Published in final edited form as:
Stat Med. 2018 January 30; 37(2): 294–301. doi:10.1002/sim.7392.

Estimating Population Effects of Vaccination using Large, Routinely Collected Data

M. Elizabeth Halloran[1,2] and **Michael G. Hudgens**[3]

[1]Center for Inference and Dynamics of Infectious Diseases, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center

[2]Department of Biostatistics, School of Public Health, University of Washington

[3]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

Abstract

Vaccination in populations can have several kinds of effects. Establishing that vaccination produces population-level effects beyond the direct effects in the vaccinated individuals can have important consequences for public health policy. Formal methods have been developed for study designs and analysis that can estimate the different effects of vaccination. However, implementing field studies to evaluate the different effects of vaccination can be expensive, of limited generalizability, or unethical. It would be advantageous to use routinely collected data to estimate the different effects of vaccination. We consider how different types of data are needed to estimate different effects of vaccination. The examples include rotavirus vaccination of young children, influenza vaccination of elderly adults, and a targeted influenza vaccination campaign in schools. Directions for future research are discussed.

Keywords

causal inference; dependent happenings; herd immunity; indirect effects; potential outcome; surveillance; vaccination

1 Introduction

Vaccination against an infectious disease in populations can have several kinds of effects. Vaccination in a population can increase the herd immunity, that is the collective immunological status of a population, providing indirect protection for the unvaccinated individuals and also enhancing the protection provided to the vaccinated individuals. In causal inference, interference is present when the treatment status of one individual can affect the potential outcomes of another individual [1], as with many vaccines.

Establishing that vaccination provides population-level effects that go beyond the direct effects in the vaccinated can have important consequences for public health policy. Formal methods have been developed for study designs and analysis that can estimate the different effects of vaccination. However, implementing formal studies in the field to evaluate the different effects of vaccination can be expensive, unethical, and/or of limited generalizability. It would be advantageous to use routinely available surveillance data and other sources of routinely collected data to estimate the different effects of vaccination.

In this paper we examine estimating different effects of vaccination from routinely collected data. The structure of the paper is the following. In the next section different effects of vaccination are defined, and some key results for estimating such effects in the causal inference framework are presented. We discuss the data requirements that allow estimating (i.e., identifiability of) the different kinds of effects. In Section 3, we present examples of evaluating different effects of vaccination from large administrative and surveillance data sets. The examples include rotavirus vaccination of young children, influenza vaccination of elderly adults, and a targeted influenza vaccination campaign in the schools. Finally, we discuss some directions for future research.

2 Vaccine Effects of Interest

2.1 Direct, indirect, total and overall effects

Halloran and Struchiner [2, 3] defined direct, indirect, total and overall effects of vaccination in the presence of interference [1]. Consider a cluster or group of individuals under two scenarios. Under the first scenario, a certain portion of individuals in the cluster is vaccinated and the rest remains unvaccinated. Under the second scenario, no one in the cluster is vaccinated. The *direct* effect of vaccination is defined by comparing in the first scenario (i) the average potential outcome when an individual receives vaccine with (ii) the average potential outcome when an individual does not receive vaccine. The *indirect* effect is defined as a contrast between (i) the average potential outcome when an individual does not receive vaccine in the first scenario and (ii) the average potential outcome when an individual does not receive vaccine in the second scenario. The *total* effect is defined as a contrast between (i) the average potential outcome when an individual receives vaccine in the first scenario and (ii) the average potential outcome when an individual does not receive vaccine in the second scenario. The *overall* effect is defined by a contrast between (i) the average potential outcome in the entire cluster under the first scenario and (ii) the average potential outcome of the entire cluster under the second scenario. Analogous indirect, total and overall effects can also be defined as contrasts across subsets within clusters, such as particular age groups. The contrasts can be on the relative risk, risk difference, or odds ratio scale. Vaccine efficacy and effectiveness are often defined on the one minus the relative risk scale [4].

2.2 Causal Estimands

A principled approach to assessing the effect of a treatment (such as a vaccine) or an exposure entails using causal inference methods. The potential outcome approach to causal inference generally assumes no interference between individuals [1]. Then if there is one treatment and a control, an individual has two potential outcomes before receiving treatment or control. The individual causal effect is the difference between two potential outcomes under treatment and control. The population average causal effect is the average of the difference between the outcomes if everyone received treatment and if everyone received control. Under an independent assignment mechanism, such as randomization, one can construct an unbiased estimator of the population average causal effect from the observed average outcomes in the two treatment groups.

Assuming interference cannot occur between individuals in different groups but allowing for the possibility there is interference within groups, i.e., there is *partial inteference* [5], Hudgens and Halloran [6] defined causal estimands for direct, indirect, total, and overall effects in the potential outcomes framework. Under partial interference, the potential outcomes for any individual may depend on the treatment assignments to other individuals in the group [3, 7, 8]. The four vaccine effects are then defined according to contrasts in average potential outcomes associated with a particular group allocation strategy and individual treatment assignment. In particular, let the allocation strategies be different levels of vaccine coverage, denoted by α and α'. For example, α may indicate that 60% of the individuals are vaccinated. The individual average potential outcome for individual vaccine assignment a, where $a = 0, 1$ denotes unvaccinated and vaccinated, and group-level vaccination coverage α is defined as the average potential outcome over all possible randomization assignments within the group. A marginal individual potential outcome at a given coverage level α is defined similarly, whereby the marginal average potential outcome does not include the assignment to vaccine or not. Group average potential outcomes are obtained by averaging over the individual average potential outcomes within a group. Population average potential outcomes of individuals with individual vaccine assignment a and group-level vaccination coverage α are defined as the average over groups. Let $\bar{y}(a, \alpha)$ be the population average potential outcome if individuals receive individual assignment a at coverage level α. Analogous group-level and population-level marginal potential outcomes are defined, with $\bar{y}(\alpha)$ denoting the marginal population average potential outcome.

The direct effect of vaccination corresponds to a contrast at a particular coverage level between the population average potential outcomes when individuals receive vaccine and when individuals do not receive vaccine. On the difference scale, the population average direct effect when vaccine coverage is α is defined as $\overline{\mathrm{DE}}(\alpha)=\bar{y}(0;\alpha)-\bar{y}(1;\alpha)$.

Indirect, total, and overall effect estimands are defined as contrasts under two different coverage levels. Indirect effects are defined as the difference in population average potential outcome when an individual is not vaccinated at two different levels of vaccine coverage. The population average indirect effect under coverages α and α' are defined as $\overline{\mathrm{IE}}(\alpha,\alpha')=\bar{y}(0;\alpha)-\bar{y}(0;\alpha')$. The population average total effect are defined as $\overline{\mathrm{TE}}(\alpha,\alpha')=\bar{y}(0;\alpha)-\bar{y}(1;\alpha')$. The overall effect of vaccination is defined as the marginal population average potential outcome under one group allocation strategy compared to another allocation strategy, defined by $\overline{\mathrm{OE}}(\alpha,\alpha')=\bar{y}(\alpha)-\bar{y}(\alpha')$. These causal estimands can also be defined on the relative risk, odds ratio, or one minus relative risk, i.e., the VE scale.

2.3 Estimators

To draw inference about the causal estimands described above, a two-stage randomized experiment can be employed [6, 9, 10, 11]. At the first stage, groups would be randomized to receive certain allocation strategies. For example, one might want to compare effects at 60% and 30% coverage with a vaccine. Some groups would be randomized to 60% coverage, the others to 30% coverage. At the second stage, individuals within groups would be randomized to receive vaccine or control. The coverage level assigned to each group would determine the probability that an individual would be randomized to vaccine or control. Inferential methods for the direct, indirect, total and overall effects for a two-stage randomized experiment have been developed, e.g., see [6, 9, 12, 13, 14].

Most studies are not randomized at two stages, but only at the individual level, the cluster level, or neither in which case the estimators described above would in general be biased or inconsistent. In the absence of randomization at the group and/or individual level, Tchetgen Tchetgen and VanderWeele [12] proposed estimators based on a generalized group-level propensity score [15], that is, the probability a *group* of individuals receives a particular *vaccination vector*. They used inverse proportional weighting (IPW) of the observed individual responses to obtain estimators. When the group-level propensity score is known, they proved the IPW estimators are unbiased under the assumptions of conditional independence and positivity. Perez-Heydrich *et al.* [16] used these IPW estimators to estimate the different effects of cholera vaccination in an individually-randomized study in Matlab, Bangladesh. The geographic location of each household was known, so they formed geographic groups using a clustering algorithm.

3 Using routinely collected data to estimate vaccine effects

In the previous section we described methods to estimate direct, indirect, total and overall effects from studies of vaccination. Vaccination with numerous vaccines occurs world-wide, and drawing inference about indirect effects of vaccination can have important public health policy implications. It would be desirable to use routinely collected data to estimate the different effects of vaccination.

Different types of routinely collected data are available for evaluating the effectiveness of vaccination. For example, insurance claims data or electronic medical records from health care service providers typically include data on services received (such as vaccination) and diagnoses (such as infection or disease). These services and diagnoses are often coded according to the *International Classification of Diseases, Ninth or Tenth Revision, Clinical Modification* (ICD-9-CM or ICD-10-CM).

Establishing vaccination status of individuals and vaccine coverage in the population can be difficult. In some cases insurance claims provide information about vaccination status of individuals and coverage levels. Some states and countries have vaccine registries where individual vaccination status can be linked to outcome data. In some settings, the number of doses bought or used or the number of people vaccinated is recorded, but not which individuals are vaccinated. In this case, vaccine coverage can be estimated but the vaccination status of any particular individual is unknown. In the United States, vaccination is funded by a number of different sources that can vary by state, county, employer, and insurance company. Thus vaccination records vary from place to place, and vaccination status of individuals may be difficult to determine.

Because of the observational nature of routinely collected data, covariate information is crucial to adjust for possible confounding or selection bias. Such covariates might include socio-economic status, urban versus rural environment, age, gender, and general level of health. Analyses to estimate the different effects of routine immunization often use information from more than one database. These databases may have different individuals, aggregate by different geographic or temporal units, or differ in completeness of reporting over time.

3.1 Direct, indirect, total, and overall effectiveness of rotavirus vaccination

Panozzo *et al.* [17] estimated the direct, indirect, total, and overall effectiveness of rotavirus vaccines in preventing gastroenteritis hospitalization in privately insured children in the US. The data were from the MarketScan Research Databases (Truven Health Analytics, Inc., Ann Arbor Michigan). In 2010, the database included about 920,000 infants, representing about 25% of the US birth cohort and 50% of the US birth cohort with commercial insurance. Panozzo *et al.* [17] utilized the MarketScan database to extract individual level information on rotavirus vaccination status, whether an individual was hospitalized for rotavirus infection or acute gastroenteritis, and covariates. On the basis of these data, Panozzo *et al.* [17] then estimated the direct, indirect, total, and overall effects of rotavirus vaccination.

Two rotavirus vaccines for infants are licensed in the US, one since February 2006, and one since April 2008. Data on infants with a live birth recorded between May 1, 2000, and April 30, 2005, (pre-vaccine era) or May 1, 2006 and April 30, 2010, (post-vaccine introduction) were extracted from the database to form an analytical cohort. Information on rotavirus gastroenteritis was extracted for infants aged 8 months up to a maximum age of 20 months. Outcomes for rotavirus gastroenteritis and acute gastroenteritis were identified using the appropriate ICD-9-CM codes. Vaccination status was determined using appropriate Current Procedural Terminology codes. Infants vaccinated after 8 months of age were excluded from the analysis. To increase the sensitivity of the vaccination status determination, infants living in any of the 13 states with state-funded rotavirus immunization programs were excluded. Infants funded by such a program would not have vaccination recorded in the private insurance database, so there would be many infants classified as unvaccinated who were in fact vaccinated. Only infants who had also received at least one dose of diphtheria, tetanus, and acellular pertussis vaccine were included, because children who failed to receive vaccines that are usually administered may differ from those who had received them.

The analysis accounted for household-level variation in rotavirus vaccine coverage, disease, and mixing behaviors by examining the number of other dependent children less than 10 years of age covered by the same insurance holder as the infant. Geographical variation was accounted for by including the region and rurality of the child’s residence as defined by the US Department of Agriculture, Economic Research Service, available on the web. The percentage of infants who had overnight hospital stays unrelated to acute gastroenteritis prior to two months of age was compared in the vaccinated and unvaccinated infants, also in the pre-vaccine era to characterize general infant health and potential differences in susceptibility to rotavirus disease.

Estimates of the vaccine effects were based on Cox proportional hazards models of the rate of rotavirus gastroenteritis or acute gastroenteritis hospitalizations. Panozzo *et al.* [17] estimated the hazard ratios in infants entering the cohort in 2007, 2008, 2009, and 2010, to obtain VE estimates by calendar year. The direct effects for each calendar year were estimated by comparing outcomes in the vaccinated and unvaccinated infants in each year from 2007 to 2010. This is analogous to estimating the direct effects at four levels of coverage. To estimate indirect, total, and overall effectiveness during each calendar year, comparisons were made to the unvaccinated infants in the pre-vaccine baseline period 2001–2005. For example, indirect VE in 2007 was estimated by comparing unvaccinated infants in the analytical cohort in 2007 with the unvaccinated infants in the baseline period 2001–2005.

After exclusions, the final analytical cohort had 905,718 children. Of those, 277,900 children were born during the prevaccine baseline period, and 627,818 were born during the rotavirus vaccine period, 476,576 of whom received rotavirus vaccine and 151,242 did not. Vaccination coverage ranged from 51% in 2007 to 86% in 2010.

Despite over 900,000 children in the analytical cohort, the annual number of events of rotavirus gastroenteritis hospitalizations during the vaccine era per vaccinated or unvaccinated cohort was relatively small, ranging from 3 to 114 per year, with 722 events in the 277,900 unvaccinated children in the pre-vaccine era 2001–2005. Direct effectiveness estimates were 87% to 92% in the four years, with 95% confidence interval lower limits in 2008 and later above 75%. Indirect effectiveness estimates ranged from 14 % (95% CI – 14, 36) in 2007 shortly after introducing rotavirus vaccination to 82% (95% CI 70, 90%) in 2010. Total effectiveness estimates ranged from 91% (95% CI 73, 97%) in 2007 to 98% (95% CI 96, 99%) in 2010. The overall effectiveness estimates ranged from 40% in 2007 to 96% in 2010. It was possible to estimate all four effects because individual level data on outcomes and vaccination status and a baseline pre-vaccine comparison group were available.

3.2 Direct effectiveness of influenza vaccination

Of particular interest and controversy has been the effectiveness of influenza vaccination in the elderly population. Kwong *et al.* [18] estimated the direct effectiveness of influenza vaccination against laboratory-confirmed influenza hospitalizations among community-dwelling elderly adults aged >65 years during the 2010–2011 influenza season using a test-negative design. The study was done in Ontario, Canada, where different databases can be linked through personal identifiers to provide the necessary individual-level data.

The test-negative design is a popular observational study design for estimating direct effectiveness of influenza vaccination [19]. In this approach, individuals presenting at a clinic or hospital with influenza-like illness are tested for influenza viruses. Those testing positive are defined as cases, and those testing negative are defined as non-cases. The vaccination status and potential confounders are ascertained for each individual. An advantage of the approach is that it uses laboratory-confirmed influenza rather than non-specific influenza-like illness as the outcome. The approach is also used widely to estimate the direct effectiveness of rotavirus vaccine [20].

However, issues regarding the test-negative design have been raised. Although the approach is assumed to control for bias due to health seeking behavior compared to using usual population-based controls, the potential for selection bias still remains [19, 21]. Direct vaccine effectiveness is typically estimated in the test-negative design using a logistic regression model to adjust for potential confounders. Thus, VE is calculated by 1 minus the estimated adjusted odds ratio. However, the odds ratio is not collapsible. That is the conditional causal odds ratio will not in general equal the marginal causal odds ratio. Thus interpretation of the VE estimates across such test-negative design studies will generally depend on which confounders are included in the regression model [21, 22].

Despite such reservations, the test-negative design has the advantage of using data that is collected routinely, so it is inexpensive and easy to conduct. It does require that individual-level data on outcome, vaccination status and covariates be available.

In [18], results of respiratory specimens tested for influenza by the Public Health Ontario (PHO) Laboratories were linked to population-based provincial health administrative data, with a nearly 98% linkage success rate. Respiratory samples were submitted to the PHO laboratories for testing for respiratory viruses from the Ontario healthcare system as part of routine clinical care and by public health departments as part of outbreak investigations. The hospitalization data were obtained from the Canadian Institute of Health Information’s Discharge Abstract Database (CIHI-DAD).

Information on receipt of influenza vaccine during the 2010–2011 season was obtained from the Ontario Health Insurance Plan (OHIP) database using physician billing claims for influenza vaccine. About 75% of elderly adults in Ontario received influenza vaccine through physicians that submitted claims to OHIP.

Information on demographic covariates was obtained from the Ontario Registered Persons Database (RPDB), which contains data on everyone with a valid Ontario health card. Covariates obtained were age, sex, rural residence indicator (communities with <10,000 residents), and neighborhood income quintile. The number of hospitalizations in the past three years, outpatient visits in the past year, and prescription medications in the past year were obtained from the CIHI-DAD, the OHIP, and the Ontario Drug Benefit (ODB) databases. Co-morbidities that might increase the risk of influenza, such as heart disease, diabetes, cancers, among others in the past three years were also obtained from the CIHI-DAD and OHIP databases.

The analysis included 569 individuals who tested positive for influenza, of whom 238 were vaccinated against influenza, and 1661 individuals who tested negative for influenza, of whom 934 were vaccinated. The crude estimate of direct vaccine effectiveness was 44% (95% CI 32, 54%). After multivariable adjustment it was essentially unchanged at 42% (95% CI 29, 53%). Numerous subgroup analyses revealed effectiveness varied by influenza type/subtype, but was similar across age groups and sex. Individual-level data were available, as in the rotavirus vaccination study [17] in the previous section, so the direct effect of vaccination could be estimated. Kwong *et al.* [18] did not estimate the indirect, total, and overall effects of vaccination, as there was no obvious unvaccinated comparator group as in [17]. However, they could have estimated the indirect, total, and overall effects at different levels of coverage based on spatially-defined groups as in [16] if the individual information was geocoded.

3.3 Overall effectiveness of an influenza vaccination campaign

Tran *et al.* [23] estimated the community-level effectiveness of a school-based influenza immunization campaign in Alachua County, Florida, compared to routine influenza immunization in surrounding counties using a combination of surveillance data sets. In these databases, individual-level influenza vaccination status cannot be linked to the individual outcomes, limiting the types of effects that can be estimated. The overall effect of the immunization campaign can be estimated because individual vaccination status is not needed. The overall effect in the whole population and the overall effect within age-groups can be estimated. The indirect effects of the campaign in the age groups not included in the school-based immunization campaign can also be estimated. A drawback of this study is that it relied on influenza-like illness that is not laboratory confirmed as well as cases that were confirmed. Under certain assumptions, such measurement error in the outcome can result in biased or inconsistent effect estimates.

In Alachua County, Florida, a major initiative was undertaken to vaccinate schoolchildren in kindergarten through 8th grade with live attenuated influenza vaccine (LAIV) [24]. A pilot program began in 2006, and a comprehensive program was launched at the beginning of the 2009–2010 school year. Local pediatricians and many other community partners supported the program. In the 2010/2011 school year, the program was expanded to include high school students. The program was carried out in the schools and called a school-located influenza vaccination (SLIV) program. Children ineligible to receive LAIV due to contraindications were referred elsewhere to be vaccinated with inactivated influenza vaccine. Weekly influenza and influenza-like illness associated outpatient visits to emergency departments and urgent care centers were reported through Florida’s Electronic Surveillance System for the Early Notification of Community-based Epidemics (FL-ESSENCE). County and age-group specific resident counts were obtained from the 2010 US Census. Influenza vaccination coverage for Alachua County was obtained from the Florida Department of Health’s Florida State Health Online Tracking System (SHOTS) Vaccine Registry. Due to local collaborations, nearly all influenza vaccinations in Alachua County were entered into the Florida SHOTS Vaccine Registry. Comparable vaccination coverage data were not available for other counties, where reporting was not required. Information on individual vaccine status was not available to be linked to the FL-ESSENCE database. Thus, the analysis estimated only the overall effectiveness of the vaccination campaign in Alachua county compared with routine vaccination in the other counties of Florida both by age groups and for all ages combines, and the indirect effectiveness of the campaign compared with routine vaccination in the age groups not eligible for the SLIV campaign. The overall effect estimates were based on contrasts of the estimated attack rates per 100,000 population in each county.

Because health-seeking behavior in Alachua County may differ from the health-seeking behavior in other counties in Florida, an approach using negative controls to adjust for potential bias was employed. Negative controls have been used as a way to detect and adjust for biases due to unmeasured confounding in observational studies [25]. The general idea is that a negative control outcome is subject to the same kinds of unmeasured confounding as the outcome of interest, but is not in the causal pathway of interest and not affected by the exposure of interest or the treatment. In Tran *et al.* [23], a negative control was gastrointestinal illness (GI) rates reported through FL-ESSENCE in Alachua County and the other counties of Florida. The assumption was that individuals with GI might have similar health care seeking behavior as individuals with influenza like illness, but it would not be influenced by the SLIV campaign, so it might be a good negative control. However, the untestable assumption in using negative controls is that the nature of the unmeasured confounding is similar in the treatment and control areas [26].

In this analysis, in the 2012/2013 influenza season, comparing Alachua County to the other 12 counties in the same administratively defined Health Region, the unadjusted estimate of the overall effect was 49% (95% CI 44, 43%). Adjusting for unmeasured confounding using GI illness as a negative control, the overall effect estimate was 32% (95% CI 26, 38%). Comparing Alachua County to all other counties in Florida, the unadjusted estimate of the overall effect was 46% (95% CI 42, 50%). Adjusting for unmeasured confounding using GI illness as a negative control, the overall effect estimate was 42% (95% CI 35, 46%).

Other estimates of the overall effectiveness of the SLIV program were obtained using a different publicly-accessible, aggregated de-identified database [26]. The Florida Agency for Healthcare Administration (AHCA) includes most, if not all, hospital inpatient, emergency department, and ambulatory surgical facilities visits in Florida. The data include information on the facility type, age, sex, and zip code of patient, the ICD-9-CM codes and dollars billed by service type. The individual-level immunization status was still not available. This analysis was based on a log linear model using the county-specific case counts and population data, assuming a Poisson data generating model. In the 2012/2013 influenza season, comparing Alachua County to the surrounding 23 counties based on drawing a symmetric spatially defined region around Alachua County, the unadjusted estimate of the overall effect was 40% (95% CI 39, 41%). Adjusting for unmeasured confounding using GI illness as a negative control, the overall effect estimate was only 18% (95% CI 5, 29%). Comparing Alachua County to the rest of Florida, the unadjusted estimate of the overall effect was 27% (95% CI 25, 29%). Adjusting for unmeasured confounding using GI illness as a negative control, the overall effect estimate was only 5% with a 95% confidence interval covering 0. Thus, the use of negative controls in the two different analyses both resulted in lower overall effectiveness estimates, but it was considerably lower in the second analysis.

4 Discussion

Using routinely collected data to estimate direct, indirect, total, and overall effects of vaccination has great advantages. Such data are relatively inexpensive and available in large quantities on numerous infectious diseases and vaccines. Randomized controlled trials or prospective observational cohort studies can be expensive to conduct, and results may not be generalizable to routine vaccination campaigns. Randomized controlled trials, whether individually randomized or cluster randomized, may be unethical if a vaccine is recommended for routine use in a population. Thus making use of readily available observational data sets is attractive.

To estimate direct effects, total effects, and indirect effects in the vaccine-eligible groups, individual data on clinical outcomes, vaccination status, and potentially relevant covariates of the individuals are needed. The private insurance claims database used by Panozzo *et al.* [17] advantageously included clinical outcomes, vaccination status, and covariates such as age, number of dependent children in same household, and information to characterize general infant health in one database. Environmental variables were drawn from other data sources. The analysis by Kwong *et al.* [18] combined data from different sources, but because of the health system in Ontario, individual information was successfully linked across the data sets. In the evaluation of the SLIV program in Alachua County, individual level information could not be linked across the publicly accessible databases, so the analyses that could be performed were limited to the overall effects in the whole population or by age group, and indirect effects in the age groups not included in the school-based immunization program.

Temporal trends present challenges. In Panozzo *et al.* [17], estimates of the indirect, total, and overall effects were based on a comparison of the pre-vaccine years 2001–2005 with each of the vaccine years 2007, 2008, 2009, 2010. Studies that naively compare disease rates before and after introduction of a vaccination program can have difficulty demonstrating that an observed decrease is due to the intervention. The decrease could have been due to other temporal trends. An interrupted time series analysis could be utilized in such settings to account for temporal trends [27].

Guidelines for REporting studies Conducted using Observational Routinely-Collected health Data (RECORD) have been developed [28]. These guidelines are also applicable to reporting studies estimating direct, indirect, total or overall effects of vaccination from routinely-collected data. However, other issues still remain. In general, extensions of causal inference methods for observational data are need for the setting where there is interference. Some progress has been made in this direction, as in [12, 16], among others, but there have been no methods developed to date on test negative designs with interference, negative controls with interference, or interrupted time series with interference. These methods are needed to provide a theoretically justified analytical framework for drawing inferences about population-level vaccine effects from various data sources. In other fields, researchers have considered combining designed studies with fewer subjects with larger surveillance studies. For example, Chatterjee *et al.* [29] develop inferential methods which incorporate individual level data from an internal study and summary level information from external big data sources. Future research could examine the utility of these methods for combining routinely collected data with data from studies implemented with the intention of estimating direct, indirect, total and overall effects of vaccination.

The different effects of vaccination programs can depend heavily on who gets vaccinated as well as the underlying contact and network structure of a population. Thus, the results of such studies, whether intentionally designed to estimate such effects or using available databases, may not be generalizable or transportable [30] from one population to another.

Study designs and inferential methods are need which are robust to ascertainment bias, underreporting, and other biases inherent in these types of data sources [31]. Having access to disparate data sources may afford opportunities to develop methods which are multiply robust.

Determining that vaccination programs have population effects which can benefit both the unvaccinated and vaccinated individuals can have important policy implications. The large and rich data sources available can be beneficially used to estimate such effects. New methods for analyzing such data sets are needed.

Acknowledgments

This work was partially funded by National Institute of Allergy and Infectious Diseases grants R37 AI032042 and R01 AI085073.

References

1. Cox, DR. Planning of Experiments. John Wiley and Sons, Inc; New York: 1958.
2. Halloran ME, Struchiner CJ. Study designs for dependent happenings. Epidemiology. 1991; 2:331–338. [PubMed: 1742381]
3. Halloran ME, Struchiner CJ. Causal inference for infectious diseases. Epidemiology. 1995; 6:142–151. [PubMed: 7742400]
4. Halloran, ME., Longini, IM., Struchiner, CJ. Design and Analysis of Vaccine Studies. Springer; New York: 2010.
5. Sobel M. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J Am Stat Assoc. 2006; 101:1398–1407.
6. Hudgens MG, Halloran ME. Towards causal inference with interference. J Am Stat Assoc. 2008; 103:832–842. [PubMed: 19081744]
7. Rubin DB. Bayesian inference for causal effects: The role of randomization. Ann Stat. 1978; 7:34–58.
8. Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci. 1990; 5:472–480.
9. Baird, S., Bohren, A., McIntosh, C., Özler, B. PIER Working Paper, 14-032:., 2014. Penn Institute for Economic Research, Department of Economics; University of Pennsylvania: Designing experiments to measure spillover effects.
10. Sinclair B, McConnell M, Green DP. Detecting spillover effects: Design and analysis of multilevel experiments. American Journal of Political Science. 2012; 56(4):1055–1069.
11. Duflo E, Saez E. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. The Quarterly Journal of Economics. 2003; 118(3):815–842.
12. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. Stat Methods Med Res. 2012; 21(1):55–75. [PubMed: 21068053]
13. Rigdon J, Hudgens MG. Randomization inference for treatment effects on a binary outcome. Statistics in Medicine. 2015; 34:924–935. [PubMed: 25471299]
14. Liu L, Hudgens MG. Large sample randomization inference of causal effects in the presence of interference. Journal of the American Statistical Association. 2014; 109:288–301. [PubMed: 24659836]
15. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies. Biometrika. 1983; 70:41–55.
16. Perez-Heydrich C, Hudgens MG, Halloran ME, Clemens JD, Ali M, Emch ME. Assessing effects of cholera vaccination in the presence of interference. Biometrics. 2014; 70(3):731–741. [PubMed: 24845800]

17. Panozzo CA, Becker-Dreps S, Pate V, Weber DJ, Funk MJ, Stürmer T, Brookhart MA. Direct, indirect, total, and overall effectiveness of the rotavirus vaccines for the prevention of gastroenteritis hospitalizations in privately insured US children, 2007–2010. American Journal of Epidemiology. 2014; 179:895–909. [PubMed: 24578359]
18. Kwong JC, Campitelli MA, Gubbay JB, Peci A, Winter A-L, et al. Vaccine effectiveness against laboratory-confirmed influenza hospitalizations in elderly adults during the 2010–2011 season. Clinical Infectious Diseases. 2013; 57:820–827. [PubMed: 23788243]
19. Sullivan SG, Tchetgen Tchetgen EJ, Cowling BJ. Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. American Journal of Epidemiology. 2016; 184:345–353. [PubMed: 27587721]
20. Schwartz LM, Halloran ME, Rowhani-Rahbar A, Neuzil KM, Victor JC. Rotavirus vaccine effectiveness in low-income settings: An evaluation of the test-negative design. Vaccine. 2017; 35:184–190. [PubMed: 27876198]
21. Westreich D, Hudgens MG. Invited commentary: Beware the test-negative design. American Journal of Epidemiology. 2016; 184:354–356. [PubMed: 27587722]
22. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statistical Science. 1999; 14:29–46.
23. Tran CH, Sugimoto JD, Pulliam JRC, Ryan KA, Myers PD, et al. School-located influenza vaccination reduces community risk for influenza and influenza-like illness emergency care visits. PLoS ONE. Dec; 2014 9(12):1–17.
24. Tran CH, McElrath J, Hughes P, Ryan K, Munden J, et al. Implementing a community-supported school-based influenza immunization program. Biosecur Bioterror. 2010; 8:331–341. [PubMed: 21054182]
25. Lipsitch M, Tchetgen Tchetgen EJ, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010; 21:383–388. [PubMed: 20335814]
26. Fisher, LH. PhD thesis. University of Washington: 2016. Modeling of Infectious Disease Surveillance Data.
27. Penfold RB, F Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Ped. 2013; 13:S38–S44.
28. Benchimol EI, Smeet L, Guttmann A, Harron K, et al. The REporting of studies Conducted using Observational Routinely collected Data (RECORD statement). PLoS Medicine. 2015; 12(10):e1001885. [PubMed: 26440803]
29. Chatterjee N, Chen YH, Maas P, Carroll RJ. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. Journal of the American Statistical Association. 2016; 111(513):107–117. [PubMed: 27570323]
30. Pearl J, Bareinboim E. External validity: from *do*-calculus to transportability across populations. Statistical Science. 2014; 29:579–595.
31. Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2016; 179(2): 319–376.