, such as the rate, rate difference, the risk, risk difference, risk ratio, or number needed to treat.

INFERENCE UNDER FULL POPULATION INFORMATION

We now discuss the quantities of interest that can be estimated under each sampling design when required auxiliary population information is available.

Classic Case-Control

In the classic case-control design, the population fraction of cases, τ, can supply the required auxiliary information.[19–21] For example, in the simple case of a single binary explanatory variable, the risk can be estimated as

$$\begin{aligned} \pi &= \Pr(Y = 1 | X, \tau) \\ &= \frac{P(X|Y = 1)\Pr(Y = 1)}{P(X)} \\ &= \frac{P(X|Y = 1)\tau}{P(X|Y = 1)\tau + P(X|Y = 0)(1 - \tau)} \end{aligned} \quad (9)$$

where $P(X|Y=1)$ and $P(X|Y=0)$ are replaced with the sample mean of X among subjects where $Y=1$ and $Y=0$,

respectively. From this expression, we can compute $\text{RR} = \Pr(Y = 1|X_\ell, \tau)/\Pr(Y = 1|X_0, \tau)$, $\text{RD} = \Pr(Y = 1|X_\ell, \tau) - \Pr(Y = 1|X_0, \tau)$, or $\text{NNT} = 1/|\text{RD}|$.

For another example, the commonly used logistic regression model (5) on case-control data yields consistent maximum likelihood estimates of the slope coefficients and only the estimated intercept β_0 requires a correction such that the quantity estimated should instead be:

$$\beta_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \tag{10}$$

where $\bar{y}$ is the mean of Y or the sampling probability of $Y=1$. Thus with exact knowledge of τ, all logit parameters can be consistently estimated (see Refs. [22–26]; see Ref. [19] for a comprehensive review). Let β_τ denote the logit coefficient vector, with the first element corrected as in Eq. 10. Then the quantities of interest are:

$$\pi = \Pr(Y = 1|X, \tau) = [1 + e^{-X\beta_\tau}]^{-1},$$

$$\text{RR} = \frac{[1 + e^{-X_\ell\beta_\tau}]^{-1}}{[1 + e^{-X_0\beta_\tau}]^{-1}},$$

$$\text{RD} = [1 + e^{-X_\ell\beta_\tau}]^{-1} - [1 + e^{-X_0\beta_\tau}]^{-1},$$

$$\text{NNT} = \frac{1}{|\text{RD}|}$$

For estimation, we can plug in the maximum likelihood estimates for β or use improved methods with lower mean square error in the presence of rare events.[26] Standard errors for any of these quantities can be computed easily by simulation.[27]

The correction in Eq. 10 applies not only to logit models, but also to any multiplicative intercept model, even including many neural networks (such as the feed-forward perceptron with a logit output function[28,29]). Statistical research in medicine still predominately uses the logit specification for parametric analysis, whereas numerous other scholarly fields have been steadily switching to more flexible functional forms such as neural networks. In other fields, scholars have recognized that the logit model makes specific assumptions that are normally without substantive justification and indeed in the vast majority of research in this field are not even addressed explicitly. This is acknowledged in medical research in its use of matching models, but not when it comes to parametric approaches. Unfortunately, when using highly restrictive parametric forms such as logit, our empirical answers are functions of our theoretical specification rather than our data. By recognizing that the correction above applies to a much broader class of models, and makes considerably less stringent assumptions, researchers using case-control designs will have

many new and far more valid parametric approaches to choose from.

Density Case-Control

To estimate quantities of interest other than rr under density case-control sampling, we require information that enables the estimation of the baseline hazard rate. We employ the baseline hazard rate estimator proposed by Ref. [30] where the key population information needed is $\tau_j = n_j/N_j$, the sampling fraction for each observed risk set R_j ($j=1,\ldots,M$). With this information, and denoting the maximum likelihood estimate of β from the conditional logit procedure as b, the baseline incidence rate λ at time t_j is estimated as:[d]

$$\lambda(t_j) = \frac{1}{\sum_{k \in R_j}(e^{X_k b})(1/\tau_j)} = \frac{1}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \tag{11}$$

and with this, we can estimate the rate, $\lambda_i(t)$, from Eq. 6:

$$\lambda_i(t_j) = \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \tag{12}$$

With the rate, we can further estimate the cumulative rate:

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \tag{13}$$

and hence the risk:

$$\pi_i = \Pr(Y = 1|X_i) = 1 - e^{-H(T_i, X_i)}$$

$$= 1 - \exp\left(-\sum_{t_j \in T_i} \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}}\right) \tag{14}$$

and any of the other quantities of interest. The factor $1/\tau_j$ serves as a weighting factor that enables us to weight up each risk set to the full risk set size.

INFERENCE WITH BAYESIAN PRIOR ASSUMPTIONS

''Inference Without Population Information'' and ''Inference Under Full Population Information'' discuss inference under conditions of ignorance and full knowledge of the relevant auxiliary population information (τ and τ_j). When this information is not known exactly, but some

---

[d]For simplicity, we use the same notation for the estimated baseline hazard as for the theoretical version. We do the same for the other quantities below as well.

prior information exists, Bayesian methods are appropriate and straightforward (although we are not aware of their use in applications in this context). Ideally, this prior information would come from registry or survey data from the target population, but similar information from closely related populations would help form reasonable priors also. The procedure then is simply to put a prior distribution on τ (in classic case-control designs) or τ_j (in density case-control designs) and draw inferences about the quantity of interest from the posterior. These inferences have all the desirable properties of Bayesian estimates. They are consistent, efficient, asymptotically normal, etc. when averaging over the prior. We strongly support their use when prior information is available and known to be valid.

Bayesian estimates also share at least two disadvantages. First, they are not calculable when the analyst is completely ignorant of τ or τ_j in some or all parts of its range because the required prior density cannot be fully specified, and, in Bayesian inference, one must have a full prior. Second, Bayesian estimates yield biased inferences when prior information is biased. This second problem is especially severe in case-control studies because the data contain *no* information about τ and τ_j, and so the prior does not become dominated by the likelihood as the sample size grows. That is, even for very large samples, inferences from case-control data depend heavily on the prior.

These disadvantages combine in unfortunate ways sometimes when analysts are unsure of the validity of their prior information. In these cases, the classic Bayesian paradigm may, in practice, have the effect of encouraging researchers to guess values for their prior, hence introducing biased information into their analyses and adversely affecting their inferences. ''Diffuse'' priors with large variances are also no solution here because in general, increasing the variance of τ or τ_j will also make the mean tend toward 0.5, which of course is a statement of knowledge, not ignorance.

We therefore pursue ''robust Bayesian'' methods below that allow full or partial ignorance to be represented accurately without biasing inferences.

INFERENCES WITHOUT FULL POPULATION INFORMATION OR ASSUMPTIONS: CLASSIC CASE-CONTROL

We now discuss inferences under classic case-control designs about π, RR, and RD without the rare events assumption or population information about τ. We begin by extending Manski's results in the situation of pure ignorance and then add methods for partial ignorance, confidence intervals, and our preferred robust Bayesian interpretation. We conclude the section with two examples.

Extending Manski's ''Ignorance'' Results

As an alternative to full knowledge of τ and the rare events assumption, Manski[2,3] studied what inferences we could make when the researcher had no knowledge of τ. It is widely known that under these circumstances, no information about risk is available [i.e., $\pi \in (0,1)$] and RR is bounded between 1 and OR. That is, RR $\in$ [min(1,OR), max(1,OR)], apart from sampling error. That this expression depends on the odds ratio is useful because of how often it is easily estimable.

Manski's advance is that he also shows that RD can be bounded, although it requires the following complicated expression. Let

$$\phi = \left[\frac{\Pr(X_0|Y = 1)\Pr(X_0|Y = 0)}{\Pr(X_\ell|Y = 1)\Pr(X_\ell|Y = 0)} \right]^{\frac{1}{2}} \tag{15}$$

and

$$\gamma = \frac{a}{a - [\phi\Pr(X_\ell|Y = 1) - \Pr(X_0|Y = 1)]}, \tag{16}$$

where $a = \phi\Pr(X_\ell|Y=0) - \Pr(X_0|Y=0)$, and

$$
\begin{aligned}
\text{RD}_\gamma = {} & \frac{\Pr(X_\ell|Y = 1)\gamma}{\Pr(X_\ell|Y = 1)\gamma + \Pr(X_\ell|Y = 0)(1 - \gamma)} \\
& - \frac{\Pr(X_0|Y = 1)\gamma}{\Pr(X_0|Y = 1)\gamma + \Pr(X_0|Y = 0)(1 - \gamma)}
\end{aligned} \tag{17}
$$

Then RD is bounded between 0 and RD_γ, apart from sampling error, where RD_γ is the value of the risk difference if τ were equal to γ.

Manski's expression for the risk difference is useful but only when it can be estimated. Unfortunately, except for very simple cases, sophisticated nonparametric methods are required to estimate each of the component probabilities in Eqs. 15–17, and ϕ, γ, and RD_γ are not easy to estimate directly; to our knowledge, they have never been estimated in a real application. We remedy this situation by showing, in Appendix A, that these equations can be simplified so that $\text{RD}_\gamma = (\sqrt{\text{OR}} - 1)/(\sqrt{\text{OR}} + 1)$ (which, surprisingly, is exactly Yule's[31] ''coefficient of colligation,'' sometimes called Yule's Y). The bounds are thus a simple function of the odds ratio:

$$\text{RD} \in \left[\min\left(0, \frac{\sqrt{\text{OR}} - 1}{\sqrt{\text{OR}} + 1} \right), \ \max\left(0, \frac{\sqrt{\text{OR}} - 1}{\sqrt{\text{OR}} + 1} \right) \right] \tag{18}$$

The advantages of our expression in Eq. 18 are not only algebraic simplicity, and the familiarity with the odds ratio among applied researchers, but also that OR can be estimated very easily without nonparametric methods in simple discrete cases, in logistic regression, and in a wide

variety of multiplicative intercept models such as neural network models. Because these neural network models have arbitrary approximation capabilities, Eq. 18 can effectively always be applied.

A Proposed "Available Information" Assumption

Applied researchers have been reluctant to adopt Manski's "ignorance" assumption, perhaps in part because the knowledge they have about τ is discarded entirely, often resulting in very wide bounds on the quantities of interest. Particularly uncomfortable for researchers is that no matter how strong the empirical relationship among the variables is, the bounds on RR always include 1 and on RD always include 0, which in both cases denote no treatment effect.

Thus existing literature effectively requires researchers to choose among three extreme assumptions: τ is essentially zero, is known exactly (possibly apart from sampling error), or is completely unknown. Our alternative approach is to elicit from researchers a range of values into which they are willing to say that τ must fall (e.g., [0.001,0.05]), which appears to be a better reflection of the nature of prior information available in applied research settings than the extremes of exact knowledge or complete ignorance. Our approach seems consistent with Manski's[2] goals for future research and, like his specific methods, does not require a fully Bayesian prior distribution. Our approach could also be applied to bring available information and probabilistic inference to the methods Manski[2] has offered in other areas.

Let π_τ, RR_τ, RD_τ, and NNT_τ denote values of the probability, relative risk, risk difference, and number needed to treat, respectively, evaluated at τ. Suppose that τ is known only to fall within the range $[\tau_0,\tau_1]$, where $0<\tau_0<\tau_1<1$. (Because, by definition, choice-based samples include at least one example of a case and one of a control, τ_0 and τ_1 are known not to equal zero or one exactly.) Then, because π and RR are monotonic in τ, their bounds are simply

$$\pi \in [\pi_{\tau_0}, \pi_{\tau_1}] \tag{19}$$

and

$$\text{RR} \in [\min(\text{RR}_{\tau_0}, \text{RR}_{\tau_1}), \max(\text{RR}_{\tau_0}, \text{RR}_{\tau_1})] \tag{20}$$

The bounds for the risk difference are more complicated because RD_τ is a parabolic function of τ, and so the bounds differ in the monotonic and nonmonotonic regions. The relationship is monotonic in regions where τ_0 and τ_1 are both greater than or both less than the value of τ that corresponds to $\text{RD}_\gamma = (\sqrt{\text{OR}} - 1)/(\sqrt{\text{OR}} + 1)$. This region corresponds to cases where the derivative of

RD_τ with respect to τ, evaluated at τ_0 and τ_1, have the same sign. (This derivative can easily be checked numerically by comparing the signs of $\text{RD}_{\tau_{0+\epsilon}} - \text{RD}_{\tau_0}$ and $\text{RD}_{\tau_{1+\epsilon}} - \text{RD}_{\tau_1}$ for a suitably small value of ϵ.) When the relationship is monotonic, the bounds are

$$\text{RD} \in [\min(\text{RD}_{\tau_0}, \text{RD}_{\tau_1}), \max(\text{RD}_{\tau_0}, \text{RD}_{\tau_1})] \tag{21}$$

and otherwise, they are

$$\text{RD} \in [\min(\text{RD}_{\tau_0}, \text{RD}_{\tau_1}, \text{RD}_\gamma), \max(\text{RD}_{\tau_0}, \text{RD}_{\tau_1}, \text{RD}_\gamma)] \tag{22}$$

The bounds for number needed to treat can be directly derived from those for the risk difference by using reciprocals of the latter.

Revisiting a Numerical Example

We illustrate our methods by extending the numerical example concerning smoking and heart disease given by Manski.[2,3] For clarity, we follow Manski in ignoring uncertainty (i.e., equating sample fractions with sampling probabilities as if $n\rightarrow\infty$) in this section (only); in the next two sections, we show how to include estimation uncertainty and compute confidence intervals. This section also demonstrates the degree to which results under our approach are sensitive to assumptions about the interval $[\tau_0,\tau_1]$ while holding constant (at zero) estimation uncertainty. (The effect of estimation uncertainty, while holding constant the interval, follows standard sampling theory.)

In Manski's example, X is a binary explanatory variable taking values 1 for smokers and 0 for nonsmokers, and Y takes on the values 1 for coronary heart disease and 0 for healthy individuals. The assumptions in his example imply that $\Pr(X=1|Y=1)=0.6$, $\Pr(X=1|Y=0)=0.49$, $\Pr(X=0|Y=1)=0.4$, and $\Pr(X=0|Y=0)=0.51$. Hence using Eq. 9, we can write the probabilities as functions of τ:

$$\Pr(Y = 1|X = 1, \tau) = \frac{0.6\tau}{0.6\tau + 0.49(1 - \tau)}$$

$$= \frac{\tau}{0.82 + 0.18\tau}$$

$$\Pr(Y = 1|X = 0, \tau) = \frac{0.4\tau}{0.4\tau + 0.51(1 - \tau)}$$

$$= \frac{\tau}{1.28 - 0.28\tau}$$

For each of the quantities of interest, we now compare the case where τ is unknown, as Manski does, to where it is known to lie in the interval [0.05, 0.15] (Manski's example implies that $\tau=0.1$, which he treats as not known). Without bounds on τ, the problem provides no information about any probability, whereas the additional

information about τ gives much more informative bounds: the probability of heart disease among smokers is $\Pr(Y=1|X=1)\in[0.06, 0.18]$, whereas among nonsmokers, it is $\Pr(Y=1|X=0)\in[0.04, 0.12]$. For relative risk, Manski's "ignorance" assumption gives RR$\in[1., 1.57]$, whereas our alternative approach implies the very tight bounds of RR$\in[1.46, 1.53]$, indicating that smoking increases the risk of heart disease between 46% and 53%. For the risk difference, the bounds given no information on τ are RD$\in[0, 0.11]$, whereas our approach yields much narrower bounds of RD$\in[0.021, 0.056]$, the increase in probability due to smoking. Because smoking has an adverse effect, we can also obtain bounds for NNT, the number needed to harm, based on that for the risk difference. With no information on τ, Manski's approach would result in NNT$\in[9, \infty]$, which is hardly informative at all, while our method gives NNT$\in[18, 48]$, a range infinitely narrower than "between 9 and infinity" and of course more substantively meaningful.

Classical Confidence Intervals

We now provide a method of computing classical confidence intervals, saving our preferred robust Bayesian interpretation for the following section. Because each end of the bounds on π, RR, and RD are measured with error, upper and lower confidence intervals could be computed and reported for each. However, the inner bounds (the upper confidence limit on the lower bound and the lower confidence limit on the upper bound) are not of interest. Thus we recommend defining a confidence interval (CI) as the range between the outer confidence limits. The actual CI coverage of the resulting interval is always at least as great as the nominal coverage.

For all methods provided above, confidence intervals can easily be computed by simulation (the delta method is also possible but difficult because of the discontinuities caused by the minimum and maximum functions). The bounds are known functions of, and derive their sampling distributions from, the estimated model parameters. Therefore the distribution of the bounds can be simulated using random draws from the sampling distributions of the parameters.[27] For example, in the logit model, the asymptotic distribution of the estimated parameters is normal with mean vector and covariance matrix estimated by the usual maximum likelihood procedures. Random draws from this distribution can then be converted into random draws from the distribution of the bounds through the relevant formulas relating the bounds and the model parameters. A 90% (for example) CI for the bounds can be obtained by sorting the m random draws of the bounds and taking the 5th and 95th percentile values as the lower and upper bounds, respectively. Our software implements these procedures. The choice of m reflects the tradeoff between accuracy and speed: larger values of m improve

accuracy and reduce speed. The required m depends on the example; 1000 will often be enough, but it is easy to verify: rerun the simulation and if anything changes in as many significant digits as is needed, increase m and try again.

A Robust Bayesian Interpretation

Our estimation procedure is not strictly Bayesian in that choosing an interval for τ is not equivalent to imposing a uniform (or any version of a "noninformative") prior density within those bounds. However, our procedure can be thought of as a special case of "robust Bayesian analysis" (e.g., Refs. [32,33]) and one that happens to be easier to apply and gives results that are considerably easier for applied researchers to use than most examples in this literature.

From this robust Bayesian perspective, the interval chosen for τ can be thought of as narrowing the choice of a prior to only a class of densities rather than a (fully Bayesian) single density. In our case, the class of priors is defined to include all densities $P(\tau)$ subject to the constraint that $\int_{\tau_0}^{\tau_1}P(\tau)d\tau=1$. The advantages of this approach are that prior elicitation is much easier, it does not force analysts to give priors when no prior information exists, and, more importantly, estimates depend only on real information and so are the same for any density within the class. If a classical Bayesian prior is inaccurate, classical Bayesian inferences will be incorrect. In contrast, our "robust Bayesian" approach will give valid inferences even if we can only narrow the prior to a class of densities rather than one particular density function. (Because data do not help in making inferences about τ, this model is an example of the type of analysis for which Berger[32] argues that robust Bayesian analysis is required.)

The cost of this approach is that the information about our quantity of interest can only be narrowed to a class of posterior densities. Fortunately, in the present case, this class can be conveniently summarized as an inequality (rather than an equality) statement regarding the credible intervals: the probability that the actual quantity of interest is within the computed interval is always at least as great as the nominal coverage.[e]

---

[e]One objection to our procedure is that assuming a zero prior probability for τ outside the interval $[\tau_0,\tau_1]$ may be unrealistic. Of course, one may simply enlarge the prior interval to include the nonzero density area, but a probabilistic version is easy to construct: first, elicit the interval endpoints, τ_0 and τ_1, and also a fraction α (e.g., 0.05) such that $1-\alpha$ fraction of the time τ falls within $[\tau_0,\tau_1]$, i.e., $\int_{\tau_0}^{\tau_1}P(\tau)d\tau=1-\alpha$. Then, outside this interval, use a portion of a (single) density to allocate the remaining probability to $[0,\tau_0)$ and $(\tau_1,1]$, so that the sum of the integrals of the two add to α. In this way, every member of the class of densities on the full interval $[0, 1]$ is proper, and robust Bayesian analysis can proceed as before. We have found these changes inconsequential in the real applications we have studied, although our software offers the option to handle this situation.

Table 1 Replication and extension of 95% CIs

Quantity of interest	Risk factor			
	IV Drug use	Smoking	Pneumonia	Cirrhosis
OR	1.44–2.70	1.81–3.64	1.01–1.88	1.01–2.49
RR	1.31–2.45	1.52–3.13	1.01–1.73	1.03–2.17
RD	0.03–0.19	0.05–0.27	0.00–0.13	0.00–0.20
$\Pr(Y=1|X=0)$	0.05–0.26	0.05–0.26	0.08–0.31	0.08–0.31
$\Pr(Y=1|X=1)$	0.10–0.38	0.13–0.47	0.09–0.40	0.10–0.50

The OR row is an exact replication and the others are extensions using the methods developed here. The last two rows give the probability of contracting bacterial pneumonia given the absence and presence of the given risk factor, respectively. (From Ref. [34].)

Thus reporting results from our methods can be as simple as from existing methods used in this field. This is a considerable advantage over other applications of robust Bayesian methods because although considerable research in statistics has focused on these approaches, applications are still fairly rare. The reason is that the class of priors turns into a class of posteriors, and it is the entire class of posteriors that constitutes the research results. Summarizing a single multidimensional posterior is already fairly difficult; summarizing a class of posteriors in multidimensional space only creates more problems. Yet in our case, the class of posteriors can be reduced for presentation purposes to a credible interval, the only disadvantage of which is that it provides an inequality rather than equality statement.

An Empirical Example

We now reanalyze data provided in Tumbarello et al.,[34] the largest case-control study ever conducted of the risk factors leading to bacterial pneumonia in HIV-infected patients. We focus on their univariate analysis of risk factors in 350 cases and 700 controls. The authors report prior knowledge of τ (the fraction of HIV-seropositive individuals in the general population who have an episode of bacterial pneumonia), based on previous studies, as falling in the interval [0.097, 0.29]. We interpret this to be a 99% prior interval and, for simplicity, assume the remaining $\alpha=0.01$ mass to be uniform in [0, 0.097) and (0.29, 0.6); our experiments (not shown) indicate that inferences change very little across many reasonable choices for the density outside the [0.097, 0.29] interval for τ.

The first row of Table 1 replicates the CI for the univariate odds ratio reported in Ref. [34] for each of four risk factors they considered. The second row of the table gives $\geq 95\%$ CIs for the risk ratio of each risk factor, using the robust Bayesian methods described in ''A Robust Bayesian Interpretation'' (with $m=1000$ simulations). As the table shows, the intervals for RR indicate

somewhat smaller effects than OR, with the most noticeable effects for IV drug use and smoking.

Perhaps more interesting are the final three rows of the table which offer information not reported in any form in the original article. For example, Table 1 shows that smoking increases the probability of bacterial pneumonia between 0.05 and 0.27 (a $\geq 95\%$ CI for the risk difference, RD). For another example, the $\geq 95\%$ CI for the base probability of an IV drug user contracting bacterial pneumonia is 0.10–0.38. These examples and the other information in Table 1 all seem like valuable information for researchers and others interested in the study and its results. The information existed in the data from this study, but they are revealed only by application of the methods offered here.

INFERENCE WITHOUT FULL POPULATION INFORMATION OR ADDITIONAL ASSUMPTIONS: DENSITY CASE-CONTROL

We now give analogous results for density case-control designs to those provided in ''Inferences Without Full Population Information or Assumptions: Classic Case-Control'' for classic case-control designs and provide informative bounds for π, RR, RD, $\lambda_i(t)$, and rd_t when no information or only partial information is available about the τ_j's in ''Inference Under Full Population Information'' (we skip rr because it does not depend on τ_j and is estimable from the conditional logit procedure). Inference from density case-control samples is complicated by the fact that more than one piece of population information is involved: there is a τ_j associated with each of the M risk sets.[f] We

---

[f]See Ref. [35] for methods for estimating the risk using only the overall cohort disease rate. However, this estimator is less efficient than the one proposed by Langholz and Ørnulf[30] and employed here. See also Ref. [6] on estimating rates using information on the crude incidence density.

elicit the minimum and maximum values of each τ_j, which we denote $\underline{\tau}_j$ and $\bar{\tau}_j$, respectively. The interval $(\underline{\tau}_j, \bar{\tau}_j)$ can change with j or can be constant over risk sets; it can be specified to include 100% of the prior density, as in ''A Proposed 'Available Information' Assumption,'' or $1-\alpha$ fraction of the density, as in ''A Robust Bayesian Interpretation.'' When we are completely ignorant over τ_j, the interval is $(0, 1)$.

To simplify notation, let $r_k=e^{x_k b}$ and $r^j=\sum_{k \in R_j} r_k$. Clearly $r_k > 0$ and $r^j > 0 \; \forall k, j$. Then the estimators for the rate, cumulative rate, and risk in Eqs. 12–14, respectively, can be rewritten as:

$$\lambda_i(t_j) = \frac{r_i \tau_j}{r^j} \tag{23}$$

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j} \tag{24}$$

and

$$\pi_i = \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)}$$

$$= 1 - \exp\left(-\sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j} \right) \tag{25}$$

respectively. We now develop bounds for the quantities of interest as functions of $\underline{\tau}_j$ and $\bar{\tau}_j$.

Risk

From Eq. 25, we have

$$\frac{\partial \pi_i}{\partial \tau_j} = \frac{e^{-H(X_i, T_i)} r_i}{r^j} > 0 \quad \forall j$$

hence the risk is a monotonically increasing function with respect to every τ_j. Denote $\underline{\pi}_i$ and $\bar{\pi}_i$ as the values of π_i with all τ_j set to $\underline{\tau}_j$ and $\bar{\tau}_j$, respectively; then the bounds for π_i are simply $\pi_i \in [\underline{\pi}_i, \bar{\pi}_i]$. For example, when we are completely ignorant about τ_j and therefore $(\underline{\tau}_j, \bar{\tau}_j]$ is $(0, 1]$, the bounds give $\pi_i \in (0, 1 - \exp(-\sum_{t_j \in R_j}(r_i/r^j)))$.

Risk Ratio

We now examine $\mathrm{RR} = \pi_1/\pi_0$, where π_i is as in Eq. 25, $i=0, 1$. We have

$$\frac{\partial \mathrm{RR}}{\partial \tau_j} = \frac{r_1(1 - \pi_1)\pi_0 - r_0(1 - \pi_0)\pi_1}{\pi_0^2 r^j} \tag{26}$$

the sign of which is determined by that of the numerator. When the numerator is positive, i.e., when

$$\mathrm{rr} = r_1/r_0 > \frac{\pi_1(1 - \pi_0)}{(1 - \pi_1)\pi_0} = \mathrm{OR} \tag{27}$$

the partial derivative is positive and so RR increases with respect to τ_j. Otherwise, the derivative is negative and RR decreases with respect to τ_j. In Appendix B, we show that Eq. 27 holds whenever $\mathrm{rr} < 1$, independent of the values of τ_j. Similarly, the sign is reversed whenever $\mathrm{rr} > 1$. Hence RR is either monotonically increasing or monotonically decreasing with respect to τ_j, for all j.

Thus when $\mathrm{rr} < 1$, RR is monotonically increasing with respect to all τ_j and is therefore bounded by $\underline{\mathrm{RR}}$ and $\overline{\mathrm{RR}}$, which are values of RR with all τ_j set to their minimum and maximum values, respectively. Otherwise, RR is monotonically decreasing and the bounds are $\overline{\mathrm{RR}}$ and $\underline{\mathrm{RR}}$. In short, RR is bounded by

$$\mathrm{RR} = \in \left[\min(\underline{\mathrm{RR}}, \overline{\mathrm{RR}}), \max(\underline{\mathrm{RR}}, \overline{\mathrm{RR}}) \right] \tag{28}$$

It is easy to see that $\lim_{\tau_j \to 0} \mathrm{RR} = r_1/r_0$, hence when no information is available for τ_j and therefore $(\underline{\tau}_j, \bar{\tau}_j]$ is $(0, 1)$, the bounds become

$$\mathrm{RR} \in (\min(r_1/r_0, \overline{\mathrm{RR}}), \max(r_1/r_0, \overline{\mathrm{RR}}))$$

where $\overline{\mathrm{RR}}$ is RR evaluated at $\tau_j = 1 \; \forall j$.

Risk Difference

The case of $\mathrm{RD} = \pi_1 - \pi_0$ is more complicated because RD is not a monotonic function of the τ_j's and $\partial \mathrm{RD}/\partial \tau_j = [r_1 e^{-H(T_1, X_\ell)} - r_0 e^{-H(T_0, X_0)}]/r^j$ can change signs depending on the values of τ_j. Under the proportional hazards model where r_i and r^j are not functions of time, however, we can reduce the analytically difficult or even intractable problem of constrained optimization in multidimensional space to a simple one in which RD is a one-dimensional function of the cumulative baseline hazard, which is a monotone function of the τ_j's.

Let $Q(\tau_j) = \sum_{j=1}^{M} \tau_j/r^j$ denote the cumulative baseline hazard rate, and note that $\partial Q(\tau_j)/\partial \tau_j = 1/r^j > 0$ for all j, so $Q(\tau_j)$ is monotonically increasing in all τ_j's and therefore bounded between $\underline{Q} = Q(\underline{\tau}_j)$ and $\bar{Q} = Q(\bar{\tau}_j)$. Now rewrite RD in terms of Q. From Eq. 24, we have $H(T_k, X_k) = r_k Q$ for $k = 0, 1$; hence

$$\mathrm{RD} = (1 - e^{-r_1 Q}) - (1 - e^{-r_0 Q})$$

$$= e^{-r_0 Q} - e^{-r_1 Q} \tag{29}$$

and

$$\frac{\partial \mathrm{RD}}{\partial \tau_j} = \frac{r_1 e^{-r_1 Q} - r_0 e^{-r_0 Q}}{r^j} \tag{30}$$

RD is not monotone in Q, but Q is a scalar and we know its bounds, which brings us to a situation mathematically similar to analyzing RD in classic case-control designs.

Let Q^* be the solution to the first-order condition $\partial \mathrm{RD}/\partial\tau_j=0$. From Eq. 30, we can solve for $Q^*=[1/(r_1-r_0)]\ln(r_1/r_0)$. Then from Eq. 29, we have

$$\mathrm{RD}(Q^*) \;=\; (r_0/r_1)^{r_0/(r_1-r_0)} \;-\; (r_0/r_1)^{r_1/(r_1-r_0)} \qquad (31)$$

To obtain the bounds for RD, we first see whether $[\underline{Q},\overline{Q}]$ contains Q^*. If it does, then

$$\mathrm{RD} \in [\min(\underline{\mathrm{RD}},\overline{\mathrm{RD}},\mathrm{RD}^*), \max(\underline{\mathrm{RD}},\overline{\mathrm{RD}},\mathrm{RD}^*)] \qquad (32)$$

where $\underline{\mathrm{RD}}=\mathrm{RD}(\underline{Q})$, $\overline{\mathrm{RD}}=\mathrm{RD}(\overline{Q})$, and $\mathrm{RD}^*=\mathrm{RD}(Q^*)$. Otherwise, the bounds are

$$\mathrm{RD} \in [\min(\underline{\mathrm{RD}},\overline{\mathrm{RD}}), \max(\underline{\mathrm{RD}},\overline{\mathrm{RD}})] \qquad (33)$$

When no information is available for τ_j, the bounds become

$$\mathrm{RD} \in [\min(0,\mathrm{RD}^*), \max(0,\mathrm{RD}^*)] \qquad (34)$$

Rate

From Eq. 23, we see that $\partial\lambda_i(t_j)/\partial\tau_j=r_i/r^j>0$; hence $\lambda_i(t_j)$ is monotonically increasing in τ_j. It is therefore bounded in $(\underline{\lambda_i(t_j)}, \overline{\lambda_i(t_j)})$, where $\underline{\lambda_i(t_j)}$ is $\lambda_i(t_j)$ evaluated at $\underline{\tau}_j$ and $\overline{\lambda_i(t_j)}$ is $\lambda_i(t_j)$ evaluated at $\overline{\tau}_j$. When we are ignorant with respect to the τ_j's, the rate is bounded as $(0, r_i/r^j)$.

Rate Difference

Because $\partial \mathrm{rd}/\partial\tau_j=(r_1-r_0)/r^j$, rd is monotonically increasing in τ_j if $r_1>r_0$ and decreasing otherwise. Hence the bound on the rate difference is

$$\mathrm{rd} \in [\min[\mathrm{rd}(\underline{\tau}_j),\mathrm{rd}(\overline{\tau}_j)], \max[\mathrm{rd}(\underline{\tau}_j),\mathrm{rd}(\overline{\tau}_j)]] \qquad (35)$$

and when ignorant of all information on τ_j, the bounds are

$$\mathrm{rd} \in [\min[0,(r_1-r_0)/r^j], \max[0,(r_1-r_0)/r^j]]$$

CONCLUSION

As is increasingly recognized, the quantity of interest in most case-control studies is not the odds ratio, but rather some version or function of a probability, risk ratio, risk difference, rate, rate ratio, or rate difference, depending on context.[6,7,36–41] We provide the methods to estimate each of these quantities from case-control studies, even if no auxiliary information, or only limited auxiliary information, is available.

Unless the odds ratio happens to approximate a parameter of central substantive interest, which is quite rare, we suggest that it should not be reported any more frequently than any other intermediate quantity in statistical calculations. We suggest instead that researchers justify their assumption regarding bounds on τ (in classic case-control studies) or τ_j (in risk set case-control studies) in the data or methods section of their work. Then, they can substitute the confidence interval (CI) now reported for the odds ratio with the CI for their chosen quantity (or quantities) of interest. For example, instead of an uninformative but presently common reporting style:

> the effect of smoking on lung cancer is positive OR=1:38 (95% CI 1.30–1.46)

researchers could give the much more interesting:

> smoking increases the risk of contracting lung cancer by a factor of between 2.5 and 3.1 (a $\geq$95% CI)

or

> smoking increases the probability of contracting lung cancer between 0.022 and 0.051 (a $\geq$95% CI)

If uncertainty exists over the appropriate bounds for the unknown quantities, we suggest using the widest bounds, conducting sensitivity analyses by showing how the CI depends on different assumptions or setting α to a value other than zero.

The methods discussed here are meant to improve presentation and increase the amount of information that can be extracted from existing models and data collections. They do not enable scholars to ignore the usual threats to inference (measurement error, selection bias, confounding, etc.) that must be avoided in any study.

APPENDIX A

SIMPLIFYING MANSKI'S BOUNDS ON THE RISK DIFFERENCE

Proving Eq. 18 requires algebra only. For simplicity, let $P_{ab}=\Pr(X_a|Y=b)$, so that $\mathrm{OR}=(P_{11}P_{00})=(P_{01}P_{10})$. Then, omitting tedious but straightforward algebra at several stages, $\phi=(\mathrm{OR}P_{01}^2/P_{11}^2)^{1/2}=\sqrt{\mathrm{OR}}P_{01}/P_{11}$, and $\gamma = \sqrt{\mathrm{OR}}/(\sqrt{\mathrm{OR}} + P_{11}/P_{10})$. Then the components of RD_γ are

$$P_{11}\gamma = \frac{\sqrt{\mathrm{OR}}P_{10}P_{11}}{\sqrt{\mathrm{OR}}P_{10}+P_{11}}, \qquad P_{01}\gamma = \frac{\sqrt{\mathrm{OR}}P_{01}P_{10}}{\sqrt{\mathrm{OR}}P_{10}+P_{11}}$$

$$P_{10}(1-\gamma) = \frac{P_{10}P_{11}}{\sqrt{\mathrm{OR}}P_{10}+P_{11}}, \quad P_{00}(1-\gamma) = \frac{P_{11}P_{00}}{\sqrt{\mathrm{OR}}P_{10}+P_{11}}$$

and so putting the terms together yields $\mathrm{RD}_\gamma = \sqrt{\mathrm{OR}}/(1+\sqrt{\mathrm{OR}}) - 1/(1+\sqrt{\mathrm{OR}}) = (\sqrt{\mathrm{OR}}-1)/(\sqrt{\mathrm{OR}}+1)$.

APPENDIX B

MONOTONICITY OF RISK RATIO UNDER DENSITY CASE-CONTROL DESIGNS

We show here that if rr<1, then rr$>[\pi_1(1-\pi_0)]/[(1-\pi_1)\pi_0]$ (the $r_1/r_0>1$ case is similar). Let $H_k = H(T_k, X_k)$, $k=0, 1$. From the definition of π_1 and π_0, $[\pi_1(1-\pi_0)]/[(1-\pi_1)\pi_0]$ can be simplified to $(e^{H_1}-1)/(e^{H_0}-1)$. Because rr$=H_1/H_0$, we only need to show that if $H_1/H_0<1$, then $H_1/H_0>(e^{H_1}-1)/(e^{H_0}-1)$ or, equivalently, $H_1(e^{H_0}-1)>H_0(e^{H_1}-1)$.

The Taylor series expansions of $e^{H_1}-1$ and $e^{H_0}-1$ at 0 give

$$e^{H_1} - 1 = H_1 + (1/2)H_1^2 + (1/3!)H_1^3 + \ldots \quad (36)$$

$$e^{H_0} - 1 = H_0 + (1/2)H_0^2 + (1/3!)H_0^3 + \ldots \quad (37)$$

and hence

$$H_1(e^{H_0} - 1) = H_1 H_0 + (1/2)H_0^2 H_1$$
$$+ (1/3!)H_0^3 H_1 + \ldots \quad (38)$$

$$H_0(e^{H_1} - 1) = H_0 H_1 + (1/2)H_1^2 H_0$$
$$+ (1/3!)H_1^3 H_0 + \ldots \quad (39)$$

The first terms in Eqs. 38 and 39 are equal, and when $H_1/H_0<1$, hence $H_1<H_0$, all other terms in Eq. 38 are greater than the corresponding terms in Eq. 39 (because both $H_0>0$ and $H_1>0$ always). Thus when $H_1/H_0<1$, $H_1(e^{H_0}-1)>H_0(e^{H_1}-1)$.

ACKNOWLEDGMENTS

REFERENCES

1. Moynihan, R.; Bero, L.; Ross-Degnan, D.; Henry, D.; Lee, K.; Watkins, J.; Mah, C.; Soumerai, S.B. Coverage by the news media of the benefits and risks of medications. N. Engl. J. Med. **June 1 2000**, *342* (22), 1645–1650.

2. Manski, C.F. *Identification Problems in the Social Sciences*; Harvard University Press, 1995; p. 31.

3. Manski, C.F. Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. In *Nonparametric Identification Under Response-Based Sampling*; Hsiao, C., Morimune, K., Powell, J., Eds.; Cambridge University Press, 1999.

4. Greenland, S. On the need for the rare disease assumption in case-control studies. Am. J. Epidemiol. **1982**, *116* (3), 547–553.

5. Deeks, J.; Sackett, D.; Altman, D. Down with odds ratios. Evid.-Based Med. **1996**, *1* (6), 164–166.

6. Greenland, S. Interpretation and choice of effect measures in epidemiologic analysis. Am. J. Epidemiol. **1987**, *125* (5), 761–768.

7. Davies, H.T.O.; Manouche, T.; Iain, C.K. When can odds ratio mislead. Br. Med. J. **March 28 1998**, *31*, 989–991.

8. Rothman, K.J.; Greenland, S. *Modern Epidemiology*, 2nd Ed.; Lippincott-Raven: Philadelphia, 1998; pp. 113, 244–245.

9. Prentice, R.L.; Breslow, N.E. Retrospective studies and failure-time models. Biometrica **1978**, *65*, 153-155.

10. Chamberlain, G. Analysis of covariance with qualitative data. Rev. Econ. Stud. **1980**, *XLVII*, 225–238.

11. Greenland, S. Modeling risk ratios from matched cohort data: An estimating equation approach. Appl. Stat. **1994**, *43* (1), 223–232.

12. Goldstein, L.; Langholz, B. Asymptotic theory for nested case-control sampling in the cox regression model. Ann. Stat. **1992**, *20* (4), 1903–1928.

13. Borgan, Ø.; Langgholz, B.; Goldstein, L. Methods for the analysis of sampled cohort data in the cox proportional hazard model. Ann. Stat. **1995**, *23*, 1749–1778.

14. Langholz, B.; Goldstein, L. Risk set sampling in epidemiologic cohort studies. Stat. Sci. **1996**, *11* (1), 35–53.

15. Langholz, B.; Thomas, D.C. Efficiency of cohort sampling designs: Some surprising results. Biometrics **1991**, *47*, 1563–1571.

16. Lubin, J.H.; Gail, M.H. Sampling strategies in nested case-control studies. Environ. Health Perspect. **1994**, *102* (Suppl. 8), 47–51.

17. Robins, J.M.; Gail, M.H.; Lubin, J.H. More on biased selection of controls for case-control analyses of cohort studies. Biometrics **1986**, *42*, 293–299.

18. Prentice, R.L. A case-cohort design for epidemiological studies and disease prevention trials. Biometrica **1986**, *73*, 1–11.

19. Greenland, S. Multivariate estimation of exposure-specific incidence from case-control studies. J. Chronic. Dis. **1981**, *34*, 445–453.

20. Neutra, R.R.; Drolette, M.E. Estimating exposure-specific disease rates from case-control studies using Bayes theorem. Am. J. Epidemiol. **1978**, *108* (3), 214–222.

21. Cornfield, J. A method of estimating comparative rates from clinical data: Application to cancer of the lung, breast and cervix. J. Natl. Cancer Inst. **1951**, *11*, 1269–1275.

22. Anderson, J.A. Separate-sample logistic discrimination. Biometrika **1972**, *59*, 19–35.

23. Prentice, R.L.; Pyke, R. Logistic disease incidence models and case-control studies. Biometrica **1979**, *63*, 403–411.

24. Mantel, N. Synthetic retrospective studies and related topics. Biometrics **1973**, *29*, 479–486.

25. Manski, C.F. The estimation of choice probabilities from choice based samples. Econometrics **November 1977**, *45* (8), 1977–1988.

26. King, G.; Zeng, L. Logistic regression in rare events data. Polit. Anal. **Spring 2001**, *9* (2), 137–163.

27. King, G.; Tomz, M.; Wittenberg, J. Making the most of statistical analyses: Improving interpretation and presentation. Am. J. Polit. Sci. **April 2000**, *44* (2), 341–355.

28. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, 1995.

29. Beck, N.; King, G.; Zeng, L. Improving quantitative studies of international conflict: A conjecture. Am. Polit. Sci. Rev. **March 1999**, *94* (1), 21–36.

30. Langholz, B.; Ørnulf, B. Estimation of absolute risk from nested case-control data. Biometrics **June 1997**, *53*, 767–774.

31. Yule, G.U. On the methods of measuring the association between two attributes. J. R. Stat. Soc. **1912**, *75*, 579–642.

32. Berger, J. An overview of robust Bayesian analysis (with discussion). Test **1994**, *3*, 5–124.

33. Insua, D.R.; Fabrizio, R. *Bayesian Analysis*; Springer Verlag, 2000.

34. Tumbarello, M.; Tacconelli, E.; de Gaetano, K.; Ardit, F.; Pirronti, T.; Claudia, R.; Ortona, L. Bacterial pneumonia in HIV–infected patients: Analysis of risk factors and prognostic indicators. J. Acquir. Immune Defic. Syndr. Human Retrovirol. **1998**, *18*, 39–45.

35. Benichou, J.; Gail, M. Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics **1995**, *51*, 182–194.

36. Nurminen, M. To use or not to use the odds ratio in epidemiologic analysis. Eur. J. Epidemiol. **1995**, *11*, 365–371.

37. Zhang, J.; Kai Yu, F. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. N. Engl. J. Med. **1998**, *280* (19), 1690–1691.

38. Davies, H.T.O.; Manouche, T.; Kinloch Iain, C. Authors reply. Br. Med. J. **October 24 1998**, *317*, 1156–1157.

39. Deeks, J. Odds ratio should be used only in case-control studies and logistic regression analyses. Br. Med. J. **October 24 1998**, *317*, 1155–1156.

40. Michael, B.; Bracken, J.C. Avoidable systematic error in estimating treatment effects must not be tolerated. Br. Med. J. **October 24 1998**, *317*, 11–56.

41. Altman, D.G.; Deeks, J.J.; Sackett, D.L. Odds ratios should be avoided when events are common. Br. Med. J. **Nov. 7 1998**, *317*, 1318.

Request Permission or Order Reprints Instantly!

Request Permission/Order Reprints

Reprints of this article can also be ordered at
http://www.dekker.com/servlet/product/DOI/101081EEBS120023390