VE estimates with administrative data

Arseniy Khvorov

October 2019

Simulation repository
https://github.com/khvorov45/ve-admin

Contents

1 Methods

1.1 Core simulation

The size of the starting population was set to 500,000. Every individual had their attributes randomly allocated in the order shown in Figure 2 and with probabilities shown in Tables 1, 2 and 3.

True vaccination status was allocated. Measurement of this status for the purposes of the simulated study was allowed to be inaccurate in order to simulate exposure misclassification. True vaccination status was used to determine which probability to use to allocate individuals to the flu-infected category. Both vaccinated and unvaccinated subjects were allocated to the non-flu infected category with the same probability. Any one subject could be infected with flu or a non-flu pathogen but not both. The remaining subjects ended up as part of the non-ARI category.

Everyone with an ARI (either infected with flu or a non-flu pathogen) was assigned to either the symptomatic or the asymptomatic group. Those who were symptomatic were assigned to the clinically assessed or unassessed groups. Being clinically assessed in this context means that they presented to a clinic with ARI illness and they were classified as an ARI case. These clinically assessed ARI cases got one probability of being tested, everyone else got another. Tests were allowed to be imperfect to simulate outcome misclassification.

1.2 Population summary

Each population was collapsed down to summary results - each individual was considered to be part of one of eight categories: administrative/surveillance vaccinated/unvaccinated case/control as shown in Figure 1.

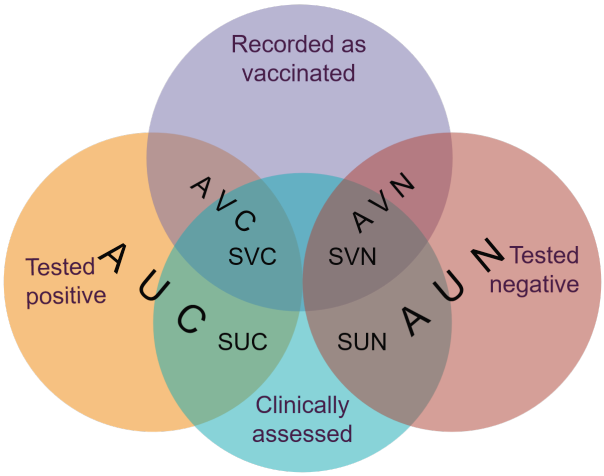

Figure 1: Assignment of an individual to appropriate categories. A - administrative, S - surveillance, V - vaccinated, U - unvaccinated, C - case, N - non-case (control).

Individuals in each of the categories were counted. These counts were representative of those that could have been obtained if a test-negative study was conducted on the population either using administrative or surveillance data. VE estimates could then be calculated as 1 - OR where $\text{OR} = \frac{\text{Odds in vaccinated}}{\text{Odds in unvaccinated}}$ where $\text{Odds} = \frac{\text{Count of cases}}{\text{Count of controls}}$.

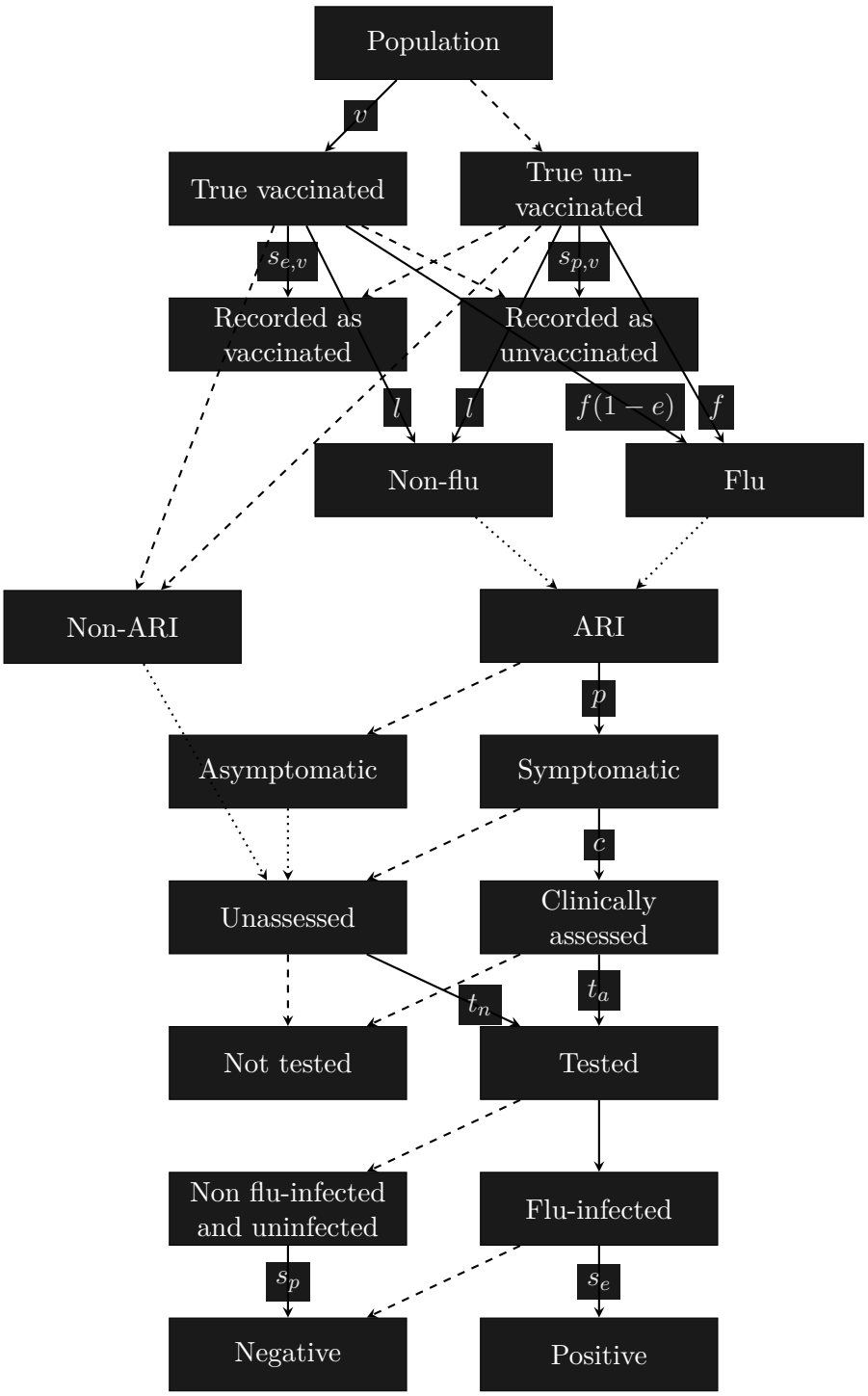

Figure 2: Simulation decision tree. v — vaccination probability, $s_{e,v}$ — sensitivity of vaccination record, $s_{p,v}$ — specificity of vaccination record, l — probability of nonflu infection, f — probability of flu infection, e — vaccine effectiveness, p — probability of infection being symptomatic, c — probability of symptomatic infection being clinically assessed, t_a — probability of clinically assessed infection being tested for flu, t_n — probability of unassessed symptomatic infection, asymptomatic infection and no infection being tested for flu, s_e — sensitivity of flu test, s_p — specificity of flu test. Full parameter key is in Table 1. Solid lines mean allocation with probabilities represented by the indicated parameters. Dashed line probabilities are complements of corresponding solid line probabilities. Dotted lines represent full-group allocation. Probability of being flu-infected for those who are tested isn't indicated because it wasn't necessary for the purposes of simulations.

1.3 Parameter Variation

Every parameter in the simulation was set at a prespecified value. Some of the parameters were set to vary (their prespecified value would have been ignored then). Setting a parameter to vary meant meant that the parameter was assigned a small set of values, set to the first of those values, a set amount of populations were simulated using that value, then it was set to the next value and so on until the simulation went through the entire set.

If multiple parameters were set to vary the simulation would have gone through all possible combinations of all parameters values.

1.4 Mixed-group simulations

If a simulation required a population to be composed of multiple groups, each group was simulated as if it were a separate population. The only mixed combination used was children/adults/elderly. The total size of each group was obtained by multiplying the total requested sample size (usually 200,000) by the specified proportion of each group in the population (this is the w parameter which would have been set to 1 if there was only one group in the population). Amounts of cases and controls were counted in each of the groups and added together to represent the counts obtained from the full population.

1.5 Additional simulations

To determine the effects of individual parameters, an additional set of simulations was performed with parameters fixed to values shown in Table 3. This set of parameter values produced unbiased VE estimates in surveillance data. Using this as a baseline, required parameters could be varied (e.g. s_p can be set below 1) to observe their effect in absence of other sources of bias. Additional simulations were also used to observe the effect some parameters have on others (e.g. how s_p set below 1 affects variation of VE estimates at different values of t_n).

1.6 Parameter estimates used

Table 1: Parameter names, meanings and values used in simulations. "Range used" shows the range of values used for variation in individual age group simulations. Tables 2 contains values and patterns used for variation in mixed group simulations. Every parameter except c represents an absolute probability (some of them only apply to subsets of the population). Only relative probability estimates could be obtained for c (by comparing presentation counts found in ASPREN data [1] to expected underlying population size derived from other parameter estimates), its values were set to 1 in individual group simulations unless it is the parameter varied. Parameter w was only relevant if the population was requested to be composed of multiple groups. Shown values are the ones used in children/adults/elderly mixed simulation. In individual simulations, w was set to 1. Estimates of v and e were derived from provided data.

Par.	Description	Range Used	Children (<15)	Adults (15-65)	Elderly (65+)	Ref.
w	Proportion of the age groups in the general population		0.189	0.657	0.154	[2]
v	Probability of being vaccinated	0.05 - 0.5	0.1	0.25	0.66	
$s_{e,v}$	Sensitivity of exposure measurement	0.9 - 1	0.9	0.95	0.98	[3, 4, 5]
$s_{p,v}$	Specificity of exposure measurement	0.5 - 1	0.9	0.8	0.7	[3, 4, 5]
e	Vaccine effectiveness	0.1 - 0.9	0.6	0.5	0.4	
f	Influenza risk in unvaccinated	0.05 - 0.15	0.15	0.08	0.05	[6]
l	Non-influenza ARI risk in vaccinated and unvaccinated	0.1 - 0.3	0.3	0.15	0.1	[7, 8]
p	Probability of the ARI being symptomatic	0.1 - 0.9	0.84	0.84	0.84	[9]
c	Relative probability of being clinically assessed as having ARI when it is symptomatic	0.1 - 0.9	0.4	0.3	1	
t_a	Tested probability for clinically assessed ARI	0.1 - 0.9	0.17	0.35	0.22	[1]
t_n	Tested probability for everyone without clinically assessed ARI	0 - 0.3	0.15	0.15	0.15	
s_e	Sensitivity of influenza test	0.5 - 1	0.86	0.86	0.86	[10]
s_p	Specificity of influenza test	0.9 - 1	0.984	0.984	0.984	[10]

Table 2: Combinations (patterns of variation) and values used in fixed variation of parameters when multiple groups were present in the population. Combinations 1 and 3 were not used for w.

Combination	Children	Adults	Elderly
All low	Low	Low	Low
All mid	Mid	Mid	Mid
All high	High	High	High
Children high	High	Low	Low
Adults high	Low	High	Low
Elderly high	Low	Low	High

Parameter	Low	Mid	High
w	0.15	0.33	0.7
v	0.05	0.3	0.5
$s_{e,v}$	0.9	0.95	1
$s_{p,v}$	0.5	0.75	1
e	0.1	0.5	0.9
f	0.05	0.1	0.15
l	0.1	0.15	0.3
p	0.1	0.5	0.9
c	0.1	0.5	0.9
t_a	0.1	0.5	0.9
t_n	0	0.15	0.3
s_e	0.5	0.75	1
s_p	0.9	0.95	1

Table 3: Parameter values used in the additional simulation set. Parameter w is missing because the additional simulations were always performed with only one parameter set in the population (equivalent to only having one age group).

Parameter	Value
v	0.5
$s_{e,v}$	1
$s_{p,v}$	1
e	0.5
f	0.3
l	0.3
p	1
c	1
t_a	1
t_n	1
s_e	1
s_p	1

2 Results and discussion

The following sections present and discuss results associated with every parameter whose variation within ranges defined in Table 1 had a perceivable effect on the bias of VE estimates.

2.1 Individual-group simulations

2.1.1 Effect of tested proportions — t_a and t_n

Changing t_a (test probability for those with clinically assessed ARI) and t_n (test probability for everyone else) only affected the estimates of VE in administrative data as shown in Figures 3 and 4 respectively. The parameter that allowed the pattern seen in Figure 3 to be replicated in the additional simulation set was influenza test specificity s_p when set below 1. No additional simulation replicated the pattern seen in Figure 4 (most obvious in the elderly group).

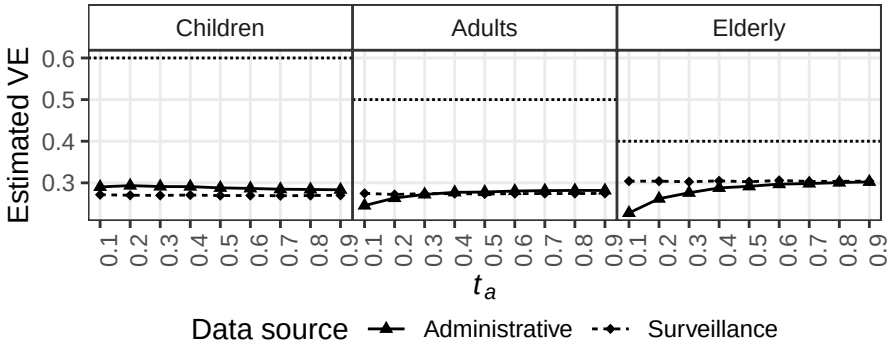

Figure 3: Effect of changing t_a (test probability for those with clinically assessed ARI) in different populations. All groups were simulated individually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

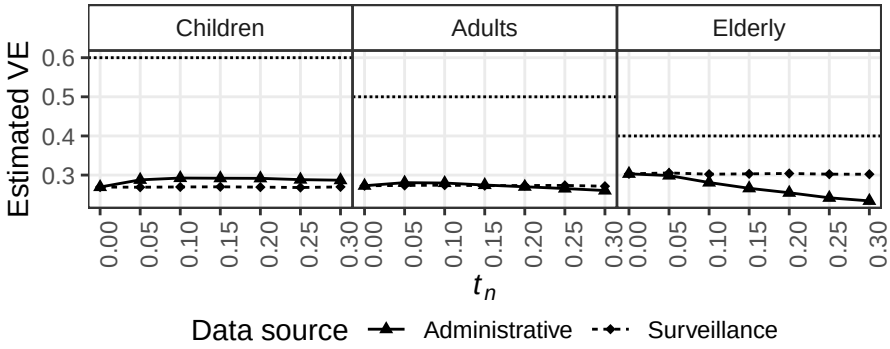

Figure 4: Effect of changing t_n (test probability for those with no infection, asymptomatic infection and not clinically assessed symptomatic infection) in different populations. All were simulated individually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

The reason why only administrative data is affected is the fact that when t_a changed, t_n remained fixed and vice versa. As t_a decreased, administrative sample became more and more different from surveillance due to there being a decreasing proportion of people with ARI in the population who get tested and a constant proportion without ARI who get tested. This compositional difference can also be created by fixing t_a and increasing t_n. The results produced are equivalent - as the administrative sample becomes more compositionally different from the surveillance sample, the administrative estimate bias may increase.

The reason why no additional simulation replicated the pattern in Figure 4 is likely the fact that additional simulations only allowed one parameter other than t_n to introduce bias. The pattern seen is likely a result of multiple parameters other than t_n introducing bias. Hence no additional simulation set had enough "accumulated" bias to replicate the pattern closely.

2.1.2 Effect of true vaccine effectiveness (e) and coverage v

Variation of true vaccine effectiveness e affected both surveillance and administrative estimates of VE — they were more biased with higher true VE values, especially so in children as shown in Figure 5.

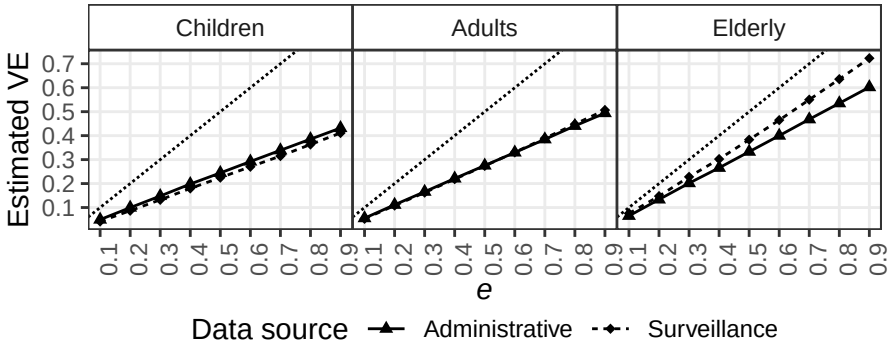

Figure 5: Effect of changing e in different populations. All were simulated individually. The dotted line is the true value. The dash-dotted line is drawn at VE estimate of 0. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

The parameter responsible for the greater bias in children was most likely the vaccinated proportion v as children had it set to the lowest value of the three groups. Since the additional simulations that had low v did not have the pattern replicated, it is likely that low v increases bias at high e only in presence of misclassification. This was confirmed when more simulations were run where e and v were both varied in different misclassification settings. Results in Figure 6 showed that the pattern is only replicated with low v and non-one $s_{p,v}$. Figure 7 shows that at vaccine coverage of 50% all groups have similar bias in VE estimates (around 10%).

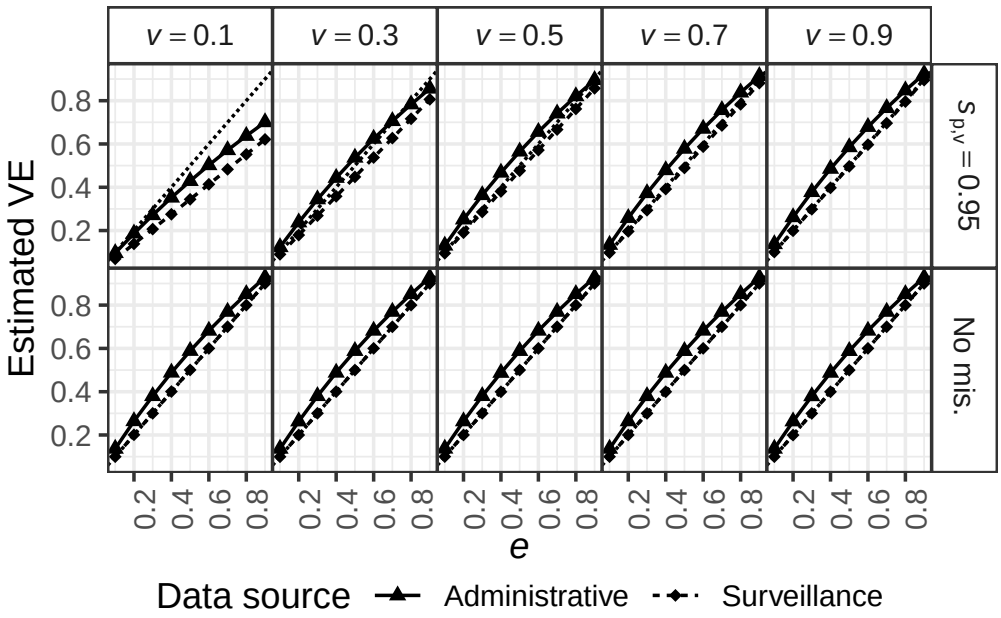

Figure 6: The effect of changing both e and v while also allowing for misclassification. The other parameters were set to values shown in Table 3. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted. The only misclassification setting shown is decreased $s_{p,v}$. The other settings (decreased $s_{e,v}$, s_p, s_e) produced results similar to those of no misclassification.

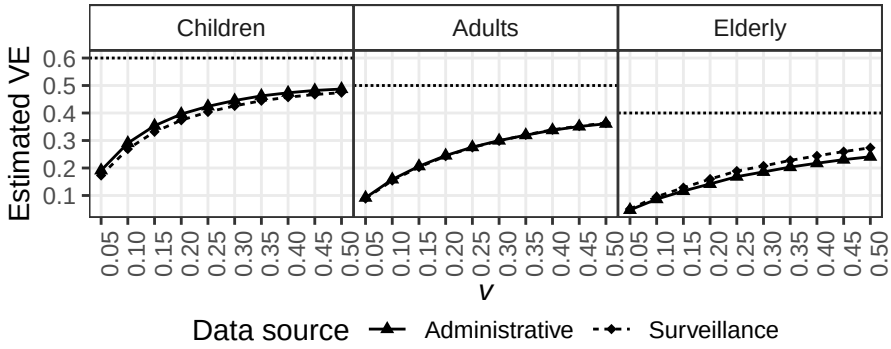

Figure 7: Effect of changing v in different populations. All were simulated individually. The dotted line is the true value. The dash-dotted line is drawn at VE estimate of 0. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

The fact that in absence of misclassification neither vaccine effectiveness nor vaccinated proportion will be expected to have an effect on the bias of a test-negative odds ratio can be shown in Table 4 and Eq. 1.

$$OR = 1 - e \tag{1}$$

When misclassification is added in a form of imperfect specificity of vaccination status measurement, bias is introduced as seen in Table 5 and Eq. 2

Table 4: Expected proportions of surveillance data. Assumptions: no misclassification, t_a= 1, t_n= 1. F - flu-infected, V - vaccinated.

	F	$\bar{F}$
V	$vf(1-e)$	vl
$\bar{V}$	$(1-v)f_u$	$(1-v)l$

Table 5: Expected proportions of surveillance data. Assumptions: no misclassification other than $s_{p,v}$, t_a= 1,t_n= 1. F - flu-infected, V - vaccinated

	F	$\bar{F}$
V	$vf(1-e)+(1-s_{p_v})(1-v)f$	$vl+(1-s_{p,v})(1-v)l$
$\bar{V}$	$(1-v)fs_{p,v}$	$(1-v)ls_{p,v}$

$$\begin{aligned} OR &= \frac{v(1-e)+(1-s_{p,v})(1-v)}{v+(1-s_{p,v})(1-v)} = \frac{v-ev+1-s_{p,v}-v+s_{p,v}v}{v+1-v-s_{p,v}+s_{p,v}v} = \frac{s_{p,v}v-s_{p,v}-ev+1}{s_{p,v}v-s_{p,v}+1} \\ &= \frac{s_{p,v}(v-1)+1-ev}{s_{p,v}(v-1)+1} = 1-\frac{ev}{s_{p,v}(v-1)+1} \end{aligned} \tag{2}$$

This OR will approach the unbiased $1-e$ as v increases. Eq. 3 shows that the bias will increase with e but only if $s_{p,v} < 1$. When $s_{p,v} = 1$ the bias is 0.

$$\begin{aligned} B &= OR_{\text{biased}} - RR_{\text{true}} = 1-\frac{ev}{s_{p,v}(v-1)+1} - (1-e) = e-\frac{ev}{s_{p,v}(v-1)+1} \\ &= e(1-\frac{v}{s_{p,v}(v-1)+1}) \end{aligned} \tag{3}$$

For a given value of e, the bias is proportional to $1-\frac{v}{s_{p,v}(v-1)+1}$ meaning that it depends on v but only if $s_{p,v} < 1$. When $s_{p,v} = 1$ the bias is 0.

The overall effect of true VE is such that as long as there is imperfect vaccination status measurement specificity, higher values of it will result in greater bias and this bias will be affected by vaccinated proportion - the smaller the proportion the greater the bias.

Other misclassification types can allow higher values of true VE to introduce bias as well but simulation results showed that it would not be as prominent as the one discussed above (associated with $s_{p,v}$), and it would not depend on vaccinated proportion which is why it would not explain the differences between age groups in Figure 5. Effect of other types of misclassification will be discussed in their own sections.

Since misclassification can always be expected, highly efficacious vaccines are likely to be underestimated to a greater degree than less efficacious vaccines, especially so when only a small proportion of the population is vaccinated.

2.1.3 Effect of specificity of vaccination status measurement ($s_{p,v}$)

Variation of this parameter affected both surveillance and administrative estimates of VE — they are more biased at lower values of specificity of vaccination status measurement. This effect was most pronounced in children and least so in the elderly. The reason for this is the fact that children had the lowest vaccinated proportion. As a result, they had the lowest number of true vaccinated cases and non-cases to which exposure misclassification would add subjects who were truly unvaccinated but classified as vaccinated. Since the true numbers in children were low, this misclassification had a greater impact on VE estimate bias than it did in the other groups.

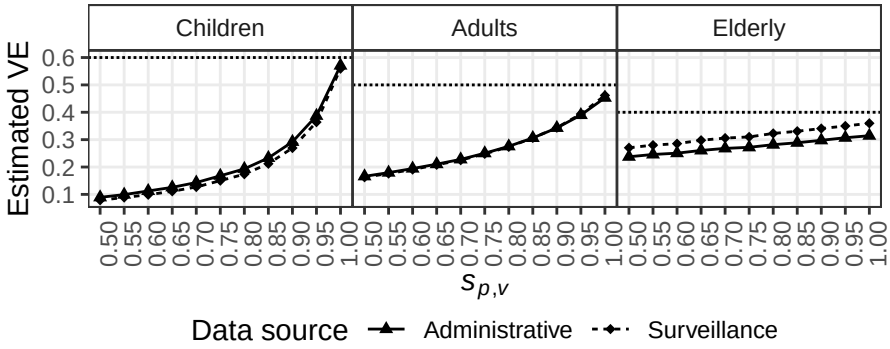

Figure 8: Effect of changing $s_{p,v}$ in different populations. All were simulated individually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

2.1.4 Effect of influenza test specificity (s_p) and influenza incidence (f)

Variation of s_p affected both surveillance and administrative estimates of VE — they are more biased at lower values. This effect was most pronounced in administrative data in the elderly as shown in Figure 9.

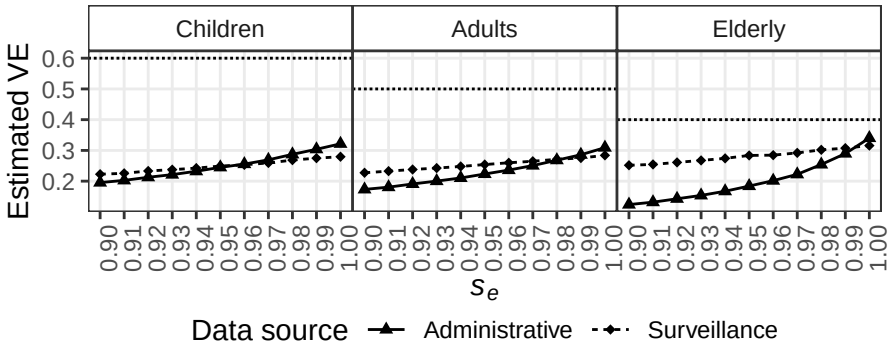

Figure 9: Effect of changing s_p in different populations. All were simulated individually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

Administrative data appears to be more sensitive to outcome misclassification. The effect is most obvious in the elderly due to them having the lowest f. To show this, the expected proportions are shown in Table 6. Eq. 4 and 5 show expected ORs in surveillance and administrative data.

Figure 10 graphs both equations at various values of s_p and f.

Table 6: Expected proportions in a population. Shown for ARI and non-ARI subjects separately. Surveillance sample would only contain ARI subjects, administrative would contain everyone. T_p - tested as flu-infected, T_n - tested as uninfected V - vaccinated, A - ARI. Assumptions: no misclassification other than s_p, t_a= 1, t_n= 1

	T_p		T_n	
	A	$\bar{A}$	A	$\bar{A}$
V	$vf(1-e)+vl(1-s_p)$	$v(1-s_p)(1-l-f(1-e))$	vls_p	$vs_p(1-l-f(1-e))$
$\bar{V}$	$(1-v)f+(1-v)l(1-s_p)$	$(1-v)(1-s_p)(1-l-f)$	$(1-v)ls_p$	$(1-v)s_p(1-l-f)$

$$OR_{\text{surveillance}} = \frac{f(1-e)+l(1-s)}{f+l(1-s)} \tag{4}$$

$$OR_{\text{administrative}} = \frac{\Big(1-s(1-f(1-e))\Big)\Big(s(1-f)\Big)}{\Big(1-s(1-f)\Big)\Big(s(1-f(1-e))\Big)} \tag{5}$$

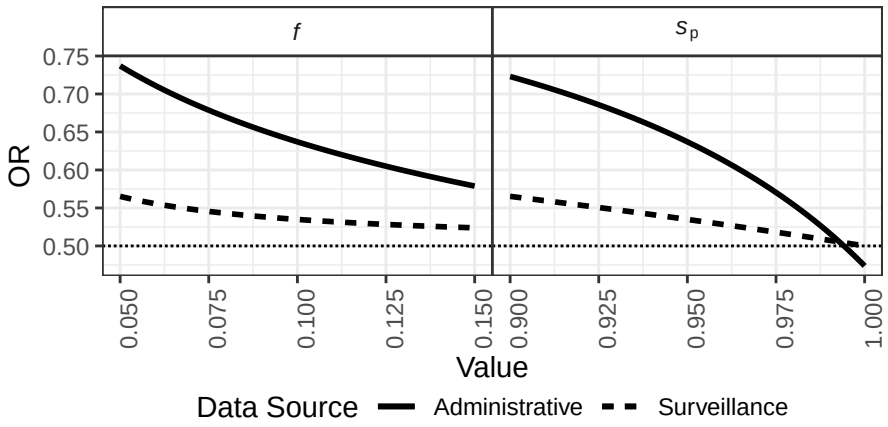

Figure 10: The impact of flu incidence and specificity. s_p was graphed at f of 0.1; f was graphed at s_p of 0.95. Both had true VE (e) set to 0.5 and l to 0.15. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data.

It can be seen from the figure that decrease in either parameter leads to bias towards the null (1). Additionally, administrative data is affected to a greater extent than surveillance. This agrees with the simulations where s_p was seen to affect administrative data more so than surveillance and the greatest impact was seen in elderly who had the lowest f.

The biasing effect of high f can be seen as a result of violating the rare disease assumption. In absence of misclassification this would introduce bias away from the null in administrative data while surveillance data would remain unbiased. However, administrative data also appears to be more susceptible to outcome misclassification. So at any value of f in absence of misclassification, administrative data will be biased away from the null, while in presence of outcome misclassification (particularly imperfect specificity) it will be more biased towards the null due to its higher susceptibility to misclassification. This can be seen in the cross-over of lines in the s_p panel of Figure 10.

2.1.5 Effect of sensitivities ($s_{e,v}$ and s_e)

Variation of these parameters within the plausible range had little effect of both surveillance and administrative estimates in every group — they are slightly more biased with at lower values as shown in Figures 11 and 12.

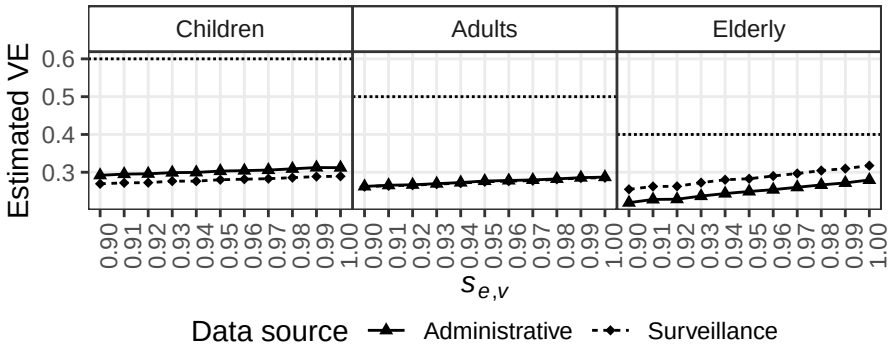

Figure 11: Effect of changing $s_{e,v}$ in different populations. All were simulated inidvidually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

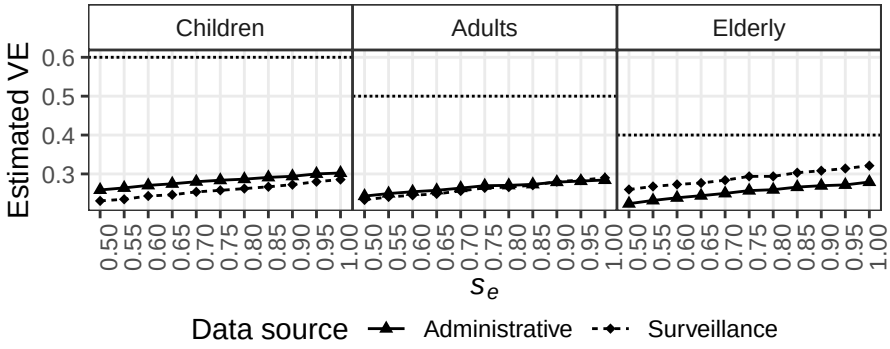

Figure 12: Effect of changing s_e in different populations. All were simulated individually. The dotted line is the true value. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted.

2.2 Mixed-group simulations

Bias is generally negative (towards the null). Administrative data is generally more biased in the same direction as surveillance.

2.2.1 Effect of t_n

Under certain conditions, the overall administrative VE estimate can become much more biased than the overall surveillance estimate or any of the group-specific estimates. This happened in the simulations whent_n was set to a high value (0.3) either in the elderly or in children and to a low value (0) in the other groups as shown in Figure 13

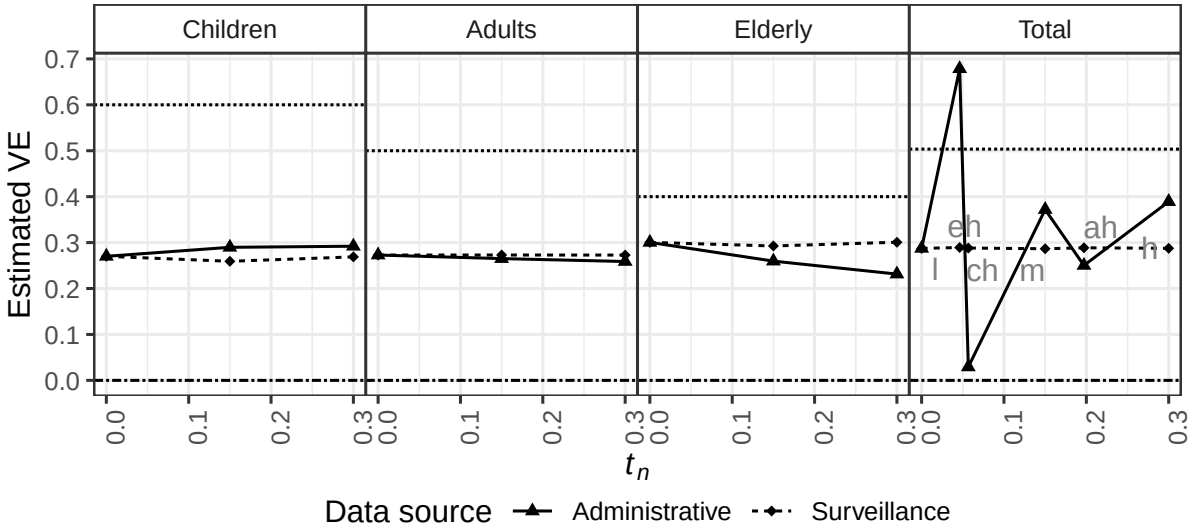

Figure 13: Effect of changing t_n in a population with multiple groups. The dotted line is the true value. The dash-dotted line is drawn at VE estimate of 0. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted. The three panels on the left show group-specific VE estimates. The right-most panel shows the VE estimates calculated from the whole sample ignoring age. Letters correspond to parameter combinations shown in Table 2: all low (l), all mid (m), all high (h), children low (cl), adults low (al), elderly low (el). The overall administrative estimate under is more biased when t_n is set to a high value in the elderly or in the children than under other combinations.

The results above show that when t_n changes among groups, the overall unadjusted estimate of VE of administrative data can swing in either direction.

The reason for this is that those who do not have an ARI but are included into the study will mostly contribute to the study's non-cases. If the majority of the ARI-free people are vaccinated, then their contribution to vaccinated controls will be greater than to vaccinated cases. This only becomes a problem for the odds ratio (and vaccine effectiveness) estimates when the groups are mixed and the estimates are coming from the overall population.

To illustrate, let's say there is a population with 3 age subgroups and non-ARI individuals are only tested in one of them. Table 7 shows expected counts and odds ratio (OR).

If the OR were to be calculated for individual age subgroups, it would be unbiased if they only contain people with ARI (groups 2 and 3). Some bias is introduced with "contamination" by healthy people. This corresponds to a traditional case-control study - control subjects are

Table 7: Expected proportions of administrative data in a population with three age groups. $h_v = 1 - (1 - e)f - l$ represents the probability of staying healthy for a vaccinated individual. $h_u = 1 - f - l$ represents the probability of staying healthy for an unvaccinated individual. Assumptions: no misclassification, $t_a = 1$, $t_n = 1$ in group 1 and 0 otherwise, disease incidence is the same in all groups, all groups have the same size.

Group	Cases vaccinated	Cases unvaccinated	Controls vaccinated	Controls unvaccinated	OR
1	$v_1(1-e)f$	$(1-v_1)f$	$v_1(l+h_v)$	$(1-v_1)(l+h_u)$	$\frac{(1-e)(l+h_u)}{(l+h_v)}$
2	$v_2(1-e)f$	$(1-v_2)f$	$v_2 l$	$(1-v_2)l$	$1-e$
3	$v_3(1-e)f$	$(1-v_3)f$	$v_3 l$	$(1-v_3)l$	$1-e$

selected from everyone who isn't a case (as opposed to everyone with an ARI who isn't a case). With low flu incidence $h_u \simeq h_v$ so the bias is small.

The overall population is where bias has the potential to swing in a more pronounced manner. Table 8 shows expected counts in the overall population (made up of subgroups 1, 2 and 3) for both surveillance and administrative data.

Table 8: Expected proportions in the overall population. The first row is a population made up of the three age groups in Table 7. The second row is the counts we would have gotten if the first subgroup in Table 7 did not have healthy people in control groups (i.e. if all groups had t_n set to 0). This corresponds to surveillance data counts.

Data type	Cases vaccinated	Cases unvaccinated	Controls vaccinated	Controls unvaccinated	OR
Administrative	$(1-e)f\sum_{i=1}^{3} v_i$	$f\sum_{i=1}^{3}(1-v_i)$	$v_1 h_v + l\sum_{i=1}^{3} v_i$	$h_u(1-v_1) + l\sum_{i=1}^{3}(1-v_i)$	Eq. 6
Surveillace	$(1-e)f\sum_{i=1}^{3} v_i$	$f\sum_{i=1}^{3}(1-v_i)$	$l\sum_{i=1}^{3} v_i$	$l\sum_{i=1}^{3}(1-v_i)$	$1-e$

Surveillance data would only include those with ARI which is why it is not affected by the "contamination" by healthy people which is experienced by administrative data hence its odds ratio is unbiased. However the OR coming from administrative data would have this form:

$$OR_{123} = \frac{(1-e)\sum_{i=1}^{3} v_i \Big[h_u(1-v_1) + l\sum_{i=1}^{3}(1-v_i)\Big]}{\sum_{i=1}^{3}(1-v_i)\Big[v_1 h_v + l\sum_{i=1}^{3} v_i\Big]} \tag{6}$$

The issue comes from the fact that group 1 can contribute different counts to numerator and denominator and therefore make the OR swing in either direction. Figure 14 demonstrates how the vaccinated proportion in subgroup 1 can affect the overall OR.

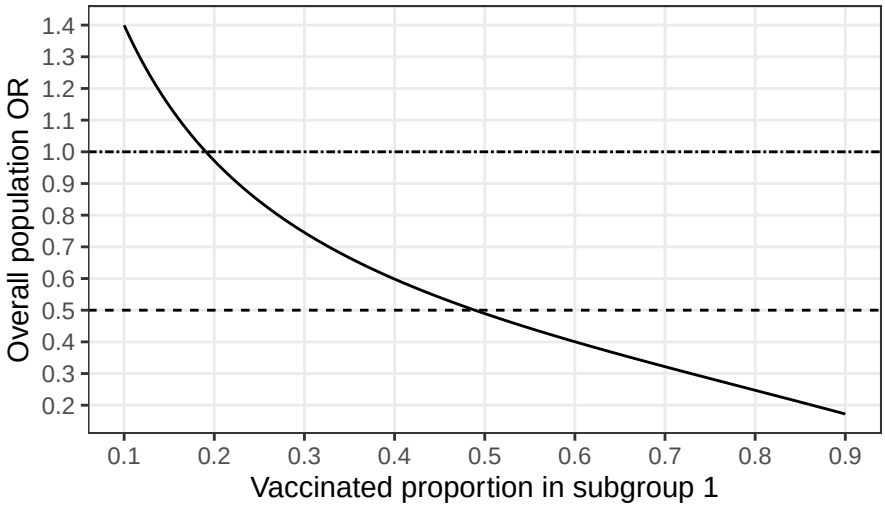

Figure 14: The impact of the vaccinated proportion in subgroup 1 on the OR of the overall population. Vaccinated proportions in the other groups are 0.5. Flu incidence is 0.025 in vaccinated and 0.05 in unvaccinated so that true OR and true VE are both 0.5. Non-flu incidence is 0.1. The dashed line is the true value. Dash-dotted line is drawn at OR of 1 (VE of 0).

At higher values of vaccinated proportion, the population OR becomes smaller leading to overestimation of VE. At lower values of vaccinated proportion, the population OR becomes larger leading to underestimation of VE.

This is the primary reason why the overall population administrative VE estimates went up so much in simulation results when elderly were the only group to have a non-0 t_n and why they went down when children were the only group to have a non-0 t_n. Elderly have a high vaccinated proportion (0.66) which is why they mostly contributed their numbers to the denominator of OR which decreased OR and increased VE estimates in the overall group. Children have a low vaccinated proportion (0.11) which is why they mainly contributed to the numerator of OR which increased OR and decreased the overall VE estimate.

The overall population administrative VE estimates were not as greatly affected when adults were the only group with a non-0 t_n. Adults had a v value of 0.25 meaning that their impact was the same as children's (VE estimates decreased) but it was not as pronounced since the additional contribution by adults was more balanced between the numerator and the denominator of the overall population OR.

2.2.2 Effect of f and v

Both the administrative and the surveillance VE estimates became greatly biased towards the null when the incidence of flu was set high in the elderly as shown in Figure 15.

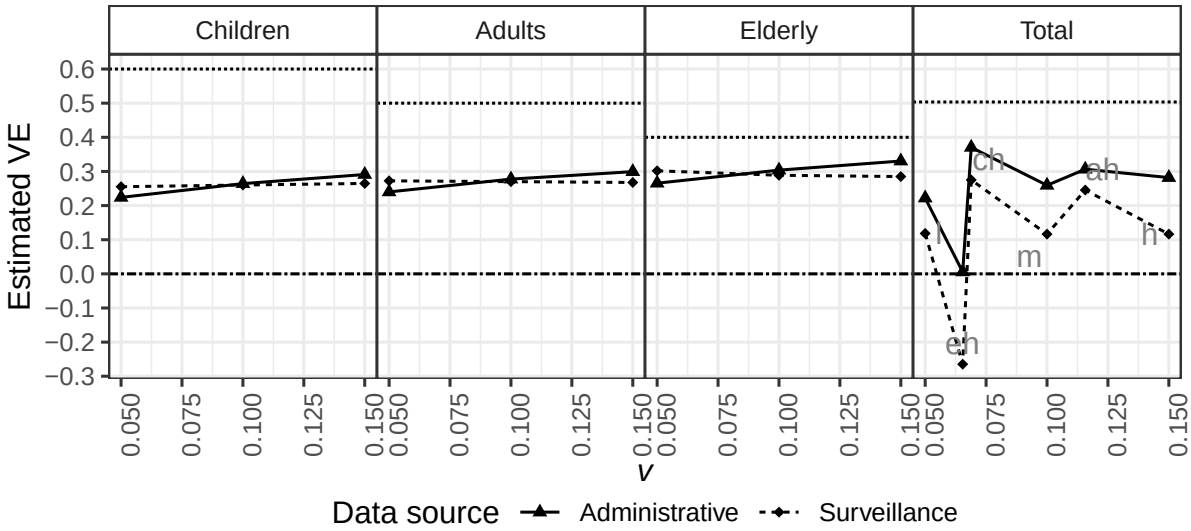

Figure 15: Effect of changing f in a population with multiple groups. The dotted line is the true value. The dash-dotted line is drawn at VE estimate of 0. The solid line is the estimated VE in administrative data, the dashed line is the estimated VE in surveillance data. Points correspond to values at which the simulations were conducted. Dotted line is drawn at VE estimate of 0. The three panels on the left show group-specific VE estimates. The right-most panel shows the VE estimates calculated from the whole sample ignoring age. Numbers correspond to parameter combinations shown in Table 2. Both surveillance and administrative overall estimates under combination 6 (elderly high) are more biased than under other combinations.

The general trend is that the overall VE estimate becomes more biased if there is a group in the population whose flu incidence and vaccinated proportion are both greatly different from the other groups.

To illustrate, let's say there is a population with three groups. Flu incidence and vaccinated proportion are the same in groups 2 and 3. Table 9 shows expected counts for surveillance data.

Table 9: Expected proportions in surveillance data in a population with 3 groups. Last row is the overall counts. Assumptions: no misclassification, the same number of people in all groups, the save vaccinated proportion and flu incidence in groups 2 and 3, everyone eligible is part of the study.

Group	Cases vaccinated	Cases unvaccinated	Controls vaccinated	Controls unvaccinated	OR
1	$v_1 f_1(1-e)$	$(1-v_1)f_1$	$v_1 l$	$(1-v_1)l$	$1-e$
2	$vf(1-e)$	$(1-v)f$	vl	$(1-v)l$	$1-e$
3	$vf(1-e)$	$(1-v)f$	vl	$(1-v)l$	$1-e$
123	$(1-e) \times (v_1 f_1 + 2vf)$	$[(1-v_1)f_1 + 2(1-v)f]$	$l(v_1+2v)$	$l[1-v_1+2(1-v)]$	Eq. 7

$$OR_{123} = (1-e)\frac{(v_1 f_1 + 2vf)(1 - v_1 + 2(1-v))}{[(1-v_1)f_1 + 2(1-v)f](v_1 + 2v)} \tag{7}$$

The bias in the overall OR will only be present if both v_1 and f_1 are different from v and f. Meaning that the group of interest has to have its flu incidence and vaccinated proportion be different from the other groups in order of this biasing effect to occur. Figure 16 shows the bias for different v_1 and f_1. Eq. 8 and 9 show that this biasing effect disappears when f_1 or v_1 are set to f or v respectively.

$$\begin{aligned} OR_{123} &= (1-e)\frac{(v_1 f + 2vf)(1 - v_1 + 2(1-v))}{[(1-v_1)f + 2(1-v)f](v_1 + 2v)} \\ &= (1-e)\frac{(v_1 + 2v)(1 - v_1 + 2(1-v))}{[1 - v_1 + 2(1-v)](v_1 + 2v)} \\ &= 1 - e \end{aligned} \tag{8}$$

$$\begin{aligned} OR_{123} &= (1-e)\frac{(v f_1 + 2v f_1)(1 - v + 2(1-v))}{[(1-v)f_1 + 2(1-v)f_1](v + 2v)} \\ &= (1-e)\frac{3[1 - v + 2(1-v)]}{3[(1-v) + 2(1-v)]} \\ &= 1 - e \end{aligned} \tag{9}$$

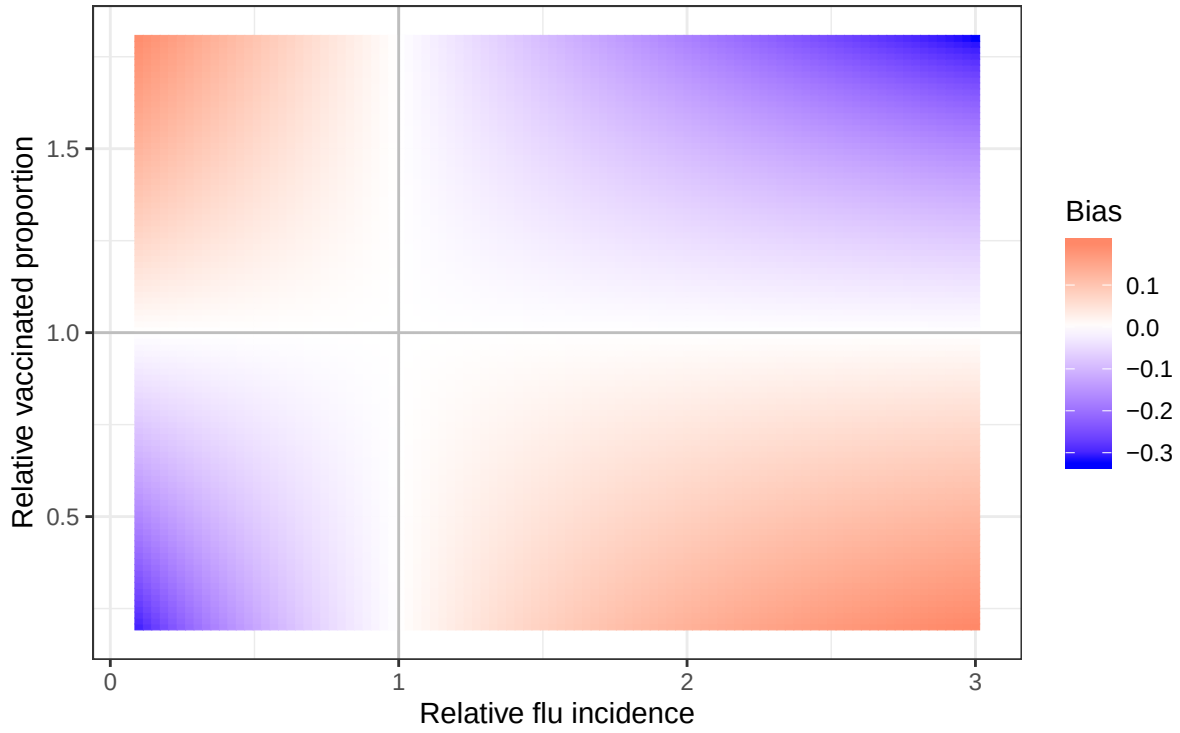

Figure 16: Bias at different values of v_1 and f_1. Shown as relative values, that is $\frac{f1}{f}$ and $\frac{v1}{v}$. Solid lines show where the bias is 0 - at relative values of 1, meaning that $f1$ and v_1 are the same as f and v.

3 Conclusion

VE estimates coming from administrative data can be expected to be more biased towards the null as compared to VE estimates of surveillance data. This is mainly due to the fact that administrative data is more susceptible to bias introduced by imperfect specificity of outcome (flu) classification. However, As long as administrative data does not contain a large amount of subjects without ARI the amount of extra bias will be small ("extra" as compared to surveillance data).

Other sources of bias affect surveillance and administrative data to a similar extent. The most prominent are true VE and specificity of vaccination status measurement. Both high true VE and low specificity will introduce bias towards the null.

Due to the fact that various parameters vary with age (e.g. vaccination status, vaccine effectiveness, flu incidence) when multiple age groups are present in the sample, many different parameter combinations are possible across the groups. Many of them will not have a great effect on the age-unadjusted overall VE estimate (i.e. it will lie in range of the age-specific estimates) but certain parameter combinations will introduce a great amount of bias to the overall estimate. As the behaviour of the unadjusted estimate is difficult to predict, it is unreliable with both surveillance and administrative data.

Bibliography

[1] Australian Sentinel Practices Research Network;. Available from: https://aspren.dmac.adelaide.edu.au/.

[2] Australian Demographic Statistics. Australian Bureau of Statistics; 2018. Available from: https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3101.0Sep2018?OpenDocument.

[3] Irving SA, Donahue JG, Shay DK, Ellis-Coyle TL, Belongia EA. Evaluation of self-reported and registry-based influenza vaccination status in a Wisconsin cohort. Vaccine. 2009;27(47):6546–6549.

[4] Donald RM, Baken L, Nelson A, Nichol KL. Validation of self-report of influenza and pneumococcal vaccination status in elderly outpatients. American Journal of Preventive Medicine. 1999;16(3):173–177.

[5] Rolnick Sj, Parker Ed, Nordin Jd, Hedblom Bd, Wei F, Kerby T, et al. Self-report compared to electronic medical record across eight adult vaccines: Do results vary by demographic factors? Vaccine. 2013;31(37):3928–3935.

[6] Tokars JI, Olsen SJ, Reed C. Seasonal Incidence of Symptomatic Influenza in the United States. Clinical Infectious Diseases. 2017;66(10):1511–1518.

[7] Influeza Surveillance Report 2017. Australian Department of Health; 2017. Available from: https://www.health.gov.au/internet/main/publishing.nsf/Content/cda-ozflu-2017.htm.

[8] Influeza Surveillance Report 2018. Australian Department of Health; 2018. Available from: https://www.health.gov.au/internet/main/publishing.nsf/Content/ozflu-surveil-2018-final.htm.

[9] Leung NHL, Xu C, Ip DKM, Cowling BJ. The fraction of influenza virus infections that are asymptomatic: a systematic review and meta-analysis. Epidemiology. 2015;26(6):862–872.

[10] Druce J, Tran T, Kelly H, Kaye M, Chibo D, Kostecki R, et al. Laboratory diagnosis and surveillance of human respiratory viruses by PCR in Victoria, Australia, 2002–2003. Journal of Medical Virology. 2004 Nov;Available from: https://www.onlinelibrary.wiley.com/doi/10.1002/jmv.20246.