Estimating Vaccine Effectiveness from Linking Population-Based Health Registries:

Some Sources of Bias

Ron Brookmeyer and Doug Morrison

Ron Brookmeyer, Ph.D., Department of Biostatistics, University of California, Los Angeles, CA 90095, [rbrookmeyer@ucla.edu](mailto:rbrookmeyer@ucla.edu)

Doug Morrison, Ph.D., Department of Epidemiology, University of California, Los Angeles, CA 90095, [dmorrison01@ucla.edu](mailto:dmorrison01@ucla.edu)

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 formula 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 *VE* 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 and 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 incomplete linking, by which we mean that 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 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, and 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 obtained from 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. The number of vaccinated persons in the vaccination registry is and the number of cases in the case registry in . The registries are linked to identify persons who appear in both registries. The linking could be based on identifiers such as name, date of birth and zip code (4). The number of individuals who appear in both the vaccination and case registries is . The population size is assumed to be 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 (3,4). The numbers , , and 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: , the number of individuals not in the vaccine registry, is defined as ; and , the number of individuals in the case registry who were not linked to a vaccine record, is defined as . Then, the estimate of the relative risk of a health condition (case) among those vaccinated relative to those unvaccinated is:

and the estimate of vaccine effectiveness is  .

We consider the impact of underreporting to registries on the bias of the estimates. Specifically, we consider independent non-differential underreporting by which we mean that: the probability a vaccinated case is reported to the vaccine registry does not depend on case status; the probability that a case is reported to the case registry does not depend on vaccination status; and reporting a vaccinated person to the vaccination registry and reporting a case to the case registry are independent events. Let be the probability that a vaccinated individual is reported to the vaccination registry and 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.

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 potential source of significant bias. Let  be the probability that the same person who is listed in both registries is correctly linked. In this paper we do not consider the error of falsely linking two different individuals; it could be argued that errors of that type are considerably less likely if an adequate number of matching identifiers are utilized.

We also consider the impact of errors in the assumed population size which in some studies has been based on U.S Census data (3,4)*.* Suppose the true population size is and let. We set out to determine the effect of errors in the population size on the bias in estimates of vaccine effectiveness.

The term (equation 1) is estimating (or more precisely, converging in probability to) *R*, which we call the apparent relative risk. In the supplementary material we show that *R* is not necessarily equal to the true relative risk ( and that the apparent vaccine effectiveness, , is not necessarily equal to the true vaccine effectiveness . We show that

where  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 ; 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 . The bias factor does not depend on underreporting to the case registry but does depends on underreporting to the vaccine registry . The bias factor also does not depend on the baseline probability of becoming a case 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 , and *f.* The formula that takes and produces an adjusted estimate of the relative risk is (see supplementary material),

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 , and *f* . Estimates of , and *f* may also be available from supplementary studies of the registries. We evaluate the performance of 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 the supplementary material). The values of the input parameters (e.g., *,* 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 (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  (column 1) providing empirical validation of equation 3.

We also examined the empirical standard deviation of 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 ( we find that tests of the null hypothesis performed at the α=.05 level would actually have a type 1 error probability nearly 1.0 because of the bias in (i.e., *R*=.861 instead of ) and its very small standard deviation.

Table 2 also demonstrates the impact of errors in . If is lower than (i.e. ), the apparent relative risk is less than and apparent is greater than . The direction of the bias is reversed if is greater than (i.e. ).

Figure 1 illustrates the biases in the apparent relative risk *R* and *VE* and their relationship with and when  and. We find that apparent *VE* can be either greater or less than *.* If , the apparent *VE* will be less than . However, if , the apparent can either be greater or less than *.*

**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 then nondifferential underreporting of vaccinated persons to the vaccination registry biases the apparent relative risk toward 1 and the apparent vaccine effectiveness toward 0. If the null hypothesis is true, , 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 does not bias the apparent relative risk or apparent vaccine effectiveness and that result holds for all values of . 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 record appear in both registries are not matched then for all values of 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 are undercounted because some persons listed in both registries are not linked together. As this result holds even when the null hypothesis is true1), if there is incomplete linking then type 1 errors of tests of the null hypothesis are inflated.

If the population size is underestimated, that is ,then for all values of the apparent relative risk will be biased downward toward 0 and the apparent vaccine effectiveness will be biased upward. The explanation is that *N* only comes into the calculation of through the term (see equation1 and Table1) and thus if *N* is too small then will also be too small biasing the apparent relative risk downward. On the other hand, if the population size is overestimated, that is then the apparent relative risk will be 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 incomplete linkage , but no error in *N* , then an effective vaccine  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 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 and Although if the vaccine is truly effective then, these two sources of error cannot make the vaccine appear harmful (that is, if then regardless of the values of and . On the other hand, if the vaccine is either ineffective or harmful ), then 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: underreporting of vaccinations attenuates the expected estimated effect sizes; 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 to quantify and adjust for multiple biases in estimates of vaccine effectiveness. The formula could be used in sensitivity analyses to evaluate the potential impact of the one or more sources of errors. Supplemental studies could also be undertaken to gauge the magnitude of some of these errors. For example, studies could be undertaken to measure the magnitude of underreporting to registries (8-10) or incomplete linkages from case investigations.

The impact of a number of other errors could be investigated beyond those considered in this paper. For example, differential underreporting or dependence in reporting to the registries could be considered. We also assumed that if are persons listed in a registry that they do indeed have the condition the registry is tracking, and more general setting could also be considered. While we considered incomplete linking, we did not consider the alternative error of falsely linking record in the registries from two different people, although that error could be could be considerably less common if a sufficient number of matching variables are used to link individuals. These more general situations will lead to even more complex relationships. The simulation framework could also be extended to evaluate bias 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Case | Non-Case |  |
| Vaccinated |  |  |  |
| Unvaccinated |  |  |  |
|  |  |  |  |

Table 1: 2x2 table of vaccination and case status in a population from linked vaccination and case registries. The marginal totals are obtained from the two registries; the entry is obtained by linking the two registries. The population size is assumed to be . All other table entries are calculated so that rows and columns sum to marginal totals.

Table 2. Simulation study of the average estimated relative risk and its standard deviation (SD) under various conditions each based on 1000 replications. *N*=11x106, , *R* is the theoretical apparent relative risk obtained from equation 2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

Table 3: Summary of impact of incomplete reporting and linking on vaccine effectiveness (*VE*) and relative risk (*R* ). Results in table are for the situation when the population size is correctly specified, (1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| True Effects  *Rtrue VEtrue* |  |  | Apparent Effect | Comment |
| <1 >0 | 1 | <1 | *Rtrue* <*R*<1, 0<*VE*<*VEtrue* | Attenuation of true effect. Underestimate true *VE* |
|  | <1 | 1 | *R* < *Rtrue* , *VE* > VEtrue | 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 , , and (equation 1) |
|  |  |  |  |  |
| 1 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 |
|  |  |  |  |  |
| >1 <0 | 1 | <1 | 1 <*R*<*Rtrue* , *VEtrue* <*VE*<0 | Vaccine appears less harmful than it is |
|  | <1 | 1 | *R*<*Rtrue* , *VE*>*VEtrue* | Vaccine appears less harmful than it is, and could even appear effective |
|  | <1 | <1 | *R*<*Rtrue*, *VE*>*VEtrue* | Vaccine appears less harmful than it is, and could even appear effective |

1 If population is underestimated (*f*<0), relative risk will be further biased downwards and *VE* overestimated beyond results in Table;

If population is overestimated (*f*>0), relative risk will be further biased upwards and *VE* underestimated beyond results in Table.

Figure 1. Relationship of apparent relative risk *R* and apparent vaccine effectiveness *VE* with andwhen Calculated from equation 2 with and .

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.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.

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

10.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.

**SUPPLEMENTARY MATERIAL**

**Justification of formula for the apparent relative risk *R* (Equation 2)**

Here we justify equation 2. The notation is:

is the probability that a vaccinated individual is reported to the vaccine registry;

is the probability that a case is reported to the case registry;

is the proportion of the population that is vaccinated;

is the probability an unvaccinated person becomes a case;

is the true relative risk that a vaccinated person becomes a case compared to unvaccinated;

is the probability that a person in both registries is correctly linked;

.

The estimate of the relative risk is

Since , it follows that , which is substituted into equation (*S*1) and we obtain

We first consider the first term on the right of equation S2 which is the estimate of the relative risk if the true population size is used. This term converges to an expression involving the cell probabilities in the 2x2 table for classifying by case and vaccination status from the linked registries (see Table *S*1) and is given by

where

The term is the probability that a person is identified to be both in the case and vaccination registry. This probability is calculated by noting that for a person to be identified to be in both the vaccinated and case registry the person must be: (1)vaccinated with probability ; (2) reported to the vaccine registry with probability ; (3) became a case with probability ;(4) reported to the case registry with probability ; (5)identified (that is linked) between the two registries with probability . It follows that:

Noting that

it follows that

Substituting equations (*S*4)-(*S*7) into equation (S3), we find

The second term on the right side of equation (*S*2) converges to

Using equation (S2) together with equations *S*8 and *S*9 we find that converges to the apparent relative risk

**Justification of formula for the adjusted relative risk (equation 3)**

The term in formula for the apparent relative risk *R* (equation *S*10), is the proportion of the population that is both vaccinated and reported to the vaccination registry and that term can be consistently estimated by the expression =. When this expression is inserted into equation (*S*10) and we solve for , the solution, the adjusted relative risk  is

**Simulation Study**

The simulation was implemented using the R programming language. A shiny app is available to perform simulations and calculate adjusted relative risks and can be accessed at <https://morrison.shinyapps.io/VaxEffApp/> and the corresponding code can be accessed at <https://github.com/d-morrison/vax.eff>.

The inputs for the simulation were motivated by a recent real-world vaccine effectiveness study among adults in New York State (4). The simulations were based on the following conditions: population size, , true vaccination probability, , vaccination reporting probability, , case reporting probability, , and case probability among unvaccinated individuals, . We considered various values for the true relative risk, , the record linkage probability, the and the error in the population size We performed 1,000 replications for each set of conditions. Each replication consisted of the following steps.

We first simulated the true number of vaccinated individuals, , using a binomial distribution with trials and success probability . Next, we simulated the number of individuals in the vaccination registry, , using a binomial distribution with trials and success probability . We simulated the number of individuals who are in both the vaccination and case registries (although some of them may not be successfully linked), , using a binomial distribution with trials and success probability . We simulated the number of individuals who appear in both registries and are linked, that is there vaccination and case records are linked together , , as a binomial distribution with trials and success probability . We simulated the number of vaccinated individuals who are not in the vaccination registry but are in the case registry, , as a binomial distribution with trials and success probability . We simulated the number of unvaccinated individuals who are in the case registry, , as a binomial distribution with trials and success probability . Finally, we derived the number of reported cases, ; the estimated population size, ; the apparent number of nonvaccinated individuals, ; the apparent number of nonvaccinated cases; ; the estimated relative risk ; and the estimated vaccine efficacy,  . In mathematical shorthand, the simulation model was thus:

After performing 1,000 replications of this simulation model, we calculated the mean and standard deviation of the resulting estimates.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Case | Non-Case |  |
| Vaccinated |  |  |  |
| Unvaccinated |  |  |  |
|  |  |  |  |

Table S1: Cell probabilities for 2x2 table resulting from classifying a population by vaccination and case status from a vaccination registry and a case registry that are linked in a population.