Supplemental Material for “Using social contact data to improve the overall effect estimate of a cluster-randomized influenza vaccination program in Senegal”

Gail E. Potter

National Institute of Allergy and Infectious Diseases, National Institutes of Health, and the Emmes Company, Rockville Maryland, USA

Nicole Bohme Carnegie

Montana State University, Bozeman Montana, USA

Jonathan D. Sugimoto

Epidemiologic Research and Information Center, Veterans Affairs Puget Sound Health Care System and Fred Hutchinson Cancer Research Center, Seattle Washington, USA

Aldiouma Diallo

Institut de Recherche pour le Développement, Niakhar Senegal

John C. Victor

PATH, Seattle Washington, USA

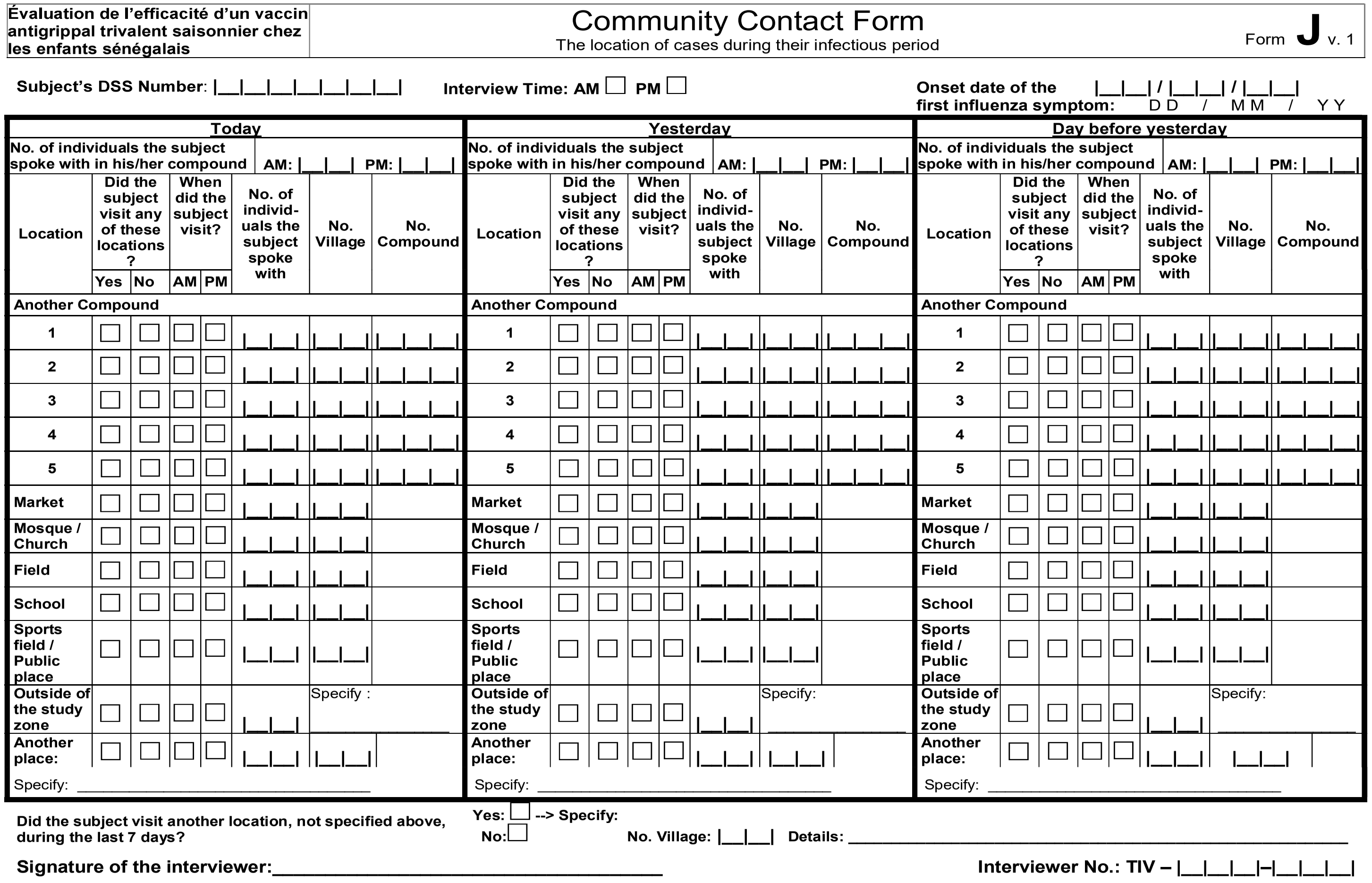
Kathleen M. Neuzil

University of Maryland School of Medicine, Baltimore Maryland, USA

M. Elizabeth Halloran

University of Washington Department of Biostatistics and Fred Hutchinson Cancer Research Center, Seattle Washington, USA

1. Contact survey form



Évaluation de l'efficacité d'un vaccin antigrippal trivalent saisonnier chez les enfants sénégalais	Community Contact Form The location of cases during their infectious period	Form **J** v. 1

Subject's DSS Number: |__|__|__|__|__|__|__|__| **Interview Time: AM** □ **PM** □ **Onset date of the first influenza symptom:** |__|__| / |__|__| / |__|__|
DD / MM / YY

Today							
No. of individuals the subject spoke with in his/her compound					**AM: |__|__|**	**PM: |__|__|**	
Location	**Did the subject visit any of these locations ?**		**When did the subject visit?**		**No. of individ-uals the subject spoke with**	**No. Village**	**No. Compound**
	Yes	**No**	**AM**	**PM**			
Another Compound							
1	□	□	□	□	|__|__|	|__|__|	|__|__|__|
2	□	□	□	□	|__|__|	|__|__|	|__|__|__|
3	□	□	□	□	|__|__|	|__|__|	|__|__|__|
4	□	□	□	□	|__|__|	|__|__|	|__|__|__|
5	□	□	□	□	|__|__|	|__|__|	|__|__|__|
Market	□	□	□	□	|__|__|	|__|__|	
Mosque / Church	□	□	□	□	|__|__|	|__|__|	
Field	□	□	□	□	|__|__|	|__|__|	
School	□	□	□	□	|__|__|	|__|__|	
Sports field / Public place	□	□	□	□	|__|__|	|__|__|	
Outside of the study zone	□	□	□	□	|__|__|	Specify : ____________	
Another place:	□	□	□	□	|__|__|	|__|__|	
Specify: ____________________________							

Yesterday							
No. of individuals the subject spoke with in his/her compound					**AM: |__|__|**	**PM: |__|__|**	
Location	**Did the subject visit any of these locations ?**		**When did the subject visit?**		**No. of individ-uals the subject spoke with**	**No. Village**	**No. Compound**
	Yes	**No**	**AM**	**PM**			
Another Compound							
1	□	□	□	□	|__|__|	|__|__|	|__|__|__|
2	□	□	□	□	|__|__|	|__|__|	|__|__|__|
3	□	□	□	□	|__|__|	|__|__|	|__|__|__|
4	□	□	□	□	|__|__|	|__|__|	|__|__|__|
5	□	□	□	□	|__|__|	|__|__|	|__|__|__|
Market	□	□	□	□	|__|__|	|__|__|	
Mosque / Church	□	□	□	□	|__|__|	|__|__|	
Field	□	□	□	□	|__|__|	|__|__|	
School	□	□	□	□	|__|__|	|__|__|	
Sports field / Public place	□	□	□	□	|__|__|	|__|__|	
Outside of the study zone	□	□	□	□	|__|__|	Specify: ____________	
Another place:	□	□	□	□	|__|__|	|__|__|	
Specify: ____________________________							

Day before yesterday							
No. of individuals the subject spoke with in his/her compound					**AM: |__|__|**	**PM: |__|__|**	
Location	**Did the subject visit any of these locations ?**		**When did the subject visit?**		**No. of individ-uals the subject spoke with**	**No. Village**	**No. Compound**
	Yes	**No**	**AM**	**PM**			
Another Compound							
1	□	□	□	□	|__|__|	|__|__|	|__|__|__|
2	□	□	□	□	|__|__|	|__|__|	|__|__|__|
3	□	□	□	□	|__|__|	|__|__|	|__|__|__|
4	□	□	□	□	|__|__|	|__|__|	|__|__|__|
5	□	□	□	□	|__|__|	|__|__|	|__|__|__|
Market	□	□	□	□	|__|__|	|__|__|	
Mosque / Church	□	□	□	□	|__|__|	|__|__|	
Field	□	□	□	□	|__|__|	|__|__|	
School	□	□	□	□	|__|__|	|__|__|	
Sports field / Public place	□	□	□	□	|__|__|	|__|__|	
Outside of the study zone	□	□	□	□	|__|__|	Specify: ____________	
Another place:	□	□	□	□	|__|__|	|__|__|	
Specify: ____________________________							

Did the subject visit another location, not specified above, during the last 7 days? **Yes:** □ **--> Specify:**
No:□ **No. Village: |__|__| Details:** ____

Signature of the interviewer:______________________________ **Interviewer No.: TIV – |__|__|__|–|__|__|__|**

Table 1. Fraction and percent of contacts reported by respondents while located in treated villages by village of residence and time of day.

Village	Treatment	AM Fraction	AM Percent	PM Fraction	PM Percent
Darou	Vaccine	149/149	100	123/123	100
Kalome Ndofane	Vaccine	1221/1244	98	959/959	100
Poudaye	Vaccine	53/65	82	47/47	100
Mokane Ngouye	Vaccine	1478/1513	98	1307/1312	100
Ngayokheme	Vaccine	6414/6489	99	5638/5682	99
Ndokh	Vaccine	135/137	99	101/103	98
Nghonine	Vaccine	303/306	99	222/231	96
Ngangarlame	Vaccine	859/953	90	533/586	91
Diohine	Vaccine	1195/1377	87	1104/1230	90
Logdir	Vaccine	158/185	85	53/80	66
Ngalagne Kop	Control	0/1052	0	0/873	0
Bary Ndondol	Control	0/545	0	0/512	0
Mboyene	Control	1/634	0	1/522	0
Toucar	Control	6/2870	0	0/2006	0
Godel	Control	0/456	0	0/449	0
Khassous	Control	0/151	0	0/96	0
Kothio	Control	3/643	0	0/440	0
Meme	Control	0/78	0	0/97	0
Poultok Diohine	Control	0/1717	0	0/1240	0
Gadiak	Control	10/466	2	10/310	3

2. Cross-village exposure summaries

This section displays analyses that informed our calculation of cross-village exposure rates. The cross-village exposure rate for a village is defined to be the percentage of contacts to people in clusters of the opposite treatment assignment. The tables in this section summarize these rates based on contacts made while the respondent was visiting other villages and do not incorporate contacts made to visitors from other villages in the respondent's own home. The tables summarize rates of contacts to treated villages by village of residence; these represent the cross-village exposure rate for control villages and one minus the cross-village exposure rate for treated villages.

Table 1 compares fractions and percentages of contacts to treated villages between morning and afternoon/evening time intervals. Cross-cluster exposure rates are similar for the two time intervals, with the main differences being Poudaye and Logdir, whose higher variability than others is likely due to the small number of overall contacts reported in those villages.

Table 2 compares fractions and percentages of contacts reported by respondents during visits to treated villages during the morning by village number and symptom status. Since numbers of asymptomatic reports are low and cross-village exposure is low, cross-village exposure is lower for asymptomatic than symptomatic participants in most villages. Table 3 shows the analogous percentages calculated based on the imputed data and shows higher levels of cross-cluster exposure for symptomatic than asymptomatic people.

Table 2. Fraction and percent of contacts reported by respondents while located in treated villages by village of residence and symptom status.

	Treatment	Asymptomatic		Symptomatic	
Village	Assignment	Fraction	Percent	Fraction	Percent
Darou	Vaccine	61/61	100	88/88	100
Ndokh	Vaccine	28/30	93	107/107	100
Ngayokheme	Vaccine	1468/1490	99	4946/4999	99
Nghonine	Vaccine	32/32	100	271/274	99
Kalome Ndofane	Vaccine	352/359	98	869/885	98
Mokane Ngouye	Vaccine	178/183	97	1300/1330	98
Ngangarlame	Vaccine	171/174	98	688/779	88
Diohine	Vaccine	555/610	91	640/767	83
Poudaye	Vaccine	6/6	100	47/59	80
Logdir	Vaccine	58/58	100	100/127	79
Ngalagne Kop	Control	0/258	0	0/794	0
Bary Ndondol	Control	0/4	0	0/541	0
Mboyene	Control	0/86	0	1/548	0
Toucar	Control	0/839	0	6/2031	0
Godel	Control	0/316	0	0/140	0
Khassous	Control	0/19	0	0/132	0
Meme	Control	0/30	0	0/48	0
Poultok Diohine	Control	0/559	0	0/1158	0
Kothiok	Control	0/157	0	3/486	1
Gadiak	Control	0/137	0	10/329	3

Table 3. Percent of contacts reported by respondents while located in treated villages by village of residence and symptom status based on multiply imputed data.

	Treatment			
Village	Assignment	All	Asymptomatic	Symptomatic
Kalome Ndofane	Vaccine	100	100	100
Ngangarlame	Vaccine	99	100	99
Diohine	Vaccine	99	100	98
Mokane Ngouye	Vaccine	99	100	99
Ngayokheme	Vaccine	99	99	99
Ndokh	Vaccine	99	99	99
Nghonine	Vaccine	98	99	98
Logdir	Vaccine	95	94	96
Darou	Vaccine	96	90	100
Poudaye	Vaccine	93	84	95
Ngalagne Kop	Control	0	0	0
Bary Ndondol	Control	0	0	0
Mboyene	Control	0	0	0
Poultok Diohine	Control	0	1	0
Toucar	Control	1	1	0
Gadiak	Control	2	3	2
Godel	Control	2	1	3
Khassous	Control	3	0	3
Kothiok	Control	3	2	4
Meme	Control	14	0	20

3. Calculation of time-to-event

We restrict our analysis to the twenty villages enrolled in the cluster-randomized trial as these villages received both active and passive surveillance while the other ten received only passive surveillance. The surveillance period for Year 1 was July 15, 2009 to May 31, 2010. These dates determined the start and end of follow-up participants with the following exceptions:

- Start of follow-up was the date participants moved to the study area if the move took place after surveillance began.
- If participants moved out of the study area or to a cluster of the opposite treatment assignment during surveillance, their end of follow-up was the move date.

Time-to-infection was calculated by subtracting the start of follow-up from the sample collection date for infected people; censoring times were calculated based on start and end of follow-up for uninfected people.

Thirteen participants were excluded from analysis because of inconsistencies in their recorded residence data. In addition, those who moved to the study area after the end of Year 1, and those who were infected before moving to the study area or before follow-up began were excluded. Because the primary analysis did not censor or exclude participants based on their residence data, our counts of participants and cases differ slightly from that paper (Diallo et al., 2019).

Time to event for Year 2 (for which surveillance covered July 15, 2010 to May 31, 2011) was calculated analogously. However, during Year 2 of the study, household-based surveillance did not occur from January 1, 2011 to February 18, 2011 due to a strike of employees performing this surveillance, so only infections reported in health posts were recorded during that time period. This could cause bias if the proportions of infections observed at home compared to in health posts different between treatment arms. During the non-strike period of Year 2, proportions of lab-confirmed symptomatic influenza infections reported during household visits were 83.07% in the control group and 87.50% in the vaccine group, respectively (Table 4). Since infections for control arm participants were reported more frequently in health posts than those for vaccine arm participants, the differential reporting could create bias in the efficacy estimate, making the vaccine appear more effective than it actually is. Inverse probability weighting was considered to correct this bias (Seaman and White, 2013). Such an approach would entail up-weighting the observed infections during the strike by $\frac{1}{0.1693} = 5.91$ in the control arm and $\frac{1}{0.1250} = 8.00$ in the vaccine arm, and down-weighting the people classified as uninfected throughout the study period (since some of these would have had infections that would have been detected during household surveillance during the strike). This approach would assume that health post visiting behavior was the same during the strike and outside of the strike. However, the data indicate that that assumption does not hold. Outside of the strike, 66% of infections reported in health posts were in the control group, but during the strike, 78% were. The larger proportion of up-weighted control group infections resulted in a weighted overall effect estimate that was higher, rather than lower, than the unweighted one. As the assumption required by inverse

Table 4. Reporting rates of lab-confirmed symptomatic infection by location within each treatment arm during Year 2, excluding the strike period

	Control	Vaccine
Percent reported in compounds	83.07	87.50
Percent reported in health posts	16.93	12.50

probability weighting did not hold, we instead censored the Year 2 data at the last day before the strike. A secondary analysis includes all of the Year 2 data.

Ties were handled by adding a random draw from a uniform distribution to tied event times (but not censoring times) as per Aalen (1989). When tied event times occurred on the last day of follow-up (along with approximately 30,000 censoring times), this approach led to non-estimable standard errors. Adding random noise to event times but not censoring times resulted in two events with times later than all of the censored times, which created a noninvertible covariance matrix. This is also unrealistic, since censored times generally mean that an entire day passed without an event, but events occur before the day is over. This computational problem was fixed by adding 1 to all censored times on the last day of follow-up, which led to very stable estimates, varying only in the 6th decimal place with different draws from the uniform. We also explored the alternate approach of adding random noise to all event times and to all censored times on the last day of follow-up. That approach produced similar effect estimates to those in this paper, but the estimates were less stable, varying in the fourth and sometimes third decimal place. The approach that we took in this paper gives the same estimates that are produced when we censor the two problematic cases on the day before they became infected.

4. Correspondence to compartmental model for infectious disease transmission

The additive hazards model applied in this paper has a natural correspondence to an SIR (Susceptible-Infected-Removed) compartmental model for disease transmission. To see this, recall that the contamination-adjusted estimator for an individual in cluster j is obtained from the following additive hazards model:

$$\lambda_j(t|M) = \beta_0(t) + \beta_M(t)m_j,$$

where m_j is the total percentage of contacts of susceptibles in cluster j that are with treated clusters.

Next, we define the following notation:

(a) $Y_k(t)$ = the number of infected people in cluster k at time t

(b) κ = the overall average contact rate

(c) η_k = the per-contact transmission probability of infectives in cluster k

(d) m_{jk} = the percentage of contacts from people in cluster j with those in cluster k

(e) α_{jk} = the rate of new infections among susceptibles in cluster j from infectives in cluster k

(f) $N_k =$ the population size of cluster k. For simplicity, we assume a fixed population size in each cluster.

The SIR compartmental model assumes that the rate of transmission from infectives in cluster k to susceptibles in cluster j is the product of the overall contact rate, the percentage of contacts from cluster j that are to people in cluster k, and the per-contact transmission probability: $\alpha_{jk} = \kappa m_{jk} \eta_k$. The hazard function of a susceptible in cluster j is found by summing these cluster-specific transmission rates, weighted by their cluster-specific proportions of infectives, across all clusters:

$$\lambda_j(t) = \sum_{k=1}^{c} \alpha_{jk} \frac{Y_k(t)}{N_k} \tag{1}$$

To simplify notation, we define $\nu_j(t) = \kappa \eta_j \frac{Y_j(t)}{N_j}$, so

$$\lambda_j(t) = \sum_{k=1}^{c} \alpha_{jk} \frac{Y_k(t)}{N_k} = \sum_{k=1}^{c} m_{jk} \nu_k(t) \tag{2}$$

The estimand of interest, which we will denote $\beta(t)$, is the population-averaged difference in hazard of infection associated with a change from 0% to 100% exposure to treatment. That is, $\beta(t) = \bar{\nu}^T(t) - \bar{\nu}^C(t)$, where $\bar{\nu}^T(t)$ is the average of $\nu(t)$ in treated clusters and $\bar{\nu}^C(t)$ is the average of $\nu(t)$ in control clusters. While $\hat{\beta}_Z(t)$ is a consistent estimator for $\beta(t)$ in the absence of contamination, Carnegie, Rui, and Wang proved that $\hat{\beta}_M(t)$ is a consistent estimator for $\beta(t)$ in the presence of measured contamination. (Carnegie et al., 2016)

The expected instantaneous rate of change of the number of infected individuals in cluster i at time t is found by summing the individual hazards of all susceptibles in cluster i. Letting $S_i(t)$ denote the number of susceptibles in cluster i at time t and substituting from (2) yields:

$$\frac{dY_i(t)}{dt} = S_i(t) \sum_{j=1}^{c} \alpha_{ij} \frac{Y_j(t)}{N_j} = \sum_{j=1}^{c} \alpha_{ij} \frac{Y_j(t) S_i(t)}{N_j},$$

which corresponds to an SIR model with no birth and or death. A similar expression for the rate of change of susceptibles is analogously derived, and generalizations such as birth, death, and the addition of an exposed state for an SEIR model are addressed in Carnegie et al. (2016).

5. Rationale for adjustment in estimated contamination estimates based on reports from visitors to the respondent's compound

We define the following notation as described in the main text:

- n_j = number of people living in cluster j
- D_i = number of contacts reported by person i
- T_i = number of contacts person i made in a location in a treated cluster.
- p_j = the proportion of contacts from cluster j to treated clusters.
- $V_{T,j}$ = the total number of contacts reported by people in any treated cluster during their visits to compounds in cluster j.

We initially estimated p_j with

$$\hat{p_j} = \frac{\sum_{i=1}^{n_j} T_i}{\sum_{i=1}^{n_j} D_i}$$

The numerator does not include contacts occurring within the respondent's own compound to visitors from other clusters, since these occurred within the respondent's assigned cluster. We can use estimates reported by visitors from clusters of the opposite assignment, rather than by respondents in cluster j, to obtain this information. When j is a control cluster, our estimator is appropriately updated by adding the percent of contacts from treated clusters to compounds in cluster j to the contamination estimate:

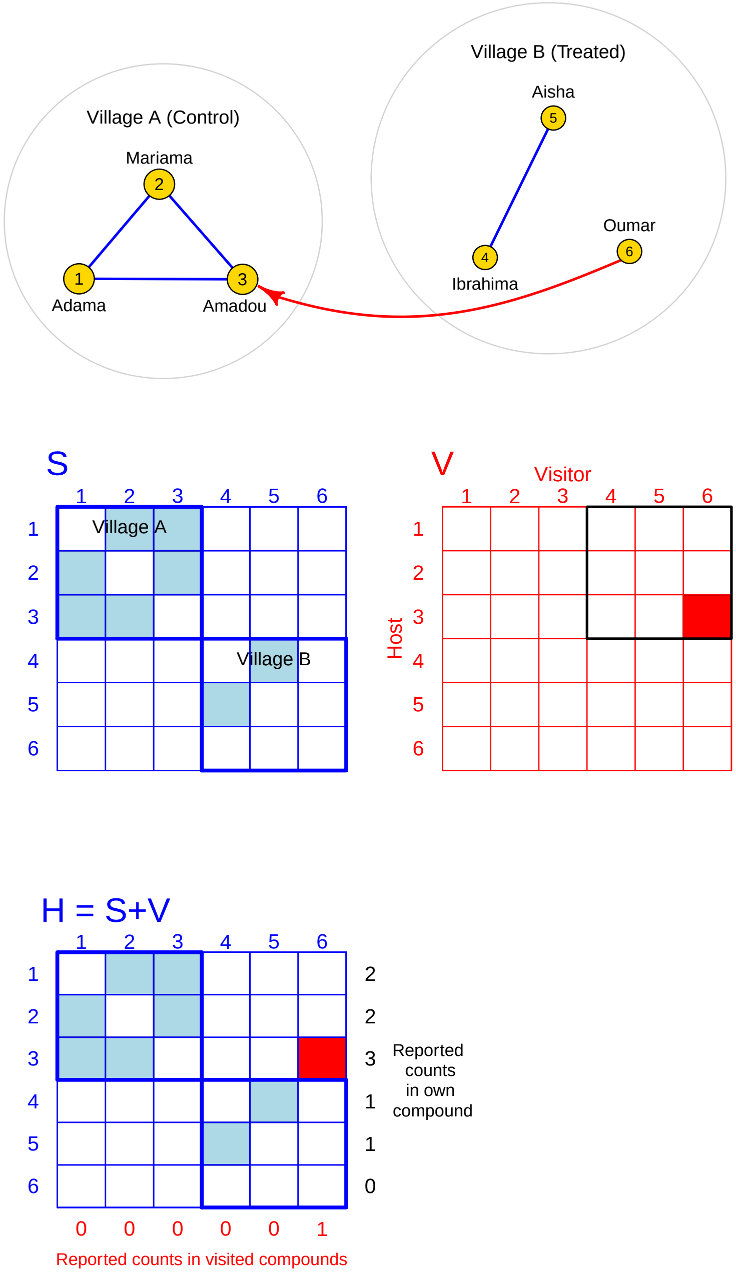
$$\hat{p_j} = \frac{\sum_{i=1}^{n_j} T_i + V_{T,j}}{\sum_{i=1}^{n_j} D_i} = \frac{\sum_{i=1}^{n_j} T_i}{\sum_{i=1}^{n_j} D_i} + \frac{V_{T,j}}{\sum_{i=1}^{n_j} D_i}$$

To understand this, we will walk the reader through a toy example of a network depicted in Figure 1, a diagram similar to that in Potter et al. (2019).

Here, A is a control village and B is a treated village, and the red arrow indicates that Oumar contacted Amadou while visiting Amadou's home in village A. For simplicity, assume all network members are surveyed. The true cross-cluster exposure value for A is 1/7 (noting that each within-village contact is reported twice); it is 1/3 for B. Also for simplicity, our example omits contacts occurring in non-home locations (e.g. market, mosque, etc.), as the proposed adjustment to our estimator does not change how these contribute to the estimates.

The depicted network is not completely observed since respondents did not report the identity of their contacts. We will define adjacency matrices to illustrate how the completely observed network relates to the recorded data. Define S to be an adjacency matrix indicating contacts to members of one's own village, so $S_{ij} = 1$ if i and j made contact and belong to the same village. S is symmetric, since if i contacted j, then j contacted i as well. Let V denote contacts reported while a member of one cluster was visiting a member of a cluster in the opposite treatment arm in the latter's compound. V is asymmetric to distinguish the host from the visitor and to align with the way these contacts were reported, and $V_{3,6} = 1$ since person 6 visited person 3 in the home of person 3. The recorded counts of contacts occurring in the respondent's own compound

Fig. 1. Toy example of a social network with corresponding adjacency matrices.



are the row sums of $H \equiv S + V$. The recorded counts of contacts while the respondent was visiting compounds in villages of the opposite treatment assignment are the column sums of V. Our preliminary approach to estimating interference (without the proposed adjustment) would calculate as follows:

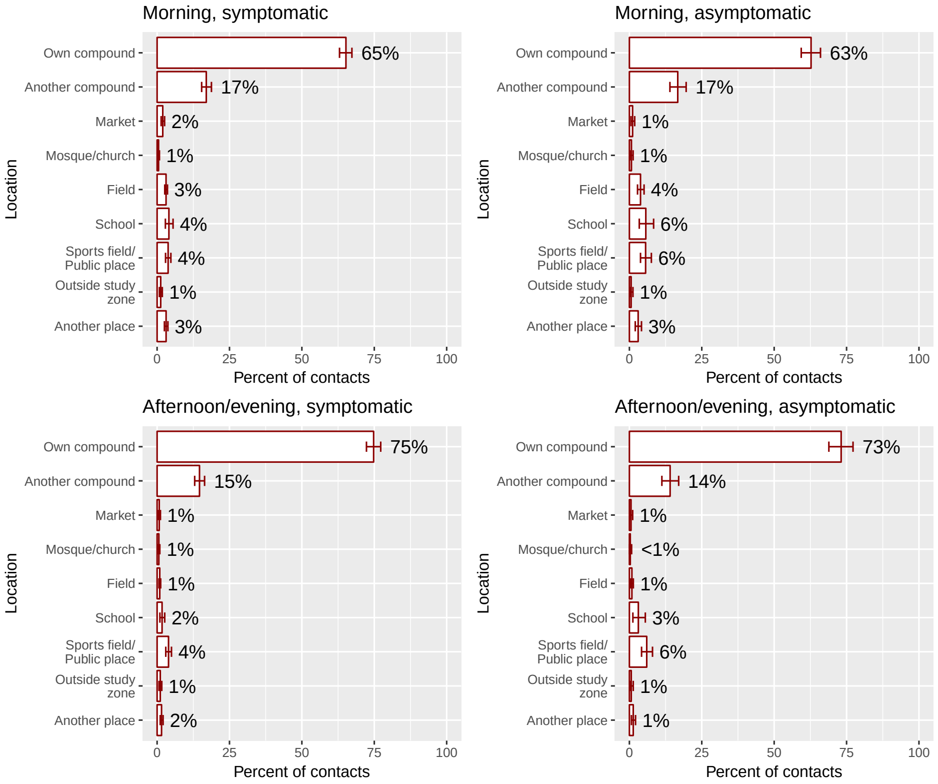
- For village A, the denominator (total number of contacts) is the sum of the row sums of rows 1, 2, 3 of H (total number of contacts reported in the respondent's home) plus the column sums of columns 1, 2, and 3 of V (total number of contacts while the respondent visited a home in a cluster of the opposite assignment), so the denominator is 7. The numerator is the sum of column sums of columns 1, 2, 3 of V, so the numerator is zero. Our cross-cluster exposure estimate is 0/7.
- An analogous approach for village B yields a cross-cluster exposure estimate of 1/3.

Our cross-cluster exposure estimate for A is incorrect since it does not account for the contamination while Oumar was visiting Amadou since it occurred in Amadou's home.

Our proposed adjustment is to subtract from the numerator of A contacts reported by members of B and occurring in compounds within cluster A. These comprise the upper right quadrant of matrix V, shown in black, whose sum is 1. Thus our adjusted estimate for cross-cluster contamination for cluster A is 1/7, the correct value. A similar adjustment for B involves the lower left quadrant of V; whose sum is zero, so the estimate for B (which was already accurate) remains the same.

We have demonstrated the reasoning for our update to the estimator assuming that all network members were surveyed. When network members are randomly sampled, the rows of S and columns of V are sampled randomly. We performed two simulation studies indicating that our proposed estimator is unbiased under random sampling. First, we randomly sampled four people (two from A and two from B) from the toy example and calculated our estimator of cross-cluster contamination for village A for one million samples. The true value is $1/7 = 0.14$, and the mean of the estimators was 0.14 and was within 0.002 of the true value. Next, we created a new, larger example, by simulating two villages, each of size 100, with 0.2 density of within-village contacts and 0.1 density of between-village contacts. The estimand of this example was calculated on the complete network (as we did for the toy example) to be 0.34. We then repeatedly sampled 20 people (10 from each village). The mean of estimators for one million samples was 0.34 and was within 0.002 of the estimand. Therefore, we believe our estimator is unbiased under random sampling. Respondents in this study were not randomly sampled, however. Our sampling process favored symptomatic people, who could be less likely to travel, but the data show no evidence for difference in travel patterns based on symptom status, as evidenced by Figure 2. This figure displays the location distribution of contacts by symptom status and time of day based on the multiply imputed data set. Standard errors were calculated by generating 500 nonparametric bootstrap resamples of each imputed data set, pooling across the imputed data sets, and then calculating the 2.5% and 97.5% quantiles for each location proportion (Schomaker and Heumann, 2018).

Fig. 2. Location distribution of contacts by symptom status and time interval.
Note: This figure has been published in (Potter et al., 2019) and is reproduced with permission of the authors.



6. R code

```
library(dplyr)
library(xtable)

#Function to compute change in cumulative incidence
#due to treatment over some number of years
cumulInc <- function(Y, inf.time, years, cens){
  Xt = solve(t(Y)%*%Y, t(Y))
  if(length(unique(time)) < length(time)) time[cens==0] <-  time[cens==0] +
runif(length(time[cens==0]),0,1)
  Xt = Xt[,inf.time<(years*365) & cens == 0]
  At = matrix(apply(Xt, 1, sum), nrow = 1)
  vcovmat = matrix(0, ncol = dim(Xt)[1], nrow = dim(Xt)[1])
  for(v in 1:dim(Xt)[2]){
    if(sum(is.na(Xt[,v]))==0)
      vcovmat = vcovmat + Xt[,v]%*%t(Xt[,v])
  }
  return(list(coefs = At, vcovmat = vcovmat))
}

#Function to update model formula with covariate terms
#including const() wrapper for time-invariance
update_form_const <- function(X, formula){
  if(is.null(dim(X))){
    return(update.formula(formula, .~. + const(X)))
  }else{
    if(is.null(names(X))){
      dimnames(X)[[2]] = paste0("X", 1:dim(X)[2])
      X = data.frame(X)
    }
    formTerms = terms(formula)
    modelTerms <- c(attr(formTerms, "term.labels"), paste0("const(", names(X), ")"))
    return(reformulate(modelTerms, response = attr(formTerms, "variables")[[2]]))
  }
}

## Function to fit additive hazards model
## returns results for randomized treatment effect
## and overall treatment effect.
fit_addHaz <- function(time, #time of event or censoring
                       cens, #indicator for censoring (1 = censored, 0 = event)
                       trt,  #randomized treatment assignment
                       X = NULL,   #matrix of additional covariates.
                       mix.pct, #percent of contacts to treated clusters
                       clust, #cluster membership
```

```
                     years, #follow-up time (in years)
                     max.time=NULL, # end of follow-up for analysis
                     plot.trt = TRUE){  #Boolean: plot time-varying coefs for treatment?

require(survival)
require(timereg)

# Resolve tied event times by adding draw from a uniform(0,1) distribution
event.times = time[which(cens==0)]

ties = which(time %in% event.times[duplicated(event.times)] & cens==0)

# If there are tied events on last day of follow-up, need to add 1 to censoring times
# on last day of follow-up
if(max(time)==max(time[ties]))
  time[which(cens==1 & time==max(time))] = time[which(cens==1 & time==max(time))]+1

if(length(unique(event.times)) < length(event.times))
  time[ties] =  time[ties] + runif(length(time[ties]),0,1)

#set up survival data
surv.data = Surv(time, 1-cens)

#get randomized treatment effect estimate and SE
if(is.null(X)){
  form = surv.data ~ trt + cluster(clust)
}else{
  form = update_form_const(X, surv.data ~ trt + cluster(clust))
}
surv.fit = aalen(form, max.time=max.time, covariance=TRUE)

#get overall treatment effect estimate and SE
Z = mix.pct
if(is.null(X)){
  form = surv.data ~ Z + cluster(clust)
}else{
  form = update_form_const(X, surv.data ~ Z + cluster(clust))
}
surv.fit.adj = aalen(form, max.time=max.time, covariance=TRUE)

if(plot.trt){
  par(mfrow = c(1,2), mgp = c(2, 0.5, 0))
  plot(surv.fit$cum[,1], surv.fit$cum[,3], type = "l", main = "Randomized Effect")
  plot(surv.fit.adj$cum[,1], surv.fit.adj$cum[,3], type = "l", main = "Overall Effect")
}
```

```
  #get change in cumulative incidence under randomized effect
  id = which(time == max(time[time <= years*365]))

  id = which (surv.fit$cum[,1]==max(surv.fit$cum[,1]))[1]

  cumIncTrt = surv.fit$cum[id, 3]
  sd_CIT = surv.fit$robvar.cum[id, 3]

  #get change in cumulative incidence under overall effect
  cumIncMix = surv.fit.adj$cum[id, 3]
  sd_CIM = surv.fit.adj$robvar.cum[id, 3]

  return(list(rand_fit = surv.fit,
              overall_fit = surv.fit.adj,
              cumIncTrt = cumIncTrt, sd_CIT = sd_CIT,
              cumIncMix = cumIncMix, sd_CIM = sd_CIM))
}

# To get variance of incidence difference for covariates at values other than 0
# x1 (and x2 if wanting different values for treated (x1) and control (x2))
# should be a vector of the same length as the covariates used in the model
incDiffX <- function(inc.res, x1, x2 = NULL){
  if(is.null(x2)) x2 <- x1

  sigma = inc.res$vcovmat
  coef1 = matrix(c(1,1,x1), nrow = 1)
  coef2 = matrix(c(1,0,x2), nrow = 1)

  sd_X <- sqrt(coef1%*%sigma%*%t(coef1)+coef2%*%sigma%*%t(coef2))
  return(sd_X)
}




est_ci = function (mod, digits=2){

  est=mod$cumIncMix
  se=sqrt(mod$sd_CIM)
  cl_mix = est - 1.96*se
  cu_mix = est + 1.96*se
  ci_mix = paste("[",round(100*cl_mix,digits), ', ', round(100*cu_mix,digits), ']',
sep='')
```

```
  estr=mod$cumIncTrt
  ser=sqrt(mod$sd_CIT)
  cl_rand = estr - 1.96*ser
  cu_rand = estr + 1.96*ser
  ci_rand = paste("[",round(100*cl_rand,digits), ’, ’, round(100*cu_rand,digits), ’]’,
   sep=’’)
  return(c(round(100*est,digits), ci_mix, round(100*estr,digits), ci_rand))
}

dat=read.csv(’analysis_dataset.csv’)

y1=dat %>% filter(!is.na(tte_year1))
set.seed(6)

mod=fit_addHaz(time=y1$tte_year1, #time of event or censoring
               cens=1-y1$infected_year1, #indicator for censoring (
               trt=factor(y1$treatment1),  # treatment assignment
               X = NULL,   #matrix of additional covariates.
               mix.pct=y1$pct1, #percent of contacts to treated clusters
               clust=y1$village1, #cluster membership
               years=1, #follow-up time (in years)
               plot.trt = FALSE)

tab=matrix(nrow=4,ncol=4)
tab[1,] = est_ci(mod)


tab

## Model results, Year 1, excluding H1N1 2009 infections


y1seas=dat %>% filter(!is.na(tte_year1_seasonal))

mod=fit_addHaz(time=y1seas$tte_year1_seasonal,
               cens=1-y1seas$infected_year1_seasonal,
               trt=factor(y1seas$treatment1),
               X = NULL,
               mix.pct=y1seas$pct1,
               clust=y1seas$village1,
               years=1,
               plot.trt = FALSE)

tab[2,]=est_ci(mod)
```

```
## Model results, Year 2

y2=dat %>% filter(!is.na(tte_year2))

mod=fit_addHaz(time=y2$tte_year2,
               cens=1-y2$infected_year2,
               trt=factor(y2$treatment2),
               X = NULL,
               mix.pct=y2$pct2,
               clust=y2$village2,
               years=1,
               plot.trt = FALSE)

tab[4,]=est_ci(mod)

## Model results, Year 2, censored before the strike:

BEGIN = as.Date("7/15/2010", "%m/%d/%Y") # Beginning of follow-up for this year
strike.begin = as.Date("1/1/2011", "%m/%d/%Y")
max.time = strike.begin - BEGIN

mod=fit_addHaz(time=y2$tte_year2,
               cens=1-y2$infected_year2,
               trt=factor(y2$treatment2),
               X = NULL,
               mix.pct=y2$pct2,
               clust=y2$village2,
               max.time=max.time,
               years=1,
               plot.trt = FALSE)

tab[3,]=est_ci(mod)


rownames(tab) = c("Year 1, all infections", "Year 1, excluding A/H1N1pdm09",
"Year 2, censored", "Year 2, all")
colnames(tab) = c("Estimate", "95% C.I.", "Estimate", "95% C.I.")
xtable(tab, rownames=FALSE)
tab
```

7. Data Cleaning

Village of residence was recorded during quarterly censuses of the Niakhar population by the Niahkar Demographic Surveillance System. Delaunay et al. (2002) If participants moved during the trial, their departure date, arrival date, and village of their new residence were recorded. Those who moved a second time had their departure date (but not residence after second move) recorded as well. The cleaning process for inconsistencies in the recorded movement data is described below:

(a) In 8 cases, the departure and arrival dates of the second move were earlier than those of the first move. For these cases, the information for second and first moves was swapped.

(b) In 46 cases, the arrival date of the second move was earlier than the arrival and departure dates of the first move, and the departure date of the second move was missing. For these cases, the information for second and first moves was swapped. After the swap, the (missing) departure date for the first move was imputed to be the arrival date of the second.

(c) In 13 cases where the departure date of the first move was missing, it was imputed to be the arrival date of the second move.

(d) In 83 cases where the arrival date of the second move was earlier than the departure date of the first, the departure date of the first was recoded to equal the arrival date of the second.

(e) In 13 cases where the departure date of the first move was earlier than the arrival date of the first move, and the arrival date of the second move was non-missing, the departure date of the first move was recorded to be the arrival date of the second.

(f) After these updates were made, there were 13 cases that did not have arrival and departure dates in sequential order (i.e., arrival 1 $\leq$ departure 1 $\leq$ arrival 2 $\leq$ departure 2); these were excluded from analysis.

(g) The movement data was recorded by storing the village, arrival date, and departure date, of the first "stay" and the second "stay", as well as an overall "village" variable. In over 99% of cases, the village variable matched that of the first stay. However, there were 168 participants for whom the overall village variable differed from that of the first stay. This is because movement data were recorded differently for this small subset of the data. For them "village" indicated the village of residence prior to the first stay rather than the village of the first stay. As such, these cases had up to three distinct residence stays recorded, which differs from the rest of the data which only had up to two distinct stays recorded. These cases were re-coded to be consistent with the rest of the data by transferring the village information (which actually describes the first distinct stay) to the variables for the first stay (so that village and village.stay.1 are consistent), transferring the information recorded for the first stay (actually the second stay) to the variables for the second, and removing information for the second (actually third) stay, as follows:

(i) village.stay.1 was re-coded to village

(ii) arrival.date.stay.1 was re-coded to birth.date

(iii) departure.date.stay.1 was re-coded to arrival.date.stay.1

(iv) village.stay.2 was re-coded to village.stay.1

(v) arrival.date.stay.2 was re-coded to the original arrival.date.stay.1 (the updated departure.date.stay.1)

(vi) departure.date.stay.2 was re-coded to departure.date.stay.1

References

Aalen, O. O., 1989. A linear regression model for the analysis of life times. Statistics in Medicine 8, 907–925.

Carnegie, N. B., Wang, R., De Gruttola, V., 2016. Estimation of the overall treatment effect in the presence of interference in cluster-randomized trials of infectious disease prevention. Epidemiologic Methods 5 (1), 57–68.

Delaunay, V., Marra, A., Levi, P., Etard, J.-F., 2002. Niakhar DSS, Senegal. INDEPTH Network. Populations and health in developing countries 1, 279–285.

Diallo, A., Diop, O. M., Diop, D., Niang, M. N., Sugimoto, J. D., Ortiz, J. R., Diarra, B., Goudiaby, D., Lewis, K. D., Emery, S. L., et al., 2019. Effectiveness of seasonal influenza vaccination in children in Senegal during a year of vaccine mismatch: A cluster-randomized trial. Clinical Infectious Diseases.

Potter, G. E., Wong, J., Sugimoto, J., Diallo, A., Victor, J. C., Neuzil, K., Halloran, M. E., 2019. Networks of face-to-face social contacts in Niakhar, Senegal. PLoS One 14 (8), e0220443.

Schomaker, M., Heumann, C., 2018. Bootstrap inference when using multiple imputation. Statistics in medicine 37 (14), 2252–2266.

Seaman, S. R., White, I. R., 2013. Review of inverse probability weighting for dealing with missing data. Statistical methods in medical research 22 (3), 278–295.