American Journal of Epidemiology **Submitted Manuscript**

Title: Estimating Vaccine Effectiveness from Linking Population-Based Health Registries: Some Sources of Bias

Authors: Ron Brookmeyer and Doug Morrison

Correspondence Address: Correspondence to Dr. Ron Brookmeyer, Department of Biostatistics, UCLA Fielding School of Public Health, Box 951772, Los Angeles, CA 90095 (e-mail: rbrookmeyer@ucla.edu

Affiliations: Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, California, United States (Ron Brookmeyer); Department of Public Health Sciences, University of California, Davis, California, United States (Doug Morrison).

Funding: N/A

Data Availability Statement: The simulation software and data are available from the authors.

Thanks: N/A

Conference presentation: N/A

Preprint Information: N/A

Disclaimer: N/A

Conflict of Interest: The authors report no conflicts.

Running Head: Vaccine Effectiveness from Linking Registries

Key words: bias, COVID-19, registries, statistics, vaccines

Abbreviations: SD, standard deviation; VE, vaccine effectiveness.

Abstract

The COVID-19 pandemic has underscored the importance of observational studies of real-world vaccine effectiveness to help answer urgent public health questions. One approach to rapidly answering questions about real-world vaccine effectiveness relies on linking data from a population-based registry of vaccinations with a population-based registry of health outcomes. Here we consider some potential sources of bias in linked registry studies including: incomplete reporting to the registries; errors in linking individuals between registries; and errors in the assumed population size of the catchment area of the registries. We show that the direction of the bias resulting from one source of error by itself is predictable. However, if multiple sources of error are present, the direction of the bias can be either upward or downward. The biases can be so strong as to make harmful vaccines appear effective. We provide explicit formulas to quantify and adjust for multiple biases in estimates of vaccine effectiveness which could be used in sensitivity analyses. While this work was motivated by COVID-19 vaccine questions, the results are generally applicable to studies that link population-based exposure registries with population-based case registries to estimate relative risks of exposures.

Randomized clinical trials provide the most reliable evidence about vaccines in controlled settings (1). The COVID-19 pandemic has also underscored the importance of observational studies of real-world vaccine effectiveness to address timely public health issues (2). Such studies help answer questions such as: Do vaccines protect against emerging viral variants which may not have been prevalent when the original clinical trials were conducted? Does vaccine effectiveness wane over time in populations? What is the effectiveness of vaccines among people who were under-represented in clinical trials?

Addressing urgent epidemiologic questions about vaccines requires conducting real world vaccine effectiveness studies essentially in real time which presents enormous logistical and study design challenges. One approach relies on linking data from a population-based registry of vaccinations with a population-based registry of health outcomes. For example, recent studies of real-world vaccine effectiveness against COVID-19 have been performed in the United States by linking state and local registries of vaccinated persons with registries of cases with a particular health outcome such as infection, hospitalization or death. These studies have provided valuable and timely information (3,4). Identifying and linking the records of the same individuals listed in both registries is typically based on a combination of matching variables such as name, date of birth or zip code of residence (4). The approach is challenging to carry out in the United States which has more than fifty separate state and local public health data systems that are not easily linkable unlike some other countries, such as the United Kingdom and Israel which have reliable networks of national interconnected data systems.

Here we consider some potential sources of bias in vaccine effectiveness studies based on linking health registry studies. One potential source of bias is underreporting to the registries. Another is errors in linking. For example, the records of a person who is in both registries are not matched and therefore we fail to identify that the records correspond to the same person. An assumption underlying some linked registry studies of vaccine effectiveness is that cases in the case registry who are not matched (or linked) to persons in the vaccination registry are unvaccinated. As we discuss in the next section, linked health registry studies also rely on estimates of the size of the population that serves as the catchment for the registries. Errors in the assumed population size could introduce significant bias.

The objective of this paper is to evaluate the magnitude and direction of some potential biases on estimates of relative risk and vaccine effectiveness from studies of linking population-based health registries. While this work was motivated by COVID-19 vaccine questions, the results are applicable more generally to studies that link population-based exposure registries with population-based case registries to estimate relative risks of exposures.

Methods

Suppose vaccinated persons in a population are reported to a vaccination registry, and cases in the population (i.e., persons with a health outcome such as infection, hospitalization or death) are reported to a case registry. For example, the study in New York State provided estimates of vaccine effectiveness for each week starting May 3, 2021 through June 19, 2021 (4). We consider the problem of estimating vaccine effectiveness in the population by calendar time (for example, in week t). The number of vaccinated persons in the vaccination registry who were vaccinated prior to week t is N_V . The number of cases in the case registry that occurred during week t is N_C. The registries are linked to identify persons who appear in both registries. The linking is based on identifiers. For example, the New York State study used individual name-based identifiers, date of birth and zip code of residence for linking between registries (4). The number of individuals who appear in both the vaccination and case registries and who were vaccinated before week *t* and became cases during week *t* is N_{VC}. Implicit in what follows is that N_V , N_C, and N_{VC} may refer to a specific calendar period (e.g. week *t*).

The population size is assumed to be N where the population refers to the catchment area of the two registries. For example, U.S Census data has been used to determine the population size N (3,4). The numbers N_V, N_C, N_{VC} and N are used to partially complete a 2x2 table for vaccination status by case status in the population. The missing data elements in the 2 x 2 table are calculated to ensure that the cells correctly sum to the row and column totals as shown in Table 1: $N_{\overline{V}}$, the number of individuals not in the vaccine registry, is defined as $N_{\overline{V}} = N - N_V$; and $N_{\overline{V}C}$, the number of individuals in the case registry who were not linked to a vaccine record, is defined as $N_{\overline{V}C} = N_C - N_{VC}$.

The estimate of the relative risk of a health condition (case) occurring in week *t* among those previously vaccinated relative to those unvaccinated, based on Table 1, is:

$$\hat{R} = \frac{N_{VC}N_{\overline{V}}}{N_V N_{\overline{V}C}} \#(1)$$

and the estimate of vaccine effectiveness is $\widehat{VE} = \left(1 - \hat{R}\right) \times 100\%$.

The data in Table 1 could be restricted to a subset of the population to produce estimates of vaccine effectiveness by subgroups such as age or gender, and in this way control for confounders. For example, a study of COVID-19 vaccine effectiveness in 13 U.S jurisdictions produced estimates by age (3). In that situation, the population size N refers to persons in the catchment area who are in a specific age subgroup. Similarly, N_V, N_C, and N_{VC} refer to the numbers of persons in the specific age subgroup who are in the vaccine registry, the case registry, and both the case and vaccine registries, respectively. Thus, equation 1 can refer to a specific subgroup at a particular calendar period (e.g., persons over the age of 65 at the week of May 3, 2021).

We consider the impact of underreporting to registries on the bias of vaccine effectiveness. Specifically, we consider non-differential underreporting by which we mean that the probability a vaccinated person is reported to the vaccine registry does not depend on the person's case status and the probability a case is reported to the case registry does not depend on vaccination status. Our development also assumes that conditional on a person's true vaccination status and case status, the event of being reported to the vaccine registry and the event of being reported to the case registry are independent.

Let r_v be the probability that a vaccinated individual is reported to the vaccination registry and r_c be the probability that a case is reported to the case registry. In this paper we assume that persons reported to the vaccination registry are truly vaccinated and persons reported to the case registry are truly cases, and that there is no misclassification in the opposite direction (e.g., we assume that unvaccinated individuals are not erroneously reported to the registry as vaccinated).

We also consider the impact of incomplete linking by which we mean failure to link the records of the same individual who is in both registries. Incomplete linking may occur because some of the matching identifiers on which linking is based were incorrectly entered in either or both registries (e.g., errors in dates of birth, zip code, or misspelling of names). Even small errors in these matching identifiers could be a source of significant bias. Let p_L be the probability that the same person who is listed in both registries is correctly linked. Here we do not consider the error of falsely linking two different individuals because we are considering the situation when sufficient number of strong matching identifiers are utilized (e.g., name, date of birth and zip-code of residence) which would reduce the size of false matches. We return to this point in the discussion section.

We also consider the impact of errors in the assumed population size N which in some studies has been based on U.S Census data (3,4). Suppose the true population size is N_{true} and let the fractional error in N be $f = (N - N_{true})/N_{true}$. We set out to determine the effect of errors in the population size on the bias in vaccine effectiveness.

The term $\hat{R}$ (equation 1) is estimating (or more precisely, converging in probability to) *R*, which we call the apparent relative risk. In Web Appendix 1 and Web Table 1 we show that *R* is not necessarily equal to the true relative risk (R_{true}) and that the apparent vaccine effectiveness, $VE = (1 - R) \times 100\%$, is not necessarily equal to the true vaccine effectiveness $VE_{true} = (1 - R_{true}) \times 100\%$. We show that

$$R = R_{true}\left[\frac{p_L(1 + f - p_V r_V)}{1 - p_V + R_{true} p_V (1 - p_L r_V)}\right] \#(2)$$

where p_V is the proportion of the population that is vaccinated. The bias factor is the term in brackets in equation 2: if the bias factor is less than 1 the apparent relative risk will be less than the true relative risk and the apparent *VE* will be greater than VE_{true}; if the bias factor is equal to 1 there will be no bias; and if the bias factor is greater than 1 the apparent relative risk will be greater than the true relative risk and the VE will be less than VE_{true}. The bias factor does not depend on the reporting probability to the case registry (r_c) but does depends on the reporting probability to the vaccine registry (r_V) . The bias factor also does not depend on the baseline probability of becoming a case (p_C) among unvaccinated persons. As discussed in the next sections, the bias factor can be either greater or less than 1 and, in some circumstances, could be sufficiently extreme to make harmful vaccines appear effective.

We can adjust the relative risk for biases from underreporting, incomplete linking and population size errors if we have the values for r_V, p_L and *f*. The formula that takes $\hat{R}$ and produces an adjusted estimate of the relative risk $\hat{R}_{adj}$ is (see Web Appendix 2),

$$\hat{R}_{adj} = \frac{\hat{R}[Nr_V - (1+f)N_V]}{(1+f)\left[p_L r_V (N - N_V) - \hat{R} N_V (1 - p_L r_V)\right]} \#(3)$$

The adjustment formula (equation 3) could be used in a sensitivity analysis to determine how the relative risk estimate would change under different assumptions about r_V, p_L and *f*. Estimates of r_V, p_L and *f* may also be available from supplementary studies of the registries. We evaluate the performance of $\hat{R}_{adj}$ by simulation in the next section.

Numerical Results

We performed a simulation study under various conditions motivated by a recent real-world vaccine effectiveness study among adults in New York State (4). We used a population size of

11,000,000 and performed 1,000 replications for each set of conditions (further details of the simulation study and a Shiny App are provided in Web Appendix 3). The values of the input parameters (e.g., R_{true}, r_V, p_L and f) were varied to investigate a range of conditions. Simulation results are shown in Table 2. The average value of the estimated relative risks $\hat{R}$ (column 6) is in excellent agreement with the apparent relative risk *R* calculated from equation 2 (column 5) for all conditions considered providing empirical validation of equation 2. The average value of the adjusted relative risk (column 8) is in excellent agreement with R_{true} (column 1) providing empirical validation of equation 3.

We also examined the empirical standard deviation of $\hat{R}$ from the 1000 simulations (column 7 of Table 2). For each set of conditions considered, the standard deviation was exceedingly small resulting from the very large population size *N* and highlights that typically the main source of error in linked studies of large population-based registry studies will be bias rather than sampling variation. Even when errors are small ($p_L = .95, r_V = .90, f = 0$), we find that tests of the null hypothesis ($H_0: R_{true} = 1$)performed at the α=.05 level would actually have a type 1 error probability nearly 1.0 because of the bias in $\hat{R}$ (i.e., *R*=.861 instead of $R_{true} = 1.0$) and it’s very small standard deviation of .015.

Table 2 also demonstrates the impact of errors in N. If N is lower than N_{true} (i.e. $f < 0$), the apparent relative risk R is less than R_{true} and apparent VE is greater than VE_{true}. The direction of the bias is reversed if N is greater than N_{true} (i.e. $f > 0$).

Figure 1 illustrates the biases in the apparent relative risk *R* and *VE* and their relationship with r_V and p_L when $R_{true} = .20, VE_{true} = 80\%$ and $f = 0$. We find that apparent *VE* can be either greater or less than VE_{true}. If $p_L = 1$, the apparent VE will be less than VE_{true}. However, if $p_L < 1$, the apparent VE can either be greater or less than VE_{true}.

Summary of Direction of Biases

In this section we summarize the direction of the biases from underreporting and linking errors. The findings follow from equation 2 and are summarized in Table 3.

First consider the impact of only one source of error by itself. If $R_{true} \neq 1$, then nondifferential underreporting of vaccinated persons to the vaccination registry ($r_V < 1$) biases the apparent relative risk toward 1 and the apparent vaccine effectiveness toward 0. If the null hypothesis is true, $R_{true} = 1$, then nondifferential underreporting of vaccinated persons to the registry does not induce bias. These results can be viewed as a special case of nondifferential misclassification of an exposure which biases the relative risk toward the null hypothesis (5,6). The analogy is that vaccinated persons are the exposed group some of whom are misclassified as unexposed (unvaccinated) because of underreporting to the registry.

Nondifferential underreporting of cases to the case registry ($r_C < 1$) does not bias the apparent relative risk or apparent vaccine effectiveness and that result holds for all values of R_{true}. This result can also be viewed as a special case of nondifferential misclassification of disease (7).

If there are linking errors between the two registries whereby some persons whose records appear in both registries but are not matched (i.e., $p_L < 1$), then for all values of R_{true} the apparent relative risk will be biased downwards toward 0 and the apparent vaccine effectiveness will be biased upwards. The explanation is that the numbers of person classified as both cases and vaccinated ($N_{VC)}$ are undercounted because some persons listed in both registries are not linked together. This result holds even when the null hypothesis is true (R_{true} =1). Thus, if there is incomplete linking ($p_L < 1$) then type 1 errors of tests of the null hypothesis are inflated.

If the population size is underestimated, that is $N < N_{true}$, then for all values of R_{true}, the apparent relative risk is biased downward toward 0 and the apparent vaccine effectiveness will be biased upward. The explanation is that *N* only comes into the calculation of $\hat{R}$ through the term $N_{\bar{V}} = N - N_V$ (see equation1 and Table1) and thus if *N* is too small then $N_{\bar{V}}$ will also be too small which biases the apparent relative risk downward. On the other hand, if the population size is overestimated, that is $N > N_{true}$, then the apparent relative risk is biased upwards and the apparent vaccine effectiveness is biased downward.

If multiple sources of error are present, the direction of the bias can be either upward or downward. For example, suppose there is underreporting of vaccinated persons to the registry ($r_V < 1$), incomplete linkage ($p_L < 1$) , but no error in *N* ($f = 0$), then an effective vaccine ($R_{true} < 1, VE_{true} > 0$) could appear either more or less effective than it really is (see line 3 of Table 3). The reason the apparent relative risk can be either higher or lower than R_{true} is because incomplete linkage pulls the relative risk downward toward 0 while underreporting of vaccinated persons pulls the relative risk in the opposite direction toward 1. The ultimate direction of the bias from these two sources of error depends on the values of r_V and p_L . Although if the vaccine is truly effective then, these two sources of error cannot make the vaccine appear harmful (that is, if $R_{true} < 1$ then $R < 1$ regardless of the values of r_V and p_L. On the other hand, if the vaccine is either ineffective or harmful ($R_{true} \geq 1$), then $R < R_{true}$ and in some circumstances R could even be less than 1 in which case an ineffective or harmful vaccine would falsely appear effective (line 9 of table 3).

Discussion

This paper evaluates biases in estimates of vaccine effectiveness from linking population-based health registries. While this work was motivated by COVID-19 vaccine questions, the results are broadly applicable to estimating relative risks of exposures from linking population-based health registries.

We found that the direction of the bias from a single source of error is predictable: nondifferential underreporting of vaccinations attenuates the expected estimated effect sizes; nondifferential underreporting of cases does not create bias; incomplete linking between the registries is expected to lead to overestimation of vaccine effectiveness; underestimation of the population size results in overestimation of vaccine effectiveness. If multiple sources of error are present, the direction of the bias can be either upward or downward, and in fact biases can be so strong as to make a harmful vaccine appear effective.

We provide an explicit formula (equation 3) to adjust for multiple biases in estimates of vaccine effectiveness. The formula depends on three parameters: a measure of the completeness of reporting of the vaccine registry r_V ; the probability of correctly linking an individual who is in both the vaccine and case registries p_L; and the fractional error in the assumed population size f. Sensitivity analyses using ranges of plausible values for the parameters in equation 3 determine if the biases are of sufficient magnitude to impact the practical public health conclusions about vaccine effectiveness. There is considerable literature to help inform values for these parameters using a variety of approaches and designs for supplemental studies. For example, approaches for measuring the completeness of reporting in public health surveillance databases and their advantages and disadvantages are reviewed in (8). Evaluations of the magnitude of underreporting in specific public health registries have been performed (9-11). Approaches to determine linkage errors are considered in (12) and include performing detailed case investigations on a subset of data to produce a gold standard subset. Errors in the population size from U.S Census data are informed by Census coverage errors (13). Methods used to measure Census coverage include demographic analysis and dual system estimation that compares Census results to the Post-Enumeration Survey (13).

The impact of errors beyond those considered in this paper could be investigated. For example, we considered non-differential underreporting, but, differential underreporting could arise if cases who are vaccinated are more likely than cases who are unvaccinated to be reported to the case registry perhaps because vaccinated individuals are more connected to health systems; in that situation, with no other biases acting, an effective vaccine would appear less effective than it really is.

There are two types of linking errors that occur in practice, one that results in missed matches and one that results in false matches. In this paper we considered only missed matches because we were motivated by the situation in which a sufficient number of strong matching variables are available and thus false matches are expected to be relatively uncommon. Probabilistic linkage analyses consider various thresholds for determining linkages and their tradeoffs. Lower thresholds for linkages would increase the probability of false matches but decrease the probability of missed matches (14). We also did not consider the possibility that persons listed in a registry may not have the condition the registry is tracking. These additional sources of error will lead to even more complex relationships about their composite effect on vaccine effectiveness. The simulation framework described in Web Appendix 3 could be used to evaluate biases resulting from a multitude of these errors.

Real world vaccine effectiveness studies help answer emerging public health questions that could not be answered by the data from the original vaccine clinical trials. Studies conducted by linking population-based health registries offer a useful approach. However, it is critically important to assess the potential biases inherent in the approach. Improvements in the reporting and linking of health registries as well as the overall quality of public health data systems will enhance the reliability of these studies.

References:

1. Dean NE, Gsell PS, Brookmeyer R, De Gruttola V, Donnelly CA, Halloran ME, Jasseh M, Nason M, Riveros X, Watson CH, Henao-Restrepo AM. Design of vaccine efficacy trials during public health emergencies. *Science Translational Medicine,* 2019 ;11(499).

2. Evans, S.J.W. and Jewell, N.P., 2021. Vaccine Effectiveness Studies in the Field. *The New England Journal of Medicine*. *N Engl J Med,* 2021; 385:650-651.

3.Scobie HM, Johnson AG, Suthar AB, Severson R, Alden NB, Balter S, Bertolino D, Blythe D, Brady S, Cadwell B, Cheng I. Monitoring incidence of covid-19 cases, hospitalizations, and deaths, by vaccination status—13 US jurisdictions, April 4–July 17, 2021. *Morbidity and Mortality Weekly Report,* 2021;70(37):1284.

4. Rosenberg ES, Holtgrave DR, Dorabawila V, Conroy M, Greene D, Lutterloh E, Backenson B, Hoefer D, Morne J, Bauer U, Zucker HA. New COVID-19 cases and hospitalizations among adults, by vaccination status—New York, May 3–July 25, 2021. Morbidity and Mortality Weekly Report. 2021 Sep 17;70(37):1306.

5. Flegal Km, Brownie C, Haas J. The Effects of Exposure Misclassification on Estimates of Relative Risk. *American Journal of Epidemiology,* 1986;123(4):736-51.

6. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology,* 1990;132(4):746-8.

7.Rothman KJ, Greenland S, eds. *Modern Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers; 1998.

8. Gibbons, C.L., Mangen, MJ.J., Plass, D. *et al.* Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health* 2014; 14(147).

9.Alter MJ, Mares A, Hadler SC, Maynard JE. The effect of underreporting on the apparent incidence and epidemiology of acute viral hepatitis. *American Journal of Epidemiology*. 1987;125:133-9.

10.Keramarou M, Evans MR. Completeness of infectious disease notification in the United Kingdom: a systematic review. *Journal of Infection*. 2012; 64:555-64.

11.Alves TH, Souza TA, Silva SD, Ramos NA, Oliveira SV. Underreporting of death by COVID-19 in Brazil's second most populous state. *Frontiers in Public Health*. 2020;8:909.

12.Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol*. 2017; 46(5):1699-1710.

13. O’Hare W.P. Methodology Used to Measure Census Coverage. In: O’Hare WP. *Differential Undercounts in the U.S. Census*. Cham, Switzerland: Springer Briefs in Population Studies; 2019: 25-38.

14. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. *Int J Epidemiol*. 2016;45(3):954-964.

Table 1: 2x2 table of vaccination and case status from linked vaccination and case registries [a]

	Case	Non-Case	
Vaccinated	N_{VC}	$N_{V\overline{C}} = N_V - N_{VC}$	N_V
Unvaccinated	$N_{\overline{V}C} = N_C - N_{VC}$	$N_{\overline{VC}} = N - N_C - N_V + N_{VC}$	$N_{\overline{V}} = N - N_V$
	N_C	$N_{\overline{C}} = N - N_C$	N

[a] N_V is number of individuals in the vaccine registry vaccinated before week *t*. N_C is the number of individuals in the case registry who became cases during week *t*; N_{VC} is number of individuals listed in both registries who were vaccinated prior to week *t* and became a case during week *t*. Population size is be N. All other table entries are calculated so that rows and columns sum to marginal totals.

Table 2. Simulation of average estimated relative risk $\hat{R}$ and adjusted relative risk $\hat{R}_{adj}$ (eq.3) and standard deviations (SD) from 1000 replications [a]

R_{true}	p_L	r_V	$fx100$	R	$Mean(\hat{R})$	$SD(\hat{R})$	$Mean(\hat{R}_{adj})$	$SD(\hat{R}_{adj})$
1.0	0.95	0.90	0	0.861	0.861	0.015	1.001	0.026
1.0	0.90	0.90	0	0.745	0.745	0.013	1.001	0.028
1.0	0.70	0.90	0	0.431	0.431	0.007	1.001	0.036
1.0	0.90	0.90	+20%	1.204	1.204	0.021	1.001	0.028
1.0	0.90	0.90	+10%	0.975	0.975	0.017	1.001	0.028
1.0	0.90	0.90	+5%	0.860	0.860	0.015	1.001	0.027
1.0	0.90	0.90	0	0.745	0.745	0.013	1.000	0.028
1.0	0.90	0.90	-5%	0.631	0.631	0.011	1.001	0.028
1.0	0.90	0.90	-10%	0.516	0.516	0.009	1.001	0.027
1.0	0.90	0.90	-20%	0.287	0.287	0.005	1.000	0.027
0.2	0.95	0.90	0	0.227	0.227	0.007	0.200	0.006
0.2	0.90	0.90	0	0.210	0.210	0.006	0.200	0.007
0.2	0.70	0.90	0	0.149	0.149	0.005	0.200	0.008
0.2	0.90	0.90	+20%	0.339	0.340	0.010	0.200	0.007
0.2	0.90	0.90	+10%	0.275	0.275	0.008	0.200	0.006
0.2	0.90	0.90	+5%	0.242	0.242	0.007	0.200	0.006
0.2	0.90	0.90	0	0.210	0.210	0.007	0.200	0.007
0.2	0.90	0.90	-5%	0.178	0.178	0.005	0.200	0.007
0.2	0.90	0.90	-10%	0.145	0.145	0.004	0.200	0.007
0.2	0.90	0.90	-20%	0.081	0.081	0.002	0.200	0.007

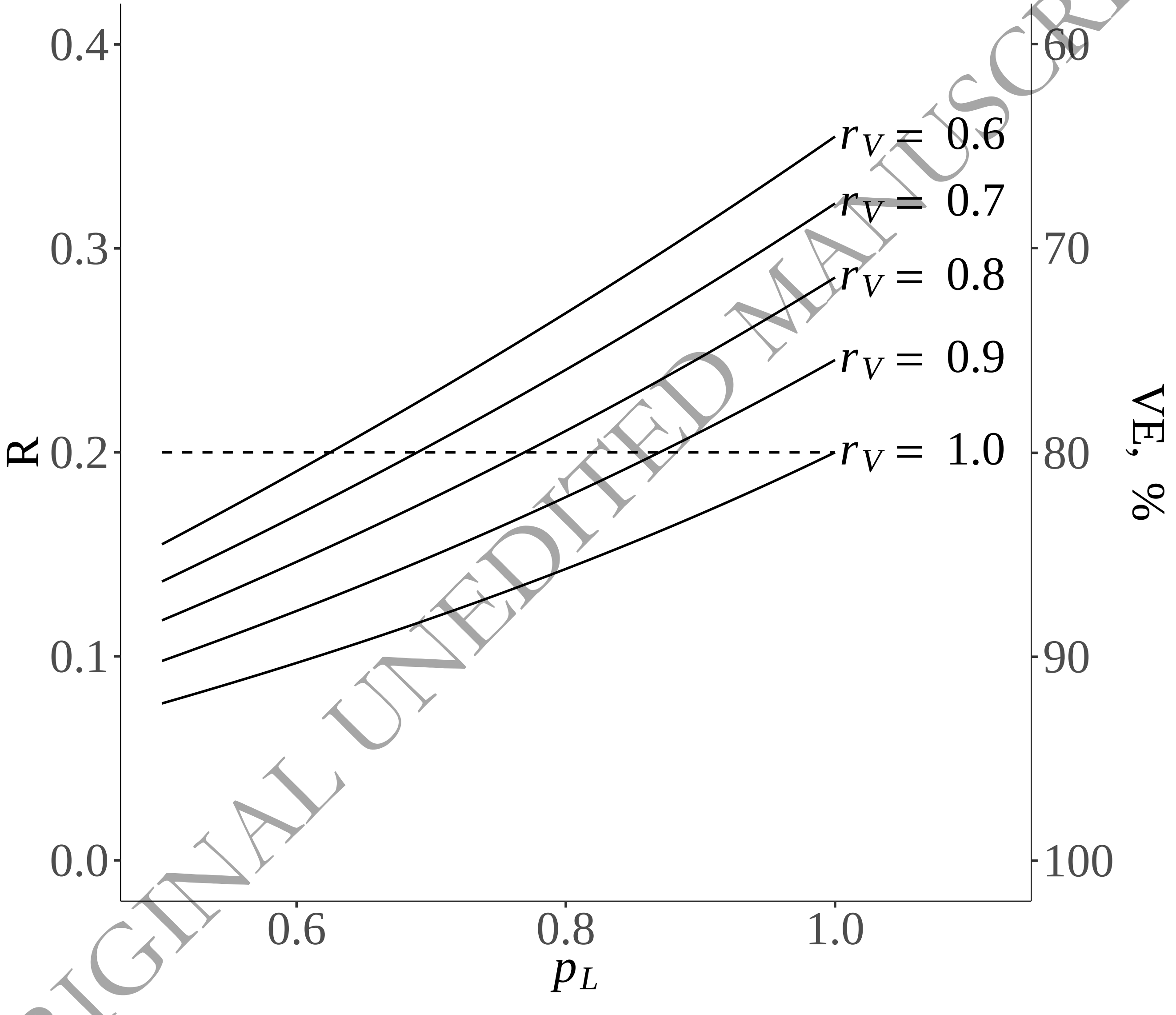
[a] R is apparent relative risk (eq. 2), r_V is vaccination reporting probability, p_L is linking probability, $fx100$ is % error in population size. N=11x10^6, $p_V = 0.75$, $r_c = 0.9, p_C = 0.0014$.

Table 3: Summary of impact of incomplete reporting and linking on vaccine effectiveness (*VE*) and apparent relative risk (*R*) [a]

p_L	r_V	Apparent Effect	Comment
$R_{true} < 1$, $VE_{true} > 0$			
1	<1	$R_{true} < R < 1$, $0 < VE < VE_{true}$	Attenuation of true effect. Underestimate true *VE*
<1	1	$R < R_{true}$, $VE > \mathrm{VE}_{\mathrm{true}}$	Exaggeration of true effect. Overestimate true *VE*
<1	<1	*R* and *VE* >, =, or, < than true values; $R<1$, $VE>0$	Direction of bias depends on of p_L, r_V, and p_V (equation 1)
$R_{true} = 1$, $VE_{true} = 0$			
1	<1	$R=1$, $VE=0$	No bias
<1	1	$R<1$, $VE>0$	Vaccine appears effective when it is not
<1	<1	$R<1$, $VE>0$	Vaccine appears effective when it is not
$R_{true} > 1$, $VE_{true} < 0$			
1	<1	$1 < R < R_{true}$, $VE_{true} < VE < 0$	Vaccine appears less harmful than it is
<1	1	$R < R_{true}$, $VE > VE_{true}$	Vaccine appears less harmful than it is, and could even appear effective
<1	<1	$R < R_{true}$, $VE > VE_{true}$	Vaccine appears less harmful than it is, and could even appear effective

[a] Results in table are for the situation when the population size is correctly specified, $N = N_{true}$ (that is, $f = 0$). p_L is linking probability, r_V is vaccination reporting probability. When population size is incorrectly specified: if population is underestimated ($f < 0$), relative risk will be further biased downwards and *VE* overestimated beyond results in Table 3; if population is overestimated ($f > 0$), relative risk will be further biased upwards and *VE* underestimated beyond results in Table 3.

Figure 1. Relationship of apparent relative risk R and apparent vaccine effectiveness VE with vaccination reporting probability r_V and linking probability p_L when $R_{true} = .20, VE_{true} = 80\%$. Calculated from equation 2 with no error in population size (i.e., $f = 0$) and with vaccination probability $p_V = 0.75$.


R
0.0
0.1
0.2
0.3
0.4
VE, %
60
70
80
90
100
p_L
0.6
0.8
1.0
$r_V = 0.6$
$r_V = 0.7$
$r_V = 0.8$
$r_V = 0.9$
$r_V = 1.0$