Early evidence for the safety of certain COVID-19 vaccines using empirical Bayesian modeling from VAERS

Chris von Csefalvay[*]

June 7, 2021

Abstract

The novel coronavirus SARS-CoV-2 has rapidly emerged as a significant threat to global public health, in particular because – as is not uncommon with novel pathogens – there is no effective pharmaceutical treatment or prophylaxis to the viral syndrome it causes. In the absence of such specific treatment modalities, the mainstay of public health response rests on non-pharmaceutical interventions (NPIs), such as social distancing. This paper contributes to the understanding of social distancing against SARS-CoV-2 by quantitatively analysing the statistical dynamics of disease propagation as a differential game, and estimating the relative costs of distancing versus not distancing, identifying marginal utility of distancing based on known population epidemiological data about SARS-CoV-2 and concluding that unless the costs of distancing vastly exceed the cost of illness per unit time, social distancing remains a dominant strategy. These findings can assist in solidly anchoring public health responses based on social distancing within a quantitative framework attesting to their effectiveness.

[*] Starschema Inc., Arlington, VA. Correspondence: csefalvayk@starschema.net.

1 Introduction

2 Methods

2.1 Data set

Data for this study was obtained from VAERS on 06 June, 2021. At the time of retrieval, the data set included reports received on or before 28 May, 2021. Data was retrieved using the CDC bulk download site.

2.2 Processing

Data was processed using R 4.1.0[1]. Upon import, data was destructured from VAERS's multi-event schema, where multiple putative AEFIs are included in a single line, to a single-event schema using `reshape2`.[2]

2.3 Metrics

One of the most widely used metrics to identify possible safety signals is the Proportional Reporting Ratio (PRR).[3] For the $m \times n$ matrix D of m adverse events and n drugs, where $D_{i,j}$ ($i \in m$, $j \in n$), the PRR of side effect i in the presence of the drug j is defined as

$$PRR_{i,j} = \frac{D_{i,j}}{D_{i,\star}} \cdot \frac{D_{\neg i,\star}}{D_{\neg i,j}}$$

The PRR commends itself by relative mathematical simplicity and ease of implementation, but is subject to a disproportional reporting bias. In other words, the PRR does not indicate whether a certain side effect is more or less frequent compared to another, or with another drug. In particular, it does not reflect relative risk. It often eludes even trained professionals that the correct interpretation of $PRR_{i,j}$ is not the relative probability that a certain adverse effect will be reported with this particular drug compared with the reference drugs. Thus, a $PRR_{anaphylaxis,j}$ of 3.0 does not indicate that anaphylaxis is three times more likely with j than any other drug. instead, it indicates that the probability of reporting anaphylaxis rather than any other event with j is three times higher than the probability of reporting anaphylaxis rather than any other event with other drugs.[4]

A better indicator of possible safety signals is the empirical Bayesian geometric mean (EBGM) or modified DuMouchel's method.[5] Since its first publication in 1999, this method has been widely used in analysing 'market basket' type problems – that is, identifying combinations of elements on each axis that occur with unusual frequency, where a Bayesian baseline is calculated through an expectation prior.[6–8]

The EBGM approach builds on the relative reporting ratio R_{rep} (occasionally also RR), defined as $\frac{N_{i,j}}{E_{i,j}}$, where Ni, j is the actual number of reported instances of the adverse effect i given the drug j. One would thus expect a value of 1.0 if no association existed, i.e. if rows and columns were independent from each other. Higher values would thus increasingly militate away from the null hypothesis and towards an association between i and j.

One of the deficiencies of the R_{rep} metric is that for low-expectancy low-occurrence issues, a single integer occurrence (which may well be entirely accidental) may, in the face of a small real valued expectancy value, result in a misleadingly high R_{rep} (e.g. $E_{i,j} = 0.05$, $N_{i,j} = 1$ yields an R_{rep} of $\frac{1}{0.05} = 20$.) DuMouchel's work expands on this by using a Poisson likelihood for actual counts, in which $N_{i,j} = Poisson(\mu_{i,j})$.[5] This affords us the ability to calculate the metric

$$\lambda_{i,j} = \frac{\mu_{i,j}}{E_{i,j}}$$

for a prior on $\lambda_{i,j}$ being drawn from a mixture of two gamma distributions. The posterior distribution of $\lambda_{i,j}$, specifically, is the mixture of two gamma distributions parametrised by the shape and scale variables

$$\alpha = \alpha_1 + n$$
$$\beta = \beta_1 + E$$

and

$$\alpha = \alpha_2 + n$$
$$\beta = \beta_2 + E$$

with the parameter $Q_{N_{i,j}}$ being the mixture fraction (i.e. the likelihood that $\lambda_{i,j}$ was drawn from the first gamma distribution of the posterior). Consequently, the posterior of λ is a probabilistic-Bayesian representation of R_{rep}

(and thus amenable to similar canons of interpretation), but with more stable results for low-expectancy low-occurrence events.

2.4 Computation

Computation was carried out using the `openEBGM`[9] package under R 4.1.0.[1] Data was stratified by gender (male, female and unknown) and age group. Age groups were aggregated into four bins: <25, 25-44, 45-64 and over 65 years of age. The Cartesian product of the two stratum variables yielded 15 strata.

For the estimation of hyperparameter vector $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, Q)$, the non-linear Newton minimisation function `stats::nlm` was used, with initialisation weights of $\alpha_1 = 0.2$, $\beta_1 = 0.1$, $\alpha_2 = 2.0$, $\beta_2 = 4.0$ and $Q = 0.333$.

The computation was carried out in two separate runs. First, the data was examined over vaccine types (VAERS variable `VAX_TYPE`), e.g. `FLU3` for all trivalent influenza vaccines and `COVID19` for all COVID-19 vaccines. Then, the same methodology, including fitting separate values for $\hat{\theta}$, was applied to the data over individual vaccines (VAERS variable `VAX_NAME`). In both cases, the same stratification was used.

In addition to the EBGM values, the mixture fraction Q_n of the posterior probability distribution was estimated using the formula described by Eqn. 6 in DuMouchel (1999).[5] Finally, the `quantBisect` function of the `openEBGM` package was used to estimate 5th and 95th percentiles, thereby providing a two-sided 10confidence margin.

3 Results

4 Discussion

blah

Competing interests

The author declares no competing interests.

Supplementary data

All simulations, code and data are available on Github and under the DOI XXXXXXX.

References

[1] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

[2] Hadley Wickham. reshape2: Flexibly reshape data: a reboot of the reshape package. *R package version*, 1(2), 2012.

[3] Stephen JW Evans, Patrick C Waller, and S Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, 10(6): 483–486, 2001.

[4] Nicholas Moore, Gillian Hall, Miriam Sturkenboom, Ron Mann, Rajaa Lagnaoui, and Bernard Begaud. Biases affecting the proportional reporting ratio (prr) in spontaneous reports pharmacovigilance databases: the example of sertindole. *Pharmacoepidemiology and drug safety*, 12(4): 271–281, 2003.

[5] William DuMouchel. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999.

[6] June S Almenoff, William DuMouchel, L Allen Kindman, Xionghu Yang, and David Fram. Disproportionality analysis using empirical bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. *Pharmacoepidemiology and Drug Safety*, 12(6):517–521, 2003.

[7] Rave Harpaz, William DuMouchel, Paea LePendu, and Nigam H Shah. Empirical bayes model to combine signals of adverse drug reactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1339–1347, 2013.

[8] Hyesung Lee, Ju Hwan Kim, Young June Choe, and Ju-Young Shin. Safety surveillance of pneumococcal vaccine using three algorithms: Disproportionality methods, empirical bayes geometric mean, and tree-based scan statistic. *Vaccines*, 8(2):242, 2020.

[9] Travis Canida and John Ihrie. openebgm: An r implementation of the gamma-poisson shrinker data mining model. *R J.*, 9(2):499, 2017.