Received: 17 May 2022 | Accepted: 10 July 2023

DOI: 10.1111/biom.13915

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A proportional incidence rate model for aggregated data to study the vaccine effectiveness against COVID-19 hospital and ICU admissions

Ping Yan[1,2] | **Muhammad Abu Shadeque Mullah**[1,3] | **Ashleigh Tuite**[4]

[1]Infectious Disease Programs Branch, Public Health Agency of Canada, Ottawa, Ontario, Canada

[2]Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada

[3]School of Epidemiology and Public Health, University of Ottawa, Ontario, Canada

[4]Dalla Lana School of Public Health, University of Toronto, Ontario, Canada

Correspondence
Ping Yan, Infectious Disease Programs Branch, Public Health Agency of Canada, Ottawa, and Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada.
Email: ping.yan55@gmail.com

Abstract

We develop a proportional incidence model that estimates vaccine effectiveness (VE) at the population level using conditional likelihood for aggregated data. Our model assumes that the population counts of clinical outcomes for an infectious disease arise from a superposition of Poisson processes with different vaccination statuses. The intensity function in the model is calculated as the product of per capita incidence rate and the at-risk population size, both of which are time-dependent. We formulate a log-linear regression model with respect to the relative risk, defined as the ratio between the per capita incidence rates of vaccinated and unvaccinated individuals. In the regression analysis, we treat the baseline incidence rate as a nuisance parameter, similar to the Cox proportional hazard model in survival analysis. We then apply the proposed models and methods to age-stratified weekly counts of COVID-19–related hospital and ICU admissions among adults in Ontario, Canada. The data spanned from 2021 to February 2022, encompassing the Omicron era and the rollout of booster vaccine doses. We also discuss the limitations and confounding effects while advocating for the necessity of more comprehensive and up-to-date individual-level data that document the clinical outcomes and measure potential confounders.

KEYWORDS

aggregated counts, conditional likelihood, incidence rate, relative risk, vaccine effectiveness

1 | INTRODUCTION

During a public health emergency such as the COVID-19 pandemic, when vaccination programs are dynamic (e.g., expanding coverage and administering additional doses), public health policy makers require timely information on vaccine effectiveness (VE) in light of the current vaccine coverage. They may also want to assess counterfactual scenarios, such as what may have happened in terms of hospitalizations and other health outcomes in the absence of vaccination programs (Ogden et al., 2022). These assessments must be ongoing and conducted quickly based on available surveillance data. To meet these needs, we propose statistical methods that leverage

aggregated count data from various public sources, such as routine case monitoring systems, vaccination registries, and demographic statistics.

Typically, VE studies are based on observational cohorts. Both prospective and retrospective longitudinal cohorts utilize detailed individual-level event history data. These data enable the estimation of risk functions by vaccination status, which define the relative risk (RR), and thus VE as one minus the RR, either as discrete probabilities or person-time incidence rates. However, since such data are often costly to collect, many studies rely on extensive linkage of multiple large administrative databases to create pseudo-cohorts.

A systematic review of COVID-19 VE (Teerawattananon et al., 2022) analyzed 42 peer-reviewed studies covering the period up to late 2021. All of these studies involved extensive data linkages using administrative data sets. For example, Lin et al. (2022) studied VE for three vaccine products (Pfizer, Moderna, and Johnson & Johnson) in an observational cohort assembled from the linkage of multiple large administrative databases in North Carolina that captured important information on vaccination events and measured waning of VE. Andrews et al. (2022) compared the effectiveness of two vaccines (Pfizer and AstraZeneca) in England against symptomatic COVID-19, hospitalization, and death. In Canada, Nasreen et al. (2022) estimated the effectiveness of mRNA vaccines (Pfizer, Moderna) and AstraZeneca against COVID-19 hospitalization and deaths using multiprovincial linked databases. Meanwhile, Buchan et al. (2021) studied the effectiveness of mixed vaccine schedules (Pfizer, Moderna, and AstraZeneca) against severe outcomes (hospitalization or death), stratified by circulating variant (Delta or Omicron) and time since last dose in Ontario, Canada, up to December 26, 2021. These linked databases allowed for the creation of observational cohorts with sufficient detail, enabling the use of test-negative case-control designs and logistic regression models, as in Andrews et al. (2022), Nasreen et al. (2022), and Buchan et al. (2021), or Cox regression models, as in Lin et al. (2022).

Individual-level data linkages can be time-consuming and raise privacy concerns. In our study, we use aggregated counts from publicly available administrative registries without the need for data linkage. These counts include COVID-19 hospital and intensive care unit (ICU) admissions, aggregated into small time intervals and stratified by vaccination status, age, and other covariates. They are as timely as the reporting of ongoing disease surveillance and monitoring. To our knowledge, no study has been conducted or published on the methods for assessing VE based on such aggregated registry data. Therefore, we have developed methods and models that differ from the logistic and Cox regressions used in the studies mentioned above.

Assuming that the occurrence of events arise from a Poisson process, we formulated a proportional incidence rate model for aggregated marginal counts. We then applied these methods to analyze weekly incidence counts of hospital and ICU admissions in Ontario between January 4, 2021, and February 20, 2022, in conjunction with information on at-risk populations categorized by age and vaccination status. A thorough discussion of confounding factors and limitations is also provided. Although the aggregate data do not contain the detailed information available in individual-level longitudinal data, such as individual vaccination status, our findings are consistent with those of the cited studies and corroborate evidence from related immunological research.

The organization of this paper is as follows: Section 2 outlines the development of our models and methods. In Section 3, we apply our statistical methods to study the efficacy of COVID-19 vaccines. Finally, we conclude with a discussion of our statistical methods and results in Section 4.

2 | METHODS

2.1 | Specifying the RR based on aggregate data

Suppose that severe outcomes resulting from SARS-CoV-2 infections arise from a counting process $\{Y(t)\}$ with an instantaneous intensity $\lambda(t) = \rho(t)N(t)$, where $\rho(t)$ is the incidence rate per person, and $N(t)$ is the size of the at-risk population consisting of individuals who are susceptible to infection at time t and may progress to being admitted to the hospital or ICU. The at-risk population is stratified by vaccination status.

In simple settings with only two vaccination groups, the unvaccinated ($j = 0$) and the vaccinated ($j = 1$), we denote the size of at-risk population in each group by $N_j(t)$ for $j = 0, 1$. We consider $\{Y(t)\}$ as a superposition of two counting processes, each with intensity functions $\lambda_0(t) = \rho_0(t)N_0(t)$ and $\lambda_1(t) = \rho_1(t)N_1(t)$, respectively. The instantaneous RR is the ratio $r(t) = \rho_1(t)/\rho_0(t)$, and the VE is defined as $VE(t) = 1 - r(t)$.

We consider data that are only available at the aggregate level, with time divided into intervals $I_k = [t_{k-1}, t_k)$, where k ranges from 1 to m. When the length of the intervals is short, it is reasonable to assume that the at-risk population sizes N_{k0} and N_{k1} remain constant during I_k so that $N_{kj} = N_j(t_{k-1})$, for $j = 0, 1$. The number of events during I_k is represented by Y_{kj} with mean values $\mu_{kj} = E[Y_{kj}]$. These mean values are given by $\mu_{k0} = \int_{t_{k-1}}^{t_k} \lambda_0(t)dt =$

$N_{k0}i_0(k)$ and $\mu_{k1} = \int_{t_{k-1}}^{t_k} \lambda_1(t)dt = N_{k1}i_1(k)$, respectively, where $i_0(k) = \int_{t_{k-1}}^{t_k} \rho_0(t)dt$ and $i_1(k) = \int_{t_{k-1}}^{t_k} \rho_1(t)dt$.

In the application of this paper, we use weekly data. Without loss of generality, we designate interval I_k as week k and refer to $i_0(k)$ as the baseline per person weekly incidence rate. The key parameter of interest is the weekly aggregated RR

$$r_k = \frac{i_1(k)}{i_0(k)},\ k = 1, \dots, m. \tag{1}$$

A crude estimate of r_k is given by the ratio

$$r_k^* = \frac{y_{k1}/N_{k1}}{y_{k0}/N_{k0}},\ k = 1, \dots, m, \tag{2}$$

where y_{kj} is the observed realization of the random variable Y_{kj}. This corresponds to the crude estimates $i_1^*(k) = y_{k1}/N_{k1}$ and $i_0^*(k) = y_{k0}/N_{k0}$, with $\mu_{k0}^* = y_{k0}$ and $\mu_{k1}^* = y_{k1}$. However, in some instances, the crude estimates may be undefined due to the lack of baseline count for a week (i.e., when $y_{k0} = 0$). Additionally, when vaccine coverage or incidence is low, both r_k^* and the crude estimates $i_1^*(k)$ and $i_0^*(k)$ tend to fluctuate considerably from week to week. Given that the infectious disease incidence rates are driven by smooth functions reflecting a complex dynamic system, it is desirable to ensure some smoothness in the estimates for $i_1(k)$ and $i_0(k)$. Bearing this in mind, we propose the following likelihood-based methods for estimating the RR.

2.2 | A direct likelihood-based approach to RR estimation

We assume that the counting process arises from a Poisson process so that the weekly data $\{Y_{kj},\ j = 0, 1; k = 1, \dots, m\}$ are Poisson random variables with $\mu_{k0} > 0$ and $\mu_{k1} > 0$. The RRs r_k are specified by a vector of parameters $\boldsymbol{\beta}$ and denoted by $r_k(\boldsymbol{\beta})$ for $k = 1, \dots, m$. The expected values are given by $\mu_{k0} = N_{k0}i_0(k)$ and $\mu_{k1}(\boldsymbol{\beta}) = N_{k1}i_0(k)r_k(\boldsymbol{\beta})$. The log-likelihood function can be arranged as

$$\begin{aligned} l(\boldsymbol{\beta}) = &\sum_{k=1}^{m} [y_{k1} \log r_k(\boldsymbol{\beta}) - y_k \log (N_{k0} + N_{k1}r_k(\boldsymbol{\beta}))] \\ &+ \sum_{k=1}^{m} [y_k \log [i_0(k)(N_{k0} + N_{k1}r_k(\boldsymbol{\beta}))] \\ &- i_0(k)(N_{k0} + N_{k1}r_k(\boldsymbol{\beta}))], \end{aligned} \tag{3}$$

through data partitioning as $\{y_{k1}, y_k = (y_{k0} + y_{k1})\}$.

The direct method for estimating RR does not require modeling the baseline incidence $i_0(k)$, because RRs do not depend on it. In (3), the likelihood function is partitioned in such as way that

(i) the first term is a conditional likelihood based on the conditional distribution of $\{Y_{k1}|Y_k = y_k\}$, which follows a binomial distribution: $Bin(y_k; N_{k1}r_k(\boldsymbol{\beta})/\{N_{k0} + N_{k1}r_k(\boldsymbol{\beta})\})$. The corresponding likelihood function is

$$\begin{aligned} l^*(\boldsymbol{\beta}) = &\sum_{k=1}^{m} [y_{k1} \log r_k(\boldsymbol{\beta}) - (y_{k0} + y_{k1}) \\ &\log (N_{k0} + N_{k1}r_k(\boldsymbol{\beta}))]. \end{aligned} \tag{4}$$

(ii) The second term is the log-likelihood based on the Poisson distribution of $Y_k = Y_{k0} + Y_{k1}$, with a mean value of $i_0(k)(N_{k0} + N_{k1}r_k(\boldsymbol{\beta}))$, which cannot differentiate $i_0(k)$ from $\boldsymbol{\beta}$.

Following the principles outlined in Kalbfleisch and Sprott (1970) and Sprott (1975), the first term is conditionally sufficient for $\boldsymbol{\beta}$ when knowledge of $i_0(k)$ is absent. Thus, this approach eliminates the baseline incidence $i_0(k)$ as a nuisance parameter. The direct estimation of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)$ can be achieved by maximizing the likelihood (4), which is equivalent to solving the unbiased estimating equations based on the score functions

$$\sum_{k=1}^{m} \frac{\partial \mu_{k1}(\boldsymbol{\beta})}{\partial \beta_l} \frac{y_{k1} - \mu_{k1}(\boldsymbol{\beta})}{V_{k1}(\boldsymbol{\beta})} = 0,\ l = 1, \dots, q, \tag{5}$$

where $\mu_{k1}(\boldsymbol{\beta}) = (y_{k0} + y_{k1})N_{k1}r_k(\boldsymbol{\beta})/\{N_{k0} + N_{k1}r_k(\boldsymbol{\beta})\}$ and $V_{k1}(\boldsymbol{\beta}) = (y_{k0} + y_{k1})N_{k0}N_{k1}r_k(\boldsymbol{\beta})/\{N_{k0} + N_{k1}r_k(\boldsymbol{\beta})\}^2$ are the mean and variance of the conditional distribution of $Y_{k1}|Y_k = y_{k0} + y_{k1}$, respectively. Solving (5) for $\beta_l, l = 1, \dots, q$, yields asymptotically *unbiased* estimates for $\boldsymbol{\beta}$ (Godambe, 1980; Godambe & Thompson, 1974). These estimates are also *optimal* as they possess the smallest asymptotic variances (Godambe, 1980).

Upon estimating $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}$, the baseline incidence rate can be estimated by

$$i_0(k|\widehat{\boldsymbol{\beta}}) = \frac{y_{k0} + y_{k1}}{N_{k0} + N_{k1}r_k(\widehat{\boldsymbol{\beta}})},\ k = 1, \dots, m \tag{6}$$

so that

$$\widehat{E}[Y_{k0}] = N_{k0}i_0(k|\widehat{\boldsymbol{\beta}}),\ \widehat{E}[Y_{k1}] = N_{k1}i_0(k|\widehat{\boldsymbol{\beta}})r_k(\widehat{\boldsymbol{\beta}}). \tag{7}$$

The incidence rates for the vaccinated group are estimated by $i_1(k|\hat{\boldsymbol{\beta}}) = i_0(k|\hat{\boldsymbol{\beta}})r_k(\hat{\boldsymbol{\beta}})$. The estimate $i_0(k|\hat{\boldsymbol{\beta}})$ in (6) is referred to as a semiparametric estimate because, although it does not rely on the parameterization of the incidence rate function, it depends on the parametric RR estimate $r_k(\hat{\boldsymbol{\beta}})$. In contrast to the crude estimate r_k^* that could be undefined if $i_0^*(k) = y_{k0}/N_{k0} = 0$ for some k, the $r_k(\hat{\boldsymbol{\beta}})$ are defined for all $k = 1, \ldots, m$. Although data y_{k1} from the vaccinated group do not provide information about $i_0(k)$ without knowledge of the RR, $i_0(k|\hat{\boldsymbol{\beta}})$ in (6) utilizes (y_{k1}, N_{k1}) by assuming that the true RRs r_k are equal to their estimated values $r_k(\hat{\boldsymbol{\beta}})$.

The population incidence in the absence of vaccination is estimated by $(N_{k0} + N_{k1})i_0(k|\hat{\boldsymbol{\beta}})$. The expressions in (7) enable the validation of model predictions for incidence data. Regression residuals for both the vaccinated and unvaccinated groups can be easily obtained (see Web Appendix A) and used to check model adequacy.

2.3 | A joint likelihood approach for estimating the RR and baseline incidence

Although the primary objective is to estimate the RR r_k, joint estimation of r_k and the baseline incidence rate $i_0(k)$ is useful, as when multiplied by the total population size, $i_0(k)$ provides important counterfactual information, such as what would have happened in a population without vaccination. We specify the baseline incidence rate for the unvaccinated population using a vector of parameters $\boldsymbol{\theta}$, represented as $i_0(k;\boldsymbol{\theta})$. By partitioning the data into $\{y_{k0}, y_{k1}\}$, the full log-likelihood function (3) can be rearranged as

$$\begin{aligned} l(\boldsymbol{\theta},\boldsymbol{\beta}) = &\sum_{k=1}^{m} [y_{k0} \log i_0(k;\boldsymbol{\theta}) - N_{k0} i_0(k;\boldsymbol{\theta})] \\ &+ \sum_{k=1}^{m} [y_{k1} \log i_0(k;\boldsymbol{\theta}) - N_{k1} i_0(k;\boldsymbol{\theta}) r_k(\boldsymbol{\beta})] \\ &+ \sum_{k=1}^{m} y_{k1} \log r_k(\boldsymbol{\beta}). \end{aligned} \tag{8}$$

The joint estimation of $(\boldsymbol{\theta}, \boldsymbol{\beta})$ can be achieved by maximizing the full likelihood, as expressed in (8). It is important to note that the direct estimate $r_k(\hat{\boldsymbol{\beta}})$ from (4) and the semiparametric estimate $i_0(k|\hat{\boldsymbol{\beta}})$ found in (6) can be viewed as maximum likelihood estimates (MLEs) in relation to (8) if a saturated model $i_0(k;\boldsymbol{\theta}) = \theta_k$ $(k = 1, \ldots, m)$ is used. However, $i_0(k|\hat{\boldsymbol{\beta}})$ fluctuates from week to week, similar to the crude incidence estimates $i_0^*(k)$, and is only defined up to week m without the capability to extrapolate into short-term forecasts.

Public health researchers often prefer smooth baseline incidence rate estimates that can be extrapolated into short-term forecasts. This requires model specification and parameterization of $i_0(k;\boldsymbol{\theta})$. If $i_0(k;\boldsymbol{\theta})$ is correctly specified, maximizing $l(\boldsymbol{\theta},\boldsymbol{\beta})$ will yield efficient estimates. However, $i_0(k;\boldsymbol{\theta})$ is governed by a complex dynamic system, making it challenging to capture all aspects of the data-generating process. Furthermore, the computational demands are prohibitive.

We propose a two-step approach. In this approach, we model $\log i_0(k;\boldsymbol{\theta})$ as a smooth function of k by using thin plate regression splines as described by Wood (2003). Our two-step approach is based on the likelihood arrangement in Equation (8), and we argue that data from the unvaccinated group $\{y_{k0}, k = 1, \ldots, m\}$ is marginally sufficient for $i_0(k;\boldsymbol{\theta})$, because in (8):

(i) the distributions of $\{y_{k0}, k = 1, \ldots, m\}$, for any specified $\boldsymbol{\theta}$, are jointly ancillary for $\boldsymbol{\beta}$;
(ii) in the absence of knowledge regarding $\boldsymbol{\beta}$, utilizing data solely from the unvaccinated group does not result in the loss of any available information about $\boldsymbol{\theta}$.

These arguments are also following the principles outlined in Kalbfleisch and Sprott (1970) and Sprott (1975).

The first step: Assuming Poisson counts $\{y_{k0}, k = 1, \ldots, m\}$, the first term in (8) represents the marginal likelihood

$$l_0(\boldsymbol{\theta}) = \sum_{k=1}^{m} [y_{k0} \log i_0(k;\boldsymbol{\theta}) - N_{k0} i_0(k;\boldsymbol{\theta})]. \tag{9}$$

Data from the unvaccinated population $\{y_{k0}\}$ are used to estimate the baseline incidence rate $i_0(k;\boldsymbol{\theta})$. We denote the estimated baseline incidence rate as $i_0^{(sp)}(k)$ and refer to it as the penalized spline function in accordance with our model specification.

The second step: We insert $i_0^{(sp)}(k)$ into the full log-likelihood function $l(\boldsymbol{\theta},\boldsymbol{\beta})$ (8) as fully specified. The likelihood function is then reduced to that for $\boldsymbol{\beta}$ only and is based on data from the vaccinated group $\{y_{k1}, k = 1, \ldots, m\}$. The log-likelihood takes the following form:

$$l(\boldsymbol{\beta}|i_0^{(sp)}(k)) = \sum_{k=1}^{m} \left[y_{k1} \log (r_k(\boldsymbol{\beta})) - N_{k1}\, i_0^{(sp)}(k)\, r_k(\boldsymbol{\beta}) \right], \tag{10}$$

which leads to the estimation of $r_k(\boldsymbol{\beta})$ by maximization.

The distributions of Y_{k0} and Y_{k1} may exhibit extra-Poisson variation. In such cases, the Poisson log-likelihoods $l_0(\boldsymbol{\theta})$ (9) and $l(\boldsymbol{\beta})$ (10) can be substituted with those derived from negative-binomial distributions Lawless, J.F. (1987). A detailed explanation of fitting thin plate regression spline models to estimate the baseline incidence rate in the first step is provided in Mullah and Yan (2022) and Wood (2004). The key points are summarized in Web Appendix B.

Concerning bias, the estimation of $r_k(\boldsymbol{\beta})$ in the second step crucially depends on the model specification of $i_0(k;\boldsymbol{\theta})$ in the first step. Misspecification of $i_0(k;\boldsymbol{\theta})$ leads to biased estimates and the amount of bias depends on the goodness-of-fit of the model $i_0(k;\boldsymbol{\theta})$ to data.

Regarding variance estimation, we recommend against deriving it naively from (10), because simply applying score functions and second-order derivatives to (10) would result in an underestimation of the standard errors (SEs). Meanwhile, the estimate $i_0^{(sp)}(k)$, ascertained through the penalized spline method, brings its own uncertainties, necessitating estimation through bootstrapping. Therefore, the bootstrapping method becomes essential for the SE estimation of the RR.

2.4 | Some further comments about the two approaches

Both approaches include the crude estimate (2) as a special case. In the direct approach, if a saturated model $r_k(\boldsymbol{\beta}) = r_k$ is employed as a piecewise constant function defined for each week k, the conditional likelihood (4) is maximized when $r_k = r_k^*$ given by (2). In the joint likelihood approach, the saturated model $i_0(k;\boldsymbol{\theta}) = \theta_k$ $(k = 1, \ldots, m)$ concurrently with $r_k(\boldsymbol{\beta}) = r_k$ also returns the crude estimation of r_k.

If the primary objective is to estimate the RR, we recommend using the direct approach. This method proves effective by utilizing both data sets, namely $\{y_{k0}\}$ and $\{y_{k1}\}$, while also accounting for the inherent randomness present in $\{Y_{k0}\}$. Moreover, it provides an appropriate estimation of the SE.

The thin plate regression splines, as implemented within the two-step approach, offer sufficient flexibility and potential for calibration. Our simulations (detailed in Web Appendix C) demonstrate that well-calibrated penalized splines yield estimated values of $\boldsymbol{\beta}$ that are very close to those obtained through the direct approach. Both estimates approximate the true parameter $\boldsymbol{\beta}$ very well. Furthermore, the confidence intervals (CIs) estimated for $\boldsymbol{\beta}$ in the two-step approach, using Wald-based bootstrapping, align well with those derived from the direct approach. In contrast, when the two-step approach employs naively estimated model-based SEs to obtain CIs for $\boldsymbol{\beta}$, the resulting intervals are relatively narrower, leading to lower coverage probabilities.

Therefore, if the goal is to simultaneously estimate both the baseline incidence and the RR, we suggest calibrating the spline functions in a way that ensures the estimated $\boldsymbol{\beta}$ from the two-step approach aligns well with that obtained from the direct approach. To account for uncertainties associated with the estimated baseline incidence, bootstrapping should be used. Similarly, the uncertainties around the estimated RR can be addressed based on the direct method, with support from bootstrapping.

2.5 | Regression analysis

To assess VE among various groups based on factors like age, preexisting health conditions, or distinct time frames, we consider a covariate vector $\boldsymbol{z}_k' = (z_{1k}, \ldots, z_{qk})$ that could be time-dependent. The RR $r_k(\boldsymbol{\beta})$ is formulated using a log-linear model

$$\log r_k(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{z}_k = \beta_0 + \beta_1 z_{1k} + \cdots + \beta_q z_{qk}. \tag{11}$$

Under the generalized linear model (11), the conditional likelihood (4) in the direct approach takes the form

$$l^*(\boldsymbol{\beta}) = \sum_{k=1}^{m} \left[y_{k1}(\boldsymbol{\beta}'\boldsymbol{z}_k) - (y_{k0} + y_{k1}) \log\left(N_{k0} + N_{k1} e^{\boldsymbol{\beta}'\boldsymbol{z}_k}\right)\right], \tag{12}$$

whereas the likelihood (10) in the second step of the two-step approach becomes

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{m} \left[y_{k1}(\boldsymbol{\beta}'\boldsymbol{z}) - N_{k1}\, i_0^{(sp)}(k)\, e^{\boldsymbol{\beta}'\boldsymbol{z}}\right]. \tag{13}$$

The MLEs of $\boldsymbol{\beta}$ corresponding to the direct method and two-step approach can be obtained by directly maximizing the likelihood functions (12) and (13), respectively, using a nonlinear optimization tool (e.g., optim or nlm in R Venebles, W.N. & Ripley, B.D. (2002)). Alternatively, MLEs can be obtained by iteratively solving the estimating equations provided in Web Appendix A. The variances of these estimates are derived using the Fisher information matrix.

Upon estimating the parameters, we can compute the predicted values for the direct estimation approach based on (7) as

$$\widehat{E}[Y_{k0}] = N_{k0}\left(\frac{y_{k0} + y_{k1}}{N_{k0} + N_{k1} e^{\widehat{\boldsymbol{\beta}}'\boldsymbol{z}_k}}\right),\ \widehat{E}[Y_{k1}]$$

$$= N_{k1}\left(\frac{y_{k0}+y_{k1}}{N_{k0}+N_{k1}e^{\hat{\beta}'z_k}}\right)e^{\hat{\beta}'z_k}.$$

Regression residuals for the direct estimation method are summarized in Web Appendix A.

2.6 | Extension to multiple vaccination statuses

Our main objective here is to assess the comparative risks associated with different vaccination statuses. We will focus solely on the direct method. We consider two distinct vaccination statuses labelled as $j = 1$ and $j = 2$. Each status is associated with a respective response, Y_{k1} and Y_{k2}, compared against a baseline represented by $j = 0$. The baseline signifies the unvaccinated group, and their corresponding response is denoted as Y_{k0}.

In the following Section 3.2, we incorporate the booster vaccine into this framework. We classify recipients of a third vaccine dose as a separate vaccination status, in contrast to those who have been fully vaccinated with two doses. This adjustment impacts the associated at-risk population sizes, leading to revised values of N_{k1}, N_{k2}, and N_{k0}, respectively.

Let y_{k1} be the count of event occurrences within time interval I_k for individuals who received two vaccine doses (from a population of size N_{k1}). The RR against the unvaccinated is denoted as r_k. Similarly, let y_{k2} be the count of event occurrences within the same time interval I_k for those who received a third (booster) dose (from a population of size N_{k2}).

The addition of a third dose to those who have already received two doses has a multiplicative effect ϕ_k, rendering the RR against the unvaccinated as $r_k\phi_k$. If ϕ_k is statistically significantly less than 1, it implies that the third dose further reduces the RR among those vaccinated with two doses compared to the unvaccinated.

In the regression analysis, covariates $\boldsymbol{z}$ may have distinct impacts on r_k and ϕ_k. We can consider a pair of log-linear models

$$\begin{cases} \log r_k(\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{q_1}\beta_j z_{jk} = \boldsymbol{\beta}'\boldsymbol{z}, \\ \log \phi_k(\boldsymbol{\gamma}) = \gamma_0 + \sum_{j=1}^{q_2}\gamma_j z_{jk} = \boldsymbol{\gamma}'\boldsymbol{z}. \end{cases} \tag{14}$$

Given the marginal incidence numbers $Y_k = Y_{k0} + Y_{k1} + Y_{k2}$, the conditional distributions are multinomial with expectations

$$E[Y_{k0}|y_k] = N_{k0}\left(\frac{y_{k0}+y_{k1}+y_{k2}}{N_{k0}+N_{k1}e^{\boldsymbol{\beta}'\boldsymbol{z}}+N_{k2}e^{\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\gamma}'\boldsymbol{z}}}\right),$$

$$E[Y_{k1}|y_k] = N_{k1}\left(\frac{y_{k0}+y_{k1}+y_{k2}}{N_{k0}+N_{k1}e^{\boldsymbol{\beta}'\boldsymbol{z}}+N_{k2}e^{\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\gamma}'\boldsymbol{z}}}\right)e^{\boldsymbol{\beta}'\boldsymbol{z}},$$

$$E[Y_{k2}|y_k] = N_{k2}\left(\frac{y_{k0}+y_{k1}+y_{k2}}{N_{k0}+N_{k1}e^{\boldsymbol{\beta}'\boldsymbol{z}}+N_{k2}e^{\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\gamma}'\boldsymbol{z}}}\right)e^{\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\gamma}'\boldsymbol{z}}.$$

The conditional likelihood (12) is extended to a function of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ as

$$l^*(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{k=1}^{m}\left[y_{k1}(\boldsymbol{\beta}'\boldsymbol{z}) + y_{k2}(\boldsymbol{\beta}'\boldsymbol{z}+\boldsymbol{\gamma}'\boldsymbol{z})\right] - \sum_{k=1}^{m} y_k \log\left(N_{k0} + \left(N_{k1}+N_{k2}e^{\boldsymbol{\gamma}'\boldsymbol{z}}\right)e^{\boldsymbol{\beta}'\boldsymbol{z}}\right). \tag{15}$$

With MLE $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, the fitted baseline incidence rate is given by

$$i_0(k|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) = \frac{y_{k0}+y_{k1}+y_{k2}}{N_{k0}+N_{k1}e^{\hat{\boldsymbol{\beta}}'\boldsymbol{z}}+N_{k2}e^{\hat{\boldsymbol{\beta}}'\boldsymbol{z}+\hat{\boldsymbol{\gamma}}'\boldsymbol{z}}}.$$

Residual estimates are given in Web Appendix A.

3 | APPLICATIONS

We obtained daily counts of vaccinated and unvaccinated individuals, as well as daily data on new hospital and ICU admissions stratified by age and vaccine status from the Ontario Case and Contact Management System (Government of Ontario, 2021; Public Health of Ontario, 2022) between January 2021 and February 2022. To determine the at-risk populations, we used age-stratified population statistics for Ontario from Q2 2021, sourced from Statistics Canada (Statistics Canada, 2021).

We aggregated daily counts into weekly totals, starting with Week 1 on January 4, 2021, and ending with Week 52 on January 2, 2022. To make the time-series easier to present, we continued the labelling of weeks in 2022 from the previous year, meaning that Week 53 was the first week of 2022. Our analysis was truncated by the end of Week 59, which was on Sunday, February 20, 2022.

The first part of this section demonstrates the performance of the statistical methods by analyzing a subset of COVID-19 hospital admission data for two 10-year age cohorts: 30–39 and 70–79 years, assuming that r_k is piecewise constant. The second part focuses on applying the statistical analysis to the entire adult population in Ontario, considering both hospital and ICU admissions as endpoint events. These analyses are stratified by age cohorts, with $\log r_k$ modeled as a linear function of time and other covariates, including the additional time effect during the Omicron era and the effect of the booster vaccine.

We do not differentiate among vaccine types or brands. We use the term "at-risk populations" to convey that individuals with varying vaccination statuses are truly susceptible each week. Unfortunately, we cannot account for those with immunity due to recent infections. A more comprehensive discussion on potential bias will be presented in Section 3.3.

3.1 | Performance of the statistical methods based on analysis of data in two age cohorts

We study two age cohorts, 30–39 years and 70–79 years, as well as two vaccination groups: vaccinated and unvaccinated. Vaccinated individuals are defined as those who have received at least two doses of COVID-19 vaccine. Data for individuals with only one dose are not analyzed, because this group was very small during the latter half of 2021, accounting for less than 5% of the whole population by Week 30 and 1% by Week 59. Of the 92% who received a minimum of one dose, 98.7% received their second or third dose by the end of the study period.

The RR $r_k(\boldsymbol{\beta})$ is modeled as a piecewise constant function across three distinct periods:

$$r_k(\boldsymbol{\beta}) = \beta_0, \text{ if } k \in \Theta_0;\ r_k(\boldsymbol{\beta}) = \beta_1, \text{ if } k \in \Theta_1;\ r_k(\boldsymbol{\beta}) = \beta_2, \text{ if } k \in \Theta_2,$$

where Θ_0 is the period from the beginning of 2021 until the last week when the two-dose coverage is less than 50% in a given age cohort; Θ_1 spans from the first week when two-dose coverage reaches 50% or more until Week 49; and Θ_2 covers the period from Week 50 of 2021 to February 20, 2022 (Week 59), which is the end of the time series for this analysis. For the 30–39 age group, Θ_1 starts at Week 29, while for the 70–79 age group, Θ_1 starts at Week 25 due to age-based vaccination prioritization in Ontario. We compare results using the two estimation approaches.

The direct estimation:
The RRs β_0, β_1, and β_2 are estimated by maximizing the conditional likelihood (4), which is separately defined for each of $k \in \Theta_0, k \in \Theta_1$ and $k \in \Theta_2$ as

$$l^*(\beta_j) = \sum_{k\in\Theta_j} \left[y_{k1} \log \beta_j - (y_{k0} + y_{k1}) \log\left(N_{k0} + N_{k1}\beta_j\right)\right].$$

The 95% CIs for β_j can be directly obtained from the conditional likelihood using the likelihood ratio statistics (Kalbfleisch, 1985; Sprott, 2000). Based on the MLEs $\widehat{\beta}_j$, the incidence rates for unvaccinated and vaccinated populations are fitted by

$$i_0(k|\widehat{\beta}_j) = \frac{y_{k0} + y_{k1}}{N_{k0} + N_{k1}\widehat{\beta}_j},\ i_1(k|\widehat{\beta}_j) = i_0(k|\widehat{\beta}_j)\widehat{\beta}_j;\ k \in \Theta_j,\ j = 0, 1, 2. \tag{16}$$

The two-step approach:
Assuming the count data for all 59 weeks among unvaccinated individuals follow negative binomial distributions, a penalized spline estimate $i_0^{(sp)}(k)$ is obtained in the first step. In the second step, the RRs β_j, $j = 0, 1, 2$, are estimated for three distinct periods by maximizing (10), which has an explicit solution given by

$$\widetilde{\beta}_j = \frac{\sum_{k\in\Theta_j} y_{k1}}{\sum_{k\in\Theta_j} N_{k1} i_0^{(sp)}(k)}.$$

We use the Wald-based bootstrapping procedure to obtain CIs. The incidence rate in the vaccinated group is calculated as $i_1^{(sp)}(k|\widetilde{\beta}_j) = i_0^{(sp)}(k)\widetilde{\beta}_j; k \in \Theta_j, j = 0, 1, 2$.

Comparison:
The results from both methods are presented in Table 1. The point estimates by the two methods are nearly identical for both age groups and across the three periods. The CIs from the two approaches are in agreement in most cases. However, some discrepancies are expected because the direct approach captures the randomness in the data $\{Y_{k0}\}$ from the unvaccinated group in the conditional likelihood (4), whereas the two step approach does not involve data $\{Y_{k0}\}$ in its "likelihood" (10), and the randomness of these data is reflected in the uncertainty of the estimate $i_0^{(sp)}(k)$ via bootstrapping.

Figure 1 displays the predicted incidence rates for both unvaccinated and vaccinated populations. The "raw data" are presented as circles in the form of the crude incidence rates y_{k0}/N_{k0} and y_{k1}/N_{k1}. The dotted lines correspond to estimates based on the direct method as calculated by (16), and closely track the crude incidence rates. The smoother solid trend lines represent $i_0^{(sp)}(k)$ and $i_1^{(sp)}(k|\widetilde{\beta}_j)$ from the two-step approach, and they agree well with the direct estimates.

3.2 | Full analysis of hospital and ICU admission data for the adult population

In the full analysis, we consider four different age groups: 20–49 years, 50–69 years, 70–79 years, and 80+ years, with two vaccination statuses: those who received only two doses of the vaccine versus those who received a

TABLE 1 Numerical comparisons of the two estimation approaches. Numbers in brackets are estimated 95% confidence intervals. The intervals correspond to the direct approach are based on the likelihood ratio statistics derived from the conditional likelihood (4). Those correspond to the two-step approach are calculated based on the Wald-based bootstrap standard errors.

	Direct estimation	Two-step approach
30–39 years		
Θ_1 : Week 1–28	$\widehat{\beta}_0 = 0.054$ (0.016,0.127)	$\widetilde{\beta}_0 = 0.054$ (0.001,0.107)
Θ_2 : Week 29–49	$\widehat{\beta}_1 = 0.021$ (0.013,0.031)	$\widetilde{\beta}_1 = 0.021$ (0.013,0.030)
Θ_3 : Week 50–59	$\widehat{\beta}_2 = 0.222$ (0.179,0.275)	$\widetilde{\beta}_2 = 0.223$ (0.190,0.256)
70–79 years		
Θ_1 : Week 1–24	$\widehat{\beta}_0 = 0.050$ (0.026,0.084)	$\widetilde{\beta}_0 = 0.051$ (0.021,0.081)
Θ_2 : Week 25–49	$\widehat{\beta}_1 = 0.067$ (0.057,0.079)	$\widetilde{\beta}_1 = 0.067$ (0.057,0.077)
Θ_3 : Week 50–59	$\widehat{\beta}_2 = 0.177$ (0.161,0.193)	$\widetilde{\beta}_2 = 0.181$ (0.164,0.198)

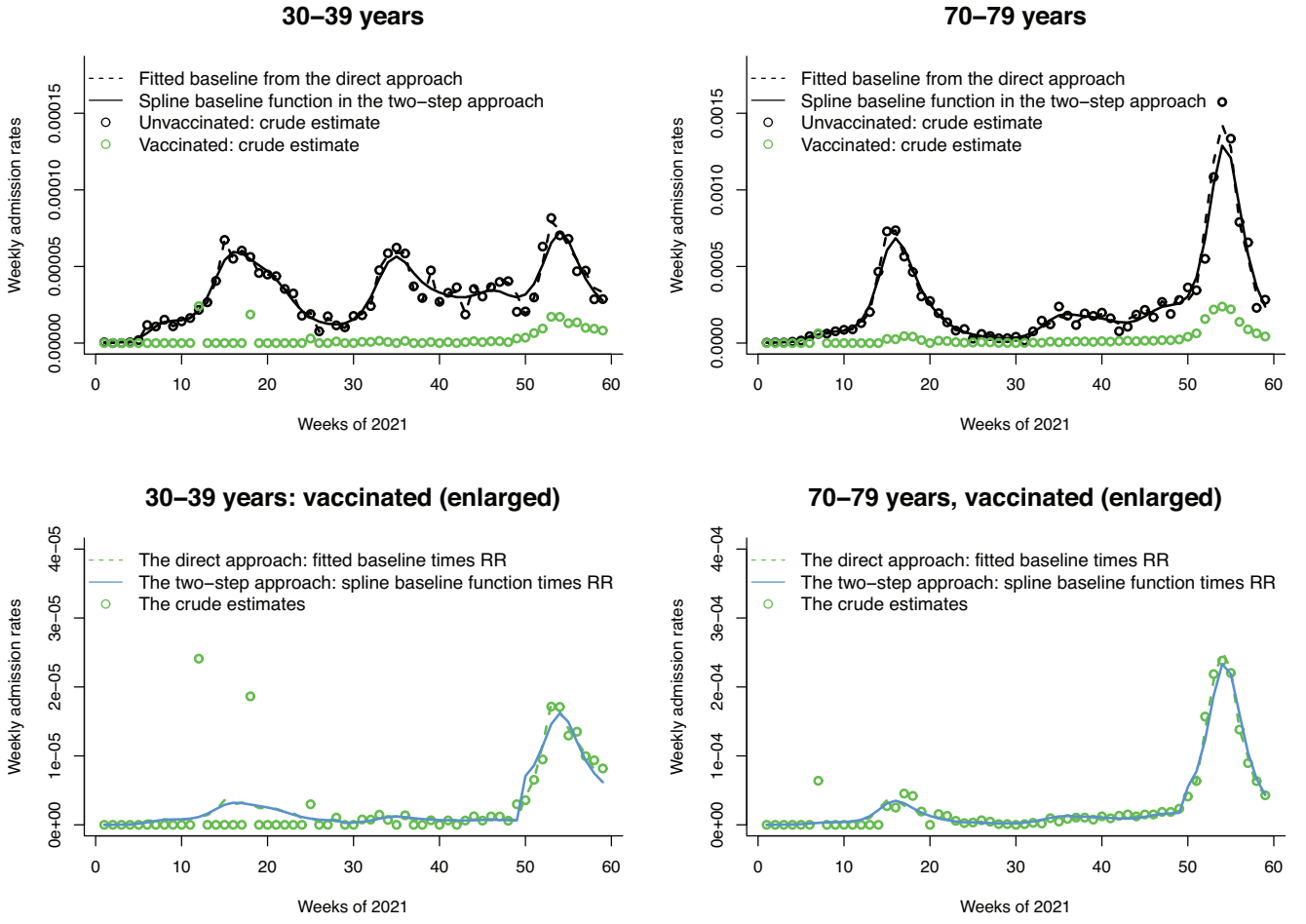

FIGURE 1 Predicted weekly hospital admission rates $i_0(k)$ and $i_1(k)$ (lines) against crude incidence rates (circles). Note that the scale of the y-axis for the 70–79 years is 10 times of that for the 30–39 years.

third booster dose. The questions to be investigated are as follows:

(1) Are there significant drifts over time in the RRs for those who received only two doses?
(2) How significant is the Omicron era in influencing changes in these RR s for those who received only two doses?
(3) What is the effect of a third (booster) dose on the time-trend of the RR?

The Omicron variant in Ontario was first identified on November 21, 2021 (Week 47) and its prevalence exceeded 90% among cases by December 21 (Week 51). In Ontario, the ramp up of the third dose began around Week 50 (December 13) when all individuals aged 50 years and older became eligible. Therefore, the period Θ_2 from Week 50 of 2021 until February 20, 2022 (Week 59) encompasses the Omicron era while coinciding with the ramping up of the third dose.

To address Question 1, we define the time-varying covariate $z_{1k} = k$. For Question 2, we define $z_{2k} = 1$ if $k = 50, \ldots, 59$, and 0 otherwise. To address Question 3, we use the same covariate z_{2k}. The regression models (14) are given by

$$\begin{cases} \log r_k = \beta_0 + \beta_1 k + \beta_2 z_{2k}, \\ \log \phi_k = \gamma z_{2k}, \end{cases} \quad k = 1, \ldots, 59, \tag{17}$$

where β_1 captures the drift, β_2 captures the Omicron effect, and γ signifies the booster effect, which addresses Question 3. The conditional likelihood (15) is given by

$$l^*(\boldsymbol{\beta}, \gamma) = \sum_{k=1}^{59} [(y_{k1} + y_{k2})(\beta_0 + \beta_1 k + \beta_2 z_{2k})] + \gamma \sum_{k=1}^{59} y_{k2} z_{2k} \tag{18}$$

$$- \sum_{k=1}^{59} y_k \log \left(N_{k0} + (N_{k1} + N_{k2} e^{\gamma z_{2k}}) e^{\beta_0 + \beta_1 k + \beta_2 z_{2k}} \right),$$

where y_{k1} is the weekly incidence counts for those who only received two doses, and y_{k2} is the weekly incidence counts for those who received the third dose. The total incidence count is $y_k = y_{k0} + y_{k1} + y_{k2}$. The at-risk populations are defined accordingly and denoted as N_{k0}, N_{k1}, and N_{k2}. The MLEs are $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\gamma})$.

We conducted separate analyses for weekly hospital admissions and weekly ICU admissions, and the results are presented in Table 2. The positive drift over time β_1 is significant in all age groups, and the significance level increases with age (i.e. smaller p-values). This finding aligns well with VE studies Lin et al. (2022), Andrews et al. (2022), and Buchan et al. (2021), which indicate faster vaccine waning among older populations.

The VE is calculated by

$$\widehat{VE}_k = \begin{cases} 1 - e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \widehat{\beta}_2 z_{2k}}, & \text{if receiving 2 doses only,} \\ 1 - e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \left(\widehat{\beta}_2 + \widehat{\gamma}\right) z_{2k}}, & \text{with 3 doses.} \end{cases} \tag{19}$$

The estimated parameters in Table 2 suggest that, beyond the existing linear time-drift of the declining VE, there was significant further reduction of VE during the Omicron era among individuals who received only two doses. However, the third dose offered good protection against hospital and ICU admissions, as illustrated in Figure 2. For Week 59 (between February 14 and February 20, 2022):

(1) the effectiveness of the third vaccine dose ($\widehat{VE}_{59}$) against hospital admissions increased from 58% (2 doses only) to 85% (3 doses) for 20- to 49-year-olds, 72% to 93% for 50- to 69-year-olds, 47% to 91% for 70- to 79-year-olds, and 11% to 81% for those aged 80 and above;
(2) the effectiveness of the booster vaccine ($\widehat{VE}_{59}$) against ICU admissions increased from 68% to 82% in the 20–49 age group, 78% to 95% in the 50–69 age group, 61% to 95% in the 70–79 age group, and 32% to 89% in individuals 80 years and older.

The model fitted baseline weekly incidence rate in the unvaccinated is given by

$$i_0(k|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = \frac{y_{k0} + y_{k1} + y_{k2}}{N_{k0} + N_{k1} e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \widehat{\beta}_2 z_{2k}} + N_{k2} e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \left(\widehat{\beta}_2 + \widehat{\gamma}\right) z_{2k}}},$$

and the model predicted incidence counts are

$$\widehat{E}[Y_{k0}] = N_{k0} i_0(k|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}),$$

$$\widehat{E}[Y_{k1}] = N_{k1} i_0(k|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \widehat{\beta}_2 z_{2k}}$$

$$\text{and } \widehat{E}[Y_{k2}] = N_{k2} i_0(k|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) e^{\widehat{\beta}_0 + \widehat{\beta}_1 k + \left(\widehat{\beta}_2 + \widehat{\gamma}\right) z_{2k}}.$$

Using the estimates from Table 2, predicted values are plotted in Figures 3 and 4. It is evident that the predicted values align well with the observed data. Moreover, the residual plots (not displayed) exhibit random dispersion around zero without discernible patterns, suggesting that the model is adequately capturing the data.

3.3 | Limitations

3.3.1 | Natural immunity due to prior encounters with infectious agents

The aggregated data analyzed in our study lack information on prior exposure to infections. As such, we could not determine the percentage of individuals with infection-induced immunity within the "at-risk" populations with sizes N_{k0}, N_{k1}, and N_{k2}. Assuming that each of these subpopulations shares an equal proportion of individuals with preexisting immunity, this proportion would be cancelled out in the score functions corresponding to the conditional likelihoods, leaving the estimated RRs unchanged.

TABLE 2 Results for weekly hospital and intensive care unit (ICU) admissions in two separate analyses, where β_1 signifies the drift over time; β_2 signifies the Omicron effect and γ signifies the booster (third dose) effect.

Weekly hospital admissions						
20–49 years				50–69 years		
	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p -value	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p-value
$\widehat{\beta}_0$	−5.378	0.605		−4.560	0.359	
$\widehat{\beta}_1$	0.045	0.015	0.0025	0.030	0.009	0.0008
$\widehat{\beta}_2$	1.869	0.245	3×10^{-14}	1.538	0.152	$< 10^{-16}$
$\widehat{\gamma}$	−1.039	0.114	2×10^{-5}	−1.398	0.065	$< 10^{-16}$
70–79 years				80+ years		
	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p -value	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p-value
$\widehat{\beta}_0$	−4.067	0.394		−3.302	0.203	
$\widehat{\beta}_1$	0.034	0.010	0.0004	0.041	0.006	4×10^{-13}
$\widehat{\beta}_2$	1.425	0.168	$< 10^{-16}$	0.796	0.140	1×10^{-8}
$\widehat{\gamma}$	−1.819	0.060	$< 10^{-16}$	−1.673	0.046	$< 10^{-16}$
Weekly ICU admissions						
20–49 years				50–69 years		
	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p -value	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p-value
$\widehat{\beta}_0$	−9.373	1.815		−6.340	0.851	
$\widehat{\beta}_1$	0.135	0.042	0.001	0.062	0.020	0.002
$\widehat{\beta}_2$	0.298	0.571	0.6	1.167	0.316	0.0002
$\widehat{\gamma}$	−0.562	0.313	0.3	−1.494	0.147	$< 10^{-16}$
70–79 years				80+ years		
	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p -value	$\widehat{\boldsymbol{\beta}}$	s.e.($\widehat{\boldsymbol{\beta}}$)	p-value
$\widehat{\beta}_0$	−5.192	1.043		−5.828	1.058	
$\widehat{\beta}_1$	0.043	0.025	0.09	0.087	0.026	0.0009
$\widehat{\beta}_2$	1.749	0.405	0.00002	0.293	0.491	0.55
$\widehat{\gamma}$	−2.076	0.161	$< 10^{-16}$	−1.820	0.185	$< 10^{-16}$

However, a substantial bias arises if the proportions of individuals with preexisting immunity vary among these subpopulations. For instance, if N_{k0} contains a higher percentage of individuals with immunity acquired from recent infections compared to N_{k1} and N_{k2}, RRs would be overestimated, and VE would be underestimated.

Our analysis may appear to indicate that vaccinations (including the third dose) in the 20- to 49-year age group have been less effective after Week 49 than in the 50- to 69-year age group and even less so compared to the 70- to 79-year age group in some settings (see Table 2 and Figure 2). However, publicly reported case surveillance data from Open Government Licence-Ontario (Web-1) and test positivity rates from the Open Laboratory Information System (Web-2) indicate that the population under 49 years of age had the highest proportion of test-positive cases from July to October 2021. Although these data do not provide a complete picture of the transmission due to unknown factors such as testing and ascertainment rates (Lawless & Yan, 2021), they seem to suggest that individuals aged 20 to 49 were predominantly affected by infections during that period. If most of these infections during this period occurred among the unvaccinated (according to anecdote) and if the circulating strains during these months resulted in at least partial immunity against the Omicron variant, which emerged in early December 2021, this may partially explain the apparent reduction in VE for this age group. However, these hypotheses can be only verified using comprehensive longitudinal cohorts that document vaccination timing, dose numbers, and infection events, including infecting variant. Unfortunately, the available data in this study do not have the required information.

3.3.2 | Differential risk-taking and adherence to other public health measures

Vaccinated and unvaccinated individuals may exhibit different attitudes towards risk-taking and compliance with supplementary public health measures, including mask-wearing and social distancing. These behaviors could be further confounded with age and may also correlate with the differential acquired immunity resulting from

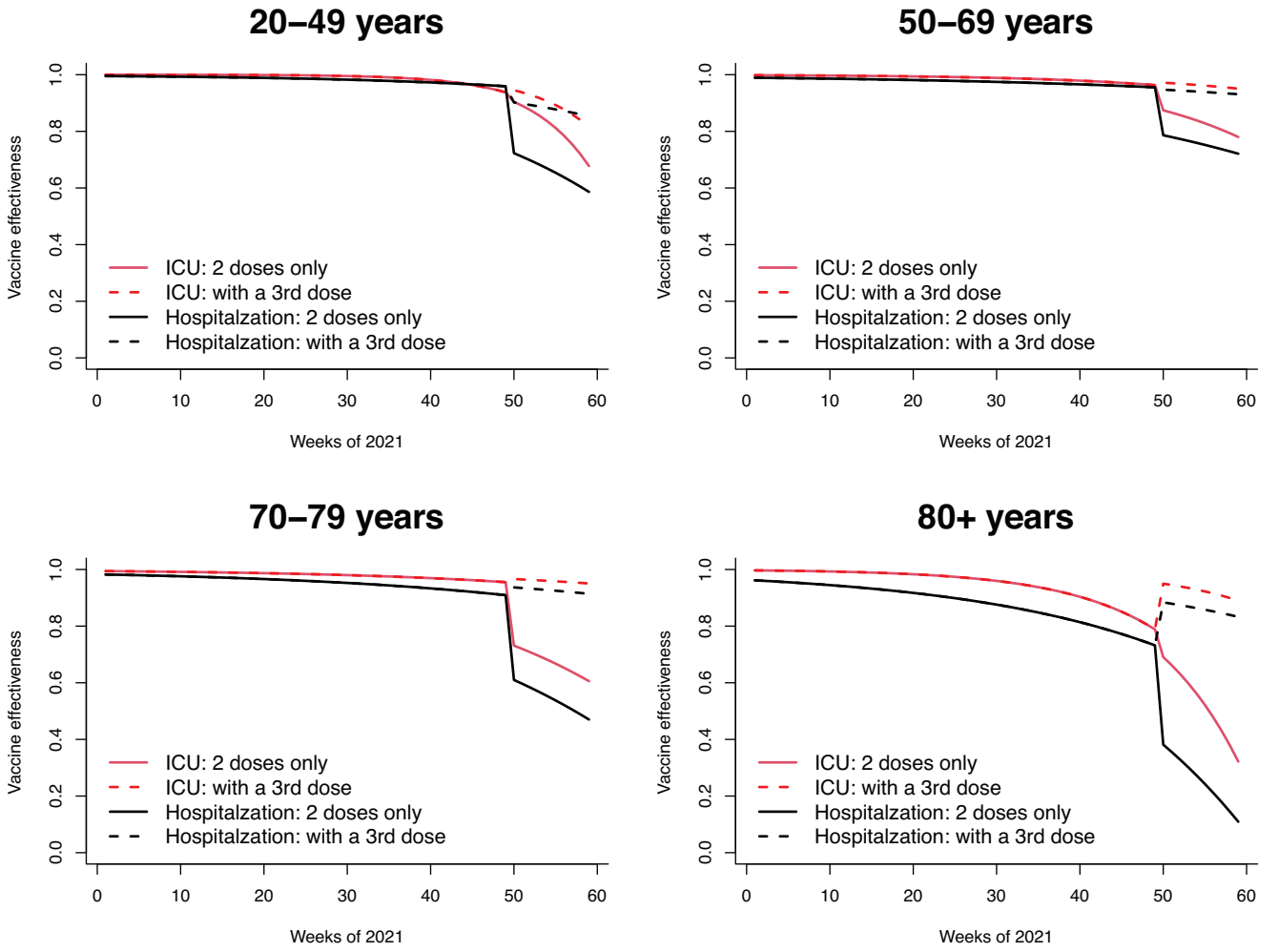

FIGURE 2 Vaccine effectiveness with respect to hospital and intensive care unit (ICU) admissions, calculated by (19) based on parameters from Table 2.

virus exposure. However, this aspect is more difficult to quantify, even within longitudinal cohorts.

3.3.3 | Incidental hospital and ICU admissions

Our model and method (described in Section 3.2) can be used to extend the response variables Y_{k1} and Y_{k2} to multiple types, not necessarily limited to vaccine types. There have been discussions concerning the distinction between hospital and ICU admission resulting directly from COVID-19 infection and "incidental admissions" (referring to patients admitted to the hospital or ICU for other medical reasons who are incidentally diagnosed with COVID-19 infection). Unfortunately, the data used in this study do not contain an indicator to distinguish the reason for admission. If a longitudinal database were available, these indicators could have been identified. If such data were available at an aggregate level, our method could be extended to analyze them.

3.3.4 | Comorbidity

Comorbidity is a crucial confounding factor that operates at multiple levels. First, it affects the trends of at-risk populations N_{k0}, N_{k1}, and N_{k2} over time, since each new vaccine rollout (such as the second, third, and fourth doses) prioritizes individuals with comorbidities before expanding to the general population. Second, it is associated with incidental hospital and ICU admissions. Third, it is correlated with age. Notably, comorbidity is a prominent feature in the 80+ years age group, as clearly demonstrated in Figure 2. Longitudinal individual-based cohort data offer superior control over this confounding factor compared to coarsely aggregated count data.

3.3.5 | Waning of acquired immunity

In Section 3.2, we intentionally avoided using the term "waning" to describe the decline of VE over time. This phenomenon cannot be accurately assessed without a

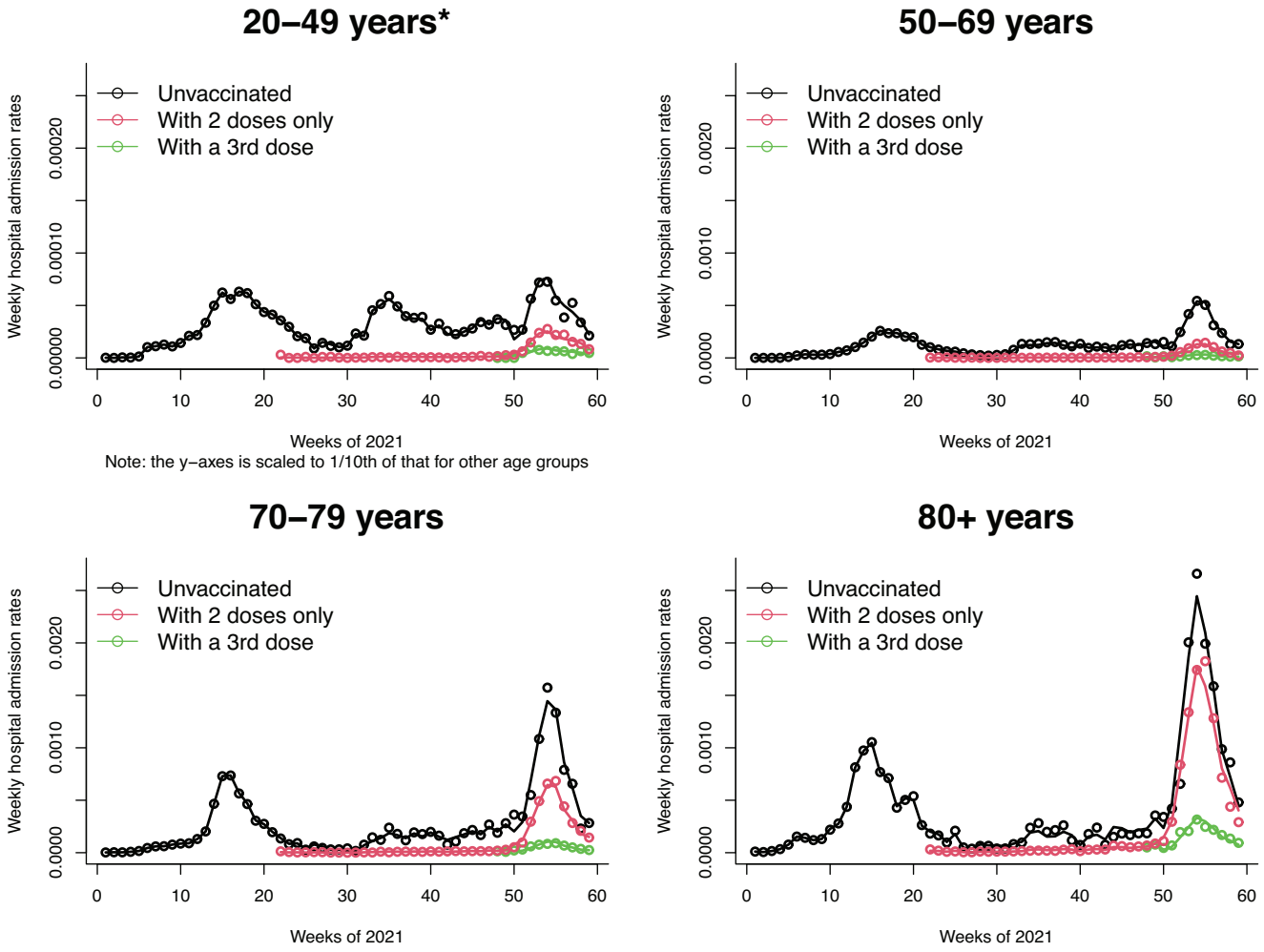

FIGURE 3 Predicted weekly hospital admission rates based on parameters in Table 2 against crude estimates based on reported data (shown as circles).

high-quality longitudinally followed cohort, because waning is a function of time that must be measured for each individual from a defined vaccination date. Moreover, it also depends on the circulating variants since the time of vaccination. Lin et al. (2022) measured the waning effect starting from the vaccination date of the first dose, while Buchan et al. (2021) used the time since the last dose and stratified their analysis by Delta and Omicron variants. Unfortunately, the aggregated data we use in our study lack such detailed information. Hence, the best we can provide is an empirical observation of the diminishing VE at the population level over time.

3.3.6 | Age

Age is inherently associated with all the factors discussed above, making it a confounding factor as well. Therefore, we deliberately refrained from incorporating age as a covariate in our regression analysis. Instead, we conducted age-stratified analyses to account for its influence.

4 | DISCUSSIONS AND CONCLUSION

There is no substitute for high-quality, individual-based longitudinal cohort data combined with an individual-based statistical model when analyzing event history processes in a population. It is also crucial to account for potential confounders. However, obtaining representative high-quality data and controlling for confounding factors pose significant challenges. Separating the "pure" VE from the influence of other public health measures proves to be a difficult task.

Considering the readily available data sources and acknowledging their limitations, we have developed novel statistical models and methods to demonstrate their application using aggregate-level COVID-19 data. Our proposed proportional incidence rate model is based on the RRs $\rho_1(t)/\rho_0(t)$ (aggregated as r_k in (1)), which determine VE. This model differs subtly from the proportional intensity model of the two counting processes with intensities $\lambda_0(t) = \rho_0(t)N_0(t)$ and $\lambda_1(t) = \rho_1(t)N_1(t)$. Given the

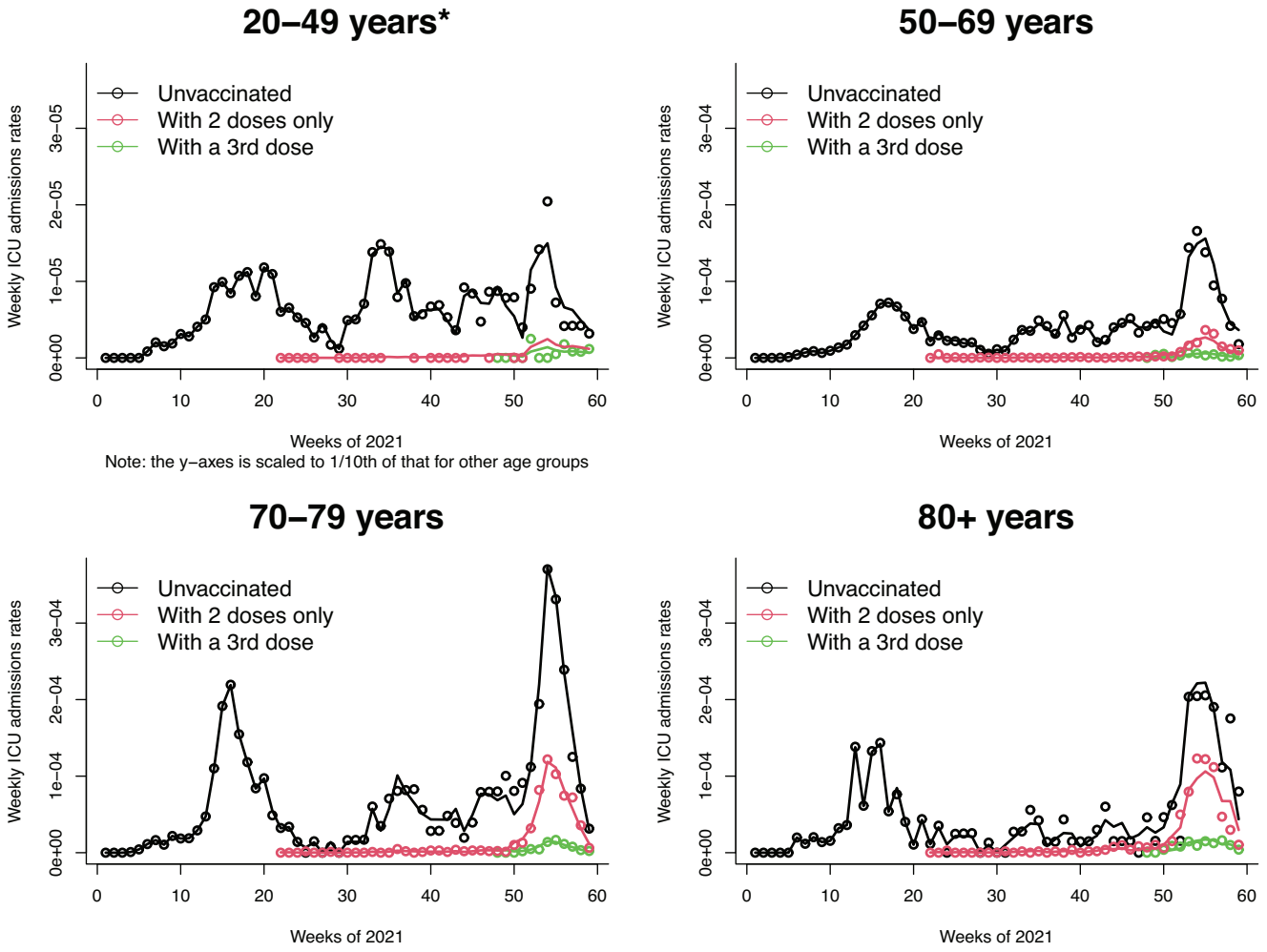

FIGURE 4 Predicted weekly intensive care unit (ICU) admission rates based on parameters in Table 2 against crude estimates based on reported data (shown as circles).

discrete (i.e., aggregated) nature of the data and the model formulation, we presented the conditional likelihoods (4), (12), and (15), treating the baseline incidence rate $i_0(k)$ as a nuisance parameter. Similar to the proportional hazards model in survival analysis, $i_0(k)$ can be estimated semi-parametrically after estimating the RR function using the conditional likelihood. Although we tailored these models and methods for a specific application, they can be extended to other applications with similar settings and data features.

We have demonstrated that event counts (e.g., hospital and ICU admissions) within a vaccinated cohort are conditionally sufficient for estimating RR (when it is the only objective), given the total numbers of these occurrences across the entire population. Additional trend information for the risk functions that define the RR can be enhanced by incorporating a separate penalized spline model. This model utilizes the aggregated counts of events in the unvaccinated group as the baseline incidence rate for the counterfactuals.

The applications presented in Section 3 not only illustrate the statistical models and methods proposed in Section 2, but also extend the study period to include the Omicron era and the introduction of third vaccine dose until February 2022. These aspects have not been addressed in the existing literature we reviewed. Our findings are in agreement with other studies, including test-negative case-control and cohort studies Teerawattananon et al. (2022), Andrews et al. (2022), Buchan et al. (2021), and Lin et al. (2022). In particular, our study and Buchan et al. (2021) covered the same population using an identical mixed vaccine schedule, yielding highly consistent results.

We extensively discussed the limitations and confounding factors in Section 3.3, emphasizing the need for improved data collection practices. This includes ensuring representativeness, comprehensive coverage, individual-level documentation of event history, and capturing potential confounding variables. Our recommendation extends beyond COVID-19 and applies to other infectious diseases,

promoting the development of a more resilient and adaptable public health infrastructure and response. Despite its limitations, this paper offers a valuable approach for examining time-related factors at an aggregate level.

ACKNOWLEDGMENTS

We received ethics approval for this study from the Research Ethics Board at the University of Toronto. We thank Professor J.F. Lawless, Department of Statistics and Actuarial Science, University of Waterloo, for insightful discussions and statistical advices.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper were collected from public domains, specifically the Ontario Case and Contact Management System (https://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/contact_mngmt/management_cases_contacts.pdf) and Statistics Canada (https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501). Regarding the former, it is important to note that the data were continuously evolving during the pandemic and updated beyond our study period. To obtain the original data used in the analysis, please contact directly Ping Yan, email: ping.yan@phac-aspc.gc.ca.

ORCID

Ping Yan https://orcid.org/0000-0001-6765-0735

REFERENCES

Andrews, N., Tessier, E., Stowe, J., Gower, C., Kirsebom, F., Simmons, R. et al. (2022) Duration of protection against mild and severe disease by Covid-19 vaccines. *New England Journal of Medicine*, 386, 340–350.

Buchan, S.A., Chung, H., Brown, K.A., Austin, P.C., Fell, D.B., Nasreen, S. et al. (2021) Estimated Effectiveness of COVID-19 Vaccines Against Omicron or Delta Symptomatic Infection and Severe Outcomes. *JAMA Netw Open*. 5(9), e2232760. https://doi.org/10.1001/jamanetworkopen.2022.32760

Godambe, V.P. & Thompson, M.E. (1974) Estimating equations in the presence of a nuisance parameter. *The Annals of Statistics*, 2(3), 568–571.

Godambe, V.P. (1980) On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika*, 67, 155–162.

Government of Ontario. (2021) *Management of cases and contacts of COVID-19 in Ontario*. https://www.health.gov.on.ca/en/pro/programs/publichealth/coronavirus/docs/contact_mngmt/management_cases_contacts.pdf (last accessed: August 2021)

Kalbfleisch, J.D. & Sprott, D.A. (1970) Application of likelihood methods to models involving large numbers of parameters. *Journal of Royal Statistical Society Series B*, 32(2), 175–208.

Kalbfleisch, J.G. (1985) *Probability and statistical inference, volume 2: statistical inference*. New York: Springer.

Lawless, J.F. (1987) Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3), 209–225.

Lawless, J.F. & Yan, P. (2021) On testing for infections during epidemics, with application to Covid-19 in Ontario, Canada. *Infectious Disease Modelling*, 6, 930–941.

Lin, D.Y., Gu, Y., Wheeler, B., Young, H., Holloway, S., Sunny, S.K. et al. (2022) Effectiveness of Covid-19 vaccines over a 9-month period in North Carolina. *New England Journal of Medicine*, 386, 933–941.

Mullah, M.A.S. & Yan, P. (2022) A semi-parametric mixed model for short-term projection of daily COVID-19 incidence in Canada. *Epidemics*, 38, 100537, https://doi.org/10.1016/j.epidem.2022.100537.

Nasreen, S., Febriani, Y., García, H.A.V., Zhang, G., Tadrous, M., Buchan, S.A., Christiaan, H., Righolt, C.H. et al. (2022) on behalf of the Canadian Immunization Research Network Provincial Collaborative Network Investigators, Effectiveness of Coronavirus Disease 2019 Vaccines Against Hospitalization and Death in Canada: A Multiprovincial, Test-Negative Design Study, Clinical Infectious Diseases, Volume 76, Issue 4, 15 February 2023, Pages 640–648, https://doi.org/10.1093/cid/ciac634

Ogden, N.H., Turgeon, P., Fazil, A., Clark, J., Gabriele-Rivet, V., Tam, T. & Ng, V. (2022) Counterfactuals of effects of vaccination and public health measures on COVID-19 cases in Canada: what could have happened? *Canada Communicable Disease Report*, 48(7/8), 292–302. https://doi.org/10.14745/ccdr.v48i78a01

Sprott, D.A. (1975) Marginal and conditional sufficiency. *Biometrika*, 62, 599–605.

Sprott, D.A. (2000) *Statistical inference in science*. New York: Springer.

Statistics Canada. (2021) Table 17-10-0005-01. *Population estimates on July 1st, by age and sex*. Available at: https://doi.org/10.25318/1710000501-eng (last accessed: Octorber 2021).

Teerawattananon, Y., Anothaisintawee, T., Pheerapanyawaranun, C., Botwright, S., Akksilp, K., Sirichumroonwit, N. et al. (2022) A systematic review of methodological approaches for evaluating real-world effectiveness of COVID-19 vaccines: advising resource-constrained settings. *PLoS ONE*, 17(1), e0261930, https://doi.org/10.1371/journal.pone.0261930.

Venebles, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*. New York: Springer.

Web-1: https://www.ontario.ca/page/open-government-licence-ontario

Web-2: https://ehealthontario.on.ca/en/standards/ontario-laboratories-information-system-standard

Wood, S.N. (2003) Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, 65, 95–114.

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

SUPPORTING INFORMATION

Web Appendices, Table, and Figures referenced in Sections 2 and 3 are available with this paper at the Biometrics website on Wiley Online Library. The R code used for data analysis has been included in Web Appendix D.

How to cite this article: Yan, P., Mullah, M.A.S. & Tuite, A. (2023) A proportional incidence rate model for aggregated data to study the vaccine effectiveness against COVID-19 hospital and ICU admissions. *Biometrics*, 1–14. https://doi.org/10.1111/biom.13915