



# The imprinting effect of covid-19 vaccines: an expected selection bias in observational studies

Susana Monge,<sup>1,2</sup> Roberto Pastor-Barriuso,<sup>1,3</sup> Miguel A Hernán<sup>4</sup>

<sup>1</sup>National Centre of Epidemiology, Institute of Health Carlos III, 28029 Madrid, Spain

<sup>2</sup>Consortium for Biomedical Research in Infectious Diseases (CIBERINFEC), Spain

<sup>3</sup>Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain

<sup>4</sup>CAUSALab and Departments of Epidemiology and Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA

Correspondence to: S Monge smonge@isciii.es (ORCID 0000-0003-1412-3012)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2023;381:e074404 <http://dx.doi.org/10.1136/bmj-2022-074404>

Accepted: 09 May 2023

Recent observational studies have found a higher risk of reinfection with the omicron variant of SARS-CoV-2 in people who received a third covid-19 booster dose. This finding has been interpreted as evidence of immune imprinting of covid-19 vaccines. This article proposes an alternative explanation: that the increased risk of reinfection in individuals vaccinated with a booster compared with no booster is the result of selection bias and is expected to arise even in the absence of immune imprinting. To clarify this alternative explanation, this article describes how previous observational analyses were an attempt to estimate the direct effect of vaccine boosters on SARS-CoV-2 reinfections—an effect that cannot be correctly estimated with observational data. Causal diagrams (directed acyclic graphs), data simulations, and analysis of real world data are used to illustrate the mechanism and magnitude of this bias, which is the result of conditioning on a collider.

Antigenic variation in the SARS-CoV-2 omicron variant and subvariants are substantial compared

with previous variants and covid-19 vaccines used until September 2022. The concerns are that past exposure to previous variants—through infection or vaccination—could alter the immunological response to an omicron related infection in such a way that the immune response to successive omicron infections would be impaired.<sup>1-3</sup> This so-called “immune imprinting hypothesis” has been used to suggest that a vaccine booster in individuals who later are infected with omicron increases the risk of a second omicron infection.<sup>4,5</sup> If this effect of immune imprinting truly exists, recommendations for additional vaccine doses may need to be re-evaluated.

The findings of recent observational studies in Qatar show an increased risk of reinfection with omicron in people vaccinated with three doses of monovalent vaccines compared with two doses,<sup>4</sup> and no increased risk of omicron reinfection in unvaccinated individuals.<sup>6</sup> These findings have been interpreted as supporting the immune imprinting hypothesis, which has raised concern among authorities in charge of vaccination policies worldwide.

Here we propose an alternative explanation to the findings of the observational studies: that the increased risk of reinfection in individuals vaccinated with a vaccine booster compared with no booster is the result of selection bias (owing to conditioning on a collider) and is expected to arise even in the absence of immune imprinting.

We start by clarifying that the causal question in immune imprinting studies is the direct effect of vaccine boosters on reinfections. A useful procedure to precisely articulate a causal question is to describe the hypothetical randomised experiment—the target trial<sup>7</sup>—that would answer it. Therefore, we first specify a target trial for the direct effect of vaccine boosters on reinfections. This target trial can be neither conducted nor, as we explain with causal diagrams, emulated with observational data. We then illustrate the bias with both simulated and real world data.

## Specification of the target trial

Consider a target trial with the following eligibility criteria for participation: aged 18 years or older; received the second dose of an mRNA vaccine at least 90 days previously but not yet received the third dose; no previous laboratory confirmed SARS-CoV-2 infection; and not part of a population with special vaccination requirements (eg, nursing home residents, healthcare workers). The intervention would have three components (fig 1). Firstly, eligible people would be randomly assigned to either immediate receipt of a booster (third dose) of an mRNA covid-19 vaccine, or no further vaccine doses. Secondly, all participants would have to remain uninfected for, say, six months

## SUMMARY POINTS

Recent observational studies found a higher risk of reinfection with omicron in people who received a third booster dose and then acquired a first omicron infection, and this finding was attributed to immune imprinting; however, this article shows that such a result is expected even if there is no imprinting. A target trial was first specified to precisely articulate the causal question: to identify the direct effect of a booster on a second omicron infection that was not mediated through the first omicron infection. Directed acyclic graphs, data simulations, and analysis of real world data were used to illustrate the mechanism and magnitude of the selection bias inherent to observational studies that aim to address such causal question. Bias arises when analyses are conditioned on a collider (a first omicron infection), which results in a non-causal association between the booster dose and the risk of a second omicron infection.

after randomisation. Thirdly, all participants would be intentionally infected with omicron at six months. The outcome of interest would be a laboratory confirmed omicron reinfection at least three months after the first omicron infection. (A variation would be to assign a random time of infection within, for example, six months after randomisation.)

Individuals would be followed from randomisation. However, because the outcome cannot occur (by definition) until nine months after randomisation, the cumulative incidence curves for both groups would stay at zero for the first nine months. Therefore, if no reinfection can truly occur during the first three months after a first infection and assuming no deaths or losses to follow-up occur, individuals can be equivalently followed from nine months after randomisation until the earliest of omicron reinfection or administrative end of follow-up (end of the study).

This target trial is not feasible because in the real world we cannot force people to become infected at a time of our choosing. But that is beside the point, because the aim of this thought experiment is to specify the causal question as unambiguously as possible, rather than to design an actual trial. If this target trial could be performed, we could use its data to quantify the (controlled) direct effect of a covid-19 booster on a second omicron infection that is not mediated through the first omicron infection by simply comparing the risk of reinfection between individuals assigned to booster and individuals assigned to no booster. If this direct effect exists, then immune imprinting has occurred.

#### Emulation of the target trial is not possible

When a target trial cannot be carried out, observational data are often used from human populations to emulate it.<sup>8-9</sup> In fact, the previous observational study<sup>4</sup> implicitly tried to emulate the target trial described in the previous section by adjusting for factors that may confound the effect of the booster on reinfection and by restricting the analysis to individuals with a first infection after the booster. Let us suppose that adjustment for confounding was successful and therefore the observational study appropriately accounted for the lack of randomised assignment of the booster. Even in that setting, the selection of individuals who had a first omicron infection is expected to introduce bias<sup>10</sup> because in the real world, an infection is expected to occur more

often among people with higher susceptibility.<sup>11</sup> Therefore, if the booster prevents infections, it is essentially guaranteed that people who received the booster and subsequently had a first infection are, on average, more susceptible to reinfection than people who did not receive the booster and subsequently had a first infection. In the absence of data on individual susceptibility, an observational study cannot unbiasedly estimate the direct effect of a booster because the restriction on first infection introduces selection bias.

To see this graphically, we use causal diagrams that are referred to as directed acyclic graphs.<sup>12-13</sup> The first graph in figure 2 represents the (randomised) target trial specified previously, and the second graph represents the observational analysis restricted to people with a first infection. Both graphs include the variables booster (yes or no), confirmed infection in period 1 (yes or no), confirmed infection in period 2 (yes or no), and an unmeasured “susceptibility” variable that represents individual characteristics that increase the risk of infection (eg, subclinical immunosuppression, occupational and behavioural factors) or of receiving a diagnosis of infection (eg, testing behaviour, access to the health system). Period 1 ends at the earliest of six months or a first infection, and period 2 starts right after the end of period 1.

In the target trial, there is no arrow from booster to infection in period 1, which, by design, occurs in all individuals, regardless of whether they were assigned to booster or no booster (fig 2). In addition, there is no arrow from susceptibility to first infection because, under the intervention of the target trial, infection is guaranteed regardless of individual susceptibility. Therefore, the unconditional association between booster and infection in period 2 is an unbiased estimator of the direct effect of booster on reinfection (the dotted arrow) because everybody had a first infection. The graph also represents a variation of the target trial in which individuals are randomly assigned to infection or no infection in period 1.

In the observational study, there are arrows into infection in period 1 from both booster, which prevents infections,<sup>14-15</sup> and susceptibility, which increases the risk of infection (fig 2). Also, when the observational study is restricted to people with infection in period 1, a box is placed around infection in period 1 to represent

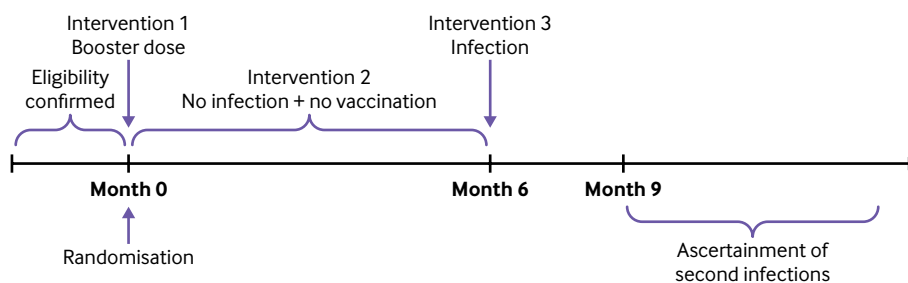


Fig 1 | Design of a hypothetical target trial to evaluate the direct effect of a covid-19 vaccine booster dose, not mediated through first infection, on risk of reinfection

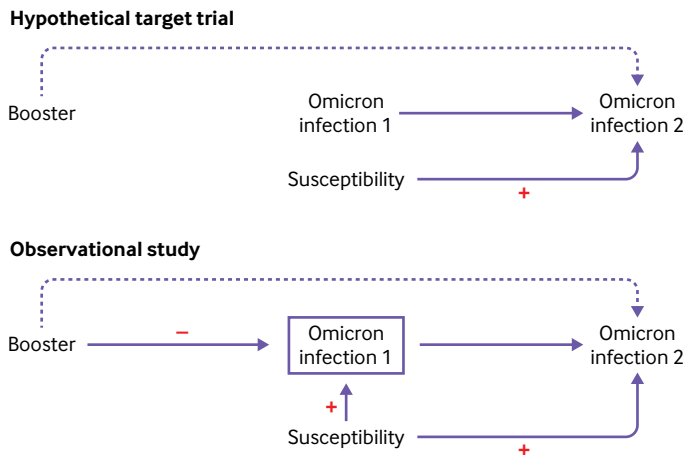


Fig 2 | Simplified causal diagrams representing a hypothetical target trial in which participants are randomly assigned to booster (yes or no) and then intentionally infected with omicron, and an observational study. The dotted arrow represents the direct effect of the booster, not mediated through first infection, on risk of reinfection. The boxed text represents the conditioning of the analysis on individuals with value “yes”

the conditioning of the analysis on individuals with value “yes” on that variable. In graph theory, infection in period 1 is regarded as a collider because it is a common effect of booster and susceptibility.<sup>16</sup> Therefore, restricting to those with infection in period 1 equal to “yes” is a form of collider stratification, which is expected to induce a non-causal association between booster and susceptibility, and, because susceptibility is associated with second infection, between booster and second infection.<sup>10 13 16</sup> That is, the association between booster and reinfection among those with a first infection combines the direct effect of booster on reinfection (the dotted arrow), if any, and the selection bias induced by conditioning on first infection.

If the booster had no direct effect (ie, if the dotted arrow did not exist) then the risk of infection in period 2 would be expected to be greater for those who received the booster than for those who did not receive the booster. This higher risk in the booster group compared with the no booster group is entirely the result of selection bias and thus has no causal interpretation as a harmful effect of the booster on reinfection. In fact, all that this increased risk indicates is that those who become infected despite receiving a booster are people who are more susceptible to reinfection.

#### Replication of the selection bias in simulated data

We designed a simplified simulation to quantify the magnitude of the selection bias under the causal directed acyclic graph for the observational study in figure 2 (without an arrow from first to second infection, for simplicity). We simulated a dataset of 10 million people with a normally distributed susceptibility variable, of whom 65% were randomly assigned to booster. We assumed that the booster decreased the probability of infection in period 1 by 50% and had no effect on the probability of infection

in period 2—that is, the booster had no direct effect on reinfection. We considered separate scenarios in which the risk of infection is increased between two and eight times per each standard deviation in susceptibility (see supplementary file for details and computer code).

For a realistic risk of infection of 10% in period 1, the odds ratio of infection in period 2 for booster compared with no booster ranged between 1.04 and 1.37, depending on the assumed distribution of susceptibility to infection in the population. As expected, restricting the observational analysis to individuals with a first omicron infection results in a higher risk of reinfection in the booster group even if the booster had zero effect on reinfection. It is all selection bias.

#### Replication of the selection bias in real world data

We conducted the observational analysis using linked individual level data from three Spanish population registries (vaccination registry (REGVACU), laboratory results registry (SERLAB), and national health system registry).<sup>15</sup>

We identified individuals eligible for the target trial starting on 1 January 2022, when omicron accounted for >90% of the circulating variants in Spain. We then assigned those who received a booster to the booster group, and for each of them we randomly chose a matched control who had not received a booster up to that week. The matching factors included sex, age ( $\pm 5$  years), province, time since primary vaccination ( $\pm 14$  days), and type of primary vaccine (BNT162b2 (Pfizer-BioNTech) or mRNA-1273 (Moderna)).

In an attempt to emulate a target trial in which all participants are infected with omicron within 10 months of booster assignment, we restricted the analysis to individuals with a laboratory confirmed SARS-CoV-2 infection in the next 10 months (and further matched eligible individuals on week of infection)—that is, we only considered for the analysis individuals with an omicron infection, and further matched each individual in the booster group with the control individual on week of infection. As mentioned, restricting to individuals with infection induces uncontrollable selection bias because of differential selection between groups that depends on the largely unknown individual susceptibility.

We then followed individuals from day 90 after infection until the earliest of a confirmed SARS-CoV-2 reinfection, death, discontinuation of registration in the national health service database, or administrative censoring at end of the study (31 October 2022). To estimate the per protocol effect, we censored at receipt of any additional vaccine dose. The cumulative incidence (risk) in each group was estimated using the Kaplan Meier method<sup>17</sup> and compared between groups through risk ratios. Bootstrapping with 500 samples was used to compute centile based 95% confidence intervals.

Of 12 749 506 initially eligible individuals, 1 704 904 experienced a first infection in the study period. Of these, 425 741 (25.0%) had received a booster dose

before the infection. We could exactly match 249 226 (58.5%) individuals with a booster to the same number of controls, with a median age of 44 years. A total of 201 266 (80.8%) matched pairs remained under follow-up 90 days after the infection and were included in the analysis. During a maximum follow-up of 211 days (mean 133 days) 1794 reinfections occurred, with a six month risk over the full period of 0.59% in the booster group and 0.54% in the control group. The risk ratio of reinfection in the booster group compared with the no booster group was 1.08 (95% confidence interval 0.97 to 1.20) at six months of follow-up (nine months post-infection), but varied between 1.03 (0.93 to 1.17) in days 0-90 of follow-up and 1.20 (0.98 to 1.45) in days 91-180. As expected, the booster was associated with a higher risk of reinfection.

### Discussion

We have used causal diagrams and simulations to explain that recent observational estimates of an apparently greater risk of reinfection with the omicron variant after a covid-19 booster vaccine dose can be fully explained by the selection bias that arises when restricting the analysis to individuals with a previous omicron infection. We further illustrated the selection bias by conducting a real world analysis of nationwide data from Spain. Our estimates of increased risk of reinfection in individuals infected after receiving a booster were compatible with our simulation results and comparable with those from previous observational studies.<sup>4</sup>

Removing the selection bias would require the measurement, and adjustment for, individual susceptibility to infection or diagnosis. Unfortunately, this information is not available. Comorbidities or health seeking behaviour are unlikely to fully capture individual susceptibility, and, in fact, studies accounting for some of these measured factors<sup>4</sup> did not provide different estimates to the ones in our study that accounted only for age, sex, location, and type of vaccine. Notably, even though the direct effect of the booster on reinfection cannot be validly estimated and thus the presence of immune imprinting cannot be assessed with the available data, the most relevant effect to guide decision making in vaccination programmes is the total effect of the booster on the risk of infection, which has been estimated in several observational studies.<sup>14 15 18</sup>

Other sources of bias may of course exist in observational studies of vaccine effectiveness, including confounding from incomplete adjustment for prognostic factors associated with vaccination, and measurement error from incomplete ascertainment of SARS-CoV-2 infection.<sup>19</sup> Here we focused on the selection bias that is expected to arise in any analyses that condition on post-vaccination infection, including analyses of both observational data and randomised trials that cannot intervene on infection itself.

### Conclusions

Analyses of observational data require a precise articulation of the causal question before the estimates

can be interpreted. Observational analyses to estimate the direct effect of a booster on the risk of reinfection (ie, imprinting) failed to specify the target trial that they were trying to emulate. As a result, an increased risk of reinfection among individuals who received a booster and had a first post-booster infection was incorrectly interpreted as showing a harmful effect of the booster. An explicitly causal approach to these questions indicates that the increased risk is mathematically expected and may be fully explained by selection bias, and observational data may not be generally used to answer these questions about imprinting.

We acknowledge the contribution of those who make it possible to have real time data on covid-19 vaccination and laboratory tests available in Spain, including professionals in the 19 autonomous communities and cities, the Vaccines Division and Health Information Systems Department of the Ministry of Health, and the National Centre of Epidemiology at the Institute of Health Carlos III.

**Patient and public involvement:** It was not appropriate or possible to involve patients or the public in the design, conduct, reporting, or dissemination plans of this research.

**Contributors:** MAH and SM conceived the study and simulations. RP and SM performed the simulations. SM performed the analyses. SM is the guarantor. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** This study received no specific funding.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at <https://www.icmje.org/disclosure-of-interest/> and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. MH is data science adviser for ProPublica, Advisory Board Member of ADIA Lab, and consultant for Cytel.

**Ethical approval:** The use of the national health service database, REGVACU, and SERLAB for the purpose of monitoring vaccine effectiveness has been approved by the research ethics committee at the Instituto de Salud Carlos III (CEI PI 98\_2020 and CEI PI 08\_2022). Informed consent was not required because this study is based on national population registries.

**Data sharing:** The databases used in the observational data analysis are owned by the Ministry of Health and the Autonomous Communities in Spain, which establish the requirements for their access and use.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

- 1 Reynolds CJ, Pade C, Gibbons JM, et al. Immune boosting by B.1.1.529 (Omicron) depends on previous SARS-CoV-2 exposure. *Science* 2022;377:eabq1841.
- 2 Röltgen K, Nielsen SCA, Silva O, et al. Immune imprinting, breadth of variant recognition, and germinal center response in human SARS-CoV-2 infection and vaccination. *Cell* 2022;185:1025-1040.e14. doi:10.1016/j.cell.2022.01.018
- 3 Hoffmann M, Behrens GMN, Arora P, et al. Effect of hybrid immunity and bivalent booster vaccination on omicron sublineage neutralisation. *Lancet Infect Dis* 2023;23:25-8. doi:10.1016/S1473-3099(22)00792-7
- 4 Chemaitelly H, Ayoub HH, Tang P, et al. COVID-19 primary series and booster vaccination and immune imprinting. *medRxiv* 2022;2022.10.31.22281756. doi:10.1101/2022.10.31.22281756.
- 5 Chemaitelly H, Ayoub HH, Tang P, et al. Long-term COVID-19 booster effectiveness by infection history and clinical vulnerability and immune imprinting: a retrospective population-based cohort study. *Lancet Infect Dis* 2023;S1473-3099(23)00058-0.
- 6 Chemaitelly H, Ayoub HH, Tang P, et al. Immune Imprinting and Protection against Repeat Reinfection with SARS-CoV-2. *N Engl J Med* 2022;387:1716-8. doi:10.1056/NEJMc2211055
- 7 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol* 2016;183:758-64. doi:10.1093/aje/kwv254
- 8 Hernán MA. Methods of Public Health Research - Strengthening Causal Inference from Observational Data. *N Engl J Med* 2021;385:1345-8. doi:10.1056/NEJMp2113319

- 9 Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA* 2022;328:2446-7. doi:10.1001/jama.2022.21383
- 10 Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615-25. doi:10.1097/01.ede.0000135174.63482.43
- 11 Hougaard P. Frailty models for survival data. *Lifetime Data Anal* 1995;1:255-73. doi:10.1007/BF00985760
- 12 Pearl J. *Causality*. Cambridge University Press, 2009. doi:10.1017/CBO9780511803161.
- 13 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37-48. doi:10.1097/00001648-199901000-00008
- 14 Abu-Raddad LJ, Chemaitelly H, Ayoub HH, et al. Effect of mRNA Vaccine Boosters against SARS-CoV-2 Omicron Infection in Qatar. *N Engl J Med* 2022;386:1804-16. doi:10.1056/NEJMoa2200797
- 15 Monge S, Rojas-Benedicto A, Olmedo C, et al, IBERCovid. Effectiveness of mRNA vaccine boosters against infection with the SARS-CoV-2 omicron (B.1.1.529) variant in Spain: a nationwide cohort study. *Lancet Infect Dis* 2022;22:1313-20.
- 16 Hernán MA, Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
- 17 Kaplan ELMP. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457-81. doi:10.1080/01621459.1958.10501452.
- 18 Barda N, Dagan N, Cohen C, et al. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. *Lancet* 2021;398:2093-100. doi:10.1016/S0140-6736(21)02249-2
- 19 Barda N. The ups and downs of observational vaccine research. *Lancet Infect Dis* 2023;S1473-3099(23)00119-6.

**Supplementary information:** Details of simulation, R script, and analysis with real world data