

THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

Research Paper No. 4

**Estimating Risk and Rate Levels, Ratios, and
Differences in Case-Control Studies**

**Gary King
Lanche Zeng**

May 2001
ISSN 0000 0000

**HARVARD BURDEN OF DISEASE UNIT
NATIONAL INSTITUTE ON AGING GRANT 1-P01-AG17625**

THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

This working paper series reports on research supported by the National Institute on Aging program grant entitled The Global Burden of Disease 2000 in Aging Populations (1-P01-AG17625). The purpose of the grant is to strengthen the methodological and empirical bases for undertaking comparative assessments of health problems, their determinants and consequences in aging populations.

Since the publication of the Global Burden of Disease Study 1990, there has been increasing interest in comparative analyses of health outcomes, determinants and consequences. A major revision of the Global Burden of Disease Study has been launched for the year 2000 with the full commitment of the World Health Organization (WHO). The Global Programme on Evidence for Health Policy at WHO has developed a Global Burden of Disease Network, which operates in parallel to the research conducted as part of the program project. The program project will strengthen the scientific basis for the large-scale undertaking led by WHO at the global, regional and national level.

The purpose of this working paper series is to present original research that emerges from the various project components of this program grant. The views expressed in these working papers are those of the author(s) and do not necessarily reflect the views of the Harvard Burden of Disease Unit, the World Health Organization nor the National Institute on Aging.

THE HARVARD BURDEN OF DISEASE UNIT

The Harvard Burden of Disease Unit was established to design, test, and implement methodologies to aid in the effective allocation of health resources. To achieve this end, the Unit conducts research in collaboration with national governments, international agencies and other researchers and policy-makers. The Unit's research has two main foci:

- to forge the theory, design, and implementation of approaches to the combined measurement of mortality and non-fatal health outcomes, in order to develop valid, reliable, comparable and comprehensive measures of population health and comparative assessments of the burden of diseases, injuries and risk factors; and
- to investigate the costs, efficacy and effectiveness of major health interventions applied in diverse settings, toward the goal of establishing a broad database on cost-effectiveness.

Harvard Burden of Disease Unit
Center for Population and Development Studies
9 Bow Street
Cambridge, MA 02138
www.hsph.harvard.edu/organizations/bdu

Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies¹

Gary King² and Lanche Zeng³

Abstract

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, 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.

¹ We thank Sander Greenland for his generosity, insight, and wisdom about the epidemiological literature, Norm Breslow and Ken Rothman for many helpful explanations, and Chuck Manski for his suggestions, provocative econometric work, and other discussions. Thanks also to Neal Beck, Rebecca Betensky, Josue Guzman, Bryan Langholz, Meghan Murray, Adrain Raferty, Ted Thompson, Jon Wakefield, Clarice Weinberg, David Williamson for helpful discussions, Ethan Katz for research assistance, and Mike Tomz for spotting an error in an earlier version. Thanks to the National Science Foundation (IIS-9874747), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institute on Aging (PO1 AG17625-01), the World Health Organization, and the Center for Basic Research in the Social Sciences for research support. Software to implement the methods in this paper is available for Stata and for Gauss from <http://Gking.Harvard.Edu>.

² Professor of Government, Harvard University, and Senior Advisor, Global Programme on Evidence for Health Policy, World Health Organization, Center for Basic Research in the Social Sciences, Harvard University, Cambridge, MA 02138

³ Department of Political Science, George Washington University, Washington, DC 20052

Moynihan et al. [1] express the conclusions of nearly all who have written about reporting standards: “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 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, and in recent years 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 HIV-infected individuals, 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>).

1 Quantities of Interest

For subject i ($i = 1, \dots, n$) in time interval $(t, t + dt)$ ($t \in (t_{0i}, t_{1i})$; for $t_{0i} < t_{1i}$), define the outcome variable Y_{it} as 1 for “events” (such as disease incidence) and 0 otherwise. The variable t usually indexes time but can denote any continuous variable. Define a k -vector of covariates and constant term as X_i .¹ Also 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). In some studies such as perinatal epidemiology, analyses of conditions with brief risk periods like acute intoxication, and in the analysis of some prevalence data, scholars only measure, or only can measure, Y_i , defined as 1 if $\int Y_{it} dt > 0$ and 0 otherwise ([4], p.548).

Quantities of interest that are generally a function of t include the *rate* (or “hazard rate” or “instantaneous rate”), $\lambda_i(t) = \lim_{(t-s) \rightarrow 0} \Pr(Y_{it} = 1 | Y_{is} = 0, \forall s < t) / (t - s)$ (which is normally specified to be a function of X_i also) and functions of the rate. For example, the *rate ratio*, $rr_t = \lambda_\ell(t) / \lambda_0(t)$, and the *rate difference*, $rd_t = \lambda_\ell(t) - \lambda_0(t)$ 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”),

$$\pi_i = \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)} = 1 - \exp\left(-\int_{t_{0i}}^{t_{1i}} \lambda_i(t) dt\right), \quad (1)$$

(where $H(T_i, X_i)$ is the *cumulative hazard rate* for individual i over time $T_i = (t_{0i}, t_{1i})$), the *risk ratio*, $RR \equiv \Pr(Y = 1 | X_\ell) / \Pr(Y = 1 | X_0)$, 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), $RD \equiv \Pr(Y = 1 | X_\ell) - \Pr(Y = 1 | X_0)$. The probability is evaluated at some values of the explanatory variables, such as X_0 or X_ℓ . The quantity RD

¹We assume for purpose of this paper that the X 's are not time dependent.

is the increase in probability, and RR is the factor by which the probability increases (an $(RR - 1) \times 100$ 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 can also be computed as a function of these quantities with the methods described below.

The other key quantity often discussed in the epidemiological literature is the *odds ratio*

$$OR \equiv \frac{\Pr(Y = 1|X_\ell)/\Pr(Y = 0|X_\ell)}{\Pr(Y = 1|X_0)/\Pr(Y = 0|X_0)} = \frac{\Pr(X_\ell|Y = 1)\Pr(X_0|Y = 0)}{\Pr(X_0|Y = 1)\Pr(X_\ell|Y = 0)}, \quad (2)$$

where the second equality holds by Bayes Theorem. The key advantage of OR is that it has been easier to estimate than the other quantities. In logit 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 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 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 [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], p.761). 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, Tavakoli, and Crombie [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 that information about all relevant quantities are always available even in case-control data. (Unfortunately, the term “relative risk” no longer seems useful since it has been co-opted to denote such diverse quantities as rr , RR, and OR.)

2 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:

$$\Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{-X_i\beta}} \quad (3)$$

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], p.113), 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 ($j = 1, \dots, M$, 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 non-parametrically (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:

$$\lambda_i(t) = \lambda(t)r(X_i, \beta) \quad (4)$$

where in most applications $r(X_i, \beta) = e^{X_i\beta}$ and the baseline hazard $\lambda(t)$ does not vary over subjects.²

3 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 $\Pr(Y_i = 1|X_j) \rightarrow 0$ 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: $\lim_{\tau \rightarrow 0} \text{OR} = \text{RR}$.

This limit result is attractive since 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 Equation 2 with their sample analogues (e.g., $\Pr(X_\ell|Y = 1)$ can be estimated by the fraction of cases for which $X_i = X_\ell$ among all observations where $Y_i = 1$). For another example, in logistic regression (3), 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 (since the elements of $X_\ell - X_0$ corresponding to those parameters are zero).

As is well known (e.g., [8], pp. 244-245), 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 ($\lim_{\tau \rightarrow 0} \text{RD} = 0$), 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

²The 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.

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 non-absorbing states or when studying prevalence.

Density Case-Control In the commonly used proportional hazard model (4), 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 get the disease, conditional on the total number of cases in the set being one:

$$\Pr(Y_{it_j} = 1 | R_j) = \frac{\Pr(Y_{it_j} = 1)}{\sum_{k \in R_j} \Pr(Y_{kt_j} = 1)} = \frac{e^{X_i \beta}}{\sum_{k \in R_j} e^{X_k \beta}}, \quad (5)$$

where i indexes the case in the risk set and the summation in each denominator is over all observations in the risk set, $k = 1, \dots, n_j$, where n_j is the number of observations, or the size, of R_j . The likelihood function then is the product of terms like (5) over all M sampled risk sets (see, e.g., [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 (5) 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 . Equation 5 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.)

Equation (5) 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 [12, 13, 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 [16, 17], and of alternative sampling designs, see [18].

With β consistently estimated by the conditional logit approach, we can compute the rate ratio using the incidence model (4) even though the baseline hazard rate $\lambda(t)$ is not estimated:

$$\text{rr} = \frac{\lambda_\ell(t)}{\lambda_0(t)} = \frac{\lambda(t)e^{X_\ell\beta}}{\lambda(t)e^{X_0\beta}} = \frac{e^{X_\ell\beta}}{e^{X_0\beta}} = e^{(X_0 - X_\ell)\beta} \quad (6)$$

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, or risk ratio.

4 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, 20, 21]. For example, in the simple case of a single binary explanatory variable, the risk can be estimated as

$$\begin{aligned} \pi = \Pr(Y = 1|X, \tau) &= \frac{\Pr(X|Y = 1) \Pr(Y = 1)}{\Pr(X)} \\ &= \frac{\Pr(X|Y = 1)\tau}{\Pr(X|Y = 1)\tau + \Pr(X|Y = 0)(1 - \tau)} \end{aligned} \quad (7)$$

where $\Pr(X|Y = 1)$ and $\Pr(X|Y = 0)$ are replaced with the sample mean of X among subjects where $Y = 1$ and $Y = 0$, respectively. From this expression, we can compute $\text{RR} = \Pr(Y = 1|X_\ell, \tau) / \Pr(Y = 1|X_0, \tau)$ or $\text{RD} = \Pr(Y = 1|X_\ell, \tau) - \Pr(Y = 1|X_0, \tau)$.

For another example, the commonly used logistic regression model (3) 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:

$$\beta_0 - \ln \left[\left(\frac{1 - \tau}{\tau} \right) \left(\frac{\bar{y}}{1 - \bar{y}} \right) \right], \quad (8)$$

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 [22, 23, 24, 25]; see [19], for a comprehensive review). Let β_τ denote the logit coefficient vector, with the first element

corrected as in Equation 8. Then the quantities of interest are $\pi = \Pr(Y = 1|X, \tau) = [1 + e^{-X\beta\tau}]^{-1}$, $\text{RR} = [1 + e^{-X_\ell\beta\tau}]^{-1}/[1 + e^{-X_0\beta\tau}]^{-1}$, and $\text{RD} = [1 + e^{-X_\ell\beta\tau}]^{-1} - [1 + e^{-X_0\beta\tau}]^{-1}$. For estimation, we can plug in the maximum likelihood estimates or use improved methods with lower mean square error [26]. Standard errors for any of these quantities can be computed easily by simulation [27].

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 [28], where the key population information needed is $\tau_j = n_j/N_j$, the sampling fraction for each observed risk set R_j ($j = 1, \dots, M$). 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:³

$$\lambda(t_j) = \frac{1}{\sum_{k \in R_j} (e^{X_k b})(1/\tau_j)} = \frac{1}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}}, \quad (9)$$

and with this we can estimate the rate, $\lambda_i(t)$, from Equation (4):

$$\lambda_i(t_j) = \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}}, \quad (10)$$

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

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \quad (11)$$

and hence the risk:

$$\pi_i = \Pr(Y = 1|X_i) = 1 - e^{-H(T_i, X_i)} = 1 - \exp\left(-\sum_{t_j \in T_i} \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}}\right) \quad (12)$$

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.

5 Inference With Bayesian Prior Assumptions

Sections 3 and 4 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

³For 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.

exactly, but some 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 too. 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, since the required prior density cannot be fully specified. Second, Bayesian estimates yield biased inferences when prior information is biased. 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. These problems are especially severe in case-control studies since 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. We therefore pursue “robust Bayesian” methods below that allow full or partial ignorance to be represented accurately without biasing inferences.

6 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.

6.1 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, $\text{RR} \in [\min(1, \text{OR}), \max(1, \text{OR})]$, 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

$$\phi = \left[\frac{\Pr(X_0|Y = 1) \Pr(X_0|Y = 0)}{\Pr(X_\ell|Y = 1) \Pr(X_\ell|Y = 0)} \right]^{\frac{1}{2}}, \quad (13)$$

$$\gamma = \frac{\phi \Pr(X_\ell|Y = 0) - \Pr(X_0|Y = 0)}{\phi \Pr(X_\ell|Y = 0) - \Pr(X_0|Y = 0) - [\phi \Pr(X_\ell|Y = 1) - \Pr(X_0|Y = 1)]}, \quad (14)$$

and

$$\text{RD}_\gamma = \frac{\Pr(X_\ell|Y = 1)\gamma}{\Pr(X_\ell|Y = 1)\gamma + \Pr(X_\ell|Y = 0)(1 - \gamma)} - \frac{\Pr(X_0|Y = 1)\gamma}{\Pr(X_0|Y = 1)\gamma + \Pr(X_0|Y = 0)(1 - \gamma)}. \quad (15)$$

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 non-parametric methods are required to estimate each of the component probabilities in Equations 13-15, 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 (1912) [29] “coefficient of colligation,” sometimes called Yule’s Y). The bounds are thus a simple function of the odds ratio:

$$\text{RD} \in \left[\min \left(0, \frac{\sqrt{\text{OR}} - 1}{\sqrt{\text{OR}} + 1} \right), \max \left(0, \frac{\sqrt{\text{OR}} - 1}{\sqrt{\text{OR}} + 1} \right) \right] \quad (16)$$

The advantages of our expression in Equation 16 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 non-parametric methods in simple discrete cases, in logistic

regression, and in a wide variety of multiplicative intercept models, even including many neural networks (such as the popular feed-forward perceptron with a logit output function [30, 31]). Since these neural network models have arbitrary approximation capabilities, Equation 16 can effectively always be applied.

6.2 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, 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], p.31) 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 , and RD_τ denote values of the probability, relative risk, and risk difference, respectively, evaluated at τ . Suppose that τ is known only to fall within the range $[\tau_0, \tau_1]$, where $0 < \tau_0 < \tau_1 < 1$. (Since, 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, since π and RR are monotonic in τ , their bounds are simply

$$\pi \in [\pi_{\tau_0}, \pi_{\tau_1}] \tag{17}$$

and

$$RR \in [\min(RR_{\tau_0}, RR_{\tau_1}), \max(RR_{\tau_0}, RR_{\tau_1})]. \tag{18}$$

The bounds for the risk difference are more complicated since RD_τ is a parabolic func-

tion of τ , and so the bounds differ in the monotonic and non-monotonic 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 $RD_\gamma = (\sqrt{OR} - 1)/(\sqrt{OR} + 1)$. 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 $RD_{\tau_0+\epsilon} - RD_{\tau_0}$ and $RD_{\tau_1+\epsilon} - RD_{\tau_1}$ for a suitably small value of ϵ .) When the relationship is monotonic, the bounds are

$$RD \in [\min(RD_{\tau_0}, RD_{\tau_1}), \max(RD_{\tau_0}, RD_{\tau_1})], \quad (19)$$

and otherwise, they are

$$RD \in [\min(RD_{\tau_0}, RD_{\tau_1}, RD_\gamma), \max(RD_{\tau_0}, RD_{\tau_1}, RD_\gamma)]. \quad (20)$$

6.3 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 non-smokers, 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 Equation 7, 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 $[\tau_0, \tau_1]$ (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 $\Pr(Y = 1|X = 1) \in [.06, .18]$, whereas among non-smokers it is $\Pr(Y = 1|X = 0) \in [.04, .12]$. 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 $RD \in [0, .11]$ whereas our approach yields much narrower bounds of $RD \in [.021, .056]$, the increase in probability due to smoking.

6.4 Classical Confidence Intervals

We now provide a method of computing classical confidence intervals, saving our preferred robust Bayesian interpretation for the following section. Since 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 due to 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.

6.5 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 “non-informative”) prior density within those bounds. However, our procedure can be thought of as a special case of “robust Bayesian analysis” (e.g., [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. (Since data do not help in making inferences about τ , this model is an example of the type of analysis for which Berger [32] argues 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.⁴

⁴One 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 non-zero 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 $[\tau_0, \tau_1]$, that is $\int_{\tau_0}^{\tau_1} P(\tau)d\tau = 1 - \alpha$. 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

6.6 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 [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 Section 6.5 (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.

7 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 Section 6 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 Section 4

found these changes inconsequential in the real applications we have studied, though our software offers the option to handle this situation.

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

Table 1: *Replication and Extension of 95% CIs in [34]. 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, respectively, of the given risk factor.*

(we skip rr since it does not depend on τ_j and is estimable from the conditional logit procedure). Inference from density case control samples are 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.⁵ We 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 Section 6.2, or $1 - \alpha$ fraction of the density, as in Section 6.5. When we are completely ignorant over τ_j , the interval is $(0, 1]$.

To simplify notation, let $r_k = e^{x_k b}$, $r^j = \sum_{k \in R_j} r_k$. Clearly $r_k > 0$ and $r^j > 0 \forall k, j$. Then the estimators for the rate, cumulative rate, and risk in (10), (11), and (12) can be rewritten as:

$$\lambda_i(t_j) = \frac{r_i \tau_j}{r^j} \quad (21)$$

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j} \quad (22)$$

and

$$\pi_i = \Pr(Y = 1|X_i) = 1 - e^{-H(T_i, X_i)} = 1 - \exp\left(-\sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j}\right) \quad (23)$$

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

⁵See [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 Borgan [28] and employed here. See also [6] on estimating rates using information on the crude incidence density.

Risk From (23) 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 $\pi_i \in (0, 1 - \exp(-\sum_{t_j \in R_j} \frac{r_i}{r^j})]$.

Risk Ratio We now examine $\text{RR} = \pi_1/\pi_0$, where π_i is as in (23), $i = 0, 1$. We have

$$\frac{\partial \text{RR}}{\partial \tau_j} = \frac{r_1(1 - \pi_1)\pi_0 - r_0(1 - \pi_0)\pi_1}{\pi_0^2 r^j}, \quad (24)$$

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

$$\text{rr} = r_1/r_0 > \frac{\pi_1(1 - \pi_0)}{(1 - \pi_1)\pi_0} = \text{OR}, \quad (25)$$

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 (25) holds whenever $\text{rr} < 1$, independent of the values of τ_j . Similarly the sign is reversed whenever $\text{rr} > 1$. Hence RR is either monotonically increasing or monotonically decreasing with respect to τ_j , for all j .

Thus, when $\text{rr} < 1$, RR is monotonically increasing with respect to all τ_j and is therefore bounded by $\underline{\text{RR}}$ and $\overline{\text{RR}}$, 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 $\overline{\text{RR}}$ and $\underline{\text{RR}}$. In short, RR is bounded by

$$\text{RR} \in [\min(\underline{\text{RR}}, \overline{\text{RR}}), \max(\underline{\text{RR}}, \overline{\text{RR}})] \quad (26)$$

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

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

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

Risk Difference The case of $RD = \pi_1 - \pi_0$ is more complicated since RD is not a monotonic function of the τ_j 's and $\partial RD / \partial \tau_j = [r_1 e^{-H(T_1, X_1)} - r_0 e^{-H(T_0, X_0)}] / r^j$ 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 $\overline{Q} = Q(\overline{\tau}_j)$. Now rewrite RD in terms of Q . From (22), we have $H(T_k, X_k) = r_k Q$ for $k = 0, 1$, hence

$$RD = (1 - e^{-r_1 Q}) - (1 - e^{-r_0 Q}) = e^{-r_0 Q} - e^{-r_1 Q} \quad (27)$$

and

$$\frac{\partial RD}{\partial \tau_j} = \frac{r_1 e^{-r_1 Q} - r_0 e^{-r_0 Q}}{r^j}. \quad (28)$$

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 $\partial RD / \partial \tau_j = 0$. From Equation (28) we can solve for $Q^* = \frac{1}{(r_1 - r_0)} \ln(r_1 / r_0)$. Then, from (27) we have

$$RD(Q^*) = (r_0 / r_1)^{r_0 / (r_1 - r_0)} - (r_0 / r_1)^{r_1 / (r_1 - r_0)} \quad (29)$$

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

$$RD \in [\min(\underline{RD}, \overline{RD}, RD^*), \max(\underline{RD}, \overline{RD}, RD^*)] \quad (30)$$

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

$$RD \in [\min(\underline{RD}, \overline{RD}), \max(\underline{RD}, \overline{RD})]. \quad (31)$$

When no information is available for τ_j , the bounds become

$$RD \in [\min(0, RD^*), \max(0, RD^*)]. \quad (32)$$

Rate From (21), 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 $\bar{\tau}_j$. When we are ignorant with respect to the τ_j 's, the rate is bounded as $(0, r_i/r^j)$.

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

$$\text{rd} \in [\min[\text{rd}(\underline{\tau}_j), \text{rd}(\bar{\tau}_j)], \max[\text{rd}(\underline{\tau}_j), \text{rd}(\bar{\tau}_j)]] , \quad (33)$$

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

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

8 A Concluding Remark on Reporting Standards

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, 36, 37, 7, 38, 39, 40, 41]. We provide the methods to estimate each of these quantities from case-control studies.

Unless the odds ratio happens to approximate a parameter of central substantive interest, we suggest that it 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 “the effect of smoking on lung cancer is positive OR = 1.38 (95 percent CI 1.30–1.46)” researchers could write “smoking increases the risk of contracting lung cancer by a factor of between 2.5 to 3.1 (a \geq 95 percent CI)” or “smoking increases the probability of contracting lung cancer between 0.022 and 0.051 (a \geq 95 percent 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.

A Simplifying Manski's Bounds on the Risk Difference

Proving Equation 16 requires algebra only. For simplicity, let $P_{ab} = \Pr(X_a|Y = b)$, so that $OR = (P_{11}P_{00})/(P_{01}P_{10})$. Then, omitting tedious but straightforward algebra at several stages, $\phi = (ORP_{01}^2/P_{11}^2)^{1/2} = \sqrt{OR}P_{01}/P_{11}$, and $\gamma = \sqrt{OR}/(\sqrt{OR} + P_{11}/P_{10})$. Then the components of RD_γ are

$$\begin{aligned} P_{11}\gamma &= \frac{\sqrt{OR}P_{10}P_{11}}{\sqrt{OR}P_{10} + P_{11}}, & P_{01}\gamma &= \frac{\sqrt{OR}P_{01}P_{10}}{\sqrt{OR}P_{10} + P_{11}}, \\ P_{10}(1 - \gamma) &= \frac{P_{10}P_{11}}{\sqrt{OR}P_{10} + P_{11}}, & P_{00}(1 - \gamma) &= \frac{P_{11}P_{00}}{\sqrt{OR}P_{10} + P_{11}}. \end{aligned}$$

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

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 $(e^{H_1} - 1)/(e^{H_0} - 1)$. Since $rr = H_1/H_0$, 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

$$e^{H_1} - 1 = H_1 + (1/2)H_1^2 + (1/3!)H_1^3 + \dots \quad (34)$$

$$e^{H_0} - 1 = H_0 + (1/2)H_0^2 + (1/3!)H_0^3 + \dots \quad (35)$$

and hence

$$H_1(e^{H_0} - 1) = H_1H_0 + (1/2)H_0^2H_1 + (1/3!)H_0^3H_1 + \dots \quad (36)$$

$$H_0(e^{H_1} - 1) = H_0H_1 + (1/2)H_1^2H_0 + (1/3!)H_1^3H_0 + \dots \quad (37)$$

The first term in (36) and (37) are equal, and when $H_1/H_0 < 1$, hence $H_1 < H_0$, all other terms in (36) are greater than the corresponding terms in (37) (since both $H_0 > 0$ and $H_1 > 0$ always). Thus, when $H_1/H_0 < 1$, $H_1(e^{H_0} - 1) > H_0(e^{H_1} - 1)$.

References

- [1] Ray Moynihan, Lisa Bero, Dennis Ross-Degnan, David Henry, Kriby Lee, Judy Watkins, Connie Mah, and Stephen B. Soumerai. Coverage by the news media of the benefits and risks of medications. *New England Journal of Medicine*, 342(22):1645–1650, June 1 2000.
- [2] Charles F. Manski. *Identification Problems in the Social Sciences*. Harvard University Press, 1995.
- [3] Charles F. Manski. Nonlinear statistical inference: Essays in honor of Takeshi Amemiya. In C. Hsiao, K. Morimune, and J. Powell, editors, *Nonparametric Identification Under Response-Based Sampling*. Cambridge University Press, 1999.
- [4] Sander Greenland. On the need for the rare disease assumption in case-control studies. *American Journal of Epidemiology*, 116(3):547–553, 1982.
- [5] J. Deeks D. Sackett and D. Altman. Down with odds ratios. *Evidence-Based Medicine*, 1(6):164–6, 1996.
- [6] Sander Greenland. Interpretation and choice of effect measures in epidemiologic analysis. *American Journal of Epidemiology*, 125(5):761–768, 1987.
- [7] Huw Talfryn Oakley Davies, Tavakoli Manouche, and Crombie Kinloch Iain. When can odds ratio mislead. *British Medical Journal*, 31:989–91, March 28 1998.
- [8] Kenneth J. Rothman and Sander Greenland. *Modern Epidemiology*. Philadelphia: Lippincott-Raven, 2nd edition edition, 1998.
- [9] R.L. Prentice and N.E. Breslow. Retrospective studies and failure-time models. *Biometrika*, 65(153–5), 1978.
- [10] Gary Chamberlain. Analysis of covariance with qualitative data. *Review of Economic Studies*, XLVII:225–238, 1980.

- [11] Sander Greenland. Modeling risk ratios from matched cohort data: An estimating equation approach. *Applied Statistics*, 43(1):223–232, 1994.
- [12] Larry Goldstein and Bryan Langholz. Asymptotic theory for nested case-control sampling in the cox regression model. *The Annals of Statistics*, 20(4):1903–1928, 1992.
- [13] Ørnulf Borgan, B. Langholz, and L. Goldstein. Methods for the analysis of sampled cohort data in the cox proportional hazard model. *The Annals of Statistics*, 23(1749-1778), 1995.
- [14] Bryan Langholz and Larry Goldstein. Risk set sampling in epidemiologic cohort studies. *Statistical Science*, 11(1):35–53, 1996.
- [15] B. Langholz and D.C. Thomas. Efficiency of cohort sampling designs: Some surprising results. *Biometrics*, 47:1563–1571, 1991.
- [16] J.H. Lubin and M.H. Gail. Sampling strategies in nested case-control studies. *Environmental Health Perspectives*, 102(suppl 8):47–51, 1994.
- [17] J.M. Robins, M.H. Gail, and J.H. Lubin. More on biased selection of controls for case-control analyses of cohort studies. *Biometrics*, 42(293-299), 1986.
- [18] R.L. Prentice. A case-cohort design for epidemiological studies and disease prevention trials. *Biometrika*, 73(1-11), 1986.
- [19] Sander Greenland. Multivariate estimation of exposure-specific incidence from case-control studies. *Journal of Chronic Disease*, 34:445–453, 1981.
- [20] Raymond R. Neutra and Margaret E. Drolette. Estimating exposure-specific disease rates from case-control studies using bayes theorem. *American Journal of Epidemiology*, 108(3):214–222, 1978.
- [21] Jerome Cornfield. A method of estimating comparative rates from clinical data: Application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, 11(1269-1275), 1951.
- [22] J.A. Anderson. Separate-sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

- [23] R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 63(403-411), 1979.
- [24] N. Mantel. Synthetic retrospective studies and related topics. *Biometrics*, 29(479-486), 1973.
- [25] Charles F. Manski. The estimation of choice probabilities from choice based samples. *Econometrics*, 45(8):1977–88, November 1977.
- [26] Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 2001, in press.
- [27] Gary King, Michael Tomz, and Jason Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, 44(2):341–355, April 2000.
- [28] Bryan Langholz and Boran Ørnulf. Estimation of absolute risk from nested case-control data. *Biometrics*, 53:767–774, June 1997.
- [29] G.U. Yule. On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, 75:579–642, 1912.
- [30] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [31] Nathaniel Beck, Gary King, and Langche Zeng. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*, 94(1):21–36, March 1999.
- [32] James Berger. An overview of robust bayesian analysis (with discussion). *Test*, 3:5–124, 1994.
- [33] David R. Insua and Ruggeri Fabrizio. *Bayesian Analysis*. Springer-Verlag, 2000.
- [34] M. Tumbarello, E. Tacconelli, K. de Gaetano, F. Ardit, T. Pirroni, R. Claudia, and L. Ortona. Bacterial pneumonia in hiv-infected patients: Analysis of risk factors and prognostic indicators. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 18(39-45), 1998.

- [35] J. Benichou and M. Gail. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics*, 51:182–194, 1995.
- [36] Markku Nurminen. To use or not to use the odds ratio in epidemiologic analysis. *European Journal of Epidemiology*, 11(365-371), 1995.
- [37] Jun Zhang and F. Kai Yu. What’s the relative risk? a method of correcting the odds ratio in cohort studies of common outcomes. *New England Journal of Medicine*, 280(19):1690–1, 1998.
- [38] Huw Talfryn Oakley Davies, Tavakoli Manouche, and Crombie Kinloch Iain. Authors reply. *British Medical Journal*, 317:1156–7, October 24 1998.
- [39] Jon Deeks. Odds ratio should be used only in case-control studies and logistic regression analyses. *British Medical Journal*, 317:1155–6, October 24 1998.
- [40] Michael B. and John C. Bracken. Avoidable systematic error in estimating treatment effects must not be tolerated. *British Medical Journal*, 317:11–56, October 24 1998.
- [41] Douglas G. Altman, Jonathan J. Deeks, and David L. Sackett. Odds ratios should be avoided when events are common. *British Medical Journal*, 317:1318, Nov. 7 1998.