Estimating Vaccine Effectiveness from Linked Public Health Registries: Impact of Reporting and Linking Errors

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 Biostatistics, University of California, Los Angeles, CA 90095, [dmorrison01@ucla.edu](mailto:dmorrison01@ucla.edu)

Randomized clinical trials provide the most reliable evidence about vaccines in controlled settings (1). The COVID-19 pandemic has underscored the importance of clinical trials as well as observational studies of real-world vaccine effectiveness (*VE*) (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 *VE* wane over time in populations? What is the effectiveness of vaccines among people who were under-represented in clinical trials?

Addressing urgent public health questions about vaccines requires conducting real world vaccine effectiveness studies essentially in real time, which presents enormous challenges. One approach that has been utilized has been to link data from a population registry of vaccinations with a population registry of health outcomes. Several recent studies of real world VE against COVID-19 have been performed in the United States by linking state or local registries of persons who were vaccinated with registries of cases who had a particular health outcome such as infection, hospitalization or death. These studies have provided valuable and timely information (3,4). The linking of individuals between two registries is typically based on a combination of names, date of birth, or zip code of residence (4). That approach presents more challenges in the United States than in certain countries such as the United Kingdom and Israel that have reliable networks of national interconnected data systems that are linkable. The public health data system in the United States is not a single network, but instead more than fifty separate state and local systems, that are not easily linkable.

There are a number of critical assumptions underlying linked registry studies. First, the studies assume that cases who are not matched (or linked) to persons in the vaccination registry are unvaccinated. Second, the studies assume that reporting of both vaccinations and cases to the registries are complete; Third, there may be errors in identifying and linking the same person who appear in both registries. Fourth, the studies assume the size of the population that serves as the catchment for the registries is known. The objective of this paper is to evaluate the impact of these sources of error including the impact of: incomplete reporting of vaccinated persons to the vaccination registry; incomplete reporting of cases to the case registry, failure to link persons who are in both registries; and errors in the assumed value of the population size. While this work was motivated by COVID-19 vaccine effectiveness questions, the results are applicable more generally to relative risk estimation based on linking population-based exposure registries with population-based case registries.

**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). Based on the linking of the registries, the number of individuals who appear in both the vaccination and case registries is . 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 population estimates has been used to determine *N* (3,4). The numbers , , and *N* are used to partially complete the 2x2 table for vaccination status by case status in the population. The missing data elements in the 2x2 table are calculated to ensure that the cells correctly sum to the row and column totals as shown in Table 1.

The estimate of the relative risk of being a case among vaccinated relative to unvaccinated is (Table 1):

and the estimate of *VE* is  .

Here we consider the potential impact of several sources of error on estimates of the relative risk. We consider independent non-differential underreporting by which we mean: 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. We do assume persons listed in the vaccination registry are truly vaccinated and person listed in the case registry are truly cases.

We also consider the error of not correctly linking an individual who is in both the vaccination and case registries. Such an error may occur because some of the matching identifiers on which linking is based were incorrectly entered in either or both registries (e.g., date of birth, zip code, name spelling). Even small errors in these matching identifiers could be an important source of this error. Let be the probability that a 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 unlikely if an adequate number of matching identifiers are utilized and they carefully chosen.

We also consider the impact of errors in the assumed population size *N.* Suppose the true population size is and let. In some studies using linked public health registries the population sizes were based on U.S Census data (3,4). We set out to determine whether or not small errors in the population size could have important impact on bias.

The estimator of the relative risk (equation 1) is estimating (or more precisely, converging 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 *VE* which is, is not necessarily equal to the true *VE* which is. We show that

where  is the proportion of the population that is vaccinated. We call the term in brackets in equation 2 the bias factor: 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 the true *V*E; 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 apparent *VE* will be less than the true *VE*.

Equation 1 reveals a number of surprising results that are not immediately obvious. First, the bias factor does not depend on underreporting to the case registry but does depends on the degree of underreporting to the vaccine registry . Second, the bias factor also does not depend on the baseline probability that an unvaccinated person is a case. Third, depending on the values of the input variables in equation 2, such as the bias factor can be either greater or less than 1. Under some circumstances ineffective or even harmful vaccines could look protective. In the next section we numerically evaluate equation 2 to identify the magnitude and direction of the biases under different conditions. We also present the results of a simulation study (see supplementary material for details of the of the simulation study) to provide further validation of equation 2 and to evaluate the standard errors of the estimated relative risk under various conditions.

**Numerical Results**

We numerically evaluated the bias in the apparent relative risk and *VE* using equation (2). First, we considered the case when there is no error in *N* (i.e., *f*=0). Figure 1 shows the apparent relative risk *R* and *VE* when We find that apparent *VE* can be either greater or less than the true *VE*. If =1, the apparent *VE* will be less than the true *VE* and the apparent relative risk will be greater than . However, if , the apparent *VE* can either be greater or less than the true *VE.*

We performed a simulation study under various conditions. The inputs for the simulation were 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. The simulation results are shown in Table 2. The average value of the estimated relative risks (column 6) is in excellent agreement with the theoretical calculation of *R* using equation 2 (column 5) which provides empirical validation of equation 2. In the first three lines of Table 2 we considered the situation when the null hypothesis of no vaccine effect is true (=1). We find that if reporting and linking is less than perfect, that is and and with no error in *N* *(f*=0), then the apparent relative risk *R* will be less than 1 and the apparent *VE* will be greater than 0. In line 7-13 of Table 2, we consider the impact of errors in *N*. If *N* is lower than (i.e. *f*<0), the apparent relative risk *R* is less than and apparent *VE* is greater than *VEtrue*. The direction of the biases is reversed if *N* is greater than (i.e. *f>*0). We also examined the empirical estimate of the standard error of from the simulations (last column of Table 2). In all situations considered the standard error was exceedingly small which results in large part from the very large population size *N*. The small standard errors highlight that the main source of concern in linked large population registry studies is bias rather than sampling variation.

**Summary of Direction of Biases**

A number of general results can be obtained from further analysis of Equation 2. Some of these results are summarized in Table 3….

**Discussion**

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 *N*. 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 specified1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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* can be >, =, or, < than their true values | Direction of bias depends on magnitudes of , , and  as predicted by 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 really is |
|  | <1 | 1 | *R*<*Rtrue* , *VE*>*VEtrue* | Vaccine appears less harmful than it is, and could even incorrectly appear beneficial |
|  | <1 | <1 | *R*<*Rtrue*, *VE*>*VEtrue* | Vaccine appears less harmful than it is, and could even incorrectly appear beneficial |

1 If population size is underestimated, relative risk will be further biased downwards and *VE* further overestimated;

If population size is overestimated relative risk will be further biased upwards and *VE* further underestimated.

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.

Additional references to consider:

Wacholder S, Hartge P, Lubin JH, Dosemeci M. Non-differential misclassification and bias towards the null: a clarification. *Occupational and Environmental Medicine,* 1995; 52(8):557.

Sorahan T, Gilthorpe MS. Non-differential misclassification of exposure always leads to an underestimate of risk: an incorrect conclusion. *Occupational and Environmental Medicine,* 1994;51(12):839.

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

Tenforde MW, Self WH, Naioti EA, Ginde AA, Douin DJ, Olson SM, Talbot HK, Casey JD, Mohr NM, Zepeski A, Gaglani M. Sustained effectiveness of Pfizer-BioNTech and Moderna vaccines against COVID-19 associated hospitalizations among adults—United States, March–July 2021. *Morbidity and Mortality Weekly Report,* 2021;70(34):1156.

Tenforde MW, Self WH, Naioti EA, Ginde AA, Douin DJ, Olson SM, Talbot HK, Casey JD, Mohr NM, Zepeski A, Gaglani M. Sustained effectiveness of Pfizer-BioNTech and Moderna vaccines against COVID-19 associated hospitalizations among adults—United States, March–July 2021. *Morbidity and Mortality Weekly Report,* 2021;70(34):1156.

Bajema KL, Dahl RM, Prill MM, Meites E, Rodriguez-Barradas MC, Marconi VC, Beenhouwer DO, Brown ST, Holodniy M, Lucero-Obusan C, Rivera-Dominguez G. Effectiveness of COVID-19 mRNA Vaccines Against COVID-19–Associated Hospitalization—Five Veterans Affairs Medical Centers, United States, February 1–August 6, 2021. *Morbidity and Mortality Weekly Report,* 2021;70(37):1294.

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.

Weinberg CA, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? *American Journal of Epidemiology,* 1994;140(6):565-71.

**SUPPLEMENTARY MATERIAL**

**Justification of 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 Substituting into equation *S*1 we have

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

In addition, we have that

And it follows that

Substituting equations *S*4-*S*7 into *S*3, we find that *S*3 can be expressed as

We now consider the second term on the right size of equation *S*2. This term converges to

Summing the results in equations *S*8 and *S*9 we find that converges to

## Proofs of Table 3

### Row 2:

Claim: if , , and , then .

Proof:

[we can divide by without flipping the inequality, because ]

### Row 3:

Claim: If , , , and , then .

Proof:

If then by Equation 2, .

If then by Equation 2, .

**Adjusted estimator:**

We can solve the following equation for as follows:

This formula requires estimates of and .

**Simulation Study**

The inputs for the simulation were motivated by a recent real-world vaccine effectiveness study among adults in New York State (4). We held the following parameters constant: population size, , true vaccination probability, , vaccination reporting probability, , case reporting probability, , and case rate among unvaccinated individuals, . We varied the true relative risk, , the record linkage probability, , and the population size overestimation factor, . 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 with vaccination records, , using a binomial distribution with trials and success probability . We simulated the number of individuals with vaccination and case records (not necessarily linked), , using a binomial distribution with trials and success probability . We simulated the number of individuals with linked vaccination and case records, , as a binomial distribution with trials and success probability . We simulated the number of vaccinated individuals without vaccination records but with case records, , as a binomial distribution with trials and success probability . We simulated the number of unvaccinated individuals with case records, , 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 of vaccination and case status from linked vaccination and case registries in a population.