International Journal of Infectious Diseases 142 (2024) 106976

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Does the healthy vaccinee bias rule them all? Association of COVID-19 vaccination status and all-cause mortality from an analysis of data from 2.2 million individual health records

Tomáš Fürst [1], Angelika Bazalová [1], Tadeáš Fryčák [1], Jaroslav Janošek [1,2,*]

[1] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacky University Olomouc, Olomouc, Czech Republic
[2] Center for Health Research, Faculty of Medicine, University of Ostrava, Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 12 December 2023
Revised 24 January 2024
Accepted 19 February 2024

Keywords:
Healthy vaccinee bias
Healthy vaccinee effect
COVID-19
All-cause mortality
Vaccination
Observation studies

ABSTRACT

Objectives: We investigated the validity of claims of the healthy vaccinee effect (HVE) in COVID-vaccine studies by analyzing associations between all-cause mortality (ACM) and COVID-19 vaccination status.
Methods: Approximately 2.2 million individual records from two Czech health insurance companies were retrospectively analyzed. Each age group was stratified according to the vaccination status (unvaccinated vs. individuals less than 4 weeks vs. more than 4 weeks from Doses 1, 2, 3, and 4 or more doses of vaccine). ACMs in these groups were computed and compared.
Results: Consistently over datasets and age categories, ACM was substantially lower in the vaccinated than unvaccinated groups regardless of the presence or absence of a wave of COVID-19 deaths. Moreover, the ACMs in groups more than 4 weeks from Doses 1, 2, or 3 were consistently several times higher than in those less than 4 weeks from the respective dose. HVE appears to be the only plausible explanation for this, which is further corroborated by a created mathematical model.
Conclusions: In view of the presence of HVE, the baseline difference in the frailty of vaccinated and unvaccinated populations in periods without COVID-19 must be taken into account when estimating COVID-19 vaccine effectiveness from observational data.

Introduction

Vaccination against COVID-19 was the key measure in battling the pandemic and its effectiveness against death and severe course has been demonstrated in a multitude of studies. However, a vast majority of these studies (apart from registration studies) were observational. It has been proposed that observational studies are subject to inherent biases, including differences in testing strategies between vaccinated and unvaccinated groups or in determining the cause of death (with COVID vs due to COVID) [1,2]. Lately, it has been suggested that the so-called "healthy vaccinee effect" (HVE) might be in play, i.e., that the vaccinated population might have been generally healthier than the unvaccinated one. As observational studies are inherently based on the assumption of the identical baseline likelihood of dying due to COVID in both groups, HVE might have biased the results of such studies toward higher vaccine effectiveness [3,4]. Unfortunately, most publicly available datasets contain only summary statistics provided by various national offices not allowing reliable stratification of the population and computing the number of person-years spent in various age and vaccination brackets, thus preventing the evaluation of the possible influence of HVE on the results of observational studies. We have recently published a paper corroborating the existence of HVE in the Czech population based on aggregate data from the biggest health insurer in the Czech Republic [5]. That paper was, however, based on aggregate data that prevented us from exploring the issue sufficiently, which led us to the acquisition of more detailed line data enabling a more detailed analysis.

In this paper, therefore, we aim to analyze the association between all-cause mortality (ACM) and vaccination status in the most vulnerable age groups in order to better understand the possible differences between the vaccinated and unvaccinated cohorts.

* Corresponding Author: Jaroslav Janošek, Center for Health Research, Faculty of Medicine, University of Ostrava, Syllabova 19, Ostrava, Czech Republic.
E-mail address: janosek@correcta.cz (J. Janošek).

https://doi.org/10.1016/j.ijid.2024.02.019

Methods

Data acquisition

Using a Freedom of Information request, we obtained data from two health insurance houses (in the Czech Republic, health insurance is compulsory and multiple health insurance houses provide the service). Each line in the dataset corresponded to a unique individual and included their sex, age, dates and types of all COVID vaccines, and (if applicable), the date of death. When trying to find out whether there is a difference between the baseline frailty of the groups, ACM is the most telling parameter as it is not burdened with any possible misclassification on the cause of death. For this reason, only ACM will be considered in this study.

Each cooperating insurance company decided to blur the data slightly to exclude the possibility of identification of any individual. The first dataset provided by the Czech Business Insurance Company (CPZP) comprises 1,362,924 individuals, i.e., approx. 13% of the Czech population. It is by no means a fully representative sample of the Czech population because the clients of CPZP tend to be younger than average. However, the size of the cohort allows interesting analyses. The time data were blurred to week and year. The other dataset provided by the Professional Insurance Company (OZP) comprises 827,475 individuals, representing another 8% of the Czech population, the time data were blurred to month and year of events. The OZP dataset will serve for validation of the results obtained from the larger CPZP dataset. In all, we have complete data on mortality and vaccination of more than 2 million individuals over the entire period of 2021 and 2022. Both the datasets are available at https://github.com/PalackyUniversity/hve.

Methods of statistical analysis

Simple methods of descriptive statistics and data visualization were used. As total counts in full datasets were analyzed, no uncertainty indicators were applicable. As the majority of deaths (86% for CPZP, and 85% for the OZP) in our datasets occurred in the age category of >60 years and this age group was also the most likely to die from COVID (over 93% of COVID-associated deaths were in this age category) [6], we further focused only on this cohort and classified it further into the age brackets of [60, 70), [70, 80), and 80+.

We decided not to distinguish between the types of vaccines as the majority (approx. 80%) of all applied doses were Comirnaty and further stratification would yield too small numbers of other vaccines for reliable analysis. However, all recipients of the Janssen vaccine (43,603, 3.2% in CPZP and 21,712, 2.6% in OZP) were excluded as its vaccination scheme differed from the others. For each dataset, the cohort was stratified into the vaccination categories "Unvaccinated" and "less than 4 weeks/more than 4 weeks" after Doses 1, 2, and 3. Distinguishing the status of the less/more than 4 weeks from a vaccine dose was based on the recommendation to apply the second dose 4 weeks after the first one and further delay needed for reaching full immunity. ACM per 100,000 person-years was calculated for all categories (a) for the entire study period, (b) for periods of high COVID intensity, and (c) for periods of low COVID intensity.

Results

ACM according to the vaccination status

Figure 1 shows the ACM computed for the entire study period, i.e., January 2021-December 2022, for both datasets. It shows a remarkable pattern with a "higher mortality triangle" formed by unvaccinated individuals and individuals more than 4 weeks after each dose. This triangle is supplemented by markedly lower ACM among those who are less than 4 weeks after any dose of vaccine. Note that the pattern is very similar in all three age brackets and both datasets. Also, note that the combined extent of the datasets (more than 20% of the Czech population from two independent sources) guarantees that the pattern is not a statistical artifact.

At first sight, the figure might suggest that vaccination works remarkably well to prevent death. However, Figure 1 shows the *all-cause* mortality, not *COVID-related* mortality. Since only approx. 14% of all deaths over the study period were COVID-related (37,000 out of 269,000 deaths) [7,8], it was impossible for the vaccine to have had such an effect on *all-cause* mortality. The findings become even more paradoxical when periods of high and low COVID intensity are analyzed separately (Figure 2).

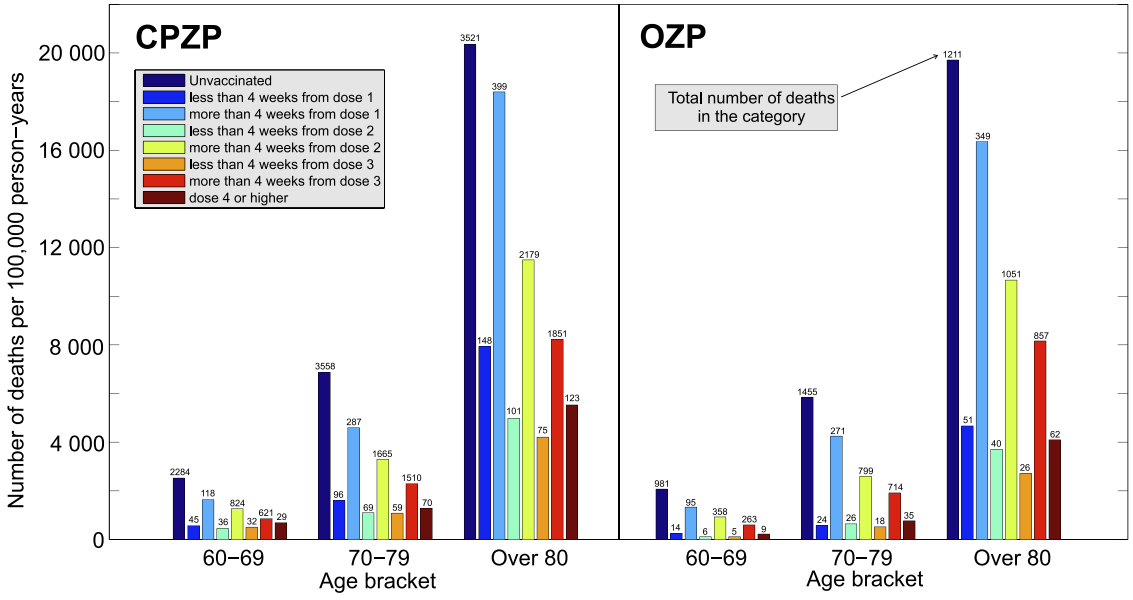

Figure 1. ACM stratified according to the vaccination status (color-coded) and age for both CPZP (left) and OZP (right) datasets over the entire study period January 2021-December 2022. In addition to all-cause mortality per 100,000 person-years, the total number of deaths in each category is shown by the figure above each bar.

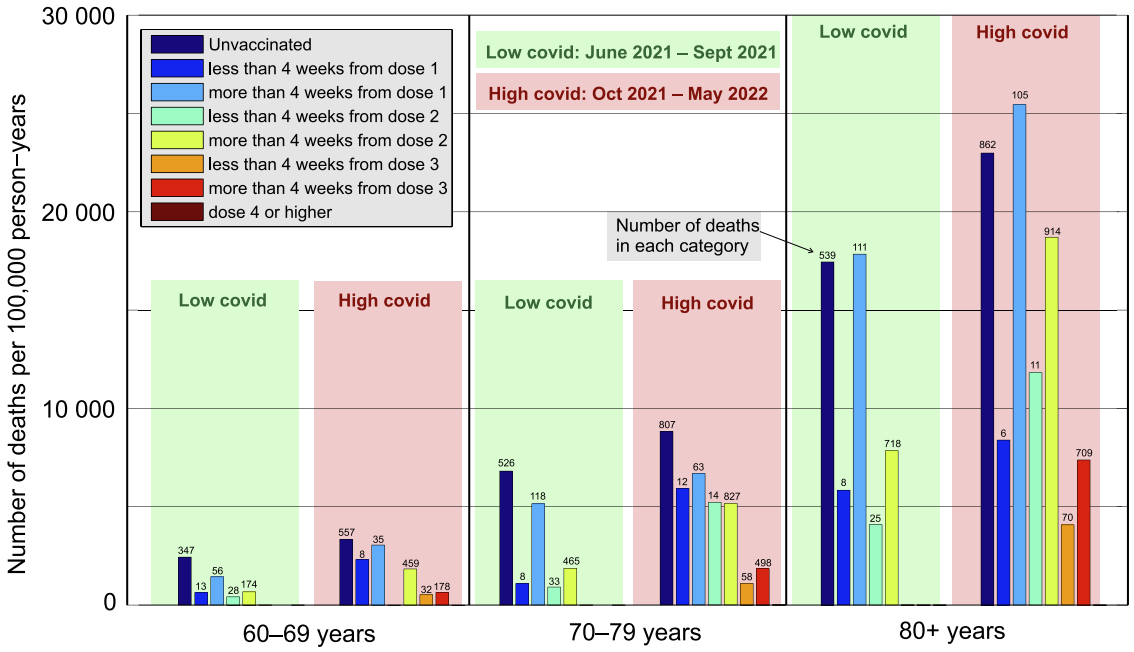

Figure 2. ACM in the CPZP cohort according to the vaccination status and age. Green panels: The period of very low COVID intensity June 2021-September 2021. Red panels: The period of high COVID intensity October 2021-May 2022. The figure above each bar indicates the total number of deaths in the respective category. Vaccination status is color-coded.

Between June 2021 and September 2021, virtually no COVID-related deaths were recorded in the Czech Republic (only approx. 0.3% of deaths were COVID-related). Thus, almost all the deaths shown in the green panels of Figure 2 were COVID-unrelated, although we can observe huge differences in ACM among groups in this period. Note the magnitude of these differences in the low-COVID period: the ACM of individuals 80+ years old who are "less than 4 weeks from Dose 2" (light green bars) is more than three times lower than that of the unvaccinated. In the case of the 70-80 age bracket, the difference is more than fivefold. When comparing the two largest groups in that period, i.e., unvaccinated (dark blue) and those with the completed primary course (yellow bars), the unvaccinated population was more than twice as likely to die as the population with the completed primary course. This apparent "vaccine efficacy" in a period when no COVID was present is likely an artifact of the HVE.

Now, let us focus on the red panels, covering the "high-COVID" period of October 2021-May 2022. In that period, the Czech Republic recorded almost 10,000 COVID-related deaths, which translates into an average of 40 deaths with/from COVID-19 per day. The vaccine effectiveness in preventing COVID-related deaths should lead to an increase in the ratio of unvaccinated:vaccinated ACM. However, the exact opposite happened. In all age groups, the ACM in the cohort with completed primary course (yellow bars) more than doubled compared to the low-COVID period, while the ACM of the unvaccinated cohort rose only by a third. A sanity check with the same analysis performed for the other dataset (OZP insurance company) yielded consistent results (see Supplement S6). This paradoxical observation can be caused by the fact that in the high-COVID period, vaccination by the third dose was underway, which, again, led to a subselection of the healthier group for booster dose vaccination, while infirm individuals concentrated in the group "primary course for more than 2 months."

Evolution of ACM over time

Much can be inferred from the evolution of ACM in time captured in Figure 3. We repeat the analysis of the CPZP data, but this time, ACM is evaluated separately for each quarter of 2021 and 2022. Only the oldest cohort is presented here, a detailed presentation of younger cohorts is shown in the Supplement Figure S5.

The ACM of those with the completed primary course of vaccination doses (orange line) starts in very low numbers, but as soon as the distribution of booster doses starts, their ACM increases and quickly reaches that of the unvaccinated population, while ACM of those who received the booster dose is minimal. Considering the proportion of COVID-related deaths in the quarters 4/21 (17%) and 1/22 (12%), this effect cannot be attributed solely to the protective effect of the vaccines as its magnitude is much greater. Rather, it further supports the notion that the "primary course" group split into the group who opted for the booster dose (the group with fewer frail individuals) and the group with a greater representation of the frail individuals who did not opt for the next dose. A more detailed analysis by month with all groups is shown in the Supplement—it shows these effects even more clearly, but is more difficult to read.

A simple model of HVE

To understand the mechanism behind the remarkable structure observed in Figures 1 and 2, and Supplement Figure S5, we prepared a simple model of how HVE would present in populational data.

Let us model a cohort of 170,000 individuals of which 17,000 die during the 104 weeks (2 years) of the follow-up. This roughly matches the group of CPZP clients who are older than 80 years (i.e., the cohort shown in Figure 3). For simplicity, the deaths are uniformly distributed throughout the 104 weeks and age is not considered in the model. The deaths are modeled first—for each individual, a bent coin is tossed with the probability of 1/10 to indicate if this individual dies within those 104 weeks (i.e., each death has a Bernoulli distribution with $P = 1/10$). The week of death is then selected randomly from the uniform distribution.

In the first run of the model (as a sanity check), the three doses of vaccines are distributed among the modeled population as follows: The first dose is given to 82% of the population. The week of

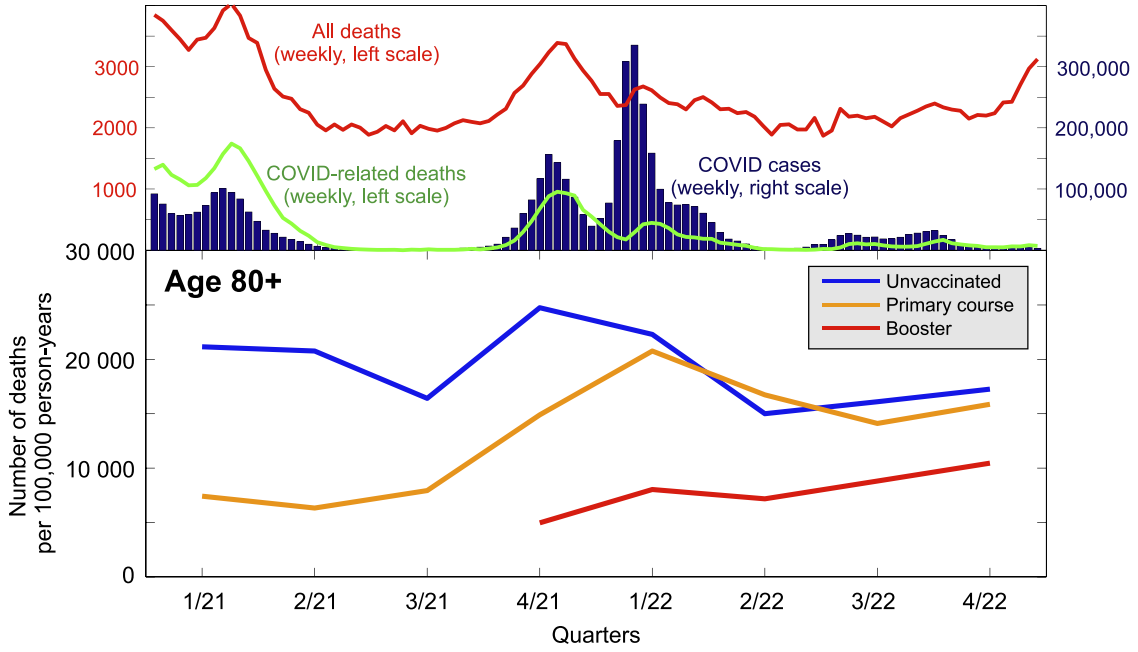

Figure 3. Bottom panel: The quarterly evolution of the ACM in the 80+ age cohort for the CPZP dataset. Upper panel: The context of the COVID epidemic in the Czech Republic. Blue bars represent the weekly numbers of positive PCR tests (right scale). The red line shows the weekly numbers of deaths from any cause, and the green line weekly numbers of COVID-related deaths (both left scale). See Supplement Figures S5 and S6 for more detail.

Dose 1 administration is derived from a normal distribution with a mean of 20 weeks and a standard deviation (SD) of 3 weeks. The second dose is given to 96% of those who obtained Dose 1. The lag between Dose 1 and Dose 2 is selected from a normal distribution with mean = 20 and SD = 3 weeks. The third dose is given to 82% of those who obtained Dose 2. The lag between Dose 2 and Dose 3 is again selected from a normal distribution with mean = 20 and SD = 3 weeks. The prevalence of the doses (82%, 96%, and 82%, respectively) roughly matches the observation from the CPZP cohort. In the event that the model assigns a dose to an individual later than at the time of their death, the dose is not administered. The distribution of the three doses in time was proposed to be as simple as possible with no intention of matching the real data. Results of this model run confirm that the vaccines are completely independent of death, i.e., the ACM in all vaccination categories is the same as illustrated in the left panel of Figure 4, which serves as a sanity check that the model is correct.

Now, let us add HVE into the model by assuming that individuals in poor health (who will die soon) have a lower probability of taking up the intervention—either because they are unable/unwilling to reach a vaccination site, or because of vaccine hesitancy. To account for this, the model was altered using the following condition: *If a vaccine dose is to be administered to an individual who will die within 26 weeks, the dose will be administered only with a reduced probability of (1-p)*, where p indicates the magnitude of HVE. This simple mechanism is implemented for all three doses and for three values of p: 0.25, 0.5, and 0.75 (see panels 2-4 in Figure 4). The interesting double-triangular pattern, similar to that observed in Figures 1 and 2, is readily visible. Note that it is very easy to implement HVE in this model because the deaths are modeled first (thus, at the time of vaccination, we already "know" who is going to die within 26 weeks).

We emphasize that the model has only two important parameters—the HVE duration (26 weeks) and the HVE magnitude (p). The size of the cohort and prevalence of the three doses were roughly matched to the study group. It is also worth noting that the observed large effects are caused by cancelling the administration of only 0.6-1.9% of doses due to HVE (depending on the HVE magnitude and population eligible for the respective dose).

Discussion

The results of the presented analysis revealed several peculiar patterns of the relationship between ACM and vaccination status. The presented data obviate that vaccination status has a profound association with ACM, which goes far beyond the possible protective effect against COVID-related death, especially in the low-COVID periods. Using a simple model, we argue that this pattern can be, to a large degree, attributed to the HVE.

Several explanations can be proposed to account for this effect. One possible explanation might suggest that long-term sequelae of COVID-19 led to excess mortality among unvaccinated individuals because they were more likely to have a history of COVID-19 infection than the vaccinated group. Nevertheless, should the long-term sequelae play a role, this (a) would have to be massive to explain the magnitude of the observed difference and (b) we should still see some excess mortality in the low-COVID period. This is, however, not the case as the total combined mortality of the vaccinated and unvaccinated populations in the low-COVID period in the entire Czech Republic remained more or less in line with previous years as corroborated by EUROSTAT data [9]. This "long-COVID" hypothesis can be, of course, responsible for a part of the observed effect but it is unlikely that it would lead to the differences in ACM of the observed magnitude.

Disregarding the low-COVID and high-COVID periods for now, we can see another evidence of HVE: the lowest ACM can be always observed in the groups vaccinated with a new dose for less than 4 weeks ("freshly vaccinated"), which indicates that the healthier part of the population took up the next dose of the vaccine, while those in poorer health concentrated in the group remaining with the unvaccinated/previous dose status.

The periods when successive vaccination campaigns for both primary and booster vaccination started coincided with high-COVID periods and it could be argued that it is difficult to accu-

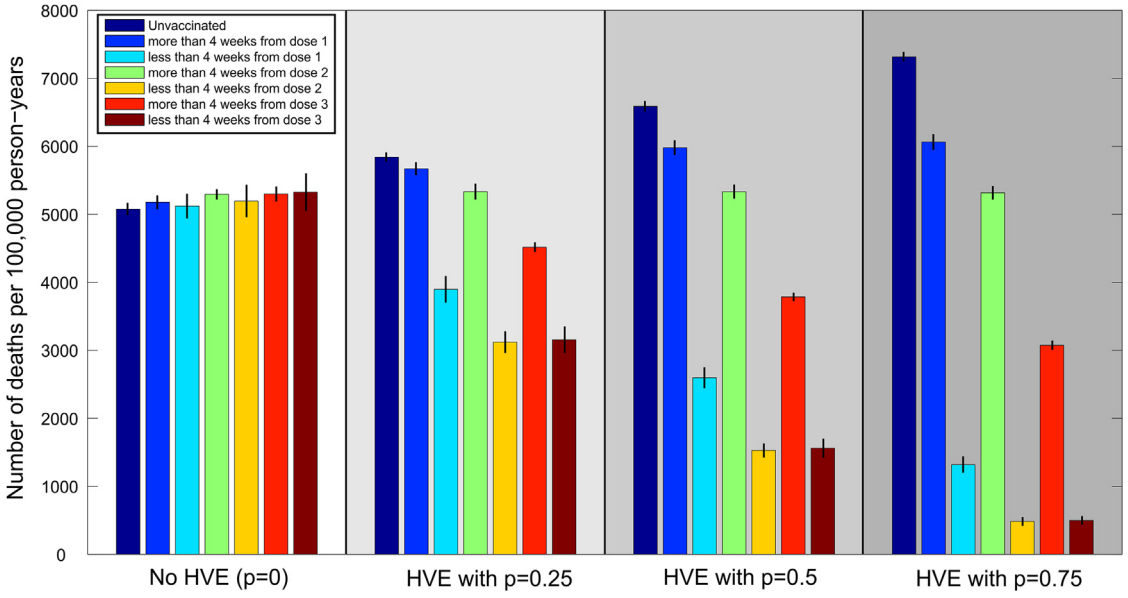

Figure 4. The results of the mathematical model of the healthy vaccinee effect (HVE). The bars show the average of 500 runs of the model, black line segments show the standard deviations. Left panel: Vaccination and all-cause mortality are independent (no HVE present). All-cause mortality in all categories is approximately the same. The bars are not identical due to the stochastic nature of the model. Panel 2 models a 25% HVE, i.e., if a vaccine dose is to be administered to an individual who will die within 26 weeks, the dose is administered only with a reduced probability of 0.75. The remaining two panels show 50% and 75% HVE, respectively. The resulting pattern qualitatively matches the observed paradoxical pattern in Figures 1 and 2.

rately distinguish between HVE and true vaccine-conferred protection. This is a rightful objection; on the other hand, the same pattern can also be seen in Dose 4 (Supplement Figure S5), which was released in a very low COVID period—the pattern of "freshly vaccinated" having the lowest ACM can be observed again. The "freshly vaccinated" groups also have the lowest mortality in low-COVID periods, even though vaccination could not have had any immediate true protective effect. Also, considering the share of COVID-related deaths on ACM, they cannot explain the observed effect size even in high-COVID periods. All of this supports the explanation that frail individuals (consistently over all doses, periods, age groups, and datasets) were less likely to take up the next dose of the vaccine.

It could be argued that HVE was highly unlikely at the beginning of vaccination when the most vulnerable groups (such as clients of elderly care homes) were prioritized for vaccination and practically all these individuals were vaccinated so no HVE was possible in these groups. This is very true; on the other hand, the proportion of these individuals in the entire population is very low and although these groups were at a disproportionally high risk of (not only) COVID-related mortality, their numbers were not sufficient to swing the all-populational trends completely. They, however, could have affected the results; this could be, for example, one of the explanations for the unvaccinated:vaccinated ACM ratio being higher in the low-COVID period than at the peak of the second COVID-19 wave (January-March 2021) when vaccination started and a vast majority of vaccinated individuals were infirm [5].

The HVE has not been much discussed in the scientific literature concerning COVID-19 vaccines so far. As far as we know, besides the only study focusing on this issue reported from Hungary [3], this issue was recently raised only in letters to the editors in the *New England Journal of Medicine* [4] and in our aforementioned letter [5]. However, to the best of our knowledge, the presented study provides the best and most robust illustration of the HVE in COVID-19 vaccination so far. The implications are huge—based on our results, we propose that the evaluation of the baseline frailty between vaccinated and unvaccinated populations (in our case, the differences observed in the low-COVID periods) should be taken into account when evaluating vaccine effectiveness in observation studies.

Limitations and strengths

Our datasets do not contain information on the cause of death, thus we do not know how many deaths in our datasets were COVID-related. On the other hand, this information is not necessary for illustrating the effects of the vaccination status on ACM. It is also necessary to note that the dependence of ACM on vaccination status found in our study does not imply that COVID-related deaths would follow the same pattern. Moreover, the HVE magnitude is probably closely associated with the percentage of unvaccinated individuals in the cohort. For these reasons, our results should not be used for direct recalculation of vaccine effectiveness against COVID-related death, especially not in other populations.

On the other hand, the use of ACM evades known biases found in observation studies dealing with COVID-19 vaccination, such as misclassification of the cause of death (due to COVID/from COVID), or uneven testing of vaccinated/unvaccinated populations.

We have no reason to believe that the two datasets contain significant errors. The data were released officially after a FOIA request. Both the datasets are very large and contain individual records—possible errors would be easy to find. Both the datasets are independent (the two companies are competitors), yet the patterns are almost identical. The data contain more than a fifth of the entire Czech population, which excludes the possibility of the observed effects being statistical artifacts. For these reasons, the data and conclusions drawn from it are highly robust.

Conclusion

On two independent datasets, we demonstrated a paradoxical pattern of strong association between COVID vaccination status and ACM, even in periods when almost no COVID-related deaths

were present in the population. Vaccinated individuals (especially those shortly after vaccination) exhibit much lower ACM than the unvaccinated, even in low-COVID periods. This pattern cannot be explained by the true effectiveness of the vaccines in preventing COVID-related deaths. We have demonstrated that the observed association can be explained by the HVE (a bias in which individuals of poorer health have a lower probability of taking up the vaccine/its further dose) and present a very simple model of HVE, which well replicates the pattern observed in the real data.

The above associations were demonstrated on two large independent datasets obtained from Czech health insurance companies that cover more than a fifth of the population. The datasets consisted of *individual* records for all clients with information on the week of birth and, if applicable, weeks of all COVID vaccination events, and of death from any cause. The data are very unlikely to contain errors or artifacts.

This study indicates that observation data on COVID-19 vaccine effectiveness must be interpreted with great caution as the baseline frailty of cohorts with different vaccination statuses may substantially differ due to HVE. Failure to account for HVE in observation studies basically invalidates any estimates of vaccine effectiveness in such studies.

Declarations of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: TF and JJ are members of the SMIS Association (Sdruzeni mikrobiologu, imunologu a statistiku/Association of the Microbiologists, Immunologists and Biostatisticians) in the Czech Republic. They have, however, never received any financial or other incentives that could bias this research and have no financial interests to disclose.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data sharing

Both datasets as well as the model are available at https://github.com/PalackyUniversity/hve.

Author contributions

ToF: Conceptualization, investigation, data analysis, visualization, data interpretation, methodology, writing—original draft, writing—review and editing; AB: Data acquisition, resources, project administration, writing—review and editing; TaF: Investigation, data analysis, software, validation, visualization, writing—review and editing; JJ: Literature search, data interpretation, writing—original draft, writing—review and editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijid.2024.02.019.

References

[1] Ioannidis JPA. Over- and under-estimation of COVID-19 deaths. *Eur J Epidemiol* 2021;**36**:581–8. doi:10.1007/s10654-021-00787-9.
[2] Fung K, Jones M, Doshi P. Sources of bias in observational studies of COVID-19 vaccine effectiveness. *J Eval Clin Pract* 2024;**30**(1):30–6. doi:10.1111/jep.13839.
[3] Pálinkás A, Sándor J. Effectiveness of COVID-19 vaccination in preventing all-cause mortality among adults during the third wave of the epidemic in Hungary: nationwide retrospective cohort study. *Vaccines (Basel)* 2022;**10**:1009. doi:10.3390/vaccines10071009.
[4] Høeg TB, Duriseti R, Prasad V. Potential "healthy vaccinee bias" in a study of BNT162b2 vaccine against COVID-19. *N Engl J Med* 2023;**389**:284–6. doi:10.1056/NEJMc2306683.
[5] Fürst T, Straka R, Janošek J. Healthy vaccinee effect—a bias not to be forgotten in observational COVID-19 vaccination studies. *Pol Arch Intern Med* 2024;**134**(2):16634. doi:10.20452/pamw.16634.
[6] Vencálek O, Beran J, Fürst T, Krátká Z, Komárek A. More analyses are needed to evaluate the effectiveness of protection by vaccines and previous infection against the omicron variant of SARS-CoV-2. *J Infect Dis* 2022;**226**(5):942–3. doi:10.1093/infdis/jiac257.
[7] Czech Statistical Office. Numbers of deaths—weekly and monthly reports (in Czech). https://www.czso.cz/csu/czso/obypz_cr; 2022. [accessed 20 September 2023]
[8] Our World in Data. Coronavirus pandemic (COVID-19). https://ourworldindata.org/coronavirus; 2023 [accessed 20 September 2023]
[9] EUROSTAT. Excess mortality—statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Excess_mortality_-_statistics; 2023 [accessed 20 September 2023]