ANTIGENIC EVOLUTION AND THE HETEROLOGOUS HUMORAL IMMUNE RESPONSE TO SEASONAL INFLUENZA VACCINATION

by

WESLEY ZANE BILLINGS

(Under the Direction of Andreas Handel)

ABSTRACT

Introduction: Annual vaccines are the primary intervention for limiting the burden of influenza disease, but vaccine effectiveness (VE) depends on many factors, including how similar the strains used for vaccine formulation are to circulating influenza strains. Thus, ensuring that the vaccine induces a broadly-reactive immune response to many possible strains of influenza is paramount for improving VE. The heterologous immune response to strains other than those used in vaccine development is understudied and difficult to measure.

Methods: We use data from a prospective, open annual influenza cohort across three study sites (Port St. Lucie, FL; Pittsburgh, PA; and Athens, GA) from Fall 2013 through Spring 2022. Each individual could repeat in subsequent years, but was not required to maintain consecutive enrollment. Every year, individuals contribute a pre-vaccination blood draw, receive a Fluzone vaccine (Sanofi) and contribute a post-vaccination blood draw at a followup visit targeted either 21 or 28 days post-vaccination. All serum samples were used for hemagglutination inhibition (HAI) assays to a wide panel of homologous and heterologous historical influenza strains. We fit Bayesian multilevel models to analyze the heterologous immune response and breadth of vaccine response.

Results: First, we quantified the causal effect of vaccine dose (standard dose vs. high dose Fluzone) on the heterologous immune response. We found that higher dose improved the response to some strains, but not all, so higher dose can induce a less broad vaccine response. In a followup study, we were able to generalize our strain-specific predictions to predictions about antigenic distance using linked genomic data. Finally, we developed an overall summary metric for breadth of vaccine response which had substantially lower variation across data from different labs than current metrics.

Conclusions: Many factors such as vaccine dose may modulate the heterologous immune response in unintuitive ways, so breadth must be carefully assessed. Breadth can be assessed using multiple measures of antigenic distance, but our novel summary metrics provide a robust measurement for understanding the breadth of response of vaccine candidates across different lab groups.

INDEX WORDS: Influenza, Influenza vaccine, High-dose vaccines, Heterologous immunity, Humoral immunity, Antigenic distance, Cohort study, Secondary data analysis

ANTIGENIC EVOLUTION AND THE HETEROLOGOUS HUMORAL IMMUNE RESPONSE TO SEASONAL INFLUENZA VACCINATION

by

WESLEY ZANE BILLINGS

B.S., Western Carolina University, 2020

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2025

ANTIGENIC EVOLUTION AND THE HETEROLOGOUS HUMORAL IMMUNE RESPONSE TO SEASONAL INFLUENZA VACCINATION

by

WESLEY ZANE BILLINGS

Major Professor:	Andreas Handel
Committee:	Amy K. Winter
	Ye Shen
	Natalie E. Dean

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2025

Dedication

For my Mother.

Acknowledgments

I thank my family and friends for their continued support.

I thank my coauthors and contributors for their excellent work.

In addition to the coauthors and contributors for each chapter, my dissertation would not have been possible without my colleagues Rachel Baugh, Tzu-Chun Chu, Caroline Harpe, Nichole McCorkle, Brian McKay, Savannah L. Miller, Meng-Hsuan Sung, Bingyi Yang, and Veronica Zarnitsyna. Thank you!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Influenza rapidly evolves to escape current vaccines through two major mechanisms: antigenic drift and shift [1, 2]. Antigenic drift is the gradual process of mutation, driven by selective pressure. Antigenic shifts are sudden and abrupt changes in influenza antigens, which occur by recombination with other strains [3].

Seasonal epidemic influenza has a substantive burden [4, 5], especially in children and older adults [6, 7]. Vaccine effectiveness (VE; a measurement of how protective the vaccine is) varies widely, and is typically lower when strains used in the seasonal vaccine do not match circulating strains [8–10]. Even when strains are correctly matched, VE rarely exceeds 50%. The current strategy of predicting circulating strains to determine the formulation of the influenza vaccine also leads to low pandemic preparedness. The spontaneous emergence of new influenza strains, as with the 2009 H1N1 pandemic [11] or the highly pathogenic H5N1 spillover cases in early 2023 [12] also demonstrate the need for a broadly protective influenza vaccine. A universal broadly-reactive vaccine, which protects against current and future influenza variants has the potential to reduce the burden of seasonal influenza and mitigate future pandemics. Unfortunately, designing influenza vaccines that induce broad protection against many relevant variants of influenza is challenging [13–15].

This dissertation aims to improve our understanding of the design of broadly-

reactive vaccines by improving the methodology used in the analysis of vaccine studies. While there are many promising next-generation broadly-reactive influenza vaccines currently being developed, the best ways to measure breadth of response and compare broadly-reactive vaccine candidates are unclear. We first conducted an analysis of a high-dose split-inactivated vaccine, an enhanced vaccination strategy targeted to protect elderly adults at risk of influenza complications. After finding that breadth of response is not uniform across many strains of influenza, we investigated the methods used for assessing response breadth and the heterologous immune response. Finally, we developed a novel method for robust comparison and evaluation of broadly-reactive vaccine candidates.

Summary of objectives

The goal of this dissertation is to improve our understanding of the heterologous humoral immune response, and the breadth of immune response to seasonal influenza vaccines. The specific aims of our study were as follows.

- **Aim 1 (Chapter 2).** Quantify the causal effect of vaccine dose on the breadth of the vaccine-induced immune response.
- **Aim 2 (Chapter 3).** Evaluate the differences between different antigenic distance metrics when measuring breadth of vaccine response.
- **Aim 3 (Chapter 4).** Develop robust metrics for the quantification of the total immune response to an influenza vaccine, incorporating both strength and breadth.

Chapter 2

The role of vaccine dose is so important in vaccine response that determining a dose that balances efficacy and side effects is a crucial part of drug approval in the United States [8, 13, 16]. High-dose (HD) influenza vaccines are approved for use

in older adults and can substantially improve the immune response for older and otherwise immunocompromised individuals [17–19] and reduce disease severity and risk of complications [20, 21]. While several studies have shown that HD vaccines induce a stronger homologous immune response (that is, an immune response to the strains used for vaccine formulation), no previous studies have examined how HD vaccination impacts the heterologous response.

Inducing strong cross-reactive responses is crucial for both epidemic and pandemic preparation, since we cannot perfectly predict which influenza strains will circulate [8, 22]. The HD vaccine could broaden the vaccine-induced immune response, or it could induce a stronger, narrower response to the vaccine antigens [23], which could have deleterious effects on the immune repertoire of the recipient that are not realized until a future season when a novel strain evolves.

We performed a secondary data analysis of a prospective human vaccination cohort, which included serological assays pre- and post-vaccination of a number of historical influenza strains using Bayesian multilevel modeling to adjust for causal confounders. We found that while HD vaccines tended to be beneficial in each season, they can induce a worse response to some historical influenza strains than the SD vaccine, potentially suggesting a narrowing of the immune response. If similar phenotypes to those historical strains emerge in future lineages of influenza, HD vaccines could induce worse protection, similar to the effect of the 2009 seasonal vaccine on pandemic swine flu immunity [24].

Chapter 3

We analyzed the effect of vaccine dose on heterologous immunity using a strain-specific model, which is not generalizable to other influenza strains. Quantifying the degree of difference between influenza strains is important for understanding the speed of genetic and antigenic evolution, and for understanding patterns of cross-reactivity in the

heterologous immune response. The antigenic distance between two influenza strains is a measurement of how cross-reactive the immune response to those two strains is expected to be [25–28]. Strains which have evolved to be highly divergent from each other should have a high antigenic distance and induce effectively independent immune responses, while strains with similar antigenic phenotypes have low antigenic distance and can induce cross-reactive immune responses.

Measuring antigenic distance is difficult, and there are many methods in the literature, none of which is clearly superior. The most prevalent method is antigenic cartography, which involves statistical dimension reduction on large panels of serological data [29]. However, methods based on analyzing the amino acid sequences of influenza proteins also appear to be effective [30–32] and do not require expensive and time consuming serological panels. While previous research has shown only moderate agreement between different measures of antigenic distance [33–35], previous analyses have only examined antigenic distance measures in a vacuum, without considering the strength of the immune response at the same time.

We modeled vaccine response using multiple different antigenic distance metrics, and we compare the predicted vaccine response instead of only comparing distance metrics in isolation. We found that despite low agreement in antigenic distance metrics, the metrics produced similar predictions of vaccine response after controlling for confounding. Our results indicate that sequence-based distance metrics are a viable alternative to antigenic cartography in the analysis of vaccine breadth, and can avoid several limitations of antigenic cartography.

Chapter 4

Finally, we aimed to use antigenic distance to improve evaluation and comparison schema for broadly-reactive vaccine candidates. Previous research on broadly-reactive vaccines use a simple method to assess response breadth: they test the immune

response to many historic strains after vaccination, and count the number of strains with a clinically meaningful response [36–38]. This method is fast and easy, but is heavily biased by the set of historical viruses used in the panel. Results cannot be compared fairly between labs which use different historical panels. Requiring all labs to use the exact same set of historical strains in every analysis would be inexpensive and logistically infeasible. Missing data imputation methods are potentially useful for estimating the response to strains which were not measured [39, 40], but these methods make strong assumptions and have not been shown to be practically useful for vaccine comparison.

The more recent development of antibody landscape analysis provides a potential alternative to missing data imputation. Antibody landscapes are created by modeling some metric of vaccine response as a function of antigenic distance. These landscapes are agnostic to the choice of immune assay or outcome [41–44] and to the metric used for antigenic distance [45, 46]. Antibody landscapes are also agnostic to the choice of method for estimating the relationship between vaccine response and antigenic distance – linear or piecewise linear models are common, but flexible models that account for covariates or nonlinear effects can also be used to estimate the antibody landscape.

We developed a set of metrics derived from the summary antibody landscape over a study sample. Incorporating antigenic distance from the antibody landscape into our summary statistics makes our metrics more robust to the choice of historical strains used for analyzing the response to a vaccine candidate, and our metrics can be fairly compared across different virus panels.

Chapter 2

High dose inactivated influenza vaccine inconsistently improves heterologous antibody responses in an elderly human cohort

Billings WZ, Ge Y, Knight JK, et al. 2025. *The Journal of Infectious Diseases*. Online ahead of print, jiaf003. Reprinted here with permission of the publisher.

Abstract

Background: Older adults often mount a weak immune response to standard inactivated influenza vaccines. To induce a stronger response and better protection, a high-dose (HD) version of the inactivated Fluzone vaccine is recommended for individuals >65 years of age. While better immunogenicity and protection against the vaccine strain has been shown, it is not known if the HD vaccine also induces a robust antibody response to heterologous strains.

Methods: We fit bayesian multilevel regression models to hemagglutination inhibition (HAI) antibody data from an influenza vaccine cohort spanning the 2013/14-2021/22 influenza seasons. We used this model to estimate the average causal effect (ACE) of obtaining the HD vaccine, relative to the SD vaccine.

Results: We show that while there is generally a benefit derived from the HD vaccine, the impact is small and inconsistent. For some strains, the HD vaccine might even result in less robust heterologous responses.

Conclusions: We suggest that further increases in dose might be worth investigating to help induce a stronger broad response.

Keywords: influenza, high-dose influenza vaccine, heterologous responses

Introduction

Influenza vaccines are an important tool for reducing the burden of seasonal influenza, but the average effectiveness is often less than 50% [47]. Standard dose (SD) split-inactivated vaccines, such as Sanofi Pasteur's Fluzone, comprise the majority of vaccines licensed in the US [48], and are formulated to contain 15 micrograms (µg) of influenza hemagglutinin (HA) antigen protein for each strain of influenza included in the vaccine [49]. Influenza vaccines often induce a weak response in elderly individuals [50, 51]. In response, Sanofi Pasteur developed a high dose (HD) formulation of the vaccine, Fluzone HD, which contains 60 µg of HA per strain in one dose [52, 53].

In elderly individuals, HD vaccines induce a stronger homologous antibody response compared to SD vaccines to the influenza strains contained in the vaccine [20, 54, 55]. Additionally, HD vaccination is associated with reduced disease severity and reduced risk of complications in elderly individuals who contract influenza after vaccination [20, 21]. HD vaccines also have the potential to elicit stronger immune responses in younger individuals [56], but younger individuals can develop protective immune responses with SD vaccine, and even fractional doses of SD vaccine [57, 58]. Therefore, the HD vaccine is currently only recommended for elderly individuals.

While several studies have shown the ability of HD vaccines to induce stronger antibody responses to the HA contained in the vaccine, whether HD vaccines induce a stronger heterologous antibody response (cross-reactive antibodies to strains not included in the vaccine formulation) is uncertain [59, 60]. Since our ability to forecast which strains of influenza will circulate in the upcoming season is imperfect, it is important for influenza vaccines to induce both homologous and heterologous responses [8, 22].

It is not obvious if HD vaccines are expected to have a positive, negative or no effect on heterologous responses. A higher dose might stimulate multiple diverse lineages of memory B-cells, and could enhance both the strength and breadth of protection. However, a higher amount of HA proteins to the homologous strains could induce a stronger homologous response, which might out-compete and dominate responses to any cross-reactive or novel

antigens, resulting in less robust heterologous responses. The potential competition among HD vaccine-induced B-cells or T-cells might lead to more narrow protection from vaccination [23], making high-risk populations more vulnerable to vaccine mismatches.

Improving our understanding of the heterologous immune response to influenza vaccination remains a critical step in developing a broadly reactive influenza vaccine. In our study, we compare the antibody response between HD and SD vaccine recipients in a vaccine cohort study, using a panel of several historical influenza A strains. We found that while the HD response was higher for most strains, the impact was not consistent and the effect size was small, suggesting that HD vaccines could be further improved.

Methods

Data source

We used data from an ongoing human vaccination cohort study, which has been previously described in detail [61–63]. The study is a prospective open cohort which allows individuals to enroll in multiple years (potentially non-consecutively), and has been conducted across three study sites. Investigators annually recruited individuals who had not yet received an influenza vaccine in the current season (the influenza season refers to the Northern Hemisphere fall and winter in the United States, typically ranging from October to May [64]). At intake, individuals provided demographic information, and investigators collected blood samples before administering a vaccine. Individuals aged 65 and older were given the choice between Fluzone SD and Fluzone HD, while individuals under 65 were given the SD vaccine. Individuals were asked to return for post-vaccination blood draws either 21 days (2013/14 through 2017/18 seasons) or 28 days (2018/19 season onward) after the initial visit. Investigators ran hemagglutination inhibition (HAI) assays the vaccine strains and several historical influenza strains for each serological sample.

For our analysis, we extracted previously deidentified records for individuals aged 65 and older from 2013/2014 influenza season through the 2021/2022 influenza season. All individuals age 65 and older who provided both pre-vaccination and post-vaccination blood

samples were included in our analysis.

Data processing

The raw data for our study were reciprocal HAI titer values produced by the HAI dilution assays described in the previous studies, along with limited demographic information for each participant (study site, season, age, sex, race, and a numeric identifier to track individuals across participation years). Since the SD vaccine was quadrivalent (containing two influenza B lineages) while the HD vaccine was trivalent (containing only a single influenza B lineage) for multiple years of the study, we elected to focus only on the heterologous response to influenza A virus strains, and excluded all data for influenza B strains. Details on the strains included in each vaccine formulation and the number of repeated individuals are shown in the supplement.

The data represent a multilevel structure, where each study site recruited individuals, each individual could participate in multiple seasons, in each season an individual had a pre-vaccination and a post-vaccination serological sample, and investigators ran a panel of HAI assays to multiple strains on each serological sample. The HAI assays had a lower limit of detection (LoD) of 10 and an upper LoD of 20480. Values below the lower LoD were coded as 5 in the raw data. No values in our dataset were at the upper LoD. Following previous studies, [65, 66] we conducted all analyses on the log scale, using the transformation

$$y = \log_2\left(\frac{\text{raw reciprocal HAI titer}}{5}\right),$$

where y is the titer variable we analyze. Using this transformation puts the titers on a scale of 0, (values below the lower LoD), $1, 2, \ldots, 12$ (values at the upper LoD).

Outcome definitions

We calculated four common outcomes used in influenza vaccine studies [53, 65]. The primary outcome we used for our study was titer increase, defined as the log (base 2) fold change between an individual's post-vaccination and pre-vaccination titer. That is, since the y

variable is already on the log scale,

$$\text{titer increase} = y_{\text{post}} - y_{\text{pre}}.$$

We present results for titer increase in the main manuscript. As a sensitivity analysis, we repeated our analysis using three additional outcomes: post-vaccination titer, seroprotection, and seroconversion. Seroprotection and seroconversion are binary outcomes, and this dichotomization leads to a loss of statistical power. However, these are commonly reported measures and can be useful for their clinical interpretation. We used standard definitions for seroprotection and seroconversion, with seroprotection defined as a post-vaccination titer greater than 1:40, that is,

$$\text{seroprotection} = I(\text{post-vaccination HAI titer} \geq 40),$$

where $I(\cdot)$ is the indicator function. The definition of seroconversion we used was seroprotection in addition to a titer increase of 2 or higher (i.e., a fold change of 4-fold or higher), that is,

$$\text{seroconversion} = \text{seroprotection} \ \times \ I(\text{titer increase} \geq 2).$$

Results for these three additional outcomes are discussed in our results, and the full details are included in the Supplement.

Statistical analyses

For an initial descriptive analysis of the study population, we calculated summary statistics for the covariates we included in our models and for each of the model outcomes previously described. Summary statistics for the outcome variables, including analyses stratified by vaccine strain and assay strain, are shown in the Supplement. We conducted a crude analysis of the difference in fold change by dose for each vaccine strain and assay strain, also shown in the Supplement.

For our main results, we fit bayesian multilevel regression models for each of the

outcomes [67, 68]. Bayesian multilevel regression can estimate the average effect of dose in the population, while allowing the effect to vary across strata and flexibly controlling for nuisance confounders. We allowed the effect of dose to vary by vaccine strain and HAI assay strain, and effects for different groups were assumed to be correlated with a heterogeneous unstructured covariance matrix. We used random intercepts for subject and study site to control for baseline differences, and used smoothing splines to control for the effects of birth year, age, pre-vaccination titer, and season. For more details, see the Supplement.

We used these models to estimate the effect of dose on each outcome of interest by calculating the posterior Average Causal Effect (ACE). To do so, we compared each individual's predicted outcome under the dose received (HD or SD) to their predicted outcome for the counterfactual dose, thereby estimating the individual causal effect (ICE) for each observation in the dataset. We estimated the ACE by summarizing the overall posterior distribution of ICEs using the mean point estimate with a 95% highest density credible interval (HDCI). We also computed conditional ACEs (cACEs) by summarizing the posterior distribution of ICEs in specific strata: within each season, within each vaccine strain, and within each assay strain and vaccine strain. In order to aid interpretation, we then base-2 exponentiated the ACEs and the limits of the HDCI. Each ACE represents the average difference in the model outcome across the full study sample between the high dose and standard dose vaccines, after controlling for confounders [69, 70]. See the Supplement for additional details on cACE calculation.

Our study is a secondary data analysis, and the data were not collected with our specific research questions in mind. Therefore, a strict hypothesis testing framework using p-values is not appropriate. Any statistical tests we conduct will have inflated false discovery rates and limited power. Therefore, we implemented models in a bayesian framework, and we focus on estimating the effect size and uncertainty as captured by credible intervals of the dose in our dataset, rather than testing hypotheses. The primary limitation of these bayesian models is the extensive computation time, and our implementation of multilevel models using Hamiltonian Monte Carlo does not suffer from the convergence issues or error inflation issues common to similar frequentist models [68, 71].

Details on model specification, outcome calculation, and details of implementation, including full data and code to reproduce results, are included in the supplement.

Implementation

We conducted all of our analyses using `R` version 4.4.2 [72]. We used the `tidyverse` suite of packages for data cleaning, manipulation, and visualization [73], along with the packages `tidybayes` [74] and `ggdist` [75, 76]. We used the packages `here` [77] and `renv` [78] for code management. We used `ggplot2` for generating all figures [79]. We used the packages `gtsummary` [80] and `flextable` [81] for making all tables. We compiled our manuscript using Quarto version 1.5.43 [82] with the R packages `knitr` [83–85], and `softbib` [86]. We implemented our Bayesian models using `brms` [87–89] and `cmdstanr` [90] with cmdstan version 2.34.1 [91] as the interface to the Stan probabilistic programming language [71, 92]. More details on implementation are included in the Supplement, along with instructions for reproducing our results. The code and data are archived on Zenodo at this link: https://doi.org/10.5281/zenodo.1266697.

Results

Data description

Our data come from a prospective open cohort study and span the influenza seasons from 2013/14 through 2021/22, and included 254 unique individuals across all study sites and seasons, who provided 668 total observations. Participants aged 65 and older were vaccinated with either Fluzone SD or HD at one of three study sites (Stuart, FL and Pittsburgh, PA, from the 2013/14 season – the 2016/17 season and the University of Georgia (UGA) in Athens, GA from the 2016/17 season – 2021/22 season). Table 2.1 shows the distribution of observations, stratified by vaccine dose. The three study sites had similar distributions of demographic characteristics at both the observation and unique participant levels (see Supplement). Since we only include individuals 65 or older in our study, the age ranges and birth cohorts were similar for both dose groups. Throughout the course of the study,

the recommended influenza vaccine composition was updated several times, encompassing 5 different H1N1 strains and 8 different H3N2 strains. A visualization of the overall trend in pre-vaccination and post-vaccination HAI titers is included in the Supplement.

Table 2.1: Number of observations in our sample, stratified by the dose, and the study site or season.

	SD, N = 234	**HD**, N = 434	**Overall**, N = 668
Study, n (%)			
FL	82 (35)	41 (9)	123 (18)
PA	73 (31)	146 (34)	219 (33)
UGA	79 (34)	247 (57)	326 (49)
Season, n (%)			
2013 - 2014	40 (17)	16 (4)	56 (8)
2014 - 2015	53 (23)	39 (9)	92 (14)
2015 - 2016	40 (17)	58 (13)	98 (15)
2016 - 2017	34 (15)	77 (18)	111 (17)
2017 - 2018	22 (9)	16 (4)	38 (6)
2018 - 2019	11 (5)	8 (2)	19 (3)
2019 - 2020	21 (9)	69 (16)	90 (13)
2020 - 2021	7 (3)	76 (18)	83 (12)
2021 - 2022	6 (3)	75 (17)	81 (12)

Strain-specific effects reveal differences in vaccine response patterns

Our models allowed the effect of dose to vary by the assay strain used for HAI assays, so we first examined the strain-specific effects of dose.

For the H1N1 vaccine component, the titer increase was higher for HD in all of the H1N1 strains except for CA/09 (Figure 2.1, where the point estimates for the heterologous responses were a mix of small positive and small negative values. MI/15 had positive point estimates for nearly all of the assay strains. Bris/18 showed a negative impact of dose for the single older strain used for testing, and positive impact for the other strains. GD/19 and Vic/19 had positive point estimates with almost all of the density of the interval estimate above 1 for the few strains they were tested against. The effect of the HD vaccine on the Chi/83 strain was negative for all three of the vaccines which it was tested against, although the negative effect was very close to 1 for the MI/15 vaccine stratum.

The effect size for our models in Figure 2.1 and Figure 2.2 and subsequent Figures is presented as a ratio of fold changes. An effect size greater than 1 (a "positive effect") indicates that the HD vaccine induced a higher average fold change than the SD vaccine, whereas an effect size less than 1 (a "negative effect") indicates that the HD vaccine induced a lower average fold change than the SD vaccine. An effect size of 1.25, for example, would indicate that, averaged across the population, the fold change from an HD vaccine would be 1.25 times larger than from an SD vaccine, assuming that the individuals affected stay the same in every other aspect. See the supplement for more details on the effect size.

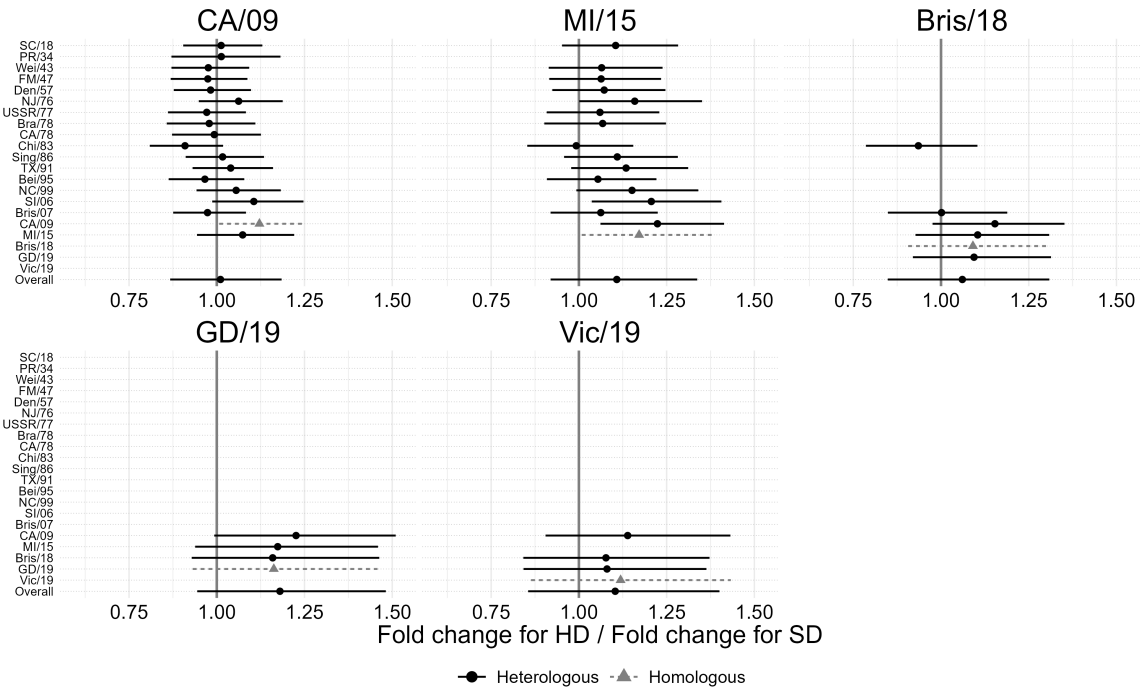

Figure 2.1: Strain-specific ratio of HD over SD antibody responses of heterologous strains following vaccination with the indicated vaccine strain. All strains shown in this figure are subtype H1N1. Values above one indicate a better response for the HD vaccine, values below one indicate a better response for the SD vaccine. For more recent vaccines, only a subset of data for heterologous responses was available.

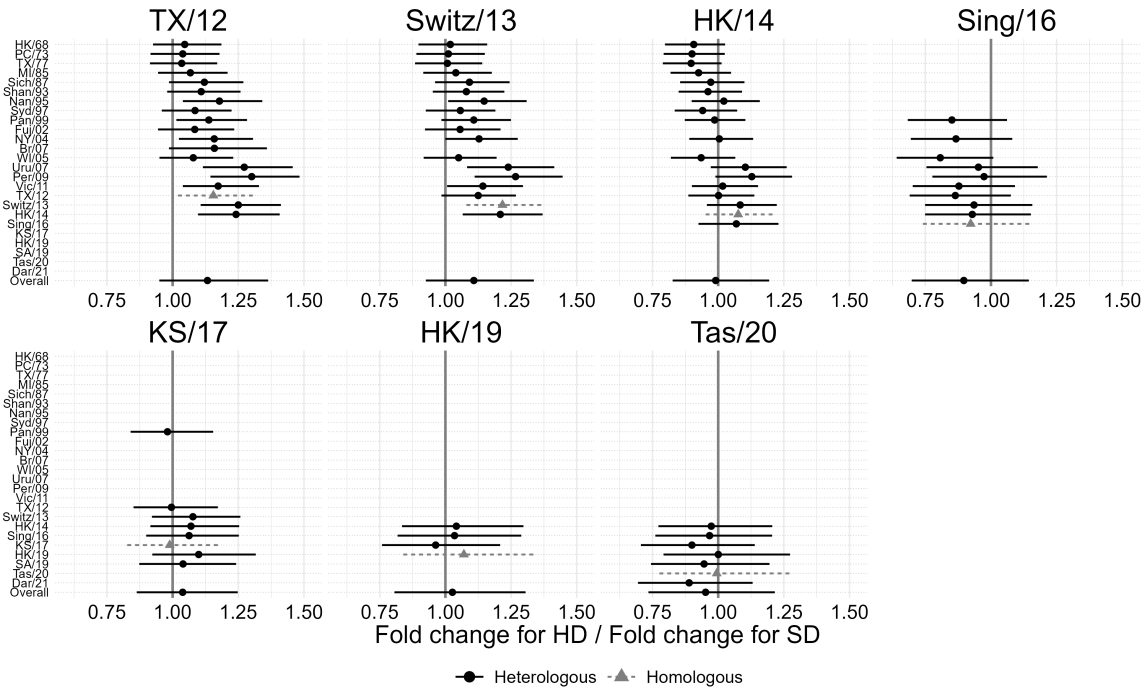

Figure 2.2: Strain-specific ratio of HD over SD antibody responses of heterologous strains following vaccination with the indicated vaccine strain. All strains shown in this figure are H3N2 subtype for both the vaccines and the assay strains. Values above one indicate a better response for the HD vaccine, values below one indicate a better response for the SD vaccine. For more recent vaccines, only a subset of data for heterologous responses was available.

Most, but not all, HD vaccine formulations elicited a stronger overall response

After examining the strain-specific impact of the HD vaccine, we computed the overall impact of the dose on each vaccine strain, by pooling together all of the posterior samples within that stratum and calculating the average effect. We found a weakly positive overall effect of the HD vaccine for all H1N1 strains Figure 2.3. (The estimates in Figure 2.3 are the same as the "overall" estimates in Figure 2.1 and Figure 2.2 for the appropriate subtype, but presented alone to allow more detailed comparisons.) While the uncertainty in our parameter estimates is large, all of the HDCI estimates were consistent with a small positive benefit from the HD vaccine. For the H3N2 strains, the majority of vaccine strains showed a benefit for the HD vaccine. However, the HK/14, Sing/16, and Tas/20 vaccines all had

negative point estimates.

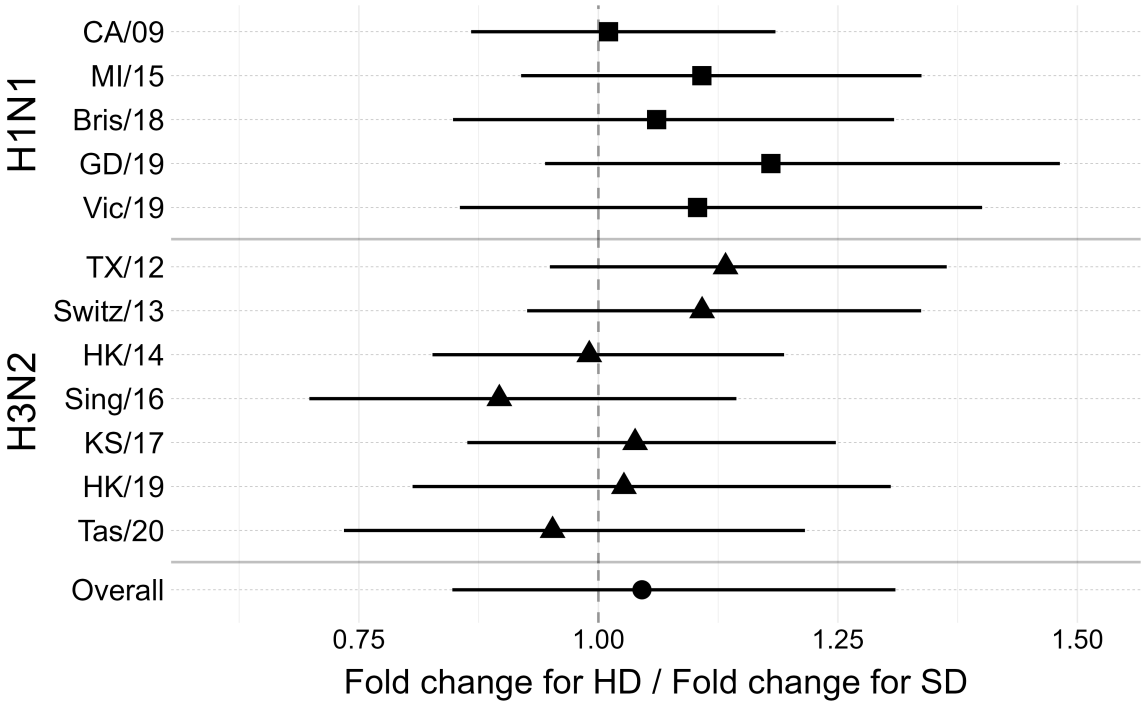

Figure 2.3: Vaccine-specific ratio of HD over SD antibody responses of following vaccination with the indicated vaccine strain. All individual responses were pooled together to calculate the cACE for each stratum defined by the vaccine component received by an individual. individual, listed on the y-axis. Values above one indicate a better response for the HD vaccine, values below one indicate a better response for the SD vaccine.

HD vaccines elicited stronger or equal responses in every influenza season

For our final analysis, we pooled the H1N1 and H3N2 responses for a given vaccine and computed an overall effect of the dose for a given vaccine in each season (ignoring the influenza B components). We found that an increase in dose had a positive but small impact for most seasons, while for a few seasons, we did not observe an impact (Figure 2.4). On a vaccine-level basis, there was no indication of an overall detrimental effect of dose.

Sensitivity analyses

We repeated the analyses shown above for the three other outcomes we defined in the Methods, i.e., post-vaccination titer, seroconversion rate, and seroprotection rate. While the

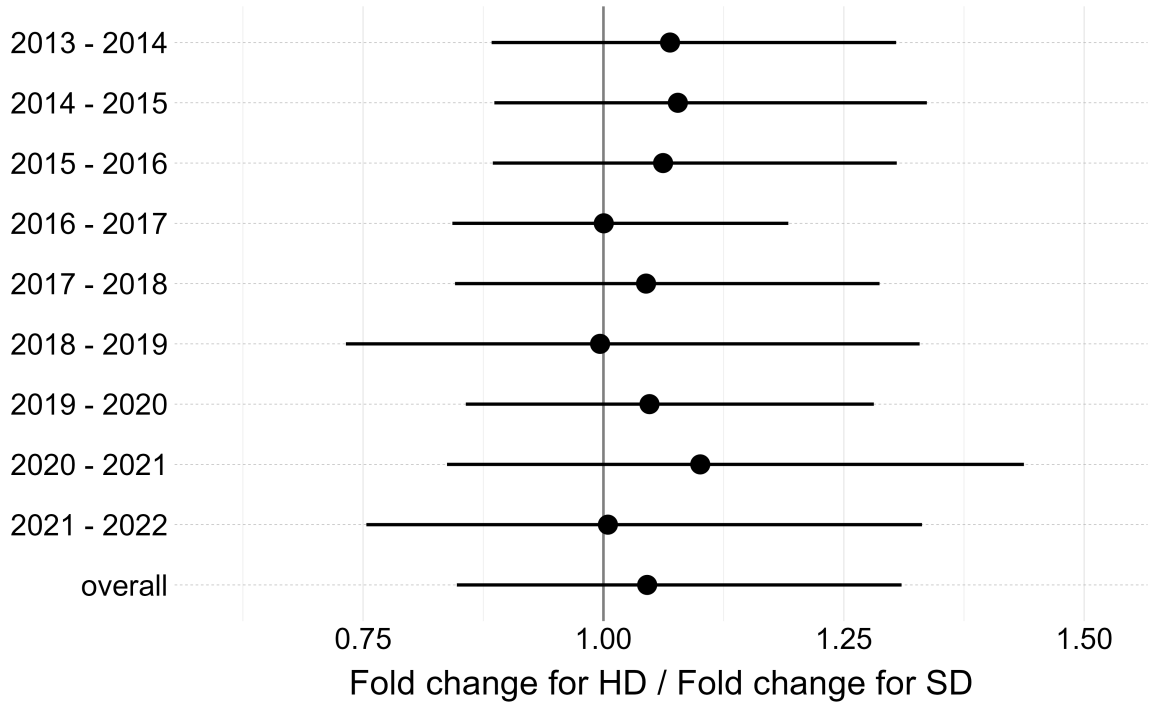

Figure 2.4: Season-specific ratio of HD over SD antibody responses of following vaccination with the indicated vaccine strain. All responses measured during a given season were pooled together to estimate the cACE with respect to the season, taking all vaccine components and historical strains into account. Values above one indicate a better response for the HD vaccine, values below one indicate a better response for the SD vaccine.

numerical estimates change slightly, the results were qualitatively similar to what we see with titer increase.

Specifically, the results for the post-vaccination titer outcome are qualitatively identical to the results for the titer increase outcome we present here. This is because the two models are mathematically very similar, and the differences between the two models largely disappear when we calculate contrasts of the type we present as our main result, but we include the results for the post-vaccination titer for completeness.

While the trends in cACE estimates for seroprotection and seroconversion are qualitatively similar, the effect sizes are closer to one for both outcomes. This is driven by the fact that the majority of individuals in our study did not achieve seroconversion or seroprotection for many strains. So even though the HD vaccine induces a better post-vaccination titer or titer increase (which could be important for preventing infection or clinical outcomes),

this additional signal is lost when we dichotomize the outcomes. For seroconversion in particular, many of the effect size estimates are close to zero because of the information loss associated with dichotomization. The full analyses of these other outcomes are shown in the supplement.

Discussion

We analyzed homologous and heterologous antibody responses to influenza A HD and SD vaccines across multiple years of influenza vaccinations.

While our estimates had high uncertainties and our secondary analyses should be interpreted as exploratory, our results preliminarily suggest a small but positive effect of the HD split-inactivated vaccine compared to the SD vaccine for inducing not only homologous (similar to prior results; see Supplement) but also better heterologous responses. However, this was not consistent and for some vaccine strains, there was a trend towards a negative impact for the HD vaccine. In general, both positive and negative effect sizes were small and most of the time the credible intervals included both a no-effect as well as positive and negative effect regions. Given that our analysis is a secondary data analysis of noisy observational data, and our observed effect sizes were small, the amount of uncertainty we observed is expected.

Our results suggest that the HD vaccine does induce both better homologous and heterologous responses for the majority of vaccine strains and thus should be a continued recommendation for older individuals.

In addition, our results also suggest that the overall impact of the HD vaccine is very modest at inducing stronger antibody responses. Since we only examined the humoral immune response, if the mechanisms related to reduced disease severity are driven by cellular immune responses [93, 94], there could be additional clinical benefits to the HD vaccine that cannot be learned from the data we used. We have not analyzed any markers of cellular immunity, and understanding how higher inoculum dosage could affect the breadth of the cellular immune response is critical information missing from our analysis. We also lack

outcome data for the individuals in our study, since there was no surveillance component. So while HD vaccination appears to be more protective against severe disease outcomes [20], these effects may not be mediated by the differences in immunogenicity we observe.

While we have a large sample size and many years collected immunogenicity data for a large panel of historical strains with coverage of major antigenic clusters, we lacked detailed data on previous infection and vaccination history. While we flexibly controlled for prior antibody titers at the time of vaccination and individual random effects, modeling of the exposure history would help us discern whether the effects of response blunting due to repeat vaccination [95–98] or the enhanced vaccine immunogenicity hypothesis [45, 99, 100], among other effects, are present in our data. In addition, vaccination history could be a useful proxy for healthcare-seeking behavior and willingness to receive an HD vaccine and thus could represent unobserved confounding in our data. However, the effect of unobserved confounding would have to be large to shift our observed causal effects enough (in either direction) to change our conclusions.

Overall, our results do not preclude a stronger beneficial effect of a higher vaccine dose. If a higher vaccine dose (above the dose in the current HD vaccine) would further enhance immunogenicity, we could potentially see a stronger effect than that is consistently different from the SD vaccine. With a stronger effect size, we could better determine which heterologous responses are boosted or diminished by increasing the dose. Since the current HD vaccine seems to be tolerated well, further studies on dose escalation and optimal dosing in elderly individuals might open avenues for improved vaccine design using the relatively cheap split-inactivated vaccine vector.

Acknowledgments

We thank the coauthors for their contributions: Yang Ge, Jessica H. Knight, Hayley Hemme, Savannah L. Miller, Amanda L. Skarlupka, Wangnan Cao, Ye Shen, Justin Bahl, Paul G. Thomas, Ted M. Ross, and Andreas Handel. Yang Ge contributed equally to this work.

We would like to thank Michael Carlock (Florida Research & Innovation Center, Cleveland

Clinic, Port St. Lucie, FL, USA) for assistance with obtaining data. Additionally, we thank Rustom Antia, Riley Drake, Cora Hirst, and Ananya Saha for many productive hiking conversations.

Conflicts of interest. None of the authors declare any conflicts of interest.

Funding sources. WZB was partially supported by the University of Georgia Graduate School as a Georgia Research Education Award Trainee. TMR is supported by the Georgia Research Alliance as an Eminent Scholar, and received partial support from NIH grant 75N93019C00052. AH received partial support from NIH grants/contracts U01AI150747, R01AI170116, 75N93019C00052 and 75N93021C00018. YS received partial support from NIH grants/contracts R35GM146612, R01AI170116 and 75N93019C00052. JB received partial support from grants/contracts CDC NU50CK000626, NIH 75N93021C00018 and NIH 75N93019C00052. PGT received partial support from NIH grant/contrant 75N93019C00052 and from the American Lebanese Syrian Associated Charities (ALSAC) at St. Jude. All other authors declare no funding sources. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement. All of the code files and data necessary to reproduce our results can be found on Zenodo (https://doi.org/10.5281/zenodo.12666976) or GitHub(https://github.com/ahgroup/billings-patient-reported-symptom-data-public).

Supplementary information. The Supplement for this manuscript is included as Appendix A.

Chapter 3

Different antigenic distance metrics generate similar predictions of influenza vaccine response breadth despite low correlation

Billings WZ, Ge Y, Skarlupka AL, et al. To be submitted to PLOS Computational Biology.

Abstract

Background: Influenza constantly evolves to escape population immunity, which makes formulating a vaccine challenging. Licensed vaccine formulations are based on strains that are expected to circulate, so if the circulating strain is different, vaccine effectiveness (VE) can be reduced. The effect on vaccine effectiveness should depend on how different the circulating strain and vaccine strain are, which is measured by antigenic distance. Antigenic distance is normally quantified using antigenic cartography of hemagglutination inhibition (HAI) titer data, which is costly and has the potential to produce biased antigenic distances. While multiple antigenic distance measurements have been developed and compared in previous literature, they have not been compared in their ability to predict immunological outcomes. We compare how predicted breadth of vaccine response varies when different metrics are used to calculate antigenic distance.

Methods: We analyzed data from a seasonal influenza vaccine cohort which collected serum samples from 2013/14 – 2017/18 at three study sites. The data include pre- and post-vaccination hemagglutination inhibition titers to the vaccine strains and a panel of heterologous strains. We used that data to calculated four different antigenic distance measures between assay strains and vaccine strains: difference in year of isolation (temporal), p-Epitope (sequence), Grantham's distance (biophysical), and antigenic cartography distance (serological). We analyzed agreement between the four metrics using Spearman's correlation and intraclass correlation (ICC). We then fit bayesian generalized additive mixed-effects models (GAMMs) to predict the effect of antigenic distance on post-vaccination titer after controlling for confounders and analyzed the pairwise difference in predictions between metrics.

Results: The four antigenic distance metrics demonstrated weak correlations for influenza subtypes A(H1N1), B (Victoria), and B (Yamagata). The exception was A(H3N2), which showed high correlations. However, we found that after accounting for pre-vaccination titer, study site, and repeated measurements across individuals, the predicted post-vaccination titers conditional on antigenic distance and subtype were nearly identical across antigenic

distance metrics, with A(H1N1) showing the only notable deviation between metrics.

Conclusions: Despite low correlation among metrics, we found that different antigenic distance metrics generate similar predictions about breadth of vaccine response. Costly titer arrays for antigenic cartography may not be needed when simple sequence-based metrics suffice for quantifying vaccine breadth.

Keywords: influenza, influenza vaccine, heterologous responses, antigenic distance, cohort study, secondary data analysis

Introduction

Influenza viruses constantly evolve over time. As host immunity induces selective pressure, new influenza strains accumulate mutations, a phenomenon called antigenic drift [1–3, 101–103]. As mutations accumulate, antigenic drift leads to vaccine escape [95, 104, 105]. Seasonal influenza vaccines are formulated based on the strains that are expected to circulate, but these predictions are sometimes incorrect and vaccine effectiveness (VE) varies annually [106]. A major determinant of VE is the similarity between the strains used for vaccine formulation and the circulating influenza strains [8, 10, 96, 107–113]. While previous analyses have analyzed the effect of binary match or mismatch, measuring the full effect of antigenic distance on vaccine response requires quantitative antigenic distance calculations [25–28, 114].

The most common method for quantifying antigenic distance between influenza strains is antigenic cartography, which relies on extensive panels of serological data [29]. Cartographic distance has proven useful in understanding influenza evolution, but validating the ability of cartography to estimate population-level protection is difficult because of the required data [33, 46, 115]. Evolution or sequence-based methods can accurately predict cartographic distance based on genetic sequences of influenza strains, but still rely on accurate serological data for calibration [34, 39, 116–122]. Furthermore, multiple cartography methods yield different maps on the same data [29, 33, 123–125]. Maps based on HAI titers also incorporate bias from HAI assays, which are often not replicable between labs [126, 127] and do not always accurately reflect differences in antigenic clusters [25–28, 114, 119, 123, 128–130]. While cartographies can be generated from alternative assays [41–43], HAI is still the most common immunological assay used for influenza and the majority of highly-cited cartographies in use are based on HAI [29, 33, 46, 131, 132].

Not all antigenic distance measurements involve serological data, however. Simpler analyses of the genetic or amino acid sequence of influenza strains can provide reliable measures of antigenic distance that correlate with vaccine effectiveness at a population level [30–32], even though they only weakly correlate with serological antigenic distance [33–35].

Notably, new antigenic lineages can be reliably differentiated by analyzing changes at just a few sites [133–135], and advanced predictive models consistently identify sequence-based biophysical and phylogenetic properties as important predictors of antigenic innovation [136–139]. Analyses of vaccine response or immunogenicity based on temporal [38, 45, 140–143] or sequence-based distances can provide information about breadth of vaccine response [30–32, 35, 116, 128, 144–146]. Taken together, these results imply that simple genetic analyses should provide important information about antigenic evolution without the need for serology. A direct comparison of antigenic distance methods is necessary to determine whether serological and genetic antigenic distance calculations can provide the same information in a practical setting.

To compare the implications of multiple antigenic distance metrics on practical outcomes, we perform a secondary data analysis of an influenza vaccine cohort with a panel of HAI measurements to historical strains for each individual. We find that, despite the low correlation in antigenic distance metrics, these different metrics make similar conclusions about vaccine response to antigenically distant strains. Our results suggest that implementing costly antigenic analyses or deep learning methods may not be necessary, as simple sequence-based measures lead to essentially equivalent predictions about vaccine response as antigenic distance varies.

Methods

Data source

The data for our study are from a prospective, ongoing human vaccination cohort study which has been described in detail previously [61, 147, 148]. Briefly, the study recruited participants at three study sites: Pittsburgh, PA, USA, and Port St. Lucie, FL, USA beginning in the 2013/14 influenza season (approximately September through March [149]) and continuing through the 2016/17 influenza season. Beginning in January 2017, the study moved to Athens, GA, USA. Participants received Fluzone (Sanofi-Pasteur) vaccines and donated two serum samples, one before being vaccinated, and one at a follow-up visit

approximately 21 days after the first visit. The study was a prospective, open cohort design where individuals could enroll in multiple years in the study, but were not required to re-enroll in every consecutive year. Individuals under 65 received a standard dose Fluzone vaccination, and individuals aged 65 and older were given the choice between standard dose (SD) and high dose (HD) Fluzone vaccines.

Researchers used each serum sample for a panel of hemagglutination inhibition (HAI) assays to the homologous strains, included in the seasonal vaccine formulation, and a panel of historical, heterologous influenza virus strains. Strains included in the historical panel represented the major clades of circulating influenza viruses. In each season, all prior vaccine strains from 2012 onward were included in the historical panel. See the Supplement for details on the Fluzone vaccine formulation and for a list of strains used in each season.

For our secondary data analysis, we extracted previously deidentified records from the 2013/14 through 2017/18 influenza seasons. We included all participants with both pre-vaccination and post-vaccination blood samples in our analysis. Our primary outcome of interest was the reciprocal post-vaccination HAI titer, which we log transformed:

$$y = \log_2 \left(\frac{\text{HAI titer}}{5} \right).$$

Note that the raw HAI titer is the reciprocal of the highest serum dilution at which agglutination occurred in the assay. We divided the raw titer by 5 before taking the log because the HAI assay had a lower limit of detection (LoD) of 10, and an upper LoD of $20,480$. Values below the LoD were coded as 5 in the dataset. After our transformation, values below the LoD had a value of 0. All observed values in our dataset were below the upper LoD. We used the same outcome definitions defined in our previous work on this dataset [150].

Each HAI assay in our dataset has an associated subtype, vaccine strain, and assay strain. Here, we use "subtype" for simplicity to describe both Influenza A subtypes (H1N1 and H3N2), and Influenza B lineages (Victoria-like and Yamagata-like). The vaccine strains associated with an HAI assay are the strains used in the Fluzone vaccine formulation in the

season when the serum sample was drawn. Each assay has three or four associated vaccine strains, depending on whether the individual who gave the serum sample received a trivalent or quadrivalent vaccine (see Supplement for details on the vaccine formulations). The assay strain for a vaccine is the strain of the actual virus added to the serum sample during the HAI assay.

We computed the pairwise antigenic distance for all strains used in the dataset (again, see the Supplement for a complete list). We used four different methods to compute the antigenic distance: temporal distance, dominant p-Epitope distance [30], Grantham's distance [151], and cartographic distance [29]. For complete details on antigenic distance calculation, see the Supplement.

Statistical analyses

We first summarized demographic information about the cohort in a descriptive analysis, stratifying by measurements, individuals, and person-years to demonstrate the multilevel structure of our data.

For our formal analysis, we built generalized additive mixed-effects models (GAMMs) and linear mixed-effects models (LMMs) [68, 152] (see the Supplement for details). We fit a separate model for each of the antigenic distance metrics. We used a Gaussian outcome family with the transformed post-vaccination titer as the outcome and adjusted for censoring in the outcome [153]. To answer our primary question, we modeled antigenic distance in two ways. For the LMM, we included a linear effect of antigenic distance which was allowed to vary by strain type. For the GAMM, we modeled antigenic distance using a flexible semiparametric spline which allows the relationship to be nonlinear, but constrained.

We fit the LMMs and GAMMs in a Bayesian framework using weakly informative priors chosen by a prior predictive simulation. We obtained posterior samples of the model parameters using the No U-Turn Sampler (NUTS) algorithm implemented by Stan [71, 92], via the `brms` [87–89] and `cmdstanr` [90] packages for `R` [72]. After obtaining the posterior samples, we calculated marginal posterior predictions for many values normalized antigenic distance [154]. We summarized the posterior prediction samples with a mean point estimate

and 95% highest density continuous interval (HDCI). We compared the GAMM and LMM for each antigenic distance metric using the leave-one-out expected log pointwise predictive density (LOO-ELPD) which is conceptually similar to model selection using cross-validation in a frequentist scenario [155, 156]. See the Supplement for extensive details on our models.

To examine the differences in predictions across each of the antigenic distance metrics, we compared the slope and intercept for LMMs and the fold change in predicted post-vaccination HAI titer for the LMM and GAMM since the GAMM has no equivalent simple parametrization (shown in the Supplement).

Implementation

We conducted our analysis with R version 4.4.1 [72] in RStudio version 2024.09.0+375 [157]. We implemented our package as a `targets` pipeline [158] based on a high performance computing (HPC) template [159]. We ran our pipeline on UGA's sapelo2 cluster, which uses the Slurm scheduling software [160, 161].

We used the packages `here` [77], `renv` [78], and `tidyverse` [73] for data curation and project management and the packages `marginaleffects` [154], `tidybayes` [74], `bayesboot` [162], `ggdist` [75, 76], and `loo` [155, 156, 163] for formal analysis. We used the packages `ggplot2` [79] and `GGally` [164] for generating figures, and the packages `gtsummary` [80] and `flextable` [81] for generating tables. We generated the manuscript using Quarto version 1.6.40 [82] along with the `R` packages `knitr` [83–85] and `softbib` [86]. We implemented our Bayesian models with the `brms` package [87–89] using the `cmdstanr` backend and cmdstan version 2.34.1 [90, 91] as the interface to the Stan programming language for bayesian modeling. The Supplement contains more exhaustive details on our methodology, including instructions for reproducing our results. Our dataset and code are archived on GitHub (https://github.com/ahgroup/billings-comp-agdist-public) and Zenodo (https://doi.org/10.5281/zenodo.15522148).

Results

Data description

Our dataset included 54,101 pairs of pre-vaccination and post-vaccination HAI titer measurements drawn from 677 individuals who contributed 1,163 person-years to the study across three different study sites. The contributions of paired measurements, person-years, and unique participants from each study site are shown in Table 3.1. In a given year, each individual contributed 4 homologous HAI assay pairs, along with a number of heterologous assay pairs, which varied by season due to the change in historical panels each year, and by individual due to random lab and assay issues. Each person-year represented in the data contributed a median of 48 HAI assay pairs (range: 8 to 52 pairs). Additional demographic information about our cohort is provided in the Supplement (summaries of race/ethnicity, sex assigned at birth, contributed person-years, age at enrollment, and pre-vaccination titer).

Table 3.1: Counts of HAI assay pairs, person-years, and unique participants contributed by each study site for the duration of the study. Note that the PA and FL study sites operated from September 2013 to December 2016 and the GA study site began operating in January 2017 (during the 2016/17 influenza season).

	Season					
	2013/14	2014/15	2015/16	2016/17	2017/18	Total
Paired HAI assays, n						
FL	2459	6597	6656	6188	0	21900
PA	2163	3716	4131	3136	0	13146
UGA	0	0	0	6815	12240	19055
Overall	4622	10313	10787	16139	12240	54101
Person years, n						
FL	60	150	128	119	0	457
PA	73	88	81	64	0	306
UGA	0	0	0	145	255	400
Overall	133	238	209	328	255	1163
New participants, n						
FL	60	113	37	31	0	241
PA	73	46	2	12	0	133
UGA	0	0	0	145	158	303
Overall	133	159	39	188	158	677

Antigenic distance metrics have low correlation for all subtypes except A(H3N2)

First, we examined the overall agreement between the different distance metrics. We analyzed agreement using the intraclass correlation (ICC), shown in Table 3.2. ICC was low for all subtypes except A(H3N2), and the credible interval included zero for all subtypes except A(H3N2), so despite the moderate point estimate for B/Yamagata with a high upper limit, there was low consistency in antigenic distance measurements across methods. For A(H3N2), we observed moderate agreement across methods. Our ICC results indicate for each subtype except A(H3N2), at least one of the antigenic distance metrics systematically disagrees from the other.

Table 3.2: Intraclass correlation (ICC) across all antigenic distance measurements, calculated separately for each subtype or lineage (strain type). The posterior distribution for each ICC was calculated as the ratio of variance components for vaccine strain and assay strain divided by the sum of all variance components, estimated with a Bayesian model. Numbers shown are the mean and 95% highest density credible interval (HDCI) of the posterior distribution of ICCs.

Strain Type	**ICC**
H1N1	0.09 (0.00, 0.24)
H3N2	0.34 (0.19, 0.52)
B-Yam	0.21 (0.00, 0.42)
B-Vic	0.03 (0.00, 0.13)

To better understand the lack of overall agreement, we computed the Spearman rank correlation between each pair of metrics (again, separately for each subtype). Figure 3.1 shows the pairwise scatterplots and correlation coefficients. The pairwise correlations between distance measurements varied widely across subtypes and combinations, indicating that low agreement was not driven by a specific metric or subtype. All distance metrics tended to correlate well for H3N2. Distance metrics correlated highly for both influenza B subtypes with the exception of the cartographic distance, which had a moderately high correlation with the other three distances for B/Yamagata and a low correlation with the other three distances for B/Victoria. The only high correlation for A(H1N1) was between Grantham and p-Epitope distance, with small correlations between the other distance metrics. Grantham

and p-Epitope distances correlated well for all strains (although it was notably lower for A(H1N1)), which we expected given the similarity between the measures. The Supplement contains a table with credible intervals for each correlation.

Predicted vaccine response breadth is similar across antigenic distance metrics, despite the low between-metric correlation

Examining the agreement and pairwise correlations between the different distance metrics is useful for understanding which metrics disagree most, but these disagreements do not necessarily translate into different predictions about vaccine response. We built LMMs and GAMMs to model the effect of antigenic distance after controlling for multiple host and assay features.

To quantify whether the effect of antigenic distance deviated strongly from a linear effect, we calculated the LOO-ELPD for the GAMM and LMM models fit with each antigenic distance metric, shown in Table 3.3. LOO-ELPD is comparable to (W)AIC or BIC, and differences in ELPD strongly supported the linear model for every antigenic distance metric. The ratio of the difference in ELPD was always much greater than its standard error, so the difference between models can be trusted for model selection. Including spline terms to account for nonlinearity did not improve the model fit.

Table 3.3: Differences in expected log pointwise predictive density (ELPD) from the best-fitting model, estimated by the leave-one-out (LOO) method for all models and all antigenic distance metrics. We fit the models separately for each antigenic distance metric, so comparisons are shown separately. The Δ ELPD is the difference in ELPD between the LMM and the GAMM, so a positive number indicates the LMM performed better than the GAMM, and a larger number means the LMM outperforms the GAMM more. We show the Δ ELPD $\pm$ its standard error, along with the ratio of the estimate to its standard error.

Metric	**Δ ELPD (LMM - GAMM)**	**Δ ELPD / SE**
Cartographic	106.32 ($\pm$19.93)	5.3
Grantham	203.47 ($\pm$24.28)	8.4
Temporal	46.30 ($\pm$11.14)	4.2
p-Epitope	416.55 ($\pm$38.76)	10.7

Figure 3.2 shows how the average post-vaccination titer predicted by the model changes

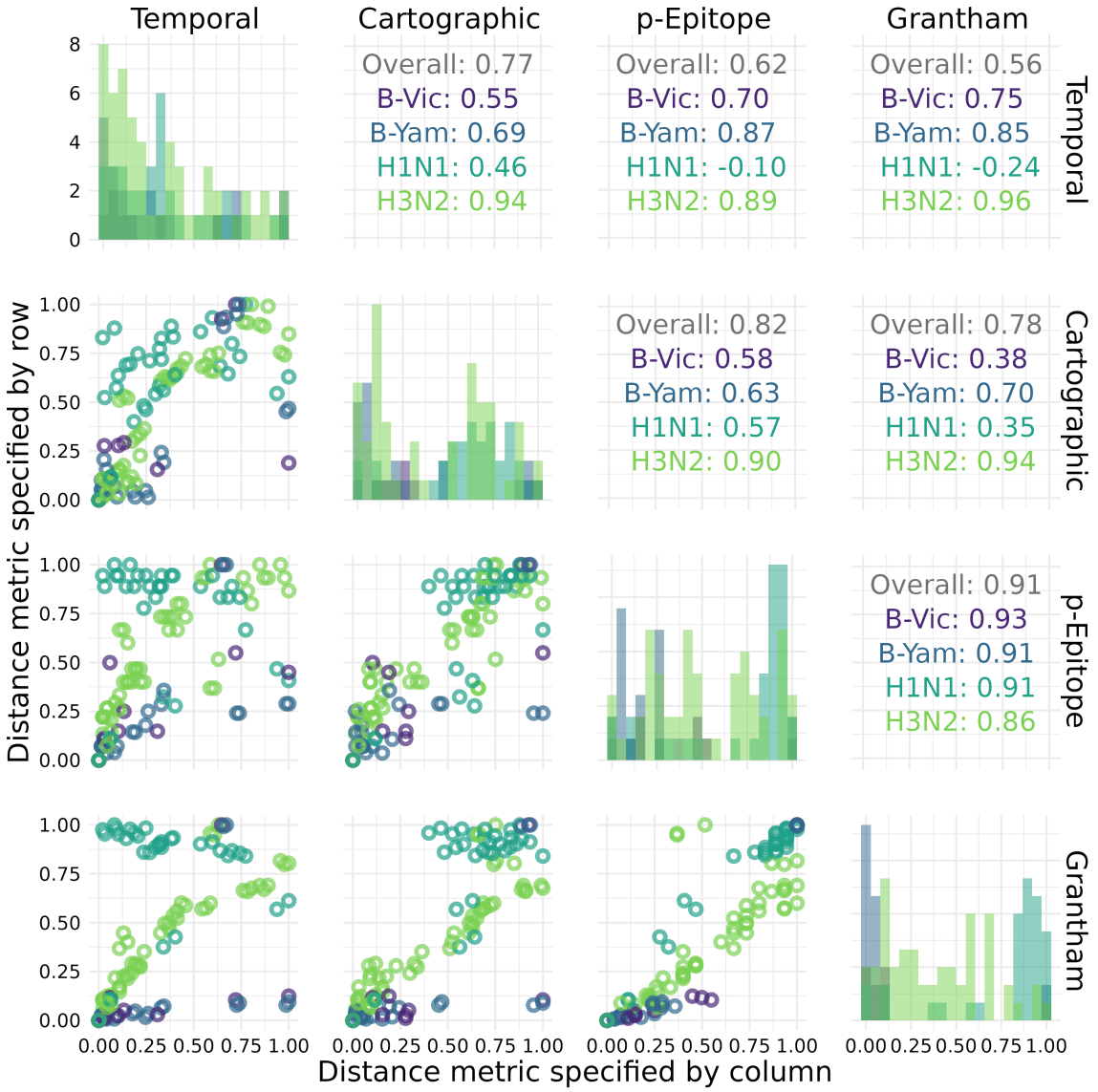

Figure 3.1: Distribution and correlation plots for each of the antigenic distance metrics. For each HAI assay in the dataset, we calculated the antigenic distance between the vaccine and assay strains with four different methods. We examined the distribution (shown along the diagonal) and the correlation between the different metrics for the same pairwise comparisons (we show pairwise scatterplots in the plots below the diagonal, and overall Spearman's correlation values in the plots above the diagonal). We include each unique combination as only one point in this plot. We calculated correlation coefficients separately for each subtype – colors in the plot indicate subtype.

along with antigenic distance for each subtype. For both influenza B lineages, the data were sparsely measured across the span of any of the antigenic distance metrics, making the GAMM predictions difficult to distinguish from the LMMs. Both influenza A subtypes showed a larger difference in predictions made by the GAMMs vs. the LMMs where the GAMMs predicted non-monotone relationships between post-vaccination titer and antigenic distance. The LMM and GAMM were most similar for cartographic distance for both A(H1N1) and A(H3N2), perhaps suggesting that cartographic distance partially accounts for nonlinear effects of antigenic distance, but There were some interesting trends in the shape of the spline curves, but the nonlinear effects for the p-Epitope and Grantham distance did not appear to match the distribution of data points well. Combined with the lack of ELPD support (Table 3.3), the spline models are likely picking up random fluctuations which may be partially driven by gaps in antigenic distance space rather than by true non-monotone signals (see the Supplement for an analysis of the gaps in antigenic distance space).

Since the linear model had better ELPD support for all metrics (Table 3.3), we focused on attempting to understand the effects in the linear model. Other than the normalized antigenic distance effect, the other effects were similar across the four models (which is what we expect). Sex and race/ethnicity had almost no effect. The effects of birth year and age are similar in magnitude, but are counterintuitively in the same direction – these effects are highly correlated and mostly cancel each other out, as a higher age leads to a lower predicted titer, while a higher birth year (indicating a lower age) also leads to a lower predicted titer. Log pre-vaccination titer had a strong positive effect on post-vaccination titer as expected. The effect of antigenic distance was negative for all four models, as we would expect, but the magnitude of the effect varied. While the point estimates were similar, the effect size for p-Eepitope was the smallest and the effect size for cartographic distance was the largest. The effect size for the cartographic distance also had the most density away from zero with an upper limit for the 95% HDCI which would still indicate a noticeable effect. Only the temporal distance model had an HDCI for the distance effect that included zero.

We also attempted to understand the variance contributions of the fixed effects of interest (Table 3.4) along with the sources of nuisance variation we included in the model by

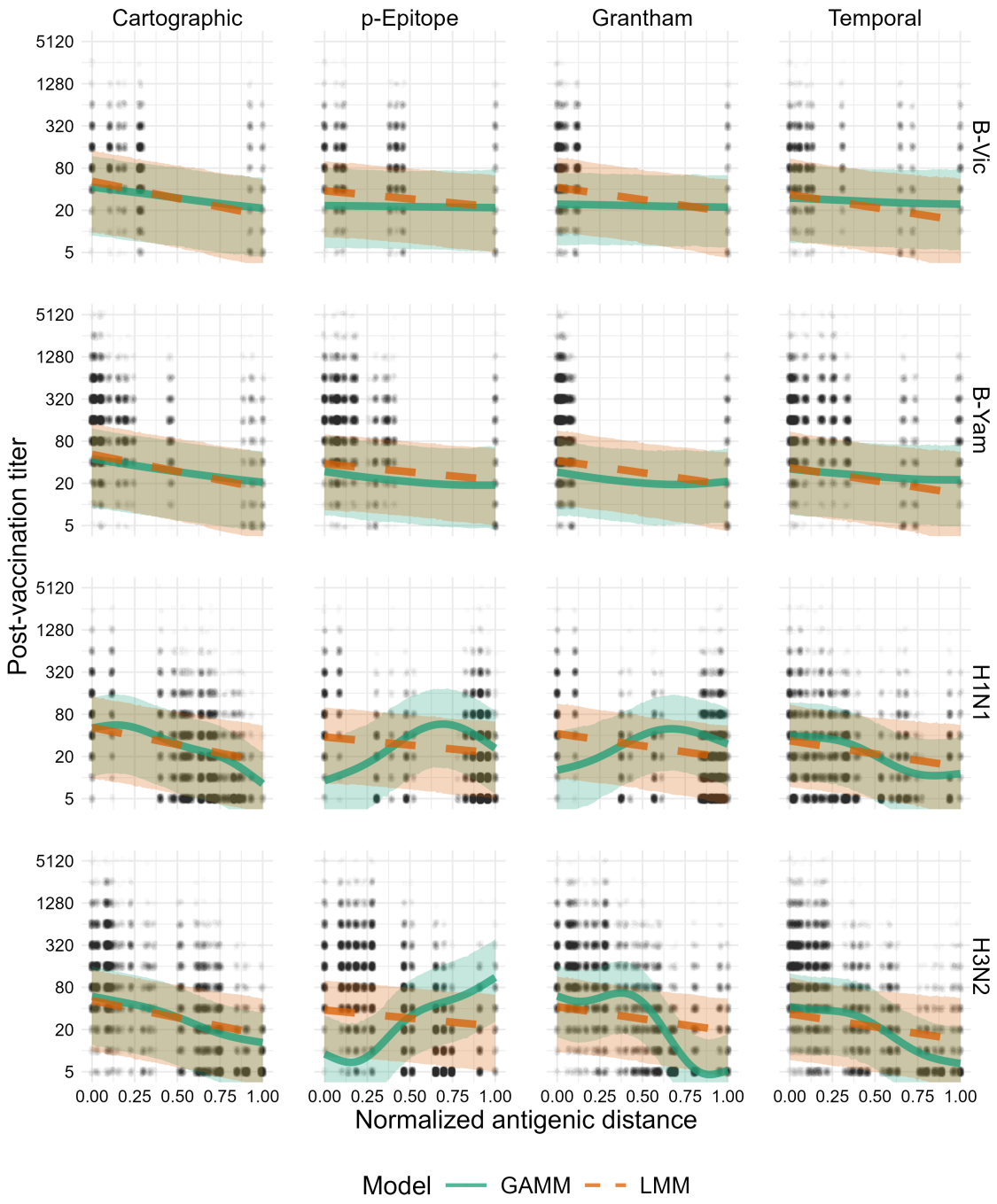

Figure 3.2: Model predictions for both the GAMM and LMM. Solid green lines and green ribbons show the mean and 95% highest density continuous interval (HDCI) for GAMM predictions. Dashed orange lines and orange ribbons show the mean and 95% HDCI for LMM predictions. Circular points show the data values. Each subplot shows the model predictions for a particular subtype (changes by row) using the model for a particular distance metric (changes by column). Outcomes shown on the plot are predicted post-vaccination titers for an average individual to an average strain (see Supplement for computational details).

decomposing the variance (Table 3.5). The fixed effects explained the most variance of the three model components in all four models. The contribution of the residual variance was nearly identical in all four models, suggesting that the random effects are more important in some models than others, without explaining any additional variance. The variance explained by the assay strain, vaccine strain, study site, and subject variance components was similar across the four models, with the most noticeably different contribution being the effect of the subtype. The subtype apparently explained more variance in the temporal and grantham distance models than in the cartographic and p-Epitope distance models, suggesting that those metrics might be more affected by differences in subtypes. Overall, the fixed effects were typically slightly more important than the random effects, but the variance explained by the random effects was still large for each model.

Predictions made by different antigenic distance metrics are similar after accounting for host factors

Finally, we directly compared estimates from the models across normalized antigenic distance metrics for each subtype (Figure 3.3). Since the LMM is easier to interpret and was supported by our ELPD analysis, we examined the slope and intercept for each subtype across the four antigenic distance metrics. The intercepts (representing the predicted post-vaccination titer to the homologous strain of the specified subtype for an individual with no pre-vaccination antibodies) were similar across all metrics regardless of the subtype. The slopes varied more, indicating that the antigenic distance had a stronger effect on predicted titer for some metrics and subtypes. For both B lineages, estimates of the slope were nearly identical across antigenic distance metrics. For A(H1N1), the cartographic distance model had a lower slope than the other three antigenic distance metrics, but the credible interval still overlapped with the credible interval for the temporal distance. For A(H3N2), the slope for the p-Epitope distance was much smaller than the other slopes (reflecting our results in @fig-gamm-plot), despite the high correlation between the antigenic distances for A(H3N2) (Figure 3.1). We can only perform a visual inspection of these overlaps, because there is no existing approach to combine posterior distributions across the four models.

Furthermore, these estimates do not take variance from the random effects in our model into account. To analyze predictions for both the LMM and GAMM, with the random effects variances included in uncertainty calculations, we directly compared predictions from the models and saw much higher overlap (shown in the Supplement), as we would expect when we include all of the variance in the data.

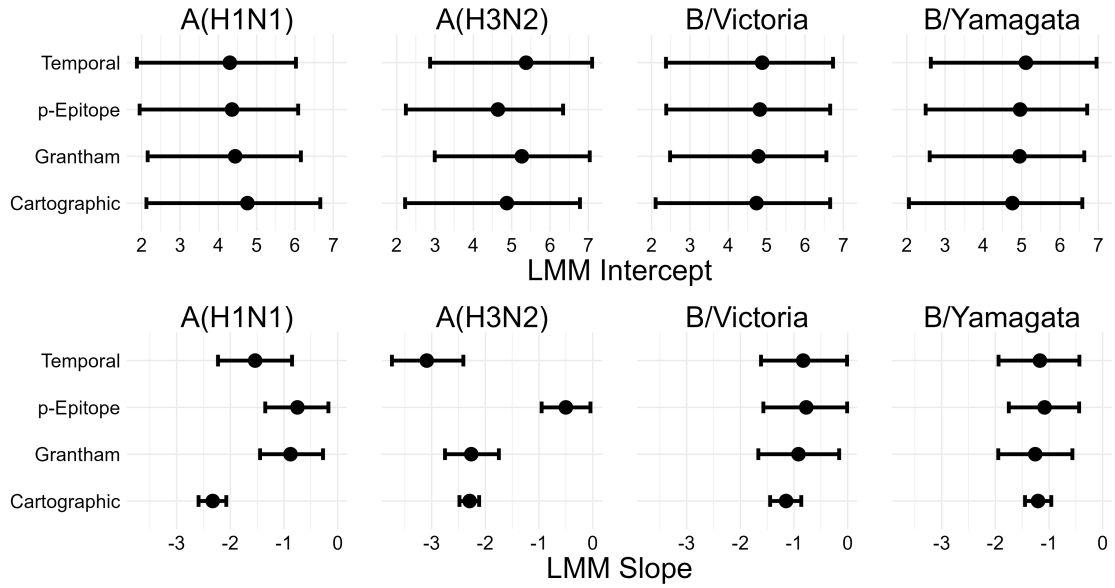

Figure 3.3: Intercept and slope estimates stratified by subtype for each LMM (one for each distance metric). Points and intervals show the mean and 95% HDCI of posterior samples of the indicated parameter. The top row of plots shows the mean and CI for estimates of the intercept, and the bottom row of plots shows the mean and CI for estimates of the slope. Columns of plots indicate which subtype the slope and intercept are for.

We compare the relative LOO-ELPD for each model in Table 3.6. Since the models are fit to the same set of predictors and data points, and the antigenic distances are normalized, the ELPDs are on the same scale and we can directly compare them. We found that all of the models had very similar performances – while the ELPDs were different between the four models, each contrast was smaller than the SE for either ELPD. For example, while the cartographic model had an ELPD around 150 points lower than the p-Epitope model, the SE for both estimates was around 470, so we cannot assume that these contrasts are meaningful differences. All of the models appeared to fit the data equally well.

Table 3.4: Coefficients for all of the fixed effects included in our primary models. We fit a separate model for each of the metrics, but the variables are standardized the same way across all four models so the coefficients are on the same scale across all models.

Metric	**Birth Year**	**Age**	**Sex**[1]	**Race**[2]	**Pre-titer**[3]	**Ag. dist.**[4]
Cartographic	-3.14 (-4.08,-2.19)	-3.46 (-4.37,-2.54)	0.01 (-0.05, 0.06)	0.03 (-0.02, 0.08)	0.78 (0.77, 0.79)	-1.62 (-2.45,-0.63)
Grantham	-3.18 (-4.13,-2.23)	-3.50 (-4.41,-2.58)	0.01 (-0.05, 0.06)	0.03 (-0.02, 0.08)	0.78 (0.78, 0.79)	-1.15 (-2.07,-0.04)
p-Epitope	-3.21 (-4.16,-2.28)	-3.53 (-4.43,-2.62)	0.01 (-0.04, 0.06)	0.03 (-0.02, 0.08)	0.78 (0.78, 0.79)	-0.77 (-1.54,-0.05)
Temporal	-3.17 (-4.11,-2.22)	-3.48 (-4.40,-2.56)	0.01 (-0.04, 0.06)	0.03 (-0.02, 0.08)	0.78 (0.78, 0.79)	-1.24 (-2.38, 0.17)

[1]Sex assigned at birth; Reference: Male (vs. female); [2]Race/Ethnicity; Reference: Non-Hispanic white (vs. other)
[3]Log scale pre-vaccination HAI titer; [4]Normalized antigenic distance

Table 3.5: Variance contributions to the total variance estimated in the model. To estimate the fixed effects variance contribution as the variance of the estimated linear predictor, while the residual variance and random effects variance contributions (all variance contributions other than the fixed effects and residual variance) are estimated as model parameters. All contributions are expressed as the nearest percent and may not sum (rowwise) to 100 due to rounding error.

				Specific random effects				
Metric	**Residual var.**	**Fixed eff.**	**All random eff.**	Subtype	Assay strain	Vaccine strain	Study site	Subject
Cartographic	12 (9, 15)	51 (38, 61)	36 (22, 51)	13 (4, 25)	1 (1, 2)	3 (1, 5)	11 (0, 30)	6 (4, 7)
p-Epitope	12 (9, 15)	49 (38, 60)	37 (24, 51)	13 (1, 28)	5 (4, 7)	3 (1, 7)	8 (0, 25)	6 (4, 7)
Grantham	12 (9, 14)	48 (37, 58)	39 (27, 53)	17 (5, 33)	4 (2, 5)	3 (1, 7)	7 (0, 24)	6 (4, 7)
Temporal	11 (8, 13)	44 (34, 54)	44 (32, 57)	23 (9, 40)	4 (2, 5)	3 (1, 7)	7 (0, 23)	5 (4, 6)

Table 3.6: Expected log pointwise predictive density (ELPD) calculated for each of the linear mixed-effects models (LMMs) using the leave-one-out (LOO) method. For each metric, we show the estimated ELPD ± its standard error. The differences between the model ELPDs were negligible.

Metric	**LMM LOO-IC**
Cartographic	151131.0 ± 471.6
p-Epitope	151277.7 ± 472.6
Grantham	151251.0 ± 472.1
Temporal	151186.8 ± 472.2

Discussion

We computed multiple antigenic distance metrics on the same set of influenza strains. Using immunogenicity data from a human cohort, we were able to compare cartographic data to sequence-based, biophysical, and temporal antigenic distance measures which have been used before for analyzing vaccine breadth. We then fit linear mixed-effects models (LMMs) and generalized additive mixed models (GAMMs) to the immunological data separately for each cohort, controlling for subtype, pre-vaccination titer, and multiple sources of random variation. By comparing the predictions and parameters from the estimated models across the four antigenic distance metrics, we were able to assess the similarity of the metrics in a more practical context.

Despite low correlations between the four antigenic distance measures for all subtypes except A(H3N2), we found that all four antigenic distance measures produced similar predictions about the heterologous vaccine response, regardless of subtype. Unexpectedly, the subtype generating the most different predictions was A(H3N2), which had the highest correlation between metrics. After we account for important confounders and other sources of variation, the differences between metrics seemed to disappear, with the exception of the unusually small slope for p-Epitope distance for influenza A(H3N2). Along with our pointwise prediction comparisons (shown in the supplement), these results suggest a systematic disagreement on the vaccine outcome scale between p-Epitope distance and other metrics for A(H3N2), which contrasts with the high pairwise correlations between p-Epitope and other metrics for this subtype. Perhaps important antigenic changes for H3N2 have

occurred outside of the immunodominant epitopes, or features like glycosylation which might be more easily captured by Grantham or cartographic distance are important [165]. Or, perhaps the difference is due to some form of noise in our study — we have no data from equivalent human cohort studies with wide heterologous panels to compare our results to, so we do not know if this result is consistent.

Our overall results could imply that the differences between antigenic distance metrics can appear large but are small compared to between-subject and between-study variability in real life, or that accounting for interindividual differences or pre-vaccination titer helps to explain the differences between metrics. We also found that a linear model was sufficient for explaining the relationship between post-vaccination titer and antigenic distance, rather than a nonlinear model which we might expect under the assumption of original antigenic sin or immune imprinting, which could imply a nonlinear effect where strains with intermediate antigenic distance from the vaccine have the lowest vaccine response [95, 166]. Notably, we even found that temporal distance tends to produce similar predictions to cartographic distance in this setting, despite the evidence for epochal antigenic evolution and emergence or circulation of multiple clusters in a single year [2, 95, 136, 167].

While we used data from a multicenter study with tens of thousands of measurements and over one thousand contributed person-years, our study still has some weaknesses. First, as a secondary data analysis, none of the data were designed with our questions in mind. While we have attempted to control for as much confounding as possible, we lack data on the exposure histories, including infections and prior vaccinations outside of the study, of individuals in our cohort which could confound our results. Our results also only apply to the split-inactivated Fluzone standard dose vaccine. Higher doses can either help or hinder heterologous responses [23, 59, 60], and in a previous study we found that the heterologous antibody response varied by Fluzone vaccine dose [150], so our results might change for other vaccine doses or formulations. A balanced design with randomized vaccine design would be preferable for understanding the impact of vaccine design on agreement between antigenic distance metrics.

We also used cartographies based on our pre-immune human data, which were generated

on the same data we analyzed. With access to multiple cartographies on the same data set or imputation techniques [40, 168] we could treat different cartographies as different antigenic distance metrics and compare cartographic distances in the same way. Our metrics also did not all cover antigenic distance space evenly as the strains in the historical panel were selected to cover a wide variety of years. However, there were several "gaps" between discrete antigenic distance values for A(H1N1) and the two B lineages, which could impact our estimates (see Supplement for details), and a broader panel with more evenly spaced strains would make our effect estimates more precise. Finally, we have no real proxy for the response to "future" strains. We could get a better predictive understanding of how the vaccine generates immune responses to future strains by testing serum samples from, say, 2016, to novel vaccine strains which have emerged since the samples were collected. Such measurements would allow us to validate the use of the historical panel as a proxy for future vaccine response. Longitudinal studies designed with long-term collection and multiplex assays in mind would be beneficial for answering similar questions about antigenic distance and vaccine breadth.

Overall, we found that simple antigenic distance metrics like Grantham's distance generated very similar predictions about vaccine breadth to distances based on antigenic cartography in our study. While some distance metrics potentially deviated, the effect was subtype specific (p-Epitope for A(H3N2) strains). While cartography is important for understanding the antigenic diversity and evolution of influenza, researchers analyzing vaccine breadth should not be afraid to use easier, potentially less biased metrics of antigenic distance.

Acknowledgments

We thank the coauthors for their contributions to this work: Yang Ge, Amanda L. Skarlupka, Savannah L. Miller, Hayley Hemme, Murphy John, Natalie E. Dean, Sarah Cobey, Benjamin J. Cowling, Ye Shen, Ted M. Ross, and Andreas Handel.

We thank William Michael Landau (Eli Lilly and Company, Indianapolis, IN, USA)

and Eric R. Scott (University of Arizona, Tuscon, AZ, USA) for their generous help with computational issues and pipeline development. Additionally, we thank Michael A. Carlock (Cleveland Clinic Florida Research & Innovation Center, Port St. Lucie, FL, USA) for assistance with obtaining data.

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. Similarly, we gratefully acknowledge all data contributors for generating the genetic sequence and metadata and sharing via NCBI GenBank and UniProt, on which this research is based. More details on the sequences we used are available in the Supplement. The sequences we used from GISAID are accessible via GISAID Identifier `EPI_SET_250609vz` and DOI: https://doi.org/10.55876/gis8.250609vz.

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

Conflicts of interest. BJC has consulted for AstraZeneca, Fosun Pharma, GlaxoSmithKline, Haleon, Moderna, Novavax, Pfizer, Roche, and Sanofi Pasteur. None of these companies had any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding sources. WZB was partially supported by the University of Georgia Graduate School as a Georgia Research Education Award Trainee. TMR is supported by the Georgia Research Alliance as an Eminent Scholar. AH received partial support from NIH grants/contracts U01AI150747, R01AI170116, and 75N93019C00052. NED received partial support from NIH contract(s)/grant(s) R01-AI139761. All other authors declare no funding sources. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement. All of the code files and data necessary to reproduce our results can be found on Zenodo (https://doi.org/10.5281/zenodo.15522148) or GitHub

(https://github.com/ahgroup/billings-comp-agdist-public).

Supplementary information. The Supplement for this manuscript is included as Appendix B. A detailed description of the methods for calculating the antigenic distances is included as Appendix D.

Chapter 4

A novel approach for robust evaluation of broadly reactive influenza vaccine candidates

Billings WZ, Skarlupka AL, Murphy J, et al. To be submitted to a peer-reviewed journal.

Abstract

Background: Developing broadly-protective vaccines is a clear path towards reducing the burden of influenza pandemics and seasonal epidemics. However, measuring the breadth of response of a vaccine or candidate entails estimating the immunogenicity of the vaccine against many different influenza strains, and then combining these measurements. The most common method is the proportion of strains to which a vaccine induces a clinically noticeable immune response, which is strongly dependent on the panel of strains used for analysis. Estimates of breadth using this method cannot fairly be compared across different labs or even different studies from the same lab using different virus panels. In our study, we develop a novel robust metric for quantifying vaccine breadth across labs and virus panels.

Methods: We used data from a prospective, open annual influenza vaccination cohort across three study sites from Fall 2013/14 through Spring 2017/18. Each individual contributed pre-vaccination and post-vaccination serum samples which were used for a panel of hemagglutination inhibition (HAI) assays against many strains of influenza, including the strains used in each season's vaccine. We computed multiple antigenic distance metrics between each HAI assay strain and the corresponding seasonal vaccine strain.

Results: First, we calculated antigenic distance metrics using three different methods. We then fit summary antibody landscapes for vaccine response vs. antigenic distance using a simple model. From these summary landscapes, we computed our novel metrics which adjust for antigenic distance and censoring in the reported titers. We also computed standard breadth metrics from the existing literature, along with adjusted metrics which are identical to the standard metrics but also control for censoring in the titers. We subsampled multiple panels of heterologous strains and sets of individuals from our cohort study to create hypothetical studies, and we computed metrics across these subsampled studies. Finally, we calculated the intraclass correlation (ICC) to measure how much each metric varied across labs with different populations and virus panels.

Conclusions: Our novel metrics are much more robust to the panel of viruses chosen by a given lab than currently used metrics for assessing vaccine breadth. Implementing

our methods is easy and does not require imputing missing strains from vaccine panels or costly logistics for coordinating which strains are used in vaccine breadth panels, although a combination of approaches would be ideal. While a coordinated solution with a multiplex assay is likely a useful next step for vaccine breadth researchers, our metrics allow for fairer comparison of vaccine components until coordinated panels are more feasible.

Keywords: influenza, vaccines, heterologous immunity, vaccine breadth, antigenic distance, universal influenza vaccine, cohort study, secondary data analysis

Introduction

Developing a broadly-reactive (or "universal") influenza vaccine is a key goal for reducing the burden of seasonal influenza and increasing pandemic preparedness [1, 8, 13, 14]. As a given influenza strain spreads, many hosts develop an immune response to that strain, placing selective pressure on the virus. Mutant lineages can acquire antigenic changes that are not recognized by hosts who have previously received vaccines, allowing for vaccine escape. The rapid evolution and vaccine escape patterns of influenza make developing a universal vaccine difficult [3, 95, 104, 105]. Many broadly-reactive vaccine candidates are under development [169], but the best method for assessing the breadth of a vaccine candidate (how broadly-reactive a vaccine candidate is) is unclear.

In previous research, the breadth of the response is quantified by measuring vaccine-induced immunogenicity against a panel of historical influenza strains. Among multiple vaccine candidates, the one that induces a clinically meaningful immune response to the highest number of strains is considered the most broadly reactive [36, 38, 140, 141, 170, 171]. This method is easy to conduct, but the selection of different virus panels across research groups makes the results from different labs hard to compare. Some methodological work has focused on methods for imputing data across different virus panels [39, 40], but there is no practical exploration of how well these imputation measures work, and they rely on a low-rank approximation which may be unreliable [125]. Requiring many different labs to use exactly the same panel of viral strains (including constant monitoring for adaptations, and ensuring all protocols are the same) is logistically not feasible and would be extremely expensive. Breadth metrics that are robust to the selection of different panels of viral strains would circumvent these logistical issues and allow for fair comparisons of broadly-reactive vaccine candidates across lab groups.

Recent methods have focused on quantitative analysis of individual antibody landscapes, rather than counting the number of seroprotection events across an immune assay panel. HAI is the most common assay used for these panels, but inherent biases of the HAI assay can influence breadth measurements [25, 26, 126, 127], so a method for quantifying vaccine

breadth should be agnostic to the specific measurement used. Other proposed immunological assays for breadth calculation include stem-binding neutralization assays [41], binding profiles [42], neuraminidase inhibition [43], and a multiplex neutralization assay [172]. Regardless of the assay one uses, an antibody landscape is a curve which depicts antibody response on the y-axis with respect to some measurement on the x-axis which attempts to order the responses by the differences between influenza strains (distance metrics). Just as there is no gold standard for the best immunological assay to use, there is no gold standard for the distance metric, and different metrics can measure different ways for two strains to be different. Previously proposed distance metrics include the year of strain isolation [45, 142], genetic differences [29, 30, 35, 130, 144], biochemical or biophysical differences [151, 173]; and distances derived from antigenic cartography [27, 38, 44, 46, 115, 174]. Previous work suggests notable differences between different antigenic distance metrics [33], but how these differences affect breadth quantification is unclear.

In our study, we propose the use of antibody landscapes from influenza vaccine cohort studies as the basis for measuring vaccine breadth. We develop methods for creating population-summary antibody landscapes, and metrics derived from the summary antibody landscape which are more robust to differences in virus panels than previously used methods. We also include a case study showing how our methods can be used to assess breadth of response between two different vaccine candidates.

Methods

Data source

For this study, we performed a secondary data analysis of a subcohort from a prospective, open design ongoing vaccine cohort study. The cohort study has been described previously [61, 147], but in brief, individuals were recruited one of two study sites (Port St. Lucie, FL, or Pittsburgh, PA) during the influenza season, donated a pre-vaccination blood draw, received a Fluzone (Sanofi) seasonal influenza vaccine, and returned at a follow-up visit to donate a post-vaccination blood draw at approximately 28 days after vaccination. Each

serum sample was used for a panel of hemagglutination inhibition (HAI) assays to the strains in the vaccine, as well as for a panel of historical strains. HAI titers (the reciprocal of the highest serum dilution which showed agglutination) in our study had a lower limit of detection of 10 and an upper limit of detection of $20,480$. We considered only the influenza A(H1N1) responses to the standard dose Fluzone vaccine in our study as a proof of concept, but our methods apply equally to any panel of heterologous assays. We calculated descriptive statistics for our cohort, including counts per season and demographic summaries.

We calculated the antigenic distance between the vaccine strain (the A(H1N1) strain used in the vaccine formulation in a given season) and the assay strain (the actual virus strain added to the serum during the HAI assay) for every HAI assay using three different metrics. We calculated the temporal antigenic distance, which is the difference in isolation year between the vaccine strain and the assay strain; the dominant p-Epitope distance [30], which is the maximum proportion of amino acid differences across the five immunodominant epitope sites on the hemagglutinin head; and the cartographic distance [29]. Antigenic cartography employs statistical dimension reduction on a matrix of serological titer data before taking Euclidean distances between strains. We created a two-dimensional cartographic map with Racmacs [175] using pre-immune human sera from all assays in our study sample, and used the lowest stress map from 100 random initializations with 100 L-BFGS optimization rounds each. To fairly compare across the three antigenic distance metrics, which have different units, we min-max normalized all of the antigenic distance values within a given season.

Vaccine strength estimation framework

Next, we developed a framework for estimating vaccine performance including both the magnitude of the response (response to the homologous vaccine strain) and breadth of response (summary of responses to heterologous strains). Our framework also allows for estimation of the total strength of a vaccine, which combines the magnitude and breadth into a one-number summary of the response induced by vaccination. We developed one metric for each of the magnitude, breadth, and total strength of the response and compared these to the metrics currently used in the literature.

Metrics in the current literature (see Introduction) rely on calculating the geometric mean titer (GMT) or seroconversion rate (SCR) across a panel of historical strains. In this "current" framework, the estimate of the magnitude of the response is the GMT to the homologous strain; the estimate of the breadth of response is the seroconversion rate across all strains in the panel; and the estimate of the total strength of the response is the GMT across all strains. See the Supplement for mathematical details on these quantities.

Our novel robust metrics are derived from a sample summary antibody landscape. That is, rather than consider all of the individual antibody landscapes in our study (there is one per person-year), we fit a model that estimates the summary antibody landscape for our study sample. Any statistical model that can estimate vaccine outcomes as a function of antigenic distance could be used here, but for simplicity we used a linear regression model. In order to fairly weight the contributions of each individual antibody landscape, we modeled the post-vaccination titer as a Gaussian outcome with a global intercept, a fixed effect for antigenic distance, and correlated varying intercepts and antigenic distance effects for each individual. For simplicity, we treated all person-years from the same individual as independent data. The Supplement contains model formulas and details on our regression model implementation.

From the population summary antibody landscape, we can derive estimates of the magnitude, breadth, and total strength of the vaccine response. We estimate the magnitude as the intercept of the linear regression line (or in general, the predicted response at an antigenic distance of zero); the breadth as the antigenic distance value where the summary antibody landscape intercepts the value 40 (the HAI titer accepted as a marker for seroconversion and seroprotection); and the total strength as the area under the curve (AUC) of the regression line between antigenic distances of 0 and 1.

Model implementation and censoring correction

We implemented our summary antibody landscape models in a bayesian framework, which allows us to easily obtain uncertainty estimates for each metric as credible intervals. In order to fairly compare the current metrics with our novel metrics, we also developed models

for calculating the current metrics in a bayesian framework to obtain comparable CIs [68, 152, 176]. We estimated the homologous GMT using a Gaussian regression model for the post-vaccination titer with an intercept only, fit only to titers from the homologous strain — in this model, the estimate of the intercept estimates the post-vaccination GMT. Similarly, we estimated the total strength GMT with a Gaussian regression model for the post-vaccination titer with an intercept only, using data from all assay strains. Finally, we estimated the seroconversion rate using a logistic regression model with an intercept only, where the outcome was seroconversion to a given strain. We defined seroconversion as a post-vaccination titer of at least 40 with a 4-fold change or greater between pre- and post-vaccination titers.

Because we estimated all of our metrics using regression models, we can apply a correction to the likelihood to adjust for the interval censoring and limit of detection issues inherent to HAI titers [153]. We estimated all six of the metrics (the current metrics and our novel metrics) using models with censoring correction and without a censoring correction, since censoring is known to artificially deflate the amount of uncertainty. In order to fairly compare metrics across each study, we min-max normalized each of the metrics within each season. Since all of the metrics have different units, normalization allows us to compare the relative spread of each metric fairly.

We estimated these metrics on the entire dataset for each season to determine if vaccine strength varied seasonally after accounting for the breadth of response, and we calculated the metrics with and without a censoring correction to gain a baseline understanding of how much censoring affects the metrics.

Robustness analysis

Next, in order to show that our novel metrics are more robust to comparisons across different historical strain panels than currently used metrics, we implemented a subsampling analysis. We created 25 subsampled "studies" which are intended to simulate 25 labs analyzing the same vaccine candidate using different virus panels. For each of the subsampled studies, we randomly sampled 9 heterologous strains from the panel of either 15 or 16 for a given season,

and added the homologous strain for that season to get a panel of 10 viruses for that study. We also randomly sampled 100 individuals to "participate" in a given study. We calculated all six metrics on each subsampled study, both with and without a censoring correction.

Once we calculated the metrics for each subsampled study, we used the intraclass correlation (ICC) as a measure of consistency across the subsamples for a given metric. The ICC measures the proportion of variance in a statistic that can be attributed to a specific grouping factor, and ranges from 0 to 1. An ICC close to 0 indicates that the differences between studies contribute very little to the variation in estimated metrics, while an ICC close to 1 indicates that the differences between studies are the major source of variation in estimated metrics. We computed the ICC for each of the six metrics, both with and without a censoring correction.

We performed this robustness analysis for each seasonal cohort in our study sample to determine if the ICC for each metric changed seasonally. We calculated each ICC using a bayesian random effects model which contained only a global intercept and a random intercept for each subsampled study. The outcome in each model was the metric of interest, and we used a Gaussian likelihood distribution. See the Supplement for detailed information on all of our models.

Implementation

We implemented all of our results in a bayesian framework for two main reasons. First, estimating the vaccine response metrics in a bayesian framework allows us to calculate CIs for each metric without resorting to bootstrapping or other solutions for frequentist confidence intervals that can be unreliable. Second, our study is a secondary data analysis, so any frequentist confidence intervals and p-values we calculate would suffer from inflated false positive rates. Thus, we do not report any p-values in our results.

Our analysis was conducted with R version 4.4.2 using RStudio version 2024.12.1.563 [177]. Our project was developed as a `targets` pipeline [158] using `renv` [78] and `here` [77] to enhance reproducibility. We deployed our pipeline using the `crew` package on the University of Georgia's Sapelo2 computing cluster, which uses the Slurm scheduling software

[160, 161]. We implemented our bayesian models with the `brms` package [87–89] using the `cmdstanr` backend [90] and `cmdstan` version 2.36.0 [91] as the interface to the Stan probabilistic programming language [178, 179]. We additionally used the `tidyverse` package suite [73] and the packages `qs2` [180], `tidybayes` [74], and `pracma` [181] for formal analysis.

We created our manuscript and supplement using Quarto version 1.6.40 [82] with the packages `knitr` [83–85] and `softbib` [86]. We made our figures using `ggplot2` [79], `patchwork` [182], and `ggdist` [75, 76] and our tables using `flextable` [81]. The Supplement contains more details on our methodology, including instructions for running our code. We archived our dataset and analysis code on Zenodo https://doi.org/10.5281/zenodo.15578876 and GitHub https://github.com/ahgroup/billings-breadth-quantification-public.

Results

Data description and antibody landscape

Our study sample consisted of several thousand pairs of HAI assays (Table 4.1), with at least 1000 in each study season. These assay pairs represent homologous and heterologous HAI assay pairs (one pair includes one pre-vaccination and one post-vaccination titer). Every individual enrolling in a given season contributed the same number of HAI assays (or was excluded from our analysis if any were missing), and each study site used the same historical panel in a given season (Table 4.2). Overall, we included $1,156$ person-years of data, representing $17,601$ pairs of HAI assays. While some individuals re-enrolled in the study in successive seasons, at the same study site each time, we treated all person-years as independent for the purpose of our analysis. Thus, $1,156$ is the total number of enrolled person-years, but the total number of enrolled unique individuals was 675 (241 from the FL study site, 131 from PA, and 303 from UGA). Summary plots of the sample titer distributions, along with more information on the heterologous strains used each season, are shown in the Supplement.

Table 4.1: Count of HAI assay pairs included in our study from each study site during each season. Numbers shown are count (cell percentage). The only season in which all three study sites operated was 2016/2017, and a dash indicates that the study site for that row was not operational during the season indicated by the column.

	Season					
Site	2013/14	2014/15	2015/16	2016/17	2017/18	Overall
PA	638 (4%)	1290 (7%)	1215 (7%)	1024 (6%)	—	4167 (24%)
FL	960 (5%)	2250 (13%)	1920 (11%)	1904 (11%)	—	7034 (40%)
UGA	—	—	—	2320 (13%)	4080 (23%)	6400 (36%)
Overall	1598 (9%)	3540 (20%)	3135 (18%)	5248 (30%)	4080 (23%)	17601 (100%)

Table 4.2: Count of individuals enrolled in our study from each study site during each season. Numbers shown are count (cell percentage). The last row of the table shows the number of assay pairs contributed by each individual who enrolled in the study and completed both study visits.

	Season					
Site	2013/14	2014/15	2015/16	2016/17	2017/18	Overall
PA	68 (6%)	86 (7%)	81 (7%)	64 (6%)	—	299 (26%)
FL	60 (5%)	150 (13%)	128 (11%)	119 (10%)	—	457 (40%)
UGA	—	—	—	145 (13%)	225 (22%)	400 (35%)
Overall	128 (11%)	236 (20%)	209 (18%)	328 (28%)	255 (22%)	1156 (100%)
Num. strains	16	15	15	16	16	18

Figure 4.1 shows our summary antibody landscape for the 2016-2017 influenza season, pooling data across the three studies. We show the antibody landscapes for the three different distance metrics, both with and without the censoring correction. Overall, the censoring correction slightly increased the intercept and made the slope more negative (making the line slightly steeper) for all three models, but the choice of antigenic distance metric had a more noticeable impact on the landscapes. The summary landscapes using the cartographic distance were steeper and had a higher intercept than the p-Epitope landscapes, which in turn had steeper slopes and higher intercepts than the temporal distance landscapes.

The differences between the antigenic distance metrics are also noticeable in a qualitatively different ordering and relative placement of certain strains. Using the temporal distance, the CA/09 strain is more similar to the USSR/77 strain than to the SC/18, which is misleading. The CA/09 strain represents the pandemic-like H1N1 lineage which is more genetically and antigenic similar to the SC/18 strain, as reflected in the p-Epitope and cartographic distance orderings respectively. Notably, the p-Epitope distance placed the three distances roughly equidistant while the cartographic distance estimates the CA/09 strain as being more different from either SC/18 or USSR/77 than those strains are to each other, which intuitively reflects the unique deletions that differentiate the CA/09 clade from other H1N1 variants.

The differences between the metrics and the corrected (for censoring) vs. uncorrected landscapes is apparent in the vaccine immunogenicity metrics for this subcohort (Table 4.3). After correcting for censoring, all of the metrics increased (the current metric for breadth, seroconversion rate, which is based on a binary outcome and therefore cannot be corrected for censoring, so it did not change). While many of the corrected and uncorrected metrics had overlapping credible intervals, the intercepts in particular (novel metrics for magnitude) showed a notable increase for all three antigenic distance metrics. The difference between censoring methods shows how failing to properly account for censoring can bias vaccine results towards the null, potentially making vaccine candidates look worse than they are. Focusing only on the metrics with censoring corrections going forward, the novel metrics also varied greatly across the three antigenic distance metrics.

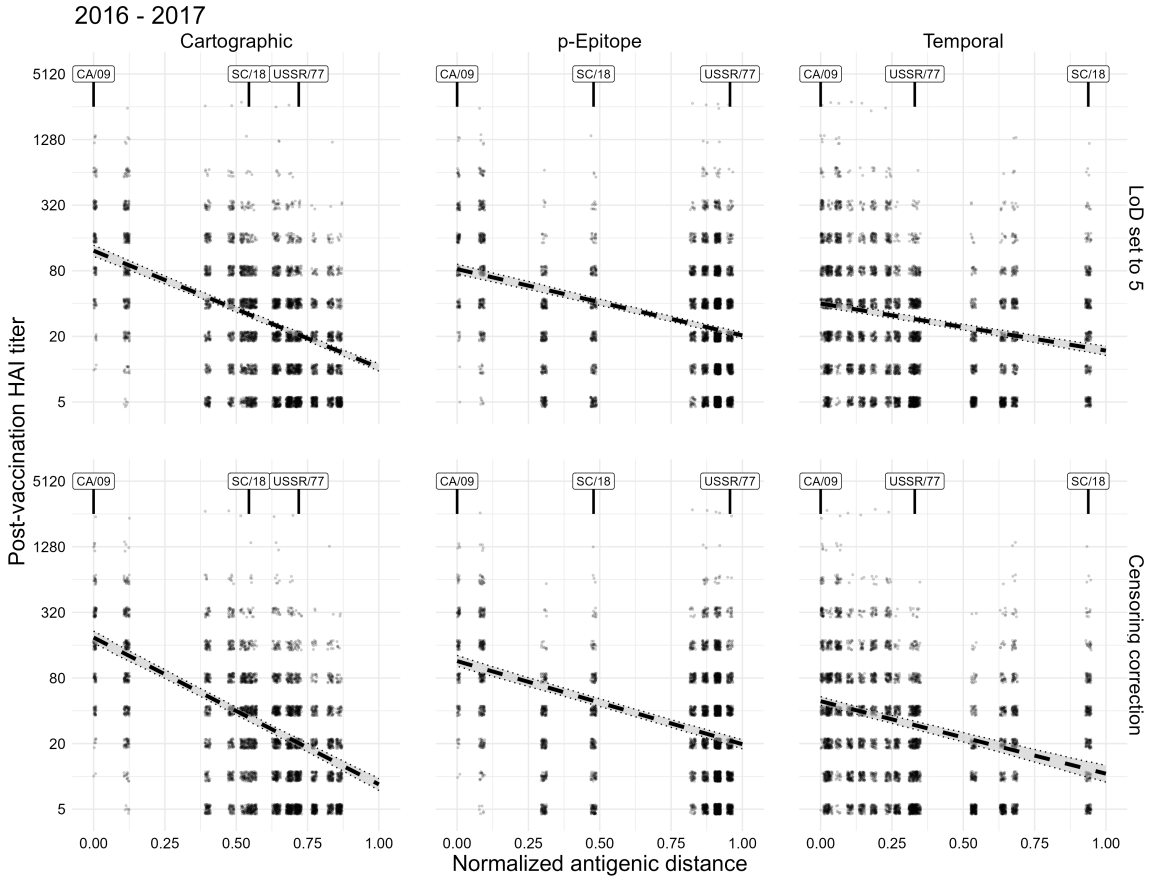

Figure 4.1: Raw data and summary antibody landscapes for the 2016 - 2017 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (CA/09 in 2016/17). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape. The strain labels show the relative positions of three different strains (CA/09, the homologous strain; SC/18, the 1918 pandemic-like strain; and USSR/77, the most different strain from the vaccine strain in the genetic comparison. Each column of plots shows a different antigenic distance metric (left: cartographic distance; center: p-Epitope sequence distance; right: temporal year-based distance). The two rows show landscapes fitted without a censoring correction (top row) where values lower than the LoD were set to 5, and the bottom row shows landscapes with the likelihood-based censoring correction.

In particular, using the temporal method (the most common method for antibody landscape analysis in the literature) underestimated all three of the metrics, because the temporal method is insufficient for capturing antigenic and genetic differences between virus strains. The p-Epitope metric yielded more optimistic results for breadth and total strength than the cartographic method, but a smaller estimate of the magnitude. From these data alone, we cannot tell which metric is more appropriate, but based on the summary landscape (@fig-landscape), the distances covered by our historical panel were different for cartographic and p-Epitope distances. While the cartographic distance had a fairly even coverage of the antigenic distance metric, there was a gap where we had no viruses with a normalized cartographic distance around 0.25. The p-Epitope metric, however, is discrete and lumps many more strains close together. Having more strains with normalized antigenic distances near 1 (which have high leverage in a regression model) may bias the intercept downward, and correspondingly affect the conditional estimate of the slope. In short, we need a panel which evenly covers antigenic space for both cartographic and p-Epitope distances before we can determine which metric produces the “correct” vaccine immunogenicity metrics.

We also noted several differences between the current set of metrics and our novel metrics. While the temporal metric underestimated immunogenicity for all three metrics relative to the current metric set, the p-Epitope metric produced a smaller estimate of the magnitude and higher estimate of the breadth and total strength. Since the novel magnitude estimate of the cartographic distance was similar to the magnitude estimate from the current metric, comparisons between current and novel metrics using the cartographic distance may be the most useful for calibrating our understanding. Both the cartographic and p-Epitope metrics produced more optimistic metrics of breadth and total strength than the current metrics or novel metrics using temporal distance, which likely indicates that using a "real" metric of genetic or antigenic distance allows for less biased estimates of vaccine breadth.

While we show only the results for the 2016/17 season here, the other seasons had similar trends and we include summary antibody landscapes and metric calculations for the other seasons in the Supplement.

Table 4.3: Current and novel vaccine immunogenicity metrics for the 2016/17 season data, shown both with and without the censoring correction. The 'Novel' metric sets were each computed using a different metric for antigenic distance. All metrics were derived from bayesian regression models and numbers shown are the posterior mean and 95% CrI.

	Censoring correction			**LoD set to 5**		
Metric Set	Magnitude	Breadth	Total Strength	Magnitude	Breadth	Total Strength
Current	5.22 (5.07, 5.39)	0.10 (0.09, 0.11)	2.64 (2.58, 2.70)	4.72 (4.55, 4.88)	0.10 (0.09, 0.11)	2.57 (2.52, 2.61)
Novel (Cartographic)	5.23 (5.06, 5.43)	0.50 (0.48, 0.52)	3.00 (2.87, 3.11)	4.61 (4.44, 4.78)	0.45 (0.43, 0.48)	2.83 (2.73, 2.92)
Novel (p-Epitope)	4.52 (4.36, 4.69)	0.60 (0.55, 0.64)	3.25 (3.13, 3.37)	4.05 (3.90, 4.20)	0.52 (0.47, 0.56)	3.04 (2.94, 3.14)
Novel (Temporal)	3.29 (3.17, 3.43)	0.14 (0.08, 0.19)	2.19 (2.05, 2.34)	3.00 (2.89, 3.11)	0.02 (0.00, 0.07)	2.28 (2.19, 2.37)

Table 4.4: Intraclass correlation coefficients (ICCs) for consistency across the subsampled studies. Each number shown is the posterior mean and 95% CrI for the ICC. An ICC closer to zero indicates that little of the variance in metric estimates is due to variability across subsamples, while an ICC closer to one indicates that variability across subsamples makes up the majority of the variation.

	Season					
Site	2013/14	2014/15	2015/16	2016/17	2017/18	Overall
PA	68 (6%)	86 (7%)	81 (7%)	64 (6%)	—	299 (26%)
FL	60 (5%)	150 (13%)	128 (11%)	119 (10%)	—	457 (40%)
UGA	—	—	—	145 (13%)	225 (22%)	400 (35%)
Overall	128 (11%)	236 (20%)	209 (18%)	328 (28%)	255 (22%)	1156 (100%)
Num. strains	16	15	15	16	16	18

Subsampling study results

To show that our novel metrics are more robust to the choice of virus panel than the current metrics for vaccine breadth, we created 25 subsampled labs from each seasonal subcohort in our study sample. Each lab had a virus panel consisting of the homologous vaccine strain for the season and 9 heterologous strains randomly sampled without replacement from the set of strains used in that season. Figure 4.2 shows the posterior distribution of each metric across the different subsamples from the 2016/17 subcohort, after correcting for censoring. The results were clearly different between the current metrics, novel metrics using cartographic distance, and novel metrics using p-Epitope distance; as well as across the metrics for magnitude, breadth, and total strength of the vaccine. For magnitude metrics, the current and novel (Cartographic) metrics distributions looked qualitatively similar, although the p-Epitope estimates were notably lower (as we observed in @fig-landscape) and there is considerably more between-group variation.

The breadth metric results were striking: the estimate of breadth using the current metric (Seroconversion rate) was very consistent around 0.10 across all subsamples – this appears to be due to a low overall rate of seroconversion in our dataset. Our novel metric showed a higher overall estimate for both cartographic and p-Epitope antigenic distances, although the results were much more consistent across subsamples for the cartographic distance than the p-epitope distance. Finally, the total strength metrics followed the same trend that we observed in the full cohort (Table 4.3), but the estimates of the novel metrics using the cartographic dataset were the most consistent across subsamples. These qualitative results are supported by the intraclass correlation (ICC) that we calculated for each of the 9 immunogenicity metrics (Table 4.4).

The current metric set had the lowest ICC among the three groups we compared for the magnitude estimate, but the highest ICC for the breadth and total strength estimates, indicating variability in the current metrics for breath and total strength was dominated by between-subsample variation. The novel metric set using the p-epitope antigenic distance had high ICCs for all three metrics. Finally, the novel metrics using the cartographic

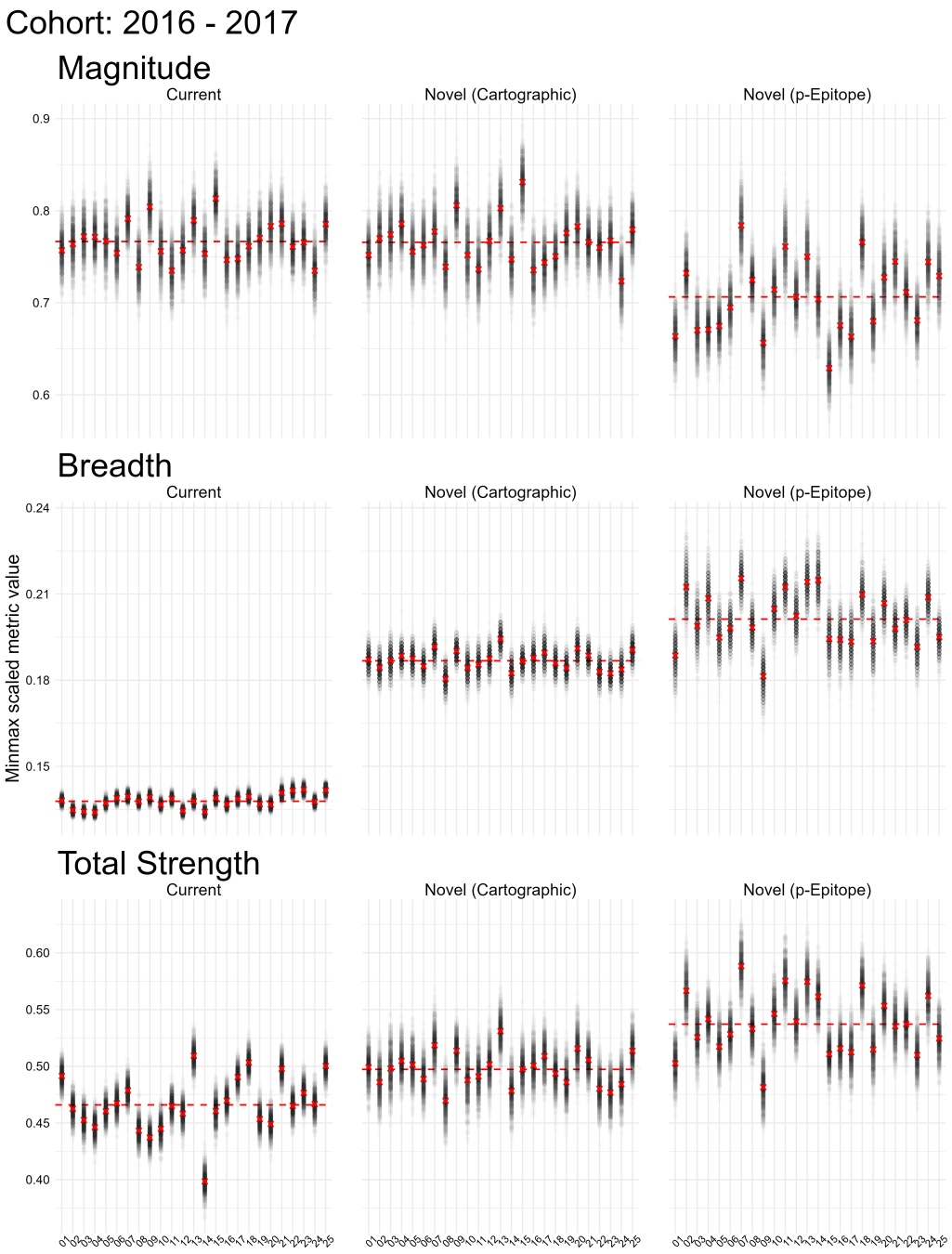

Figure 4.2: Estimated immunogenicity metrics for each simulated lab drawn from the 2016/17 subcohort data. We computed the novel metrics using both the cartographic distance and the p-Epitope distance for each subsample. The black circles show samples from the posterior distribution of each metric. The red dotted line shows the overall mean metric estimate across the subsample, and the red x for each subsample shows the mean metric estimate for that subsample. We show only 1000 posterior samples for each subsample/metric to avoid unnecessary overplotting.

antigenic distance had a comparable ICC to the current metric set for the breadth metrics, and the lowest ICC for breadth and total strength. In particular, the ICC for total strength using the cartographic distance was much lower than the current metrics, indicating that substantially less variability in our novel metrics, when the cartographic distance is used for estimation, is due to variation across subsamples. The results for the other seasons were similar and are shown in the Supplement.

Fluzone HD case study

Finally, as a case study for using our novel metrics in context, we compared participants aged 65 and older between two vaccine groups in the UGAFluVac data. Participants 65 and older could elect to receive Fluzone standard dose (SD; the rest of our study data only uses Fluzone SD) or Fluzone high dose (HD). Table 4.5 shows the number of individuals receiving HD vaccines in the study sample, and Table 4.6 shows the number of individuals aged 65 and up who received SD vaccines in the study sample.

We calculated our metrics for the magnitude, breadth, and total strength on both the SD and HD participants for each season of data (Figure 4.3). While the SD and HD vaccines performed similarly in the 2013/14 and 2016/17 seasons across all three metrics, when we compared results across all three seasons, the HD vaccine was superior in all three metrics. The credible intervals are large due to the high amount of heterogeneity and relatively small amount of older individuals in our study sample.

Table 4.5: Count of participants in each season who received the HD vaccine at each study site in our study sample. Numbers shown are counts (cell percent).

	Season					
Site	2013/14	2014/15	2015/16	2016/17	2017/18	Overall
PA	18 (9%)	39 (19%)	36 (17%)	55 (26%)	—	148 (71%)
FL	—	—	22 (11%)	19 (9%)	—	41 (20%)
UGA	—	—	—	145 (13%)	225 (22%)	400 (35%)
Overall	18 (9)	39 (19%)	58 (28%)	77 (37%)	16 (8%)	208 (100%)

Table 4.6: Count of participants aged 65 and up in each season who received the SD vaccine at each study site in our study sample. Numbers shown are counts (cell percent).

	Season					
Site	2013/14	2014/15	2015/16	2016/17	2017/18	Overall
PA	20 (22%)	19 (10%)	27 (14%)	8 (4%)	—	74 (39%)
FL	20 (22%)	35 (18%)	13 (7%)	14 (7%)	—	82 (43%)
UGA	—	—	—	23 (6)	22 (12%)	34 (18%)
Overall	40 (21)	54 (28%)	40 (21%)	34 (18%)	22 (12%)	190 (100%)

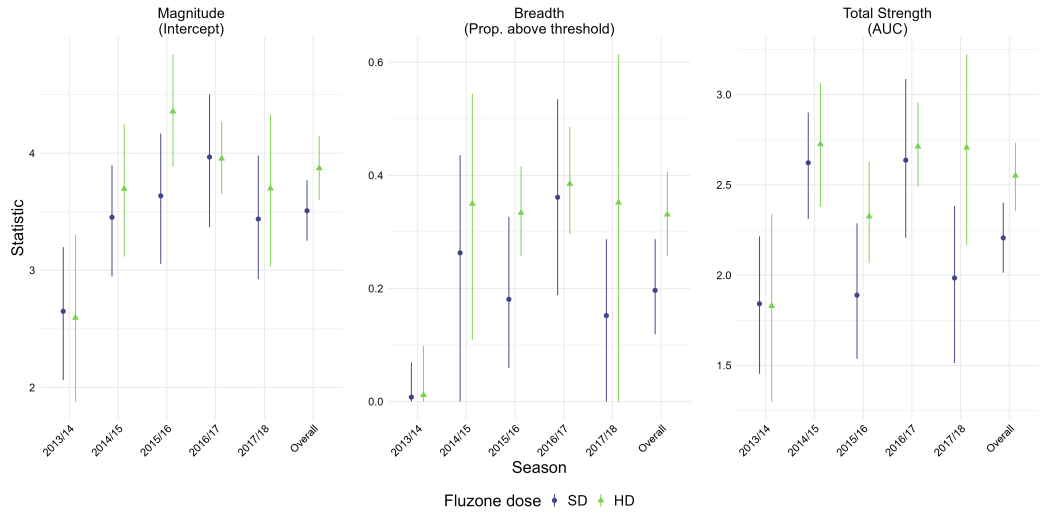

Figure 4.3: Estimated metrics derived from the summary antibody using the cartographic distance. Using individuals 65 and older from each seasonal subcohort, we calculated the metrics for the two vaccine dose groups separately. Points and lines show the mean and 95% CrI from the bayesian models estimating the metrics; color and shape indicate vaccine dose.

Discussion

Our novel methods for evaluating broadly-reactive vaccine candidates are more robust to differences in virus panels than currently used metrics. In addition, our incorporation of antigenic distance allows our metrics to be used for further analyses which provide more information about the vaccine-induced immune response than metrics based only on geometric mean titers and seroconversion rates. We found that current methods for evaluating breadth and the total strength of the vaccine response (incorporating breadth) underestimate the true breadth of vaccine response. Methods based on the seroconversion rate across strains and methods that use year of isolation (or difference between years of isolation) as a proxy for antigenic distance are the most common methods used for analyzing breadth in the literature, but neither works well.

Our metrics appear to be more robust and informative than the most common methods in current use, but rely strongly on the choice of an antigenic distance metric. However, he metrics we calculate here do not need to account for many sources of intraindivdual variability

and confounders, and uncertainty is only induced by the uncertainty in estimating model parameters from the data, so we have a much higher resolution with which to see differences in antigenic distance metrics. The major issue with choosing the correct antigenic distance metric is that there is no ontologically correct choice – different antigenic distance metrics serve different purposes. Our analysis on only A(H1N1) strains suggests that cartographic distance works well for estimating our metrics, likely because it is continuous (rather than discrete with a limited number of possible values like p-Epitope), and because the historical panel of viruses used in our study had good coverage in cartographic distance space. However, there are many issues with HAI assays [25] and results might vary across cohorts or with A(H3N2) viruses where HAI assays can be more troublesome. Since cartographic distance is widely accepted, we feel that recommending the use of cartographic distance for calculating our robust metrics is not overly hasty, although there is to date no major comparison of cartographic distances computed from ferret and human sera, or differences in cartographic maps across human cohorts.

We also do not assess the major alternative strategy for solving the problem of differences in virus panels. Methods for imputing systematically missing strain data across different panels also seem promising [39, 40], although we have not attempted to evaluate them in our subsampling context. Our robust metrics can be used alongside imputation methods, in situations where imputing titers across strain panels is reliable. A future multi-center study using the same influenza strains, vaccine candidates, and protocols (preferably with a high-throughput multiplex assay) at all study sites is necessary to truly evaluate the robustness of any procedure for measuring response breadth. Such a future study should also plan to collect sera, store serum samples long term, and run HAI assays for "future" strains that will evolve after the serum samples are collected. Since we cannot feasibly run protection or efficacy trials to assess vaccine breadth, analyzing which vaccines induce a strong response against these "future" strains will be the best surrogate to fairly assess any metric that estimates vaccine breadth.

In short, our novel metrics can be used to quantify breadth of vaccine response, and compare broadly reactive vaccine candidates more robustly across labs using different virus

panels. Understanding the best antigenic distance method to use for quantification is difficult, but using cartographic distance still provides researchers with much more information than current methods, despite the limitations of cartography. We hope that in the future our strategy for analyzing antibody landscapes to understand breadth of vaccine response can be combined with other promising methods and study designs to further the development of a universal influenza vaccine.

Acknowledgments

We thank the coauthors for their contributions to this work: Amanda L. Skarlupka, Murphy John, Savannah L. Miller, Hayley Hemme, Ted M. Ross, and Andreas Handel.

We thank Michael A. Carlock (Cleveland Clinic, Port St. Lucie, FL, USA) for assistance with obtaining data. In addition, we thank the developer of `targets`, William Michael Landau (Eli Lilly and Company, Indianapolis, IN, USA) for assistance with developing our computational pipeline.

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. Similarly, we gratefully acknowledge all data contributors for generating the genetic sequence and metadata and sharing via NCBI GenBank and UniProt, on which this research is based. More details on the sequences we used are available in the Supplement. The sequences we used from GISAID are accessible via GISAID Identifier `EPI_SET_250609vz` and DOI: https://doi.org/10.55876/gis8.250609vz.

Conflicts of interest. None of the authors declare any conflicts of interest.

Funding sources. WZB was partially supported by the University of Georgia Graduate

School as a Georgia Research Education Award Trainee. TMR is supported by the Georgia Research Alliance as an Eminent Scholar, and received partial support from NIH grant 75N93019C00052. AH received partial support from NIH grants/contracts 01AI150747, R01AI170116, and 75N93019C00052. All other authors declare no funding sources. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement. All of the code files and data necessary to reproduce our results can be found on Zenodo (https://doi.org/10.5281/zenodo.15578876) or GitHub(https://github.com/ahgroup/billings-breadth-quantification-public).

Supplementary information. The Supplement for this manuscript is included as Appendix C. A detailed explanation of the methods used to calculate antigenic distance is included as Appendix D. Finally, a detailed walkthrough of the methods for computing both the current and novel metrics in a Bayesian regression framework is included as Appendix E.

Chapter 5

Conclusion

In this dissertation, we aimed to improve our understanding of the heterologous humoral immune response to seasonal influenza vaccination. We found an inconsistent effect of dose on the vaccine-induced immune response in our cohort study, where high dose vaccination had a deleterious effect on the immune response to some heterologous strains. In an attempt to better understand the breadth of the immune response, we wanted to model the immune response with antigenic distance instead of using strain-specific effects. We found that there is no gold-standard metric for antigenic distance in the literature, but multiple metrics generated similar predictions of heterologous immunity when we compared them in a multilevel modeling approach. Finally, we applied antigenic distance to develop a robust framework for evaluating vaccine breadth.

Chapter 2

Overall, we found that high dose vaccines improve the heterologous immune response in an older cohort. However, our results suggested an inconsistent effect of the high dose vaccine. The immune response to some of the more distant heterologous strains was lower in the high dose group compared to the standard dose group, which could indicate a narrowing of the response.

While our analysis of the average causal effect conditional on the season suggested a

small positive benefit for every season in our study sample, we cannot brush aside these concerns about strains with worse responses to the HD vaccine. If a strain evolves that is more similar to those strains with worse HD responses, vaccine recipients could be at higher risk in future influenza seasons. This could result in a situation similar to the effect of the 2009 seasonal influenza vaccine on the response to the pandemic-like 2009 influenza strain. While the HD vaccine should continue to be recommended due to its beneficial effects on disease severity and lower risk of complications, further studies of the vaccine dose and careful surveillance are necessary. Monitoring vaccine response breadth should be a critical component of surveillance instead of monitoring only the VE against any circulating strains.

Our results also potentially suggest that higher vaccine doses could be beneficial. The split-inactivated vaccine is relatively cheap to produce, and high doses tend to be well-tolerated with few side effects. If pilot trials indicate that higher doses could be beneficial, we might see a stronger signal for higher dose vaccines. In some seasons where the circulating strain is virulent, a narrowing of the response to older strains could be an acceptable trade-off for a sufficiently beneficial effect against the homologous strain or similar strains.

Our results agree with the current literature on the benefits of the HD split-inactivated vaccine in older populations, but careful monitoring of the heterologous response is necessary to avoid unforeseen drops in protection.

Chapter 3

Our analysis suggests that different antigenic distance metrics are only moderately correlated in isolation, but the degree of difference between these metrics is insubstantial compared to the magnitude of interindividual variations and the effect of host factors on the immune response. We analyzed the reliability of multiple metrics of antigenic distance, and found that they were not reliable measures of the same underlying construct. Instead, these antigenic distance metrics are likely capturing different aspects of the differences between viral strains. However, when we used these different antigenic distance metrics as predictors of the post-vaccination immune response and controller for other important factors and

nuisance variation, the predicted vaccine outcome (conditional on normalized antigenic distance) was nearly identical across models.

We found that sources of nuisance variation and other causal inputs to the vaccine outcome, such as pre-vaccination titer, play a large role in predicting the heterologous immune response. Since different antigenic distance metrics generated similar predicted immune responses, we believe that these other sources of variation are more impactful on predictions than the differences between antigenic distance metrics. There are known issues with every antigenic distance metric, including cartography, which is often considered the gold standard despite the known biases that can occur when using HAI to assess antigenic distance. Sequence-based metrics are much cheaper than cartography and avoid biases inherent in certain immunological assays, and our results suggest that, in the context of modeling the heterologous vaccine response, these metrics both make similar predictions despite the known limitations of each method.

The major limitation of our study is the use of historical strain responses as a proxy for the breadth of vaccine response. An ideal longitudinal study for understanding antigenic distance and the breadth of vaccine response would collect serum samples and continuously test those serum samples against new strains that have evolved since the samples were collected. Such a study would be logistically difficult and expensive, but new advances in multiplex assay methodology may make such studies easier in the future.

Chapter 4

We were able to develop new metrics for evaluating breadth of vaccine response that are more robust to the choice of virus panel than currently used metrics. Accounting for censoring and incorporating antigenic distance into our metrics allowed us to calculate metrics for vaccine immunogenicity that take the heterologous response into account and correct for bias, preventing overly pessimistic vaccine evaluations. In an analysis of subsampled studies from our study data, our novel metrics were more consistent across subsamples than the conventional metrics by a notable margin.

However, the choice of antigenic distance metrics strongly affected our novel metrics, which was surprising given the results of chapter 3. While differences in antigenic distance metrics might be drowned out in situations with many covariates and variance components, our metrics only take into account uncertainty in statistical parameters, and do not include uncertainty from any of the variance components in our multilevel data structure. In this setting, the choice of antigenic distance appears to matter more. The results we saw might also be explained by the coverage of each metric in distance space. The p-epitope metric is discrete and our historical panel covered the p-epitope distance space more sparsely than the continuous cartographic distance space. Future studies should carefully consider the coverage of historical panels across multiple antigenic distance metrics to ensure they can fairly compare metrics.

While our novel metrics are more robust, they suffer from the same limitation as our results in chapter 3. We do not have an ultimate comparison group we can examine to tell which set of metrics gives us the true answer. While metrics that have less across-lab variability are better at giving us the correct estimate of that metric, we need a study that measures protection against future strains or synthetic strains to truly understand what metrics are best at identifying vaccines that we expect to be broadly reactive.

References

1. Kim, H., Webster, R. G. & Webby, R. J. Influenza Virus: Dealing with a Drifting and Shifting Pathogen. *Viral Immunology* **31,** 174–183. ISSN: 1557-8976. doi:10.1089/vim.2017.0141 (Mar. 2018).

2. Petrova, V. N. & Russell, C. A. The Evolution of Seasonal Influenza Viruses. *Nature Reviews Microbiology* **16,** 47–60. ISSN: 1740-1534. doi:10.1038/nrmicro.2017.118. (2025) (Jan. 2018).

3. Krammer, F. *et al.* Influenza. *Nature Reviews Disease Primers* **4,** 1–21. ISSN: 2056-676X. doi:10.1038/s41572-018-0002-y. (2021) (June 2018).

4. Bolek, H., Ozisik, L., Caliskan, Z. & Tanriover, M. D. Clinical Outcomes and Economic Burden of Seasonal Influenza and Other Respiratory Virus Infections in Hospitalized Adults. *Journal of Medical Virology* **95,** e28153. ISSN: 1096-9071. doi:10.1002/jmv.28153 (Jan. 2023).

5. Sullivan, S. G. & Cowling, B. J. Reconciling Estimates of the Global Influenza Burden. *The Lancet. Respiratory Medicine* **7,** 8–9. ISSN: 2213-2619. doi:10.1016/S2213-2600(18)30511-3 (Jan. 2019).

6. Li, L. *et al.* Influenza-Associated Excess Mortality by Age, Sex, and Subtype/Lineage: Population-Based Time-Series Study With a Distributed-Lag Nonlinear Model. *JMIR public health and surveillance* **9,** e42530. ISSN: 2369-2960. doi:10.2196/42530 (Jan. 2023).

7. Lozano, R. *et al.* Global and Regional Mortality from 235 Causes of Death for 20 Age Groups in 1990 and 2010: A Systematic Analysis for the Global Burden of Disease Study 2010. *Lancet (London, England)* **380,** 2095–2128. ISSN: 1474-547X. doi:10.1016/S0140-6736(12)61728-0 (Dec. 2012).

8. Erbelding, E. J. *et al.* A Universal Influenza Vaccine: The Strategic Plan for the National Institute of Allergy and Infectious Diseases. *The Journal of Infectious Diseases* **218,** 347–354. ISSN: 1537-6613. doi:10.1093/infdis/jiy103 (July 2018).

9. Belongia, E. A. *et al.* Variable Influenza Vaccine Effectiveness by Subtype: A Systematic Review and Meta-Analysis of Test-Negative Design Studies. *The Lancet. Infectious Diseases* **16,** 942–951. ISSN: 1474-4457. doi:10.1016/S1473-3099(16)00129-8 (Aug. 2016).

10. Okoli, G., Racovitan, F., Abdulwahid, T., Righolt, C. & Mahmud, S. Variable Seasonal Influenza Vaccine Effectiveness across Geographical Regions, Age Groups and Levels of Vaccine Antigenic Similarity with Circulating Virus Strains: A Systematic Review and Meta-Analysis of the Evidence from Test-Negative Design Studies after the 2009/10 Influenza Pandemic. *Vaccine* **39,** 1225–1240. ISSN: 0264410X. doi:10.1016/j.vaccine.2021.01.032. (2024) (Feb. 2021).

11. Trifonov, V., Khiabanian, H. & Rabadan, R. Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus. *New England Journal of Medicine* **361,** 115–119. ISSN: 0028-4793. doi:10.1056/NEJMp0904572. (2023) (July 2009).

12. CDC. *H5N1 Update: Two Human H5N1 Cases in Cambodia* https://www.cdc.gov/flu/avianflu/human-cases-cambodia.htm. Feb. 2023. (2023).

13. Wei, C.-J. *et al.* Next-Generation Influenza Vaccines: Opportunities and Challenges. *Nature Reviews Drug Discovery* **19,** 239–252. ISSN: 1474-1784. doi:10.1038/s41573-019-0056-x. (2021) (Apr. 2020).

14. Vogel, O. A. & Manicassamy, B. Broadly Protective Strategies Against Influenza Viruses: Universal Vaccines and Therapeutics. *Frontiers in Microbiology* **11,** 135. ISSN: 1664-302X. doi:10.3389/fmicb.2020.00135 (2020).

15. Henry, C., Palm, A.-K. E., Krammer, F. & Wilson, P. C. From Original Antigenic Sin to the Universal Influenza Virus Vaccine. *Trends in Immunology* **39,** 70–79. ISSN: 1471-4981. doi:10.1016/j.it.2017.08.003 (Jan. 2018).

16. Benest, J., Rhodes, S., Quaife, M., Evans, T. G. & White, R. G. Optimising Vaccine Dose in Inoculation against SARS-CoV-2, a Multi-Factor Optimisation Modelling Study to Maximise Vaccine Safety and Efficacy. *Vaccines* **9,** 78. ISSN: 2076-393X. doi:10.3390/vaccines9020078 (Jan. 2021).

17. Cowling, B. J. *et al.* Comparative Immunogenicity of Several Enhanced Influenza Vaccine Options for Older Adults: A Randomized, Controlled Trial. *Clinical Infectious Diseases* **71,** 1704–1714. ISSN: 1058-4838. doi:10.1093/cid/ciz1034. (2024) (Oct. 2020).

18. DiazGranados, C. A. *et al.* Efficacy of High-Dose versus Standard-Dose Influenza Vaccine in Older Adults. *The New England Journal of Medicine* **371,** 635–645. ISSN: 1533-4406. doi:10.1056/NEJMoa1315727 (Aug. 2014).

19. Chang, L.-J. *et al.* Safety and Immunogenicity of High-Dose Quadrivalent Influenza Vaccine in Adults ≥65 years of Age: A Phase 3 Randomized Clinical Trial. *Vaccine* **37,** 5825–5834. ISSN: 1873-2518. doi:10.1016/j.vaccine.2019.08.016 (Sept. 2019).

20. Chaves, S. S. *et al.* High-Dose Influenza Vaccine Is Associated With Reduced Mortality Among Older Adults With Breakthrough Influenza Even When There

Is Poor Vaccine-Strain Match. *Clinical Infectious Diseases* **77,** 1032–1042. ISSN: 1058-4838. doi:10.1093/cid/ciad322. (2023) (Oct. 2023).

21. Lee, J. K. H. *et al.* Efficacy and Effectiveness of High-Dose versus Standard-Dose Influenza Vaccination for Older Adults: A Systematic Review and Meta-Analysis. *Expert Review of Vaccines* **17,** 435–443. ISSN: 1744-8395. doi:10.1080/14760584.2018.1471989 (May 2018).

22. Paules, C. I., Sullivan, S. G., Subbarao, K. & Fauci, A. S. Chasing Seasonal Influenza — The Need for a Universal Influenza Vaccine. *New England Journal of Medicine* **378,** 7–9. ISSN: 0028-4793. doi:10.1056/NEJMp1714916. (2021) (Jan. 2018).

23. Angeletti, D. & Yewdell, J. W. Understanding and Manipulating Viral Immunity: Antibody Immunodominance Enters Center Stage. *Trends in Immunology* **39,** 549–561. ISSN: 1471-4906. doi:10.1016/j.it.2018.04.008. (2020) (July 2018).

24. Cowling, B. J. *et al.* Protective Efficacy of Seasonal Influenza Vaccination against Seasonal and Pandemic Influenza Virus Infection during 2009 in Hong Kong. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **51,** 1370–1379. ISSN: 1537-6591. doi:10.1086/657311 (Dec. 2010).

25. Hensley, S. E. *et al.* Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift. *Science (New York, N.Y.)* **326,** 734–736. ISSN: 1095-9203. doi:10.1126/science.1178258 (Oct. 2009).

26. Hensley, S. E. Challenges of Selecting Seasonal Influenza Vaccine Strains for Humans with Diverse Pre-Exposure Histories. *Current Opinion in Virology. Antivirals and Resistance / Virus Evolution* **8,** 85–89. ISSN: 1879-6257. doi:10.1016/j.coviro.2014.07.007. (2025) (Oct. 2014).

27. Wang, Y., Tang, C. Y. & Wan, X.-F. Antigenic Characterization of Influenza and SARS-CoV-2 Viruses. *Analytical and Bioanalytical Chemistry* **414,** 2841–2881. ISSN: 1618-2650. doi:10.1007/s00216-021-03806-6. (2025) (Apr. 2022).

28. Parker, L. *et al.* Effects of Egg-Adaptation on Receptor-Binding and Antigenic Properties of Recent Influenza A (H3N2) Vaccine Viruses. *The Journal of General Virology* **97,** 1333–1344. ISSN: 1465-2099. doi:10.1099/jgv.0.000457 (June 2016).

29. Smith, D. J. *et al.* Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **305,** 371–376. doi:10.1126/science.1097211 (July 2004).

30. Gupta, V., Earl, D. J. & Deem, M. W. Quantifying Influenza Vaccine Efficacy and Antigenic Distance. *Vaccine* **24,** 3881–3888. ISSN: 0264410X. doi:10.1016/j.vaccine.2006.01.010. (2023) (May 2006).

31. Pan, K., Subieta, K. C. & Deem, M. W. A Novel Sequence-Based Antigenic Distance Measure for H1N1, with Application to Vaccine Effectiveness and the Selection of Vaccine Strains. *Protein Engineering, Design and Selection* **24,** 291–299. ISSN: 1741-0126. doi:10.1093/protein/gzq105. (2025) (Mar. 2011).

32. Pan, Y. & Deem, M. W. Prediction of Influenza B Vaccine Effectiveness from Sequence Data. *Vaccine* **34,** 4610–4617. ISSN: 0264-410X. doi:10.1016/j.vaccine.2016.07.015. (2025) (Aug. 2016).

33. Bedford, T. *et al.* Integrating Influenza Antigenic Dynamics with Molecular Evolution. *eLife* **3,** e01914. ISSN: 2050-084X. doi:10.7554/eLife.01914. (2024) (Feb. 2014).

34. Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A. & Shraiman, B. I. Prediction, Dynamics, and Visualization of Antigenic Phenotypes of Seasonal Influenza Viruses. *Proceedings of the National Academy of Sciences* **113,** E1701–E1709. doi:10.1073/pnas.1525578113. (2024) (Mar. 2016).

35. Anderson, C. S., McCall, P. R., Stern, H. A., Yang, H. & Topham, D. J. Antigenic Cartography of H1N1 Influenza Viruses Using Sequence-Based Antigenic Distance Calculation. *BMC bioinformatics* **19,** 51. ISSN: 1471-2105. doi:10.1186/s12859-018-2042-4. (2024) (Feb. 2018).

36. Allen, J. D. & Ross, T. M. Evaluation of Next-Generation H3 Influenza Vaccines in Ferrets Pre-Immune to Historical H3N2 Viruses. *Frontiers in Immunology* **12,** 707339. ISSN: 1664-3224. doi:10.3389/fimmu.2021.707339 (2021).

37. Dugan, H. L. *et al.* Preexisting Immunity Shapes Distinct Antibody Landscapes after Influenza Virus Infection and Vaccination in Humans. *Science Translational Medicine* **12,** eabd3601. ISSN: 1946-6242. doi:10.1126/scitranslmed.abd3601 (Dec. 2020).

38. Hinojosa, M. *et al.* Impact of Immune Priming, Vaccination, and Infection on Influenza A(H3N2) Antibody Landscapes in Children. *The Journal of Infectious Diseases* **224,** 469–480. ISSN: 0022-1899. doi:10.1093/infdis/jiaa665. (2025) (Aug. 2021).

39. Huang, L. *et al.* Matrix Completion with Side Information and Its Applications in Predicting the Antigenicity of Influenza Viruses. *Bioinformatics* **33,** 3195–3201. ISSN: 1367-4803. doi:10.1093/bioinformatics/btx390. (2025) (Oct. 2017).

40. Einav, T. & Cleary, B. Extrapolating Missing Antibody-Virus Measurements across Serological Studies. *Cell Systems* **13,** 561–573.e5. ISSN: 2405-4720. doi:10.1016/j.cels.2022.06.001 (July 2022).

41. Einav, T., Creanga, A., Andrews, S. F., McDermott, A. B. & Kanekiyo, M. Harnessing Low Dimensionality to Visualize the Antibody–Virus Landscape for Influenza. *Nature Computational Science* **3,** 164–173. ISSN: 2662-8457. doi:10.1038/s43588-022-00375-1. (2023) (Feb. 2023).

42. Azulay, A., Cohen-Lavi, L., Friedman, L. M., McGargill, M. A. & Hertz, T. Mapping Antibody Footprints Using Binding Profiles. *Cell Reports Methods* **3,** 100566. ISSN: 2667-2375. doi:10.1016/j.crmeth.2023.100566. (2023) (Aug. 2023).

43. Catani, J. P. P. *et al.* The Antigenic Landscape of Human Influenza N2 Neuraminidases from 2009 until 2017. *eLife* **12** (eds Neher, R. A. & Diamond, B.) RP90782. ISSN: 2050-084X. doi:10.7554/eLife.90782. (2025) (May 2024).

44. Fonville, J. M. *et al.* Antigenic Maps of Influenza A(H3N2) Produced With Human Antisera Obtained After Primary Infection. *The Journal of Infectious Diseases* **213,** 31–38. ISSN: 0022-1899. doi:10.1093/infdis/jiv367. (2024) (Jan. 2016).

45. Auladell, M. *et al.* Influenza Virus Infection History Shapes Antibody Responses to Influenza Vaccination. *Nature Medicine* **28,** 363–372. ISSN: 1546-170X. doi:10.1038/s41591-022-01690-w (Feb. 2022).

46. Fonville, J. M. *et al.* Antibody Landscapes after Influenza Virus Infection or Vaccination. *Science (New York, N.Y.)* **346,** 996–1000. ISSN: 1095-9203. doi:10.1126/science.1256427 (Nov. 2014).

47. Centers for Disease Control and Prevention. *Past Seasons Vaccine Effectiveness Estimates* https://www.cdc.gov/flu/vaccines-work/past-seasons-estimates.html. Aug. 2021. (2021).

48. Centers for Disease Control and Prevention (CDC). *Table. Influenza Vaccines — United States, 2023–24 Influenza Season* https://www.cdc.gov/flu/professionals/acip/2022-2023/acip-table.htm. Aug. 2023. (2024).

49. Sanofi Pasteur. *FluZone Quadrivalent [Package Insert]* tech. rep. (2023). (2024).

50. Grohskopf, L. A. Prevention and Control of Seasonal Influenza with Vaccines: Recommendations of the Advisory Committee on Immunization Practices, United States, 2021–22 Influenza Season. *MMWR. Recommendations and Reports* **70.** ISSN: 1057-59871545-8601. doi:10.15585/mmwr.rr7005a1. (2021) (2021).

51. Robertson, C. A. *et al.* Fluzone® High-Dose Influenza Vaccine. *Expert Review of Vaccines* **15,** 1495–1505. ISSN: 1744-8395. doi:10.1080/14760584.2016.1254044 (Dec. 2016).

52. Centers for Disease Control and Prevention (CDC). Licensure of a High-Dose Inactivated Influenza Vaccine for Persons Aged >or=65 Years (Fluzone High-Dose) and Guidance for Use - United States, 2010. *MMWR. Morbidity and mortality weekly report* **59,** 485–486. ISSN: 1545-861X (Apr. 2010).

53. Falsey, A. R., Treanor, J. J., Tornieporth, N., Capellan, J. & Gorse, G. J. Randomized, Double-Blind Controlled Phase 3 Trial Comparing the Immunogenicity of High-Dose and Standard-Dose Influenza Vaccine in Adults 65 Years of Age and Older. *The Journal of Infectious Diseases* **200,** 172–180. ISSN: 0022-1899, 1537-6613. doi:10.1086/599790. (2024) (July 2009).

54. Wilkinson, K. *et al.* Efficacy and Safety of High-Dose Influenza Vaccine in Elderly Adults: A Systematic Review and Meta-Analysis. *Vaccine* **35,** 2775–2780. ISSN: 1873-2518. doi:10.1016/j.vaccine.2017.03.092 (May 2017).

55. Samson, S. I. *et al.* Immunogenicity of High-Dose Trivalent Inactivated Influenza Vaccine: A Systematic Review and Meta-Analysis. *Expert Review of Vaccines* **18,** 295–308. ISSN: 1476-0584. doi:10.1080/14760584.2019.1575734. (2024) (Mar. 2019).

56. Gouma, S. *et al.* Comparison of Human H3N2 Antibody Responses Elicited by Egg-Based, Cell-Based, and Recombinant Protein-Based Influenza Vaccines Dur-

ing the 2017-2018 Season. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **71,** 1447–1453. ISSN: 1537-6591. doi:10.1093/cid/ciz996 (Sept. 2020).

57. Lunny, C. *et al.* Original Research: Safety and Effectiveness of Dose-Sparing Strategies for Intramuscular Seasonal Influenza Vaccine: A Rapid Scoping Review. *BMJ Open* **11.** doi:10.1136/bmjopen-2021-050596. (2024) (2021).

58. Schnyder, J. L. *et al.* Fractional Dose of Intradermal Compared to Intramuscular and Subcutaneous Vaccination - A Systematic Review and Meta-Analysis. *Travel Medicine and Infectious Disease* **37,** 101868. ISSN: 1477-8939. doi:10.1016/j.tmaid.2020.101868. (2024) (Sept. 2020).

59. Couch, R. B. *et al.* Safety and Immunogenicity of a High Dosage Trivalent Influenza Vaccine among Elderly Subjects. *Vaccine* **25,** 7656–7663. ISSN: 0264410X. doi:10.1016/j.vaccine.2007.08.042. (2024) (Nov. 2007).

60. Hilleman, M. R. ANTIBODY RESPONSE IN VOLUNTEERS TO ASIAN INFLUENZA VACCINE. *Journal of the American Medical Association* **166,** 1134. ISSN: 0002-9955. doi:10.1001/jama.1958.02990100022005. (2024) (Mar. 1958).

61. Nuñez, I. A. *et al.* Impact of Age and Pre-Existing Influenza Immune Responses in Humans Receiving Split Inactivated Influenza Vaccine on the Induction of the Breadth of Antibodies to Influenza A Strains. *PloS One* **12,** e0185666. ISSN: 1932-6203. doi:10.1371/journal.pone.0185666 (2017).

62. Carlock, M. A. *et al.* Impact of Age and Pre-Existing Immunity on the Induction of Human Antibody Responses against Influenza B Viruses. *Human Vaccines & Immunotherapeutics* **15,** 2030–2043. ISSN: 2164-5515. doi:10.1080/21645515.2019.1642056. (2021) (Sept. 2019).

63. Abreu, R. B., Clutter, E. F., Attari, S., Sautto, G. A. & Ross, T. M. IgA Responses Following Recurrent Influenza Virus Vaccination. *Frontiers in Immunology* **11.** ISSN: 1664-3224. doi:10.3389/fimmu.2020.00902. (2021) (2020).

64. CDC. *National, Regional, and State Level Outpatient Illness and Viral Surveillance* \url{https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html}. 2021. (2021).

65. Beyer, W. E. P., Palache, A. M., Lüchters, G., Nauta, J. & Osterhaus, A. D. M. E. Seroprotection Rate, Mean Fold Increase, Seroconversion Rate: Which Parameter Adequately Expresses Seroresponse to Influenza Vaccination? *Virus Research. Proceedings of the First European Influenza Conference* **103,** 125–132. ISSN: 0168-1702. doi:10.1016/j.virusres.2004.02.024. (2022) (July 2004).

66. Ranjeva, S. *et al.* Age-Specific Differences in the Dynamics of Protective Immunity to Influenza. *Nature Communications* **10,** 1660. ISSN: 2041-1723. doi:10.1038/s41467-019-09652-6. (2024) (Apr. 2019).

67. Faraway, J. J. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition* ISBN: 978-1-315-38272-2. (2021) (2016).

68. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* 2nd ed. ISBN: 978-0-367-13991-9 (Taylor and Francis, CRC Press, Boca Raton, 2020).

69. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81,** 945–960. ISSN: 0162-1459. doi:10.2307/2289064. JSTOR: 2289064. (2024) (1986).

70. Naimi, A. I. & Whitcomb, B. W. Defining and Identifying Average Treatment Effects. *American Journal of Epidemiology* **192,** 685–687. ISSN: 0002-9262, 1476-6256. doi:10.1093/aje/kwad012. (2024) (May 2023).

71. Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual* 2024.

72. R Core Team. *R: A Language and Environment for Statistical Computing* Manual (Vienna, Austria, 2025).

73. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4,** 1686. ISSN: 2475-9066. doi:10.21105/joss.01686. (2025) (Nov. 2019).

74. Kay, M. *tidybayes: Tidy Data and Geoms for Bayesian Models* Manual (2023). doi:10.5281/zenodo.1308151.

75. Kay, M. ggdist: Visualizations of Distributions and Uncertainty in the Grammar of Graphics. *IEEE Transactions on Visualization and Computer Graphics* **30,** 414–424. doi:10.1109/TVCG.2023.3327195 (2024).

76. Kay, M. *ggdist: Visualizations of Distributions and Uncertainty* Manual (2024). doi:10.5281/zenodo.3879620.

77. Müller, K. *Here: A Simpler Way to Find Your Files* Manual (2020).

78. Ushey, K. & Wickham, H. *Renv: Project Environments* Manual (2025).

79. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4 (Springer-Verlag New York, 2016).

80. Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A. & Larmarange, J. Reproducible Summary Tables with the Gtsummary Package. *The R Journal* **13,** 570–580. doi:10.32614/RJ-2021-053 (2021).

81. Gohel, D. & Skintzos, P. *Flextable: Functions for Tabular Reporting* Manual (2024).

82. Allaire, J., Teague, C., Scheidegger, C., Xie, Y. & Dervieux, C. *Quarto* Feb. 2024. doi:10.5281/zenodo.5960048.

83. Xie, Y. *Knitr: A General-Purpose Package for Dynamic Report Generation in R* Manual (2024).

84. Xie, Y. *Dynamic Documents with R and Knitr* 2nd ed. (Chapman and Hall/CRC, Boca Raton, Florida, 2015).

85. Xie, Y. in *Implementing Reproducible Computational Research* (eds Stodden, V., Leisch, F. & Peng, R. D.) (Chapman and Hall/CRC, 2014).

86. Arel-Bundock, V. & McCrain, J. Software Citations in Political Science. *PS: Political Science & Politics* **April,** 1–4. doi:10.1017/S1049096523000239 (2023).

87. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80,** 1–28. doi:10.18637/jss.v080.i01 (2017).

88. Bürkner, P.-C. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* **10,** 395–411. doi:10.32614/RJ-2018-017 (2018).

89. Bürkner, P.-C. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* **100,** 1–54. doi:10.18637/jss.v100.i05 (2021).

90. Gabry, J., Češnovar, R., Johnson, A. & Bronder, S. *Cmdstanr: R Interface to 'CmdStan'* Manual (2024).

91. Stan Development Team. *CmdStan User's Guide* 2025. (2025).

92. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76,** 1–32. ISSN: 1548-7660. doi:10.18637/jss.v076.i01. (2021) (Jan. 2017).

93. Pleguezuelos, O., Robinson, S., Fernandez, A., Stoloff, G. A. & Caparrós-Wanderley, W. Meta-Analysis and Potential Role of Preexisting Heterosubtypic Cellular Immunity Based on Variations in Disease Severity Outcomes for In-

fluenza Live Viral Challenges in Humans. *Clinical and Vaccine Immunology : CVI* **22,** 949. doi:10.1128/CVI.00101-15. (2024) (July 2015).

94. Nguyen, T. H. O. *et al.* Immune Cellular Networks Underlying Recovery from Influenza Virus Infection in Acute Hospitalized Patients. *Nature Communications* **12,** 2691. ISSN: 2041-1723. doi:10.1038/s41467-021-23018-x. (2021) (May 2021).

95. Oidtman, R. J. *et al.* Influenza Immune Escape under Heterogeneous Host Immune Histories. *Trends in microbiology* **29,** 1072–1082. ISSN: 0966-842X. doi:10.1016/j.tim.2021.05.009. (2024) (Dec. 2021).

96. Jones-Gray, E., Robinson, E. J., Kucharski, A. J., Fox, A. & Sullivan, S. G. Does Repeated Influenza Vaccination Attenuate Effectiveness? A Systematic Review and Meta-Analysis. *The Lancet. Respiratory Medicine* **11,** 27–44. ISSN: 2213-2619. doi:10.1016/S2213-2600(22)00266-1 (Jan. 2023).

97. Moritzky, S. A. *et al.* The Negative Effect of Preexisting Immunity on Influenza Vaccine Responses Transcends the Impact of Vaccine Formulation Type and Vaccination History. *The Journal of Infectious Diseases* **227,** 381–390. ISSN: 0022-1899. doi:10.1093/infdis/jiac068. (2023) (Feb. 2022).

98. McLean, H. Q., Levine, M. Z., King, J. P., Flannery, B. & Belongia, E. A. Serologic Response to Sequential Vaccination with Enhanced Influenza Vaccines: Open Label Randomized Trial among Adults Aged 65–74 Years. *Vaccine* **39,** 7146–7152. ISSN: 0264410X. doi:10.1016/j.vaccine.2021.10.072. (2023) (Dec. 2021).

99. Bi, Q. *et al.* Reduced Effectiveness of Repeat Influenza Vaccination: Distinguishing among within-Season Waning, Recent Clinical Infection, and Subclinical Infection. *The Journal of Infectious Diseases,* jiae220. ISSN: 1537-6613. doi:10.1093/infdis/jiae220 (Apr. 2024).

100. Fox, A. *et al.* Opposing Effects of Prior Infection versus Prior Vaccination on Vaccine Immunogenicity against Influenza A(H3N2) Viruses. *Viruses* **14,** 470. ISSN: 1999-4915. doi:10.3390/v14030470. (2023) (Mar. 2022).

101. van de Sandt, C. E., Kreijtz, J. H. C. M. & Rimmelzwaan, G. F. Evasion of Influenza A Viruses from Innate and Adaptive Immune Responses. *Viruses* **4,** 1438–1476. ISSN: 1999-4915. doi:10.3390/v4091438. (2025) (Sept. 2012).

102. Morens, D. M., Taubenberger, J. K. & Fauci, A. S. The 2009 H1N1 Pandemic Influenza Virus: What Next? *mBio* **1,** e00211–10. ISSN: 2150-7511. doi:10.1128/mBio.00211-10. (2025) (Sept. 2010).

103. Paules, C. I. & Fauci, A. S. Influenza Vaccines: Good, but We Can Do Better. *The Journal of Infectious Diseases* **219,** S1–S4. ISSN: 0022-1899. doi:10.1093/infdis/jiy633. (2025) (Apr. 2019).

104. Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans. *Science* **314,** 1898–1903. doi:10.1126/science.1132745. (2025) (Dec. 2006).

105. Zinder, D., Bedford, T., Gupta, S. & Pascual, M. The Roles of Competition and Mutation in Shaping Antigenic and Genetic Diversity in Influenza. *PLOS Pathogens* **9,** e1003104. ISSN: 1553-7374. doi:10.1371/journal.ppat.1003104. (2025) (Jan. 2013).

106. Centers for Disease Control and Prevention (CDC). *National, Regional, and State Level Outpatient Illness and Viral Surveillance* https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. (2024).

107. Smith, D. J., Forrest, S., Ackley, D. H. & Perelson, A. S. Variable Efficacy of Repeated Annual Influenza Vaccination. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 14001–14006. ISSN: 0027-8424 1091-6490. doi:10.1073/pnas.96.24.14001. (2024) (Nov. 1999).

108. Skowronski, D. M. *et al.* Serial Vaccination and the Antigenic Distance Hypothesis: Effects on Influenza Vaccine Effectiveness During A(H3N2) Epidemics in Canada, 2010-2011 to 2014-2015. *The Journal of infectious diseases* **215,** 1059–1099. ISSN: 1537-6613 0022-1899. doi:10.1093/infdis/jix074. (2023) (Apr. 2017).

109. Lim, W. W. *et al.* The Impact of Repeated Vaccination on Relative Influenza Vaccine Effectiveness among Vaccinated Adults in the United Kingdom. *Epidemiology and Infection* **150,** e198. ISSN: 1469-4409. doi:10.1017/S0950268822001753 (Nov. 2022).

110. Xie, H. *et al.* H3N2 Mismatch of 2014–15 Northern Hemisphere Influenza Vaccines and Head-to-head Comparison between Human and Ferret Antisera Derived Antigenic Maps. *Scientific Reports* **5,** 15279. ISSN: 2045-2322. doi:10.1038/srep15279. (2025) (Oct. 2015).

111. Sandbulte, M. R. *et al.* Discordant Antigenic Drift of Neuraminidase and Hemagglutinin in H1N1 and H3N2 Influenza Viruses. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 20748–20753. ISSN: 1091-6490 0027-8424. doi:10.1073/pnas.1113801108. (2025) (Dec. 2011).

112. Smith, G. J. D. *et al.* Origins and Evolutionary Genomics of the 2009 Swine-Origin H1N1 Influenza A Epidemic. *Nature* **459,** 1122–1125. ISSN: 1476-4687. doi:10.1038/nature08182. (2025) (June 2009).

113. Morimoto, N. & Takeishi, K. Change in the Efficacy of Influenza Vaccination after Repeated Inoculation under Antigenic Mismatch: A Systematic Review and Meta-Analysis. *Vaccine* **36,** 949–957. ISSN: 1873-2518. doi:10.1016/j.vaccine.2018.01.023 (Feb. 2018).

114. Gambaryan, A. S., Robertson, J. S. & Matrosovich, M. N. Effects of Egg-Adaptation on the Receptor-Binding Properties of Human Influenza A and B

Viruses. *Virology* **258,** 232–239. ISSN: 0042-6822. doi:10.1006/viro.1999.9732. (2025) (June 1999).

115. Hay, J. A., Laurie, K., White, M. & Riley, S. Characterising Antibody Kinetics from Multiple Influenza Infection and Vaccination Events in Ferrets. *PLoS Computational Biology* **15,** e1007294. ISSN: 1553-734X. doi:10.1371/journal.pcbi.1007294. (2023) (Aug. 2019).

116. Sun, H. *et al.* Using Sequence Data To Infer the Antigenicity of Influenza Virus. *mBio* **4,** e00230–13. ISSN: 2150-7511. doi:10.1128/mBio.00230-13. (2025) (July 2013).

117. Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting Evolution from the Shape of Genealogical Trees. *eLife* **3,** e03568. ISSN: 2050-084X. doi:10.7554/eLife.03568. (2025) (Nov. 2014).

118. Han, L. *et al.* Graph-Guided Multi-Task Sparse Learning Model: A Method for Identifying Antigenic Variants of Influenza A(H3N2) Virus. *Bioinformatics* **35,** 77–87. ISSN: 1367-4803. doi:10.1093/bioinformatics/bty457. (2025) (Jan. 2019).

119. Li, L. *et al.* Multi-Task Learning Sparse Group Lasso: A Method for Quantifying Antigenicity of Influenza A(H1N1) Virus Using Mutations and Variations in Glycosylation of Hemagglutinin. *BMC Bioinformatics* **21,** 182. ISSN: 1471-2105. doi:10.1186/s12859-020-3527-5. (2025) (May 2020).

120. Harvey, W. T. *et al.* A Bayesian Approach to Incorporate Structural Data into the Mapping of Genotype to Antigenic Phenotype of Influenza A(H3N2) Viruses. *PLoS computational biology* **19,** e1010885. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1010885 (Mar. 2023).

121. Jia, Q., Xia, Y., Dong, F. & Li, W. MetaFluAD: Meta-Learning for Predicting Antigenic Distances among Influenza Viruses. *Briefings in bioinformatics* **25.** ISSN: 1477-4054 1467-5463. doi:10.1093/bib/bbae395 (July 2024).

122. Wang, L. *et al. Using Pathogen Genomics to Predict Antigenic Changes of Influenza H3N2 Virus* Aug. 2024. doi:10.21203/rs.3.rs-4788510/v1. (2025).

123. Cai, Z., Zhang, T. & Wan, X.-F. A Computational Framework for Influenza Antigenic Cartography. *PLoS computational biology* **6,** e1000949. ISSN: 1553-7358 1553-734X. doi:10.1371/journal.pcbi.1000949 (Oct. 2010).

124. Barnett, J. L., Yang, J., Cai, Z., Zhang, T. & Wan, X.-F. AntigenMap 3D: An Online Antigenic Cartography Resource. *Bioinformatics (Oxford, England)* **28,** 1292–1293. ISSN: 1367-4811 1367-4803. doi:10.1093/bioinformatics/bts105 (May 2012).

125. Arhami, O. & Rohani, P. *TOPOLOW: A MAPPING ALGORITHM FOR ANTIGENIC CROSS-REACTIVITY AND BINDING AFFINITY ASSAY RESULTS* Feb. 2025. doi:10.1101/2025.02.09.637307. (2025).

126. Waldock, J. *et al.* Assay Harmonization and Use of Biological Standards To Improve the Reproducibility of the Hemagglutination Inhibition Assay: A FLUCOP Collaborative Study. *mSphere* **6,** 10.1128/msphere.00567–21. doi:10.1128/msphere.00567-21. (2025) (July 2021).

127. Zacour, M. *et al.* Standardization of Hemagglutination Inhibition Assay for Influenza Serology Allows for High Reproducibility between Laboratories. *Clinical and Vaccine Immunology* **23,** 236–242. doi:10.1128/CVI.00613-15. (2025) (Mar. 2016).

128. Adabor, E. S. & Ndifon, W. Bayesian Inference of Antigenic and Non-Antigenic Variables from Haemagglutination Inhibition Assays for Influenza Surveillance.

Royal Society Open Science **5,** 180113. ISSN: 2054-5703. doi:10.1098/rsos.180113. (2023) (July 2018).

129. Ndifon, W. New Methods for Analyzing Serological Data with Applications to Influenza Surveillance. *Influenza and Other Respiratory Viruses* **5,** 206–212. ISSN: 1750-2659. doi:10.1111/j.1750-2659.2010.00192.x. (2025) (2011).

130. Forghani, M. & Khachay, M. Convolutional Neural Network Based Approach to in Silico Non-Anticipating Prediction of Antigenic Distance for Influenza Virus. *Viruses* **12,** 1019. ISSN: 1999-4915. doi:10.3390/v12091019. (2025) (Sept. 2020).

131. Cai, Z., Zhang, T. & Wan, X.-F. Antigenic Distance Measurements for Seasonal Influenza Vaccine Selection. *Vaccine* **30,** 448–453. ISSN: 1873-2518 0264-410X. doi:10.1016/j.vaccine.2011.10.051. (2025) (Jan. 2012).

132. Vijaykrishna, D. *et al.* The Contrasting Phylodynamics of Human Influenza B Viruses. *eLife* **4** (ed Neher, R. A.) e05055. ISSN: 2050-084X. doi:10.7554/eLife.05055. (2025) (Jan. 2015).

133. Koel, B. F. *et al.* Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science* **342,** 976–979. doi:10.1126/science.1244730. (2025) (Nov. 2013).

134. Harvey, W. T. *et al.* Identification of Low- and High-Impact Hemagglutinin Amino Acid Substitutions That Drive Antigenic Drift of Influenza A(H1N1) Viruses. *PLoS pathogens* **12,** e1005526. ISSN: 1553-7374. doi:10.1371/journal.ppat.1005526 (Apr. 2016).

135. Mögling, R. *et al.* Neuraminidase-Mediated Haemagglutination of Recent Human Influenza A(H3N2) Viruses Is Determined by Arginine 150 Flanking the Neuraminidase Catalytic Site. *The Journal of General Virology* **98,** 1274–1281. ISSN: 1465-2099. doi:10.1099/jgv.0.000809. (2025) (June 2017).

136. Castro, L. A., Bedford, T. & Meyers, L. A. Early Prediction of Antigenic Transitions for Influenza A/H3N2. *PLOS Computational Biology* **16,** e1007683. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1007683. (2024) (Feb. 2020).

137. Borkenhagen, L. K., Allen, M. W. & Runstadler, J. A. Influenza Virus Genotype to Phenotype Predictions through Machine Learning: A Systematic Review. *Emerging Microbes & Infections* **10,** 1896–1907. ISSN: 2222-1751. doi:10.1080/22221751.2021.1978824. (2025) (Dec. 2021).

138. Forna, A. *et al.* Sequence-Based Detection of Emerging Antigenically Novel Influenza A Viruses. *Proceedings of the Royal Society B: Biological Sciences* **291,** 20240790. doi:10.1098/rspb.2024.0790. (2024) (Aug. 2024).

139. Lefrancq, N. *et al.* Learning the Fitness Dynamics of Pathogens from Phylogenies. *Nature* **637,** 683–690. ISSN: 1476-4687. doi:10.1038/s41586-024-08309-9. (2025) (Jan. 2025).

140. Boyoglu-Barnum, S. *et al.* Quadrivalent Influenza Nanoparticle Vaccines Induce Broad Protection. *Nature* **592,** 623–628. ISSN: 1476-4687. doi:10.1038/s41586-021-03365-x. (2025) (Apr. 2021).

141. Li, Z.-N. *et al.* Antibody Landscape Analysis Following Influenza Vaccination and Natural Infection in Humans with a High-Throughput Multiplex Influenza Antibody Detection Assay. *mBio* **12,** 10.1128/mbio.02808–20. doi:10.1128/mbio.02808-20. (2025) (Feb. 2021).

142. Yang, B. *et al.* Life Course Exposures Continually Shape Antibody Profiles and Risk of Seroconversion to Influenza. *PLOS Pathogens* **16,** e1008635. ISSN: 1553-7374. doi:10.1371/journal.ppat.1008635. (2025) (July 2020).

143. Carlock, M. A. & Ross, T. M. A Computationally Optimized Broadly Reactive Hemagglutinin Vaccine Elicits Neutralizing Antibodies against Influenza B

Viruses from Both Lineages. *Scientific Reports* **13,** 15911. ISSN: 2045-2322. doi:10.1038/s41598-023-43003-2. (2025) (Sept. 2023).

144. Nachbagauer, R. *et al.* Defining the Antibody Cross-Reactome Directed against the Influenza Virus Surface Glycoproteins. *Nature Immunology* **18,** 464–473. ISSN: 1529-2916. doi:10.1038/ni.3684. (2025) (Apr. 2017).

145. Anderson, C. S. *et al.* Implementing Sequence-Based Antigenic Distance Calculation into Immunological Shape Space Model. *BMC Bioinformatics* **21,** 256. ISSN: 1471-2105. doi:10.1186/s12859-020-03594-3. (2023) (Dec. 2020).

146. Lee, M.-S. & Chen, J. S.-E. Predicting Antigenic Variants of Influenza A/H3N2 Viruses - Volume 10, Number 8—August 2004 - Emerging Infectious Diseases Journal - CDC. *Emerging Infectious Diseases* **10,** 1385–1390. doi:10.3201/eid1008.040107. (2025) (2004).

147. Carlock, M. A., Allen, J. D., Hanley, H. B. & Ross, T. M. Longitudinal Assessment of Human Antibody Binding to Hemagglutinin Elicited by Split-Inactivated Influenza Vaccination over Six Consecutive Seasons. *PloS One* **19,** e0301157. ISSN: 1932-6203. doi:10.1371/journal.pone.0301157. (2024) (2024).

148. Abreu, R. B., Kirchenbaum, G. A., Sautto, G. A., Clutter, E. F. & Ross, T. M. Impaired Memory B-cell Recall Responses in the Elderly Following Recurrent Influenza Vaccination. *PLOS ONE* **16** (ed Huber, V. C.) e0254421. ISSN: 1932-6203. doi:10.1371/journal.pone.0254421. (2023) (Aug. 2021).

149. Centers for Disease Control and Prevention (CDC). *CDC Seasonal Flu Vaccine Effectiveness Studies* https://web.archive.org/web/20250311230637/https://www.cdc.gov/flu-vaccines-work/php/effectiveness-studies/index.html. Mar. 2025. (2025).

150. Billings, W. Z. *et al.* High-Dose Inactivated Influenza Vaccine Inconsistently Improves Heterologous Antibody Responses in an Older Human Cohort. *The*

Journal of Infectious Diseases, jiaf003. ISSN: 0022-1899. doi:10.1093/infdis/jiaf003. (2025) (Jan. 2025).

151. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185,** 862–864. ISSN: 0036-8075. doi:10.1126/science.185.4154.862. (2025) (Sept. 1974).

152. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* ISBN: 978-0-521-86706-1 (Cambridge University Press, Cambridge ; New York, 2007).

153. Breen, R. *Regression Models: Censored, Sample Selected, or Truncated Data* ISBN: 978-1-4129-8561-1. doi:10.4135/9781412985611. (2023) (SAGE Publications, Inc., 1996).

154. Arel-Bundock, V., Greifer, N. & Heiss, A. How to Interpret Statistical Models Using marginaleffects for R and Python. *Journal of Statistical Software* **111,** 1–32. doi:10.18637/jss.v111.i09 (2024).

155. Vehtari, A. *et al. Loo: Efficient Leave-One-out Cross-Validation and WAIC for Bayesian Models* 2024.

156. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* **27,** 1413–1432. doi:10.1007/s11222-016-9696-4 (2017).

157. RStudio Team. *RStudio: Integrated Development Environment for R* Manual (Boston, MA, 2024).

158. Landau, W. M. The Targets R Package: A Dynamic Make-like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. *Journal of Open Source Software* **6,** 2959. ISSN: 2475-9066. doi:10.21105/joss.02959. (2025) (Jan. 2021).

159. Scott, E. R. *Targets-Uahpc: A Template to Use the {targets} R Package with UA-HPC* Zenodo. Apr. 2025. doi:10.5281/zenodo.15232875. (2025).

160. Yoo, A. B., Jette, M. A. & Grondona, M. *SLURM: Simple Linux Utility for Resource Management* in *Job Scheduling Strategies for Parallel Processing* (eds Feitelson, D., Rudolph, L. & Schwiegelshohn, U.) (Springer, Berlin, Heidelberg, 2003), 44–60. ISBN: 978-3-540-39727-4. doi:10.1007/10968987_3.

161. Jette, M. A. & Wickberg, T. *Architecture of the Slurm Workload Manager* in *Job Scheduling Strategies for Parallel Processing* (eds Klusáček, D., Corbalán, J. & Rodrigo, G. P.) (Springer Nature Switzerland, Cham, 2023), 3–23. ISBN: 978-3-031-43943-8. doi:10.1007/978-3-031-43943-8_1.

162. Bååth, R. *Bayesboot: An Implementation of Rubin's (1981) Bayesian Bootstrap* Manual (2025).

163. Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. Using Stacking to Average Bayesian Predictive Distributions. *Bayesian Analysis.* doi:10.1214/17-BA1091 (2017).

164. Schloerke, B. *et al. GGally: Extension to 'Ggplot2'* Manual (2024).

165. Zost, S. J. *et al.* Contemporary H3N2 Influenza Viruses Have a Glycosylation Site That Alters Binding of Antibodies Elicited by Egg-Adapted Vaccine Strains. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 12578–12583. ISSN: 1091-6490. doi:10.1073/pnas.1712377114. (2025) (Nov. 2017).

166. Cobey, S. Vaccination against Rapidly Evolving Pathogens and the Entanglements of Memory. *Nature Immunology* **25,** 2015–2023. ISSN: 1529-2916. doi:10.1038/s41590-024-01970-2. (2024) (Nov. 2024).

167. Bedford, T. *et al.* Global Circulation Patterns of Seasonal Influenza Viruses Vary with Antigenic Drift. *Nature* **523,** 217–220. ISSN: 1476-4687. doi:10.1038/nature14460. (2024) (July 2015).

168. Stacey, H. *et al. Leveraging Pre-Vaccination Antibody Titers across Multiple Influenza H3N2 Variants to Forecast the Post-Vaccination Response* Aug. 2024. doi:10.1101/2024.08.01.24311325. (2024).

169. Taaffe, J. *et al.* Advancing Influenza Vaccines: A Review of next-Generation Candidates and Their Potential for Global Health Impact. *Vaccine* **42,** 126408. ISSN: 1873-2518. doi:10.1016/j.vaccine.2024.126408 (Dec. 2024).

170. Dugan, H. L., Henry, C. & Wilson, P. C. Aging and Influenza Vaccine-Induced Immunity. *Cellular Immunology* **348,** 103998. ISSN: 1090-2163. doi:10.1016/j.cellimm.2019.103998 (Feb. 2020).

171. Jang, H. & Ross, T. M. Hemagglutination Inhibition (HAI) Antibody Landscapes after Vaccination with H7Nx Virus like Particles. *PloS One* **16,** e0246613. ISSN: 1932-6203. doi:10.1371/journal.pone.0246613 (2021).

172. Loes, A. N. *et al.* High-Throughput Sequencing-Based Neutralization Assay Reveals How Repeated Vaccinations Impact Titers to Recent Human H1N1 Influenza Strains. *Journal of Virology* **98,** e0068924. ISSN: 1098-5514. doi:10.1128/jvi.00689-24. (2024) (Oct. 2024).

173. Dang, C. C., Le, Q. S., Gascuel, O. & Le, V. S. FLU, an Amino Acid Substitution Model for Influenza Proteins. *BMC Evolutionary Biology* **10,** 99. ISSN: 1471-2148. doi:10.1186/1471-2148-10-99. (2024) (Apr. 2010).

174. Kucharski, A. J., Lessler, J., Cummings, D. A. T. & Riley, S. Timescales of Influenza A/H3N2 Antibody Dynamics. *PLOS Biology* **16,** e2004974. ISSN: 1545-7885. doi:10.1371/journal.pbio.2004974. (2023) (Aug. 2018).

175. Wilks, S. *Racmacs: Antigenic Cartography Macros* Manual (2023).

176. Gelman, A. *et al. Bayesian Workflow* Nov. 2020. doi:10.48550/arXiv.2011.01808. arXiv: 2011.01808 [stat]. (2025).

177. Posit team. *RStudio: Integrated Development Environment for R* Manual (Boston, MA, 2025).

178. Stan Development Team. *Stan Reference Manual* 2025. (2025).

179. Stan Development Team. *Stan User's Guide* 2025. (2025).

180. Ching, T. *Qs2: Efficient Serialization of R Objects* Manual (2025).

181. Borchers, H. W. *Pracma: Practical Numerical Math Functions* Manual (2023).

182. Pedersen, T. L. *Patchwork: The Composer of Plots* Manual (2024).

183. DiazGranados, C. A. *et al.* High-Dose Trivalent Influenza Vaccine Compared to Standard Dose Vaccine in Elderly Adults: Safety, Immunogenicity and Relative Efficacy during the 2009–2010 Season. *Vaccine* **31,** 861–866. ISSN: 0264410X. doi:10.1016/j.vaccine.2012.12.013. (2024) (Jan. 2013).

184. DiazGranados, C. A. *et al.* Efficacy and Immunogenicity of High-Dose Influenza Vaccine in Older Adults by Age, Comorbidities, and Frailty. *Vaccine* **33,** 4565–4571. ISSN: 0264410X. doi:10.1016/j.vaccine.2015.07.003. (2024) (Aug. 2015).

185. Lee, J. K. H. *et al.* Efficacy and Effectiveness of High-Dose Influenza Vaccine in Older Adults by Circulating Strain and Antigenic Match: An Updated Systematic Review and Meta-Analysis. *Vaccine* **39 Suppl 1,** A24–A35. ISSN: 1873-2518. doi:10.1016/j.vaccine.2020.09.004 (Mar. 2021).

186. Dunning, A. J. *et al.* Correlates of Protection against Influenza in the Elderly: Results from an Influenza Vaccine Efficacy Trial. *Clinical and Vaccine Immunology* **23,** 228–235. doi:10.1128/CVI.00604-15. (2023) (Mar. 2016).

187. Li, A. P. Y. *et al.* Immunogenicity of Standard, High-Dose, MF59-adjuvanted, and Recombinant-HA Seasonal Influenza Vaccination in Older Adults. *npj Vaccines* **6,** 25. ISSN: 2059-0105. doi:10.1038/s41541-021-00289-5. (2023) (Feb. 2021).

188. Izurieta, H. S. *et al.* Comparative Effectiveness of High-Dose versus Standard-Dose Influenza Vaccines in US Residents Aged 65 Years and Older from 2012 to 2013 Using Medicare Data: A Retrospective Cohort Analysis. *The Lancet Infectious Diseases* **15,** 293–300. ISSN: 14733099. doi:10.1016/S1473-3099(14)71087-4. (2023) (Mar. 2015).

189. Izurieta, H. S. *et al.* Comparative Effectiveness of Influenza Vaccines Among US Medicare Beneficiaries Ages 65 Years and Older During the 2019-2020 Season. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **73,** e4251–e4259. ISSN: 1537-6591. doi:10.1093/cid/ciaa1727 (Dec. 2021).

190. Robison, S. G. & Thomas, A. R. Assessing the Effectiveness of High-Dose Influenza Vaccine in Preventing Hospitalization among Seniors, and Observations on the Limitations of Effectiveness Study Design. *Vaccine* **36,** 6683–6687. ISSN: 0264410X. doi:10.1016/j.vaccine.2018.09.050. (2023) (Oct. 2018).

191. Hsiao, A. *et al.* Recombinant or Standard-Dose Influenza Vaccine in Adults under 65 Years of Age. *New England Journal of Medicine* **389,** 2245–2255. ISSN: 0028-4793, 1533-4406. doi:10.1056/NEJMoa2302099. (2024) (Dec. 2023).

192. Doyle, J. D. *et al.* Relative and Absolute Effectiveness of High-Dose and Standard-Dose Influenza Vaccine Against Influenza-Related Hospitalization Among Older Adults—United States, 2015–2017. *Clinical Infectious Diseases* **72,** 995–1003. ISSN: 1058-4838. doi:10.1093/cid/ciaa160. (2021) (Mar. 2021).

193. Paudel, M. *et al.* Relative Vaccine Efficacy of High-Dose versus Standard-Dose Influenza Vaccines in Preventing Probable Influenza in a Medicare Fee-for-Service Population. *Vaccine* **38,** 4548–4556. ISSN: 0264410X. doi:10.1016/j.vaccine.2020.05.020. (2023) (June 2020).

194. Naleway, A. L. *et al.* Immunogenicity of High-Dose Egg-Based, Recombinant, and Cell Culture–Based Influenza Vaccines Compared With Standard-Dose Egg-Based Influenza Vaccine Among Health Care Personnel Aged 18–65 Years in 2019–2020. *Open Forum Infectious Diseases* **10,** ofad223. ISSN: 2328-8957. doi:10.1093/ofid/ofad223. (2024) (June 2023).

195. Schmader, K. E. *et al.* Immunogenicity of Adjuvanted versus High-Dose Inactivated Influenza Vaccines in Older Adults: A Randomized Clinical Trial. *Immunity & Ageing* **20,** 30. ISSN: 1742-4933. doi:10.1186/s12979-023-00355-7. (2024) (July 2023).

196. Coleman, B. L., Sanderson, R., Haag, M. D. M. & McGovern, I. Effectiveness of the MF59-adjuvanted Trivalent or Quadrivalent Seasonal Influenza Vaccine among Adults 65 Years of Age or Older, a Systematic Review and Meta-Analysis. *Influenza and Other Respiratory Viruses* **15,** 813–823. ISSN: 1750-2659. doi:10.1111/irv.12871. (2024) (2021).

197. Arevalo, P., McLean, H. Q., Belongia, E. A. & Cobey, S. Earliest Infections Predict the Age Distribution of Seasonal Influenza A Cases. *eLife* **9** (eds Cooper, B. S., Ferguson, N. M., Cooper, B. S. & Baguelin, M.) e50060. ISSN: 2050-084X. doi:10.7554/eLife.50060. (2021) (July 2020).

198. Vieira, M. C. *et al.* Lineage-Specific Protection and Immune Imprinting Shape the Age Distributions of Influenza B Cases. *Nature Communications* **12,** 4313. ISSN: 2041-1723. doi:10.1038/s41467-021-24566-y. (2024) (July 2021).

199. Gostic, K. M., Ambrose, M., Worobey, M. & Lloyd-Smith, J. O. Potent Protection against H5N1 and H7N9 Influenza via Childhood Hemagglutinin Imprinting. *Science* **354,** 722–726. ISSN: 0036-8075, 1095-9203. doi:10.1126/science.aag1322. (2021) (Nov. 2016).

200. Gostic, K. M. *et al.* Childhood Immune Imprinting to Influenza A Shapes Birth Year-Specific Risk during Seasonal H1N1 and H3N2 Epidemics. *PLOS Pathogens* **15,** e1008109. ISSN: 1553-7374. doi:10.1371/journal.ppat.1008109. (2021) (Dec. 2019).

201. Wood, S. N. Thin-Plate Regression Splines. *Journal of the Royal Statistical Society (B)* **65,** 95–114 (2003).

202. Wood, S. N. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* **99,** 673–686 (2004).

203. Wood, S. N. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society (B)* **73,** 3–36 (2011).

204. Wood, S., N., Pya & S"afken, B. Smoothing Parameter and Model Selection for General Smooth Models (with Discussion). *Journal of the American Statistical Association* **111,** 1548–1575 (2016).

205. Wood, S. *Generalized Additive Models: An Introduction with R* 2nd ed. (Chapman and Hall/CRC, 2017).

206. Pedersen, E. J., Miller, D. L., Simpson, G. L. & Ross, N. Hierarchical Generalized Additive Models in Ecology: An Introduction with Mgcv. *PeerJ* **7,** e6876. ISSN: 2167-8359. doi:10.7717/peerj.6876. (2024) (May 2019).

207. Lewandowski, D., Kurowicka, D. & Joe, H. Generating Random Correlation Matrices Based on Vines and Extended Onion Method. *Journal of Multivariate Analysis* **100,** 1989–2001. ISSN: 0047-259X. doi:10.1016/j.jmva.2009.04.008. (2024) (Oct. 2009).

208. *Prior Choice Recommendations · Stan-Dev/Stan Wiki* https://github.com/stan-dev/stan. (2021).

209. Betancourt, M. *A Conceptual Introduction to Hamiltonian Monte Carlo* July 2018. doi:10.48550/arXiv.1701.02434. arXiv: 1701.02434 [stat]. (2022).

210. Cooke, P. Statistical Inference for Bounds of Random Variables. *Biometrika* **66,** 367–374. ISSN: 0006-3444. doi:10.1093/biomet/66.2.367. (2024) (Aug. 1979).

211. Loh, W.-Y. Estimating an Endpoint of a Distribution with Resampling Methods. *The Annals of Statistics* **12,** 1543–1550. ISSN: 0090-5364, 2168-8966. doi:10.1214/aos/1176346811. (2024) (Dec. 1984).

212. Sheather, S. J. & Jones, M. C. A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **53,** 683–690. ISSN: 0035-9246. JSTOR: 2345597. (2024) (1991).

213. Liu, F. *et al.* Age-Specific Effects of Vaccine Egg Adaptation and Immune Priming on A(H3N2) Antibody Responses Following Influenza Vaccination. *The Journal of Clinical Investigation* **131.** ISSN: 0021-9738. doi:10.1172/JCI146138. (2024) (Apr. 2021).

214. Yang, D. & Dalton, J. E. A Unified Approach to Measuring the Effect Size between Two Groups Using SAS®. *SAS Global Form 2012* (2012).

215. Saul, B. *Smd: Compute Standardized Mean Differences* Manual (2024).

216. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd ed. (Routledge, London, England, May 2013).

217. Sawilowsky, S. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods* **8.** ISSN: 1538 - 9472. doi:10.22237/jmasm/1257035100 (Nov. 2009).

218. Gelman, A. *et al.* Bayesian Data Analysis (2021).

219. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. *Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC* June 2021. doi:10.48550/arXiv.1903.08008. arXiv: 1903.08008. (2024).

220. Wood, S. N. *Generalized Additive Models: An Introduction with R, Second Edition* 2nd ed. ISBN: 978-1-315-37027-9. doi:10.1201/9781315370279 (Chapman and Hall/CRC, New York, May 2017).

221. Hamming, R. W. Error Detecting and Error Correcting Codes. *The Bell System Technical Journal* **29,** 147–160. ISSN: 0005-8580. doi:10.1002/j.1538-7305.1950.tb00463.x. (2025) (Apr. 1950).

222. Burke, D. F. & Smith, D. J. A Recommended Numbering Scheme for Influenza A HA Subtypes. *PLoS ONE* **9,** e112302. ISSN: 1932-6203. doi:10.1371/journal.pone.0112302. (2025) (Nov. 2014).

223. National Library of Medicine (US), National Center for Biotechnology Information. *GenBank* Bethesda, MD, 1982.

224. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Research* **44,** D67–D72. ISSN: 0305-1048. doi:10.1093/nar/gkv1276. (2025) (Jan. 2016).

225. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53,** D609–D617. ISSN: 1362-4962. doi:10.1093/nar/gkae1010. (2025) (Jan. 2025).

226. Shu, Y. & McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality. *Eurosurveillance* **22,** 30494. ISSN: 1560-7917. doi:10.2807/1560-7917.ES.2017.22.13.30494. (2025) (Mar. 2017).

227. Elbe, S. & Buckland-Merrett, G. Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health. *Global Challenges* **1,** 33–46. ISSN: 2056-6646. doi:10.1002/gch2.1018. (2025) (2017).

228. Reid, A. H., Fanning, T. G., Hultin, J. V. & Taubenberger, J. K. Origin and Evolution of the 1918 "Spanish" Influenza Virus Hemagglutinin Gene. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 1651–1656. ISSN: 0027-8424. doi:10.1073/pnas.96.4.1651 (Feb. 1999).

229. Smeenk, C. A. & Brown, E. G. The Influenza Virus Variant A/FM/1/47-MA Possesses Single Amino Acid Replacements in the Hemagglutinin, Controlling Virulence, and in the Matrix Protein, Controlling Virulence as Well as Growth. *Journal of Virology* **68,** 530–534. ISSN: 0022-538X. doi:10.1128/JVI.68.1.530-534.1994 (Jan. 1994).

230. Lee, M.-S. & Yang, C.-F. Cross-Reactive H1N1 Antibody Responses to a Live Attenuated Influenza Vaccine in Children: Implication for Selection of Vaccine Strains. *The Journal of Infectious Diseases* **188,** 1362–1366. ISSN: 0022-1899. doi:10.1086/379045 (Nov. 2003).

231. Barman, S. *et al.* Pathogenicity and Transmissibility of North American Triple Reassortant Swine Influenza A Viruses in Ferrets. *PLoS pathogens* **8,** e1002791. ISSN: 1553-7374. doi:10.1371/journal.ppat.1002791 (2012).

232. Garten, R. J. *et al.* Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. *Science (New York, N.Y.)* **325,** 197–201. ISSN: 1095-9203. doi:10.1126/science.1176225 (July 2009).

233. Bauer, K., Schrader, C., Suess, J., Wutzler, P. & Schmidtke, M. Neuraminidase Inhibitor Susceptibility of Porcine H3N2 Influenza A Viruses Isolated in Germany between 1982 and 1999. *Antiviral Research* **75,** 219–226. ISSN: 0166-3542. doi:10.1016/j.antiviral.2007.03.007 (Sept. 2007).

234. Xu, X., Kilbourne, E. D., Hall, H. E. & Cox, N. J. Nonimmunoselected Intrastrain Genetic Variation Detected in Pairs of High-Yielding Influenza A (H3N2) Vaccine and Parental Viruses. *The Journal of Infectious Diseases* **170,** 1432–1438. ISSN: 0022-1899. doi:10.1093/infdis/170.6.1432 (Dec. 1994).

235. Mohr, P. G., Deng, Y.-M. & McKimm-Breschkin, J. L. The Neuraminidases of MDCK Grown Human Influenza A(H3N2) Viruses Isolated since 1994 Can Demonstrate Receptor Binding. *Virology Journal* **12,** 67. ISSN: 1743-422X. doi:10.1186/s12985-015-0295-3 (Apr. 2015).

236. Krystal, M. *et al.* Sequential Mutations in Hemagglutinins of Influenza B Virus Isolates: Definition of Antigenic Domains. *Proceedings of the National Academy of Sciences of the United States of America* **80,** 4527–4531. ISSN: 0027-8424. doi:10.1073/pnas.80.14.4527 (July 1983).

237. Yamashita, M., Krystal, M., Fitch, W. M. & Palese, P. Influenza B Virus Evolution: Co-Circulating Lineages and Comparison of Evolutionary Pattern with Those of Influenza A and C Viruses. *Virology* **163,** 112–122. ISSN: 0042-6822. doi:10.1016/0042-6822(88)90238-3 (Mar. 1988).

238. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. Msa: An R Package for Multiple Sequence Alignment. *Bioinformatics (Oxford, England)* **31,** 3997–3999. ISSN: 1367-4811. doi:10.1093/bioinformatics/btv494. (2024) (Dec. 2015).

239. Edgar, R. C. MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinformatics* **5,** 113. ISSN: 1471-2105. doi:10.1186/1471-2105-5-113. (2025) (Aug. 2004).

240. McDonald, N. J., Smith, C. B. & Cox, N. J. Antigenic Drift in the Evolution of H1N1 Influenza A Viruses Resulting from Deletion of a Single Amino Acid in the Haemagglutinin Gene. *Journal of General Virology* **88,** 3209–3213. ISSN: 1465-2099. doi:10.1099/vir.0.83184-0. (2025) (2007).

241. McCullers, J. A., Wang, G. C., He, S. & Webster, R. G. Reassortment and Insertion-Deletion Are Strategies for the Evolution of Influenza B Viruses in Nature. *Journal of Virology* **73,** 7343–7348. ISSN: 0022-538X. doi:10.1128/jvi.73.9.7343-7348.1999. (2025) (Sept. 1999).

242. Shu, B. *et al.* Detection and Discrimination of Influenza B Victoria Lineage Deletion Variant Viruses by Real-Time RT-PCR. *Eurosurveillance* **25,** 1900652. ISSN: 1025-496X. doi:10.2807/1560-7917.ES.2020.25.41.1900652. (2025) (Oct. 2020).

243. Yang, B. *et al.* Breadth of Influenza A Antibody Cross-Reactivity Varies by Virus Isolation Interval and Subtype. *Nature Microbiology,* 1–12. ISSN: 2058-5276. doi:10.1038/s41564-025-02033-4. (2025) (June 2025).

244. van der Loo, M. P. The Stringdist Package for Approximate String Matching. *The R Journal* **6,** 111–122. doi:10.32614/RJ-2014-011 (2014).

245. Deem, M. W. & Pan, K. The Epitope Regions of H1-subtype Influenza A, with Application to Vaccine Efficacy. *Protein engineering, design & selection : PEDS* **22,** 543–546. ISSN: 1741-0134 1741-0126. doi:10.1093/protein/gzp027. (2025) (Sept. 2009).

246. Muñoz, E. T. & Deem, M. W. Epitope Analysis for Influenza Vaccine Design. *Vaccine* **23,** 1144–1148. ISSN: 0264-410X. doi:10.1016/j.vaccine.2004.08.028. (2025) (Jan. 2005).

247. Sun, W. *et al.* Antibody Responses toward the Major Antigenic Sites of Influenza B Virus Hemagglutinin in Mice, Ferrets, and Humans. *Journal of Virology* **93,** e01673–18. ISSN: 0022-538X. doi:10.1128/JVI.01673-18. (2024) (Jan. 2019).

248. Verhoeyen, M., Van Rompuy, L., Jou, W. M., Huylebroeck, D. & Fiers, W. Complete Nucleotide Sequence of the Influenza B/Singapore/222/79 Virus Hemagglutinin Gene and Comparison with the B/Lee/40 Hemagglutinin. *Nucleic Acids Research* **11,** 4703–4712. ISSN: 0305-1048. doi:10.1093/nar/11.14.4703. (2025) (July 1983).

249. Krystal, M., Elliott, R. M., Benz, E. W., Young, J. F. & Palese, P. Evolution of Influenza A and B Viruses: Conservation of Structural Features in the Hemagglutinin Genes. *Proceedings of the National Academy of Sciences of the United States of America* **79,** 4800–4804. ISSN: 0027-8424. doi:10.1073/pnas.79.15.4800. (2025) (Aug. 1982).

250. Schliep, K. P. Phangorn: Phylogenetic Analysis in R. *Bioinformatics (Oxford, England)* **27,** 592–593. ISSN: 1367-4811. doi:10.1093/bioinformatics/btq706 (Feb. 2011).

251. Schliep, K., Potts, A. J., Morrison, D. A. & Grimm, G. W. Intertwining Phylogenetic Trees and Networks. *Methods in Ecology and Evolution* **8,** 1212–1220 (2017).

252. Yang, Z. Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites. *Molecular Biology and Evolution* **10,** 1396–1401. ISSN: 0737-4038. doi:10.1093/oxfordjournals.molbev.a040082 (Nov. 1993).

253. Gu, X., Fu, Y. X. & Li, W. H. Maximum Likelihood Estimation of the Heterogeneity of Substitution Rate among Nucleotide Sites. *Molecular Biology and Evolution* **12,** 546–557. ISSN: 0737-4038. doi:10.1093/oxfordjournals.molbev.a040235 (July 1995).

254. Susko, E., Field, C., Blouin, C. & Roger, A. J. Estimation of Rates-across-Sites Distributions in Phylogenetic Substitution Models. *Systematic Biology* **52,** 594–603. ISSN: 1063-5157. doi:10.1080/10635150390235395 (Oct. 2003).

255. Felsenstein, J. Taking Variation of Evolutionary Rates between Sites into Account in Inferring Phylogenies. *Journal of Molecular Evolution* **53,** 447–455. ISSN: 0022-2844. doi:10.1007/s002390010234 (2001).

256. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32,** 268–274. ISSN: 0737-4038. doi:10.1093/molbev/msu300. (2025) (Jan. 2015).

257. Paradis, E. & Schliep, K. Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. *Bioinformatics (Oxford, England)* **35,** 526–528. doi:10.1093/bioinformatics/bty633 (2019).

258. Cox, M. A. A. & Cox, T. F. in *Handbook of Data Visualization* (eds Chen, C.-h., Härdle, W. & Unwin, A.) 315–347 (Springer, Berlin, Heidelberg, 2008). ISBN: 978-3-540-33037-0. doi:10.1007/978-3-540-33037-0_14. (2025).

259. Lapedes, A. & Farber, R. The Geometry of Shape Space: Application to Influenza. *Journal of Theoretical Biology* **212,** 57–69. ISSN: 0022-5193. doi:10.1006/jtbi.2001.2347. (2023) (Sept. 2001).

260. Allaire, J. *et al. Quarto* July 2024. doi:10.5281/zenodo.5960048.

261. Morgan, M. & Ramos, M. *BiocManager: Access the Bioconductor Project Package Repository* Manual (2024).

262. Landau, W. M. *Crew: A Distributed Worker Launcher Framework* Manual (2025).

263. Landau, W. M., Levin, M. G. & Furneaux, B. *Crew.Cluster: Crew Launcher Plugins for Traditional High-Performance Computing Clusters* Manual (2025).

264. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *Dplyr: A Grammar of Data Manipulation* Manual (2023).

265. Wickham, H. *Forcats: Tools for Working with Categorical Variables (Factors)* Manual (2023).

266. Firke, S. *Janitor: Simple Tools for Examining and Cleaning Dirty Data* Manual (2024).

267. Bengtsson, H. *Parallelly: Enhancing the 'parallel' Package* Manual (2025).

268. Wickham, H. & Henry, L. *Purrr: Functional Programming Tools* Manual (2025).

269. Wickham, H., Hester, J. & Bryan, J. *Readr: Read Rectangular Text Data* Manual (2024).

270. Wickham, H. & Bryan, J. *Readxl: Read Excel Files* Manual (2023).

271. Henry, L. & Wickham, H. *Rlang: Functions for Base Types and Core R and 'tidyverse' Features* Manual (2024).

272. Allaire, J. *et al. Rmarkdown: Dynamic Documents for R* Manual (2024).

273. Xie, Y., Allaire, J. & Grolemund, G. *R Markdown: The Definitive Guide* ISBN: 978-1-138-35933-8 (Chapman and Hall/CRC, Boca Raton, Florida, 2018).

274. Xie, Y., Dervieux, C. & Riederer, E. *R Markdown Cookbook* ISBN: 978-0-367-56383-7 (Chapman and Hall/CRC, Boca Raton, Florida, 2020).

275. Wickham, H. *Stringr: Simple, Consistent Wrappers for Common String Operations* Manual (2023).

276. Landau, W. M. *Tarchetypes: Archetypes for Targets* Manual (2021).

277. Müller, K. & Wickham, H. *Tibble: Simple Data Frames* Manual (2023).

278. Wickham, H., Vaughan, D. & Girlich, M. *Tidyr: Tidy Messy Data* Manual (2024).

279. Almende B.V. and Contributors & Thieurmel, B. *visNetwork: Network Visualization Using 'vis.Js' Library* Manual (2022).

280. Viboud, C. *et al.* Beyond Clinical Trials: Evolutionary and Epidemiological Considerations for Development of a Universal Influenza Vaccine. *PLoS pathogens* **16,** e1008583. ISSN: 1553-7374. doi:10.1371/journal.ppat.1008583 (Sept. 2020).

281. Durviaux, S. *et al.* Genetic and Antigenic Typing of Seasonal Influenza Virus Breakthrough Cases from a 2008-2009 Vaccine Efficacy Trial. *Clinical and Vaccine Immunology* **21,** 271–279. doi:10.1128/CVI.00544-13. (2024) (Mar. 2014).

Appendices

Appendix A

Supplementary material for Chapter 2

Reproducibility instructions

In order to reproduce our results, you should first download the archived repo from Zenodo (here: https://doi.org/10.5281/zenodo.12666976) or clone the Git repository (hosted on GitHub here: https://github.com/ahgroup/Billings-2024-HD-Heterologous). (You can also download the repository as a zipped folder from the GitHub page). If you use different software or package versions than what we used, or run the results in a different order, you may get errors or inconsistent results.

We ran the analysis on a Windows 10 Enterprise 64-bit (build 19045) machine with 64 GB RAM and a 36-core processor. Any statements we make about the execution time of code will vary across machines, especially if the hardware is different from these specs.

Before you can reproduce our results you will need to install the following software requirements.

- `R` version 4.4.1, available from https://cran.r-project.org/.
- RTools 4.4, also available from CRAN.
- The RStudio IDE, available from https://posit.co/download/rstudio-desktop/. We used version 2023.12.1+402 Ocean Storm (desktop).
- Quarto version 1.5.24, available from https://quarto.org/.
- Version 1.0.7 of the `renv` package for R, available from `https://cran.r-project.org/`

`web/packages/renv/index.html`.

With the software installed, follow these instructions to reproduce our results.

1. Open the `SD-HD-flu-vaccine.Rproj` file in RStudio.
2. Once `renv` initializes, run the command `renv::restore()` in the Console to begin installing the required packages. If you encounter issues, you may also run `renv::deactivate()` and install the packages manually. However, if you do not use `renv` or use different package versions, the following instructions might not work as intended.
3. **(Optional)** If you want to re-run the Bayesian models, you need to install `cmdstan` at this step. If `cmdstanr` was installed successfully using `renv`, follow the CmdStanR quick start guide at https://mc-stan.org/cmdstanr/articles/cmdstanr.html, starting from the section titled *Installing CmdStan*. Installing `cmdstan` can be challenging—please read the instructions carefully. If you encounter issues with the path or installation, you may need to:
   - Open a new R GUI or RStudio session as administrator (on Windows),
   - Manually install version `v1.0.8` of `cmdstanr`, and
   - Re-run the installation and path setup steps.

   If problems persist, the Stan discourse forum (https://discourse.mc-stan.org/) is a good place to seek help.
4. You should now be able to run the code files in the `code` directory. These are designed to run in the following order, although some steps may be skipped since we provide output files:
   - `02-Data-Summary.R`: Recreates many of the summary tables using the provided input files. This script runs quickly.
   - `03-Model-Fitting.R`: Specifies `brms` models and runs HMC sampling for Bayesian regression models. **This script is very time-consuming.** You do not need to run it to reproduce the model results, as we provide the fitted model files.
   - `04-Posterior-Summaries.R`: Computes the (c)ACE estimates from fitted models. Can be run without executing `03`. **Execution time: approximately one**

hour. Time estimates are printed after the first set of (c)ACEs is computed. This script reconstructs model fit files.

- `05-Model-Results.R`: Processes (c)ACE estimates into manuscript figures. It uses the `all-cates-combined.Rds` file generated by script 04. With cleaned data and this file, all figures can be reproduced.
- `06-Supplementary-Analyses.R`: Contains additional calculations for the Supplement. The first part (DAG and tables) runs with only the cleaned data. We recommend running script `04` first to ensure model fit files are available.
- The `common-functions` directory includes various helper functions. Running these on their own is not particularly useful unless all required packages are installed correctly.

5. After running all code files, you can render unformatted versions of the manuscript and supplement by building the following files in RStudio:
   - `products/manuscript/manuscript.qmd`
   - `products/manuscript/supplement.qmd`

   When opened in RStudio with Quarto installed, a `Render` button will appear in the GUI to facilitate this process.

Note: The script `01-Data-Processing.R` generates the finalized data set used in the Supplement. We do not provide the input file `clean-data.Rds` due to data sensitivity concerns. Instead, we provide the output files in the `data/processed` directory.

Computational Note: The Bayesian models are computationally intensive compared to frequentist alternatives. Script `03` required several days to run on a relatively powerful computer and will take substantial time even on modern hardware. We advise re-running the model fitting step only for readers who wish to validate the results.

Additional Background References

We also consulted additional references on the immunogenicity benefits [18–20, 53–55, 59, 183–187] and clinical benefits [20, 21, 188–193] of the Fluzone HD vaccine relative to the SD

vaccine. These references are included here due to citation limits in the main text imposed by JID. Some studies that included younger adults did not find a difference between the HD and SD vaccines [194].

Notably, in the main text we do not discuss comparisons with vaccines other than the split-inactivated SD vaccine, but the HD vaccine also appears to perform similarly to other enhanced vaccines such as those with adjuvants [187, 195, 196]. Unfortunately, we do not have access to data from study groups that received other enhanced vaccine candidates, but comparing the effects of, among others, adjuvanted, and high-dose vaccines on the heterologous antibody response could reveal intriguing differences in the underlying immunology of the different formulations.

Expanded Methods

Causal model for confounding

Since we used observational data rather than clinical trial data to estimate the effect of vaccine dose on immunological responses, we needed to adopt a causal model to control for confounding. A confounder is any other factor which can affect both the treatment an individual receives (i.e., which dose they got) and the outcome. We represented our causal assumptions using a directed acyclic graph (DAG), shown in Figure A.1.

In order to show the DAG nicely, the variable names are abbreviated by single letters. The letters in Figure A.1 correspond to the following variables in our data (Table A.1).

The observed confounders we included in our model were age, calendar time, and birth year. Age could potentially be a confounder since elderly individuals are more encouraged to get the high dose vaccine, and older individuals tend to have attenuated immune responses [37]. Birth year was included as a confounder for the same reason – older individuals could be more likely to receive HD vaccines, and the cohort effect of birth year imprinting affects vaccine response independently of age [197–200]. Finally, we included the season as the calendar year to account for differences in vaccines between seasons. Some vaccine formulations are more immunogenic than others, which varies by season. Additionally,

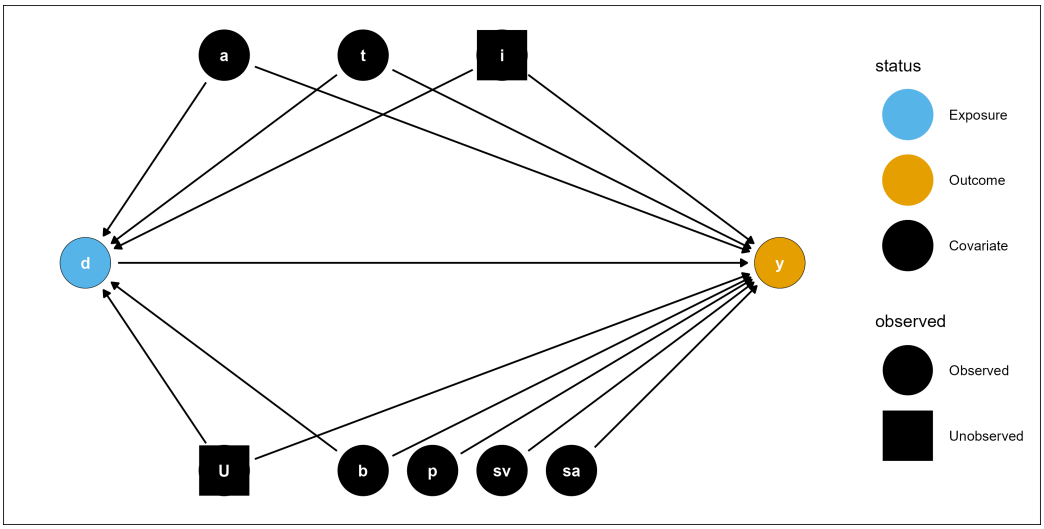

Figure A.1: The DAG we adopted as our causal model. Nodes indicate variables and arrows follow the direction of causality, i.e., an arrow from X to Y indicates that X is a cause of Y.

we included this as a confounder because the production and administration rate of HD vaccines could change by season, although we have no data with which to test this hypothesis. However, including controlling for a variable which is not actually a confounder does not include bias in the model results. (Unless collider bias is induced, which we do not believe is the case here as there are no relevant variables that have both season and dose or season and vaccine response as common causes. Any such variables are temporally separated from the treatment and response.)

Table A.1: Abbreviations used for variable names in Figure A.1.

Abbreviation	**Variable**
`d`	Vaccine dose
`y`	Immunological outcome
`a`	Age
`t`	Season
`i`	Other individual effects
`U`	Other unobserved confounders
`b`	Birth year
`p`	Pre-vaccination titer
`sv`	Strain included in vaccine
`sa`	Strain used for assay

We represent unobserved confounding in our DAG as the variable U. There are likely

many confounders, like individual variables driving vaccine choice, which we cannot account for because they were not collected as part of the study we used. We attempted to control for unobserved confounding as best as possible by using a random effects model structure which can absorb part of the effect of unobserved confounders by modeling between-individual variability (which we represent as i in the DAG). Not all unobserved confounding effects can be absorbed by individual random effects, but some, such as for demographic characteristics like sex and race, potentially can be.

Even though not all of the variables shown are confounders (p, sv, and sa are only causes of the outcome in our DAG), we include sv and sa so we can obtain stratum-specific effects for those variables. Controlling for p does not open any backdoor paths (under our assumed causal model), so since p is a cause of the outcome we can include p in the model to potentially improve the efficiency of our estimates.

Model Formula and Explanation

The full mathematical equations for our model are shown in the next section. Here, we briefly explain the model we used and its implementation in `brms`.

The model formula for our hierarchical models was chosen based on *a priori* covariate information from our causal model, along with constraints induced by the estimability of random effects. We elected not to include interaction terms or any other nuisance covariates due to the complexity of the model. Notably, our model is unlikely to converge under frequentist maximum likelihood estimation, such as by the `nlme` R package or similar methods. The random effects in the model are overdetermined, which leads to near-zero (boundary) estimates of random effect covariance terms, preventing maximum likelihood convergence. However, having similar random effects does not prevent the NUTS algorithm implemented by Stan/`brms` from exploring the implied posterior distribution.

We specified our models in `brms` using the following model formula:

```
outcome ~ dose +
    s(birth_year_c, k = 5) + s(age_c, k = 5) +
    s(log_pretiter, k = 5) + s(year_c, k = 5, by = study) +
    (1 | id) + (1 | study) +
    (1 + dose | strain_type) + (1 + dose | strain_type:strain_name) +
    (1 + dose | vaccine_name) + (1 + dose | vaccine_name:strain_type)
```

The `brms` model syntax is explained more fully in the documentation, but we will briefly explain the model formula here. The `outcome ~` specification declares that `outcome` is the dependent variable, and everything after the `~` is a predictor. The term `dose` specifies a fixed effect, which by default uses indicator (dummy) encoding.

All terms of the form `s(variable_name, k = 5)` specify smoothing splines. The spline basis is constructed by the `mgcv` package before being passed to Stan, and is parametrized similarly to frequentist models. We used penalized thin-plate regression splines with $k = 5$ basis functions. For integer-valued predictors, a low value of k provides enough flexibility while avoiding overparameterization. The option `by = study` fits separate smoothing splines of `year_c` for each level of the `study` variable, accommodating their differing time ranges.

Random effects terms appear in parentheses with vertical bars (`|`). A term like `(1 | g)` models a varying intercept for each level of grouping variable `g`, while `(1 + v | g)` allows both a random intercept and a random slope for `v`, correlated across levels of `g`. Our model includes random intercepts for individuals (`id`) and studies (`study`), allowing heterogeneous baselines.

We also include random intercepts for `strain_type` (either H1N1 or H3N2) and random slopes for `dose` nested within `strain_name` (specific virus strains). This allows H1N1 strains to share information independently of H3N2 strains. Similarly, the vaccine-related terms allow for variation by strain in the vaccine formulation, not just in the assay.

In short, our model captures correlation across repeated measurements, across strains and subtypes, and across individuals. The effect of `dose` is allowed to vary across these subgroups, following the same correlation structure.

The structure of our random effects is partially nested but not fully crossed, a natural

result of the influenza vaccine update schedule. We allow correlations among random effects where meaningful.

Finally, all variables with the `_c` suffix have been mean-centered to improve numerical estimation. All transformations (e.g., logs and centering) were performed before passing data to `brms` or Stan.

Model likelihood and priors

For the post-vaccination titer and titer increase outcomes, we used a Gaussian (Normal distribution) likelihood function for the model. Letting the outcome be y, we assumed that

$$y[i] \sim \text{Normal}\left(\mu[i], \sigma^2\right)$$
$$\sigma \sim t^{+}(3, 0, 3)$$
$$i = i, \ldots, N$$

where σ is the residual variance, and $\mu[i]$ is described by the 'brms' equation above, which builds a model for the conditional mean of $y[i]$ given the predictor data. Here, N is the number of data points passed to the model and the index variable i indexes the current data record.

The model for the mean and the priors we used for the Gaussian outcomes are shown on the following page.

$$
\begin{aligned}
\mu[i] = \alpha + \beta \cdot \text{HD}[i] + f[1]\left(\text{birth\_year\_c}[i]\right) + f[2]\left(\text{age\_c}[i]\right) + \\
f[3]\left(\log \text{pre\_titer}[i]\right) + f[4, \text{study}[i]]\left(\text{year\_c}[i]\right) + \\
u[1, \text{id}[i]] + u[2, \text{study}[i]] + \\
b[1, 0, \text{strain\_type}[i]] + b[1, 1, \text{strain\_type}[i]] \cdot \text{HD}[i] + \\
b[2, 0, \text{strain\_type}[i] \cdot \text{strain\_name}[i]] + \\
b[2, 1, \text{strain\_type}[i] \cdot \text{strain\_name}[i]] \cdot \text{HD}[i] + \\
b[3, 0, \text{vaccine\_name}[i]] + b[3, 1, \text{vaccine\_name}[i]] \cdot \text{HD}[i] + \\
b[4, 0, \text{strain\_type}[i] \cdot \text{vaccine\_name}[i]] + \\
b[4, 1, \text{strain\_type}[i] \cdot \text{vaccine\_name}[i]] \cdot \text{HD}[i]
\end{aligned}
$$

$$
f[j](\cdot) = \sum_{k=1}^{5} \gamma[j,k] \cdot \phi[j,k](\cdot); \quad j = 1, 2, 3
$$

$$
f[4, \text{study}[i]](\text{year\_c}[i]) = \sum_{k=1}^{5} \eta[k, \text{study}[i]] \cdot \phi[4,k](\text{year\_c}[i])
$$

$$
\alpha \sim \mathcal{N}(0, 5)
$$

$$
\beta \sim \mathcal{N}(0, 5)
$$

$$
\gamma[j,k] \sim \mathcal{N}(0, \tau[j]) \quad j = 1, 2, 3; \; k = 1, \ldots, 5
$$

$$
\tau[s] \sim t^{+}(3, 0, 3)
$$

$$
\eta[k, \text{study}[i]] \sim \mathcal{N}(0, \zeta[\text{study}[i]]) \quad k = 1, \ldots, 5
$$

$$
\zeta[\cdot] \sim t^{+}(3, 0, 3)
$$

$$
u[r, \cdot] \sim \mathcal{N}(0, \omega[r]) \quad r = 1, 2
$$

$$
\omega[r] \sim t^{+}(3, 0, 1)
$$

$$
\begin{pmatrix} b[q, 0, \cdot] \\ b[q, 1, \cdot] \end{pmatrix} \sim \text{MVN}\Big(\mathbf{0}, (\text{diag}(\psi[q])L[q])(\text{diag}(\psi[q])L[q])^{T}\Big) \quad q = 1, \ldots, 4
$$

$$
\psi[q] \sim t^{+}(3, 0, 1)
$$

$$
L[q] \sim \text{LKJ}(2)
$$

Note that instead of using subscript notation in our model, because of the large number of nested subscripts and the use of index-variable coding, we elected to use brackets to show indexing. For example, the notation $L[q]$ in our model is identical to the common notation L_q, but for index variables such as in the equation for $\mu[i]$, the bracket notation avoids small text which is often difficult to read. Additionally, we use the centered dot symbol $(\cdot)$ to indicate when there are many valid arguments that would all have the exact same right hand side (RHS) in a formula. For example, $\zeta(\cdot)$ indicates that all subscripts for ζ follow the same independent distributions, and writing out another index variable would only serve to make the formula more confusing.

In our model, the $\phi(\cdot)$ functions represent basis functions for thin plate splines. Internally, `brms` uses `mgcv` to create the spline basis design matrix, which uses a low-rank (in our case, rank $k = 5$) approximation of the eigenvalue matrix to create the thin plate spline, which allows for similar performance to the full-rank eigendecomposition while substantially saving computational time by elementing dimensions with small eigenvalues which contribute little to the fitting process. Since our data values are integer-valued (year, birth year, pretiter, and age are all integer values), the maximum k rank we can choose is equal to the number of unique values recorded in each variable, but we chose $k = 5$ in order to preserve the first five components after eigendecomposition because this balances performance with loss of data [201–206].

The distributions we refer to are: $\mathcal{N}(\cdot, \cdot)$ for the Normal distribution, parametrized in terms of the mean and variance; $t^+(\nu, \cdot, \cdot)$ for the location-scale **half** Student's t distribution with ν degrees of freedom, which can take on only positive values; $MVN(\cdot, \cdot)$ for the multivariate normal distribution parametrized in terms of a mean vector and a covariance matrix; and $LKJ(\cdot)$ for the Lewandowski-Kurowicka-Joe distribution for Cholesky factors [207].

To specify the covariance structures for our correlated varying effects, we used the Cholesky factor parametrization, which is more numerically stable and computationally efficient than other methods like individually parametrizing each term of the covariance matrix or specifying an inverse-Wishart prior [179, 208]. We represent the covariance matrix,

say Σ, using a Cholesky factor and scale parameter as follows:

$$\begin{aligned}\Sigma &= \text{diag}(\sigma)\Omega_{\Sigma}\text{diag}(\sigma) \\ &= \text{diag}(\sigma)(LL^T)\text{diag}(\sigma) \\ &= (\text{diag}(\sigma)L)\,(\text{diag}(\sigma)L)^T\,.\end{aligned}$$

Here, L is the Cholesky factor of Ω. We can then specify a strictly positive prior on σ, the scale parameter, and an LKJ prior on L. Specify, we use an LKJ(2) prior. An LKJ(1) indicates that all correlations are equally likely, while parameter values less than one indicate that strong correlations are more likely while parameter values greater than one indicate that strong correlations are less likely. Therefore, using a parameter value of 2 presumes slightly weaker correlations *a priori*, but large and weak correlations can still be learned from the data.

For the seroconversion and seroprotection outcomes, we used a Bernoulli likelihood with a logit link function. That is, we assumed that

$$\begin{aligned}y_i &\sim \text{Bernoulli}(p_i), \\ p_i &= \text{logit}^{-1}\left(\mu_i\right),\end{aligned}$$

where μ_i is again described by the right-hand side of the `brms` formula.

While the structure of the model for the mean is the same for the logistic regression models, we have to reduce the width of the priors because our outcome is fitted on the logit scale for the binary outcomes. Using these relatively narrow priors actually gives a more uniform prior distribution of effects, because of the nonlinear transformation of the linear predictor, which severely deflates low values and severely inflates high values. There is no residual variance to estimate in the binary outcome models.

The mean structure and priors used for the logistic regression models are shown on the following page.

$$
\begin{aligned}
\mu[i] = \alpha + \beta \cdot \text{HD}[i] + f[1]\left(\text{birth\_year\_c}[i]\right) + f[2]\left(\text{age\_c}[i]\right) + \\
f[3]\left(\log \text{pre\_titer}[i]\right) + f[4, \text{study}[i]]\left(\text{year\_c}[i]\right) + \\
u[1, \text{id}[i]] + u[2, \text{study}[i]] + \\
b[1, 0, \text{strain\_type}[i]] + b[1, 1, \text{strain\_type}[i]] \cdot \text{HD}[i] + \\
b[2, 0, \text{strain\_type}[i] \cdot \text{strain\_name}[i]] + \\
b[2, 1, \text{strain\_type}[i] \cdot \text{strain\_name}[i]] \cdot \text{HD}[i] + \\
b[3, 0, \text{vaccine\_name}[i]] + b[3, 1, \text{vaccine\_name}[i]] \cdot \text{HD}[i] + \\
b[4, 0, \text{strain\_type}[i] \cdot \text{vaccine\_name}[i]] + \\
b[4, 1, \text{strain\_type}[i] \cdot \text{vaccine\_name}[i]] \cdot \text{HD}[i]
\end{aligned}
$$

$$
f[j](\cdot) = \sum_{k=1}^{5} \gamma[j,k] \cdot \phi[j,k](\cdot); \quad j = 1, 2, 3
$$

$$
f[4, \text{study}[i]](\text{year\_c}[i]) = \sum_{k=1}^{5} \eta[k, \text{study}[i]] \cdot \phi[4,k](\text{year\_c}[i])
$$

$$
\alpha \sim \mathcal{N}(0, 1)
$$

$$
\beta \sim \mathcal{N}(0, 1)
$$

$$
\gamma[j,k] \sim \mathcal{N}(0, \tau[j]) \quad j = 1, 2, 3; \; k = 1, \ldots, 5
$$

$$
\tau[s] \sim t^{+}(3, 0, 1)
$$

$$
\eta[k, \text{study}[i]] \sim \mathcal{N}(0, \zeta[\text{study}[i]]) \quad k = 1, \ldots, 5
$$

$$
\zeta[\cdot] \sim t^{+}(3, 0, 1)
$$

$$
u[r, \cdot] \sim \mathcal{N}(0, \omega[r]) \quad r = 1, 2
$$

$$
\omega[r] \sim t^{+}(3, 0, 1)
$$

$$
\begin{pmatrix} b[q, 0, \cdot] \\ b[q, 1, \cdot] \end{pmatrix} \sim \text{MVN}\Big(\mathbf{0}, (\text{diag}(\psi[q])L[q])(\text{diag}(\psi[q])L[q])^{T}\Big) \quad q = 1, \ldots, 4
$$

$$
\psi[q] \sim t^{+}(3, 0, 1)
$$

$$
L[q] \sim \text{LKJ}(2)
$$

The details on notation and priors are the same as for the Gaussian model. The only changes are for the outcome distribution, the lack of a residual unexplained variance parameter to estimate, and the narrower variances specified for the prior distributions.

We chose our priors with three major criteria in mind.

1. The priors should reflect true constraints on parameters, but should not impose arbitrary constraints. So variance parameters should have strictly positive priors, correlations should be bounded between -1 and 1, etc. The parameter space of each prior is unbounded other than actual, scientific limitations of parameters.
2. The priors should be skeptical and regularizing. The priors for all of our effects have a mode of zero – we have effectively presupposed that there are no true effects, and any effects need to be learned from the data. This is similar to the idea of falsification in a frequentist framework – we need the data to convince us that an effect is present, rather than presupposing an effect and working to find it.
3. The priors should allow for large effects if the data support large effects.

Due to our third point, for variance parameters, we have chosen half Student's t priors. These are a compromise between half-Gaussian (or half-Normal) priors, which have most of their mass in the bulk of the distribution, and normal tails, and the Cauchy distribution, which has fat tails and more easily allows for large effects. The Cauchy distribution is poorly behaved and hard to sample from because of its pathological properties (e.g. it has no mean and infinite variance, which certainly does not reflect our prior beliefs about observable effects). The degrees of freedom parameter for the Student's t prior controls how "Cauchy-like" the distribution is, i.e., how fat the tails are. A half-t distribution with $\nu = 3$ degrees of freedom (the value we have chosen for all of our priors) provides the fattest tails that are easy to sample from without pathological problems, therefore allowing all of our parameters to become large if the data support large parameters. This principle also reflects why we chose an LKJ(2) prior for correlation matrices, as explained previously.

Model Fitting

We implemented our models in `brms` [87–89], an R package that interfaces with the `cmdstanr` package [90] and the `cmdstan` backend for the Stan probabilistic programming language [91, 179]. Stan is a programming language designed for efficient implementation of Hamiltonian Monte Carlo (HMC), a state-of-the-art sampling algorithm for Bayesian inference with continuous parameters [209]. HMC improves upon limitations of traditional MCMC methods such as random walk Metropolis-Hastings or Gibbs sampling.

We fit all models using 16 parallel chains, with 500 warmup iterations and 1250 sampling iterations per chain, resulting in $20,000$ post-warmup samples per model. We increased the `adapt_delta` parameter to 0.99 to reduce the occurrence of divergent transitions. A fixed seed was used for each HMC run to ensure reproducibility, but otherwise we used the default `cmdstan` control parameters.

cACE Calculation

Our primary post-estimation quantity of interest was the (conditional) Average Causal Effect, or (c)ACE. The cACE represents the average difference in model predictions under two counterfactual conditions. Let y_i be the observed outcome for individual i, and let t_i denote the treatment assignment for that individual, where $t_i \in \{\text{SD}, \text{HD}\}$. Using the posterior predictive distribution from our fitted model, we estimate the two counterfactual potential outcomes for each individual, denoted by $\hat{y}_i(\text{HD})$ and $\hat{y}_i(\text{SD})$. The individual causal effect (ICE) for individual i is then given by

$$\tau_i = \hat{y}_i(\text{HD}) - \hat{y}_i(\text{SD}).$$

Note that the ICE estimate *is not* a consistent estimator of the true treatment effect for each individual, but merely an intermediate step in computing a consistent estimator of the ACE. We estimate the Average Causal Effect (ACE) by aggregating these ICEs across the population. If the ACE is positive, this implies that the HD vaccine elicits a stronger immune response on average in our study sample.

In a Bayesian framework, we have k posterior samples of τ_i for each individual (in our case, $k = 20,000$). These samples can be pooled to estimate the posterior distribution of the ACE, or aggregated within subgroups (strata) to estimate conditional ACEs (cACEs), where the conditioning is on characteristics such as strain type or vaccine composition. For example, we can compute the cACE for assays conducted on samples from individuals who received vaccines containing CA/09-like virus components.

We summarize each (c)ACE using the posterior mean and the 95% Highest Density Continuous Interval (HDCI). These summaries were computed using the `ggdist` R package [75, 76], and we computed HDCIs using a CDF-bounded kernel density estimator with a Gaussian kernel via the reflection method [210, 211]. We used the Sheather-Jones Direct Plug-In method to choose the bandwidth [212], and evaluated densities over a grid of 4,096 points. The resulting density is trimmed to the range of the posterior samples.

Effect size transformation from cACE calculation

The cACE (as described in a previous section) is calculated by taking a difference of predicted model outcomes. Since the model predictions are in units of log titer measurements (regardless of whether the outcome is post-vaccination titer or titer increase), the cACE is expressed in log titer units. In order to better communicate the effect size, we transformed the cACE. As before, let $\hat{y}_i(\text{SD})$ be the predicted treatment effect for individual i if that individual had received an SD vaccine, and let $\hat{y}_i(\text{HD})$ be the predicted treatment effect for individual i if that individual had received an HD vaccine. The estimated ICE is

$$\hat{\tau}_i = \hat{y}_i(\text{HD}) - \hat{y}_i(\text{SD}),$$

which represents the estimated benefit that individual i would receive from the HD vaccine. Our model generates a posterior distribution of the estimated ICE for each individual

The (c)ACE over the study sample is estimated as

$$\widehat{ACE}_i = E\left(\hat{\tau}_i\right) = E\left(\hat{y}_i(\text{HD}) - \hat{y}_i(\text{SD})\right),$$

and we compute this by pooling together the posterior distributions of the ICEs and summarizing them. However, each ICE is in log2-titer units, and so the ACE is also in log2-titer units. To facilitate interpretation, we can exponentiate (a monotone strictly increasing transformation) the estimated ACE to obtain an estimate in more interpretable titer units.

The transformed ACE, which we (arbitrarily) denote as $\widehat{\varphi}_i$ is then

$$\widehat{\varphi}_i = 2^{\tau_i} = 2^{E(\hat{y}_i(\text{HD}) - \hat{y}_i(\text{SD}))},$$

and is in HAI titer units. This number represents the average treatment effect as a ratio of fold changes, for the results presented in the main text. It can be interpreted analogously for the other three outcomes we used as sensitivity analyses in a later section of this Supplement.

Analysis of confounders on treatment choice

In the study our data were collected in, participants over age 65 were offered the choice between SD and HD vaccine, so the treatment (vaccine dose) was not assigned randomly as it would be in a clinical trial. There are many factors that could affect which treatment group an individual was likely to choose, including prior knowledge about the HD vaccine, batch effects in receiving shipments of HD vaccines at the clinical study site, demographic characteristics, risk-taking behavior, prior adverse events from vaccination, and others. While we cannot assess many of these characteristics, we did assess the demographic characteristics that were available to us to inform our selection of confounders in the model.

Table A.2 shows the results of our analysis. We used Bayesian logistic regression to calculate the odds ratio for receiving the HD vaccine (relative to receiving the SD vaccine). Each model used the dose as the outcome (with HD coded as an event, that is, SD was coded as 0 and HD was coded as 1) and we included a single categorical predictor in each model. The ORs are calculated by exponentiating the mean beta coefficient from 2000 post-warmup Hamiltonian Monte Carlo samples, implemented using the `brms` package. We used the same priors and setup as our main logistic regression model, which are explained in detail in the

model fitting section. The credible interval shown is the highest density continuous interval over the post-warmup samples.

Our results (Table A.2) show that several demographic and study-specific variables affected the odds of receiving an HD vaccine. More HD vaccines were available at the UGA study site than the PA and FL study sites, and in general HD vaccine uptake increased as the study continued, although the effect was not strong (the credible interval was wide and had a large amount of posterior density on either side of 1). Patients who were assigned female at birth (AFAB) and assigned male at birth (AMAB) had nearly equal odds of receiving HD vaccines, as did White patients and patients of color. Finally, older patients were somewhat more likely to receive HD vaccination, although estimating the OR was unstable for the oldest groups since there were very few patients over age 80. Since our study was conducted over a small temporal scale, the results for birth year are very similar to the results for age at enrollment. These odds ratios matched our causal hypotheses about age, birth year, study site, and calendar time (represented by influenza season) and help justify our inclusion of these terms in the model as confounders. Since sex assigned at birth and race/ethnicity did not appear to be associated with which dose was received, we elected not to include these variables in the model.

Historical and vaccine strain information

Throuhgout our manuscript and supplement, we used the abbreviated names of each strain throughout the paper in order to simplify tables and graphics. The complete strain names along with the abbreviated names are shown for H1N1 strains in Table A.3 and for H3N2 strains in Table A.4.

The strains included in influenza vaccines are reviewed annually based on estimates of the ability of a lab strain to elicit an immune response and antigenic similarity to strains which are predicted to circulate. If major changes in circulating strains are expected, the vaccine composition is updated to hopefully induce better immunity to the expected strains. The vaccine composition for each year is shown in Table A.5.

Table A.2: Odds ratio (OR) for receiving HD vaccine. I.e., each OR is the odds of receiving the HD vaccine in the relevant group compared to the reference group for each variable. Each model was a univariable (simple) logistic regression model where the dose was the outcome (a patient who received SD was coded as 0 and a patient who received HD was coded as 1) and one single categorical covariate was included in the model. OR was calculated by exponentiating the relevant slope coefficient from the logistic regression model, and thus an OR higher than one indicates a group with higher odds of receiving the HD vaccine than the reference group for a given variable. A dash indicates the reference group for a given variable, which means an OR could not be calculated. CI = Confidence Interval.

Variable	**OR (95% CI)**
Study	
UGA	—
PA	0.64 (0.44 to 0.94)
FL	0.16 (0.10 to 0.25)
Season	
2013 - 2014	—
2014 - 2015	1.88 (0.88 to 3.80)
2015 - 2016	3.71 (1.90 to 7.77)
2016 - 2017	5.84 (2.88 to 12.2)
2017 - 2018	1.85 (0.72 to 4.32)
2018 - 2019	1.82 (0.61 to 5.97)
2019 - 2020	8.48 (4.00 to 18.7)
2020 - 2021	29.6 (11.2 to 80.5)
2021 - 2022	34.1 (12.2 to 97.4)
Sex assigned at birth	
Female	—
Male	0.94 (0.68 to 1.32)
Age at enrollment	
65 - 70	—
71 - 75	1.15 (0.77 to 1.70)
76 - 80	1.56 (0.88 to 2.87)
81 - 85	0.57 (0.14 to 1.99)
Birth year	
1930 - 1935	—
1936 - 1940	1.99 (0.77 to 5.64)
1941 - 1945	2.22 (0.80 to 5.80)
1946 - 1950	2.30 (0.92 to 5.90)
1951 - 1955	3.24 (1.23 to 8.56)
1956 - 1960	7.29 (0.54 to 116)
Race/Ethnicity	
White or Caucasian	—
Black or Afircan American	0.87 (0.55 to 1.39)
Other	0.97 (0.42 to 2.07)

Table A.3: Abbreviatied strain names used in figures and tables, along with complete strain names. This table shows all H1N1 strains in our study.

Strain name	**Short name**
A/H1N1/South Carolina/1/1918	SC/18
A/H1N1/Puerto Rico/8/1934	PR/34
A/H1N1/Weiss/43	Wei/43
A/H1N1/Fort Monmouth/1/1947	FM/47
A/H1N1/Denver/1957	Den/57
A/H1N1/New Jersey/8/1976	NJ/76
A/H1N1/Ussr/90/1977	USSR/77
A/H1N1/Brazil/11/1978	Bra/78
A/H1N1/California/10/1978	CA/78
A/H1N1/Chile/1/1983	Chi/83
A/H1N1/Singapore/6/1986	Sing/86
A/H1N1/Texas/36/1991	TX/91
A/H1N1/Beijing/262/1995	Bei/95
A/H1N1/New Caledonia/20/1999	NC/99
A/H1N1/Solomon Islands/3/2006	SI/06
A/H1N1/Brisbane/59/2007	Bris/07
A/H1N1/California/07/2009	CA/09
A/H1N1/Michigan 45/2015	MI/15
A/H1N1/Brisbane/02/2018	Bris/18
A/H1N1/Guangdong-Maonan/SWL1536/201	GD/19
A/H1N1/Victoria/2570/2019	Vic/19

The panel of historical assays used in each year is shown in Table A.6 for the H1N1 subtype strains, and in Table A.7 for the H3N2 subtype strains. Note that the strains for the historical panel are chosen by our clinical research collaborators. We did not control which strains were used in each year of the study, and we used all strains available to us for analysis without omitting any. For the historical panels, strains were chosen to represent a broad spectrum of historical clades, along with representing all contemporary lineages used for vaccine development. For example, H1N1 strains cover the 1918-like clade, the 1976 Fort Dix-like strain, the pre-2009 lineage, and the 2009 pandemic-like viruses. H3N2 strains were chosen roughly as new clades emerged through time.

Additionally, the influenza viruses used for HAI assays were propagated in eggs, and may have egg-derived mutations. Some H3N2 strains perform poorly [56, 213] in HAI essays due to these mutations, but we used all of the heterologous assay available.

Table A.4: Abbreviatied strain names used in figures and tables, along with complete strain names. This table shows all H3N2 strains in our study.

Strain name	**Short name**
A/H3N2/Hong Kong/8/1968	HK/68
A/H3N2/Port Chalmers/1/1973	PC/73
A/H3N2/Texas/1/1977	TX/77
A/H3N2/Mississippi/1/1985	MI/85
A/H3N2/Sichuan/2/1987	Sich/87
A/H3N2/Shangdong/9/1993	Shan/93
A/H3N2/Nanchang/933/1995	Nan/95
A/H3N2/Sydney/5/1997	Syd/97
A/H3N2/Panama/2007/1999	Pan/99
A/H3N2/Fujian/411/2002	Fuj/02
A/H3N2/New York/55/2004	NY/04
A/H3N2/Brisbane/10/2007	Br/07
A/H3N2/Wisconsin/67/2005	WI/05
A/H3N2/Uruguay/716/2007	Uru/07
A/H3N2/Perth/16/2009	Per/09
A/H3N2/Victoria/361/2011	Vic/11
A/H3N2/Texas/50/2012	TX/12
A/H3N2/Switzerland/9715293/2013	Switz/13
A/H3N2/Hong Kong/4801/2014	HK/14
A/H3N2/Singapore/infimh-16-0019/2016	Sing/16
A/H3N2/Kansas/14/2017	KS/17
A/H3N2/Hong Kong/2671/2019	HK/19
A/H3N2/South Australia/34/2019	SA/19
A/H3N2/Tasmania/503/2020	Tas/20
A/H3N2/Darwin/9/2021	Dar/21

Table A.5: Composition of the Fluzone vaccine during each influenza season. The strains used were matched to ACIP/CDC recommendations, and were the same for both the SD and HD vaccine formulations since we only considered Influenza A in this study.

Season	**H1N1**	**H3N2**
2013 - 2014	CA/09	TX/12
2014 - 2015	CA/09	TX/12
2015 - 2016	CA/09	Switz/13
2016 - 2017	CA/09	HK/14
2017 - 2018	MI/15	HK/14
2018 - 2019	MI/15	Sing/16
2019 - 2020	Bris/18	KS/17
2020 - 2021	GD/19	HK/19
2021 - 2022	Vic/19	Tas/20

Table A.6: Number of assays performed using each component of the historical panel for a given season. Over the different seasons, strains were added and removed from the historical panel, indicated by the zeros in the table. The H1N1 strains are shown in this table.

	Season									
Strain	13/14	14/15	15/16	16/17	17/18	18/19	19/20	20/21	21/22	Total
SC/18	42	92	98	111	38	0	0	0	0	381
PR/34	42	0	0	0	0	0	0	0	0	42
Wei/43	42	92	98	111	38	0	0	0	0	381
FM/47	42	92	98	111	38	0	0	0	0	381
Den/57	42	92	98	111	38	0	0	0	0	381
NJ/76	42	92	98	111	38	0	0	0	0	381
USSR/77	42	92	98	111	38	0	0	0	0	381
Bra/78	42	0	0	111	38	0	0	0	0	191
Chi/83	42	92	98	111	38	19	90	0	0	490
Sing/86	56	92	98	111	38	19	0	0	0	414
TX/91	56	92	98	111	38	19	0	0	0	414
Bei/95	56	92	98	111	38	19	0	0	0	414
NC/99	55	92	98	111	38	19	0	0	0	413
SI/06	56	92	98	111	38	19	0	0	0	414
Bris/07	56	92	98	111	38	19	90	0	0	504
CA/09	56	92	98	111	38	19	90	83	81	668
CA/78	0	92	98	0	0	0	0	0	0	190
MI/15	0	0	0	111	38	19	90	83	0	341
Bris/18	0	0	0	0	0	0	90	83	81	254
GD/19	0	0	0	0	0	0	90	83	81	254
Vic/19	0	0	0	0	0	0	0	0	81	81

Table A.7: Number of assays performed using each component of the historical panel for a given season. Over the different seasons, strains were added and removed from the historical panel, indicated by the zeros in the table. The H3N2 strains are shown in this table.

	Season									
Strain	13/14	14/15	15/16	16/17	17/18	18/19	19/20	20/21	21/22	Total
HK/68	56	92	98	111	38	0	0	0	0	395
PC/73	56	92	98	111	38	0	0	0	0	395
TX/77	56	92	98	111	38	0	0	0	0	395
MI/85	56	92	97	111	38	0	0	0	0	394
Sich/87	42	92	97	111	38	0	0	0	0	380
Shan/93	42	92	98	111	38	0	0	0	0	381
Nan/95	42	92	98	111	38	0	0	0	0	381
Syd/97	56	92	98	111	38	0	0	0	0	395
Pan/99	56	92	98	111	38	19	90	0	0	504
Fuj/02	42	92	98	0	0	0	0	0	0	232
NY/04	56	92	98	111	38	19	0	0	0	414
Br/07	42	0	0	0	0	0	0	0	0	42
WI/05	56	92	98	111	38	19	0	0	0	414
Per/09	56	92	98	111	38	19	0	0	0	414
Vic/11	56	92	98	111	38	19	0	0	0	414
TX/12	56	92	98	111	38	19	90	0	0	504
Switz/13	36	91	98	111	38	19	90	0	0	483
Uru/07	0	92	98	111	38	19	0	0	0	358
HK/14	0	91	98	111	38	19	90	83	81	611
Sing/16	0	0	0	0	38	19	90	83	81	311
KS/17	0	0	0	0	0	0	90	83	81	254
HK/19	0	0	0	0	0	0	90	83	81	254
SA/19	0	0	0	0	0	0	90	0	81	171
Tas/20	0	0	0	0	0	0	0	0	81	81
Dar/21	0	0	0	0	0	0	0	0	81	81

Descriptive analyses

We conducted a descriptive analysis of the outcomes, stratified by vaccine dose. In order to determine how different the effect of dose was across the different vaccines and assay strains, we further conducted stratified analyses.

Table A.8 shows the number of unique individuals who were recruited at each study site, along with summaries of their age at first enrollment in the study and birth year. The ages at enrollment and birth years were very similar across the three study sites.

Table A.8: Number of unique individuals who were recruited at each study site, along with summaries of the age at first enrollment and birth year of participants at each study site and overall. Age is age at first enrollment and numbers shown are median (range).

Characteristic	FL, N = 52	PA, N = 83	UGA, N = 119	Overall, N = 254
Age	68 (65 - 80)	68 (65 - 82)	68 (65 - 85)	68 (65 - 85)
Birth year	1946 (1933 - 1951)	1945 (1932 - 1951)	1950 (1934 - 1956)	1948 (1932 - 1956)

The study was collected at two different sites (PA and FL) from 2013/14 through 2016/17, but moved to the UGA site in January 2017. The demographic information stratified by study site is shown in Table A.9. The FL study site gave fewer HD vaccinations, but there were no noticeable differences in the age or birth cohort of individuals from the three study sites.

In addition, several inviduals returned to the study site in multiple years. Table A.10 shows the number of recurring individuals at each study site, with how many times a participant returned to the study. The PA and FL study sites ran for four years (2013/14 influenza season through 2016/17 influenza season) while the UGA study site had six years of data we could use for our secondary analysis (2016/17 influenza season through 2021/22 influenza season) and is ongoing.

Since individuals were not required to choose the same vaccine (SD or HD) they received at their previous visit, some individuals received different doses at subsequent visits, which is accounted for in our statistical analyses. From Table A.10, we can see that the majority of individuals did not switch vaccine doses at subsequent visits, and out of all of the times

Table A.9: Demographics of the study sample stratified by the three study sites and all influenza seasons, along wiht the number of people who received each vaccine dose. The only season when all three study sites recruited individuals was 2016/17.

Characteristic	FL, N = 123	PA, N = 219	UGA, N = 326	Overall, N = 668
Season, n (%)				
2013 - 2014	20 (16)	36 (16)	0 (0)	56 (8)
2014 - 2015	35 (28)	57 (26)	0 (0)	92 (14)
2015 - 2016	35 (28)	63 (29)	0 (0)	98 (15)
2016 - 2017	33 (27)	63 (29)	15 (5)	111 (17)
2017 - 2018	0 (0)	0 (0)	38 (12)	38 (6)
2018 - 2019	0 (0)	0 (0)	19 (6)	19 (3)
2019 - 2020	0 (0)	0 (0)	90 (28)	90 (13)
2020 - 2021	0 (0)	0 (0)	83 (25)	83 (12)
2021 - 2022	0 (0)	0 (0)	81 (25)	81 (12)
Dose, n (%)				
SD	82 (67)	73 (33)	79 (24)	234 (35)
HD	41 (33)	146 (67)	247 (76)	434 (65)

an individual switched vaccine doses, almost all of them were switching from HD to SD vaccines. While a history of adverse events or side effects could account for switching from HD to SD vaccines, the likely explanation is the availability of HD vaccination at the study site.

While we do not have access to detailed shipping and receiving logs along with individual dates of vaccination to confirm this, the study site ordered fewer HD vaccines than SD vaccines and it is likely than some individuals who would have preferred an HD vaccine received an SD vaccine since no HD vaccines were available. Most individuals are likely to prefer receiving an SD vaccine while at the study site rather than returning to receive an HD vaccine at a later date. To our knowledge, there are no studies, either empirical or computational, showing whether receiving an SD vaccination instead of HD vaccination is preferable (in terms of intraseason waning immunity and overall protection from influenza disease) to waiting to receive an HD vaccine.

Table A.10: Summary statistics of how many individuals participated in the study in multiple years at each study site. Individuals were not required to choose the same vaccine (SD or HD) at subsequent visits, so we also show summary statistics for the number of times individuals switched which vaccine they received at subsequent visits.

Count	UGA, N = 119	FL, N = 52	PA, N = 83	Overall, N = 254
Number of total visits by same individual, n (%)				
1	27 (23)	19 (37)	16 (19)	62 (24)
2	23 (19)	7 (13)	20 (24)	50 (20)
3	46 (39)	14 (27)	25 (30)	85 (33)
4	8 (6.7)	12 (23)	22 (27)	42 (17)
5	7 (5.9)	0 (0)	0 (0)	7 (2.8)
6	8 (6.7)	0 (0)	0 (0)	8 (3.1)
Number of SD vaccinations for same individual, n (%)				
0	74 (62)	1 (1.9)	36 (43)	111 (44)
1	27 (23)	32 (62)	27 (33)	86 (34)
2	9 (7.6)	12 (23)	15 (18)	36 (14)
3	5 (4.2)	2 (3.8)	4 (4.8)	11 (4.3)
4	2 (1.7)	5 (9.6)	1 (1.2)	8 (3.1)
5	1 (0.8)	0 (0)	0 (0)	1 (0.4)
6	1 (0.8)	0 (0)	0 (0)	1 (0.4)
Number of HD vaccinations for same individual, n (%)				
0	19 (16)	30 (58)	11 (13)	60 (24)
1	24 (20)	3 (5.8)	33 (40)	60 (24)
2	23 (19)	19 (37)	13 (16)	55 (22)
3	42 (35)	0 (0)	17 (20)	59 (23)
4	4 (3.4)	0 (0)	9 (11)	13 (5.1)
5	7 (5.9)	0 (0)	0 (0)	7 (2.8)
Number of times an individual switched from HD to SD, n (%)				
0	93 (78)	31 (60)	49 (59)	173 (68)
1	25 (21)	21 (40)	34 (41)	80 (31)
2	1 (0.8)	0 (0)	0 (0)	1 (0.4)
Number of times an individual switched from SD to HD, n (%)				
0	116 (97)	52 (100)	79 (95)	247 (97)
1	3 (2.5)	0 (0)	4 (4.8)	7 (2.8)

Pre and post-vaccination titer figures

In order to better conceptualize the data we use for our models, we have included plots of the pre and post-vaccination titers. Due to technical limitations with how LaTeX processes the layout for large images, the raw data for each vaccine strain is included in a separate figure.

In each of the following figures, every person-year of data is represented as a line segment and a pair of points, one for the recorded pre-vaccination titer and one for the recorded post-vaccination titer. The line is a visual aid to help gauge the relative titer increase for that person-year, and to more easily understand for which strains the relevant vaccine candidate induced an immunogenic response at the cohort level. Each subplot in each figure corresponds to one assay strain, showing all person-years of data where the noted vaccine was administered and a response was measured to the noted assay strain.

The following figure numbers correspond to the listed vaccine strains. Please reference Table A.6 and Table A.7 for the complete strain names.

- Figure A.2: Ca/09 (H1N1) vaccine strain;
- Figure A.3: MI/15 (H1N1) vaccine strain;
- Figure A.4: Bris/18 (H1N1) vaccine strain;
- Figure A.5: GD/19 (H1N1) vaccine strain;
- Figure A.6: Vic/19 (H1N1) vaccine strain;
- Figure A.7: TX/12 (H3N2) vaccine strain;
- Figure A.8: Switz/13 (H3N2) vaccine strain;
- Figure A.9: HK/14 (H3N2) vaccine strain;
- Figure A.10: Sing/16 (H3N2) vaccine strain;
- Figure A.11: KS/17 (H3N2) vaccine strain;
- Figure A.12: HK/19 (H3N2) vaccine strain; and
- Figure A.13: Tas/20 (H3N2) vaccine strain.

Figure A.2: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing CA/09-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against CA/09 vaccinated individual samples are shown.

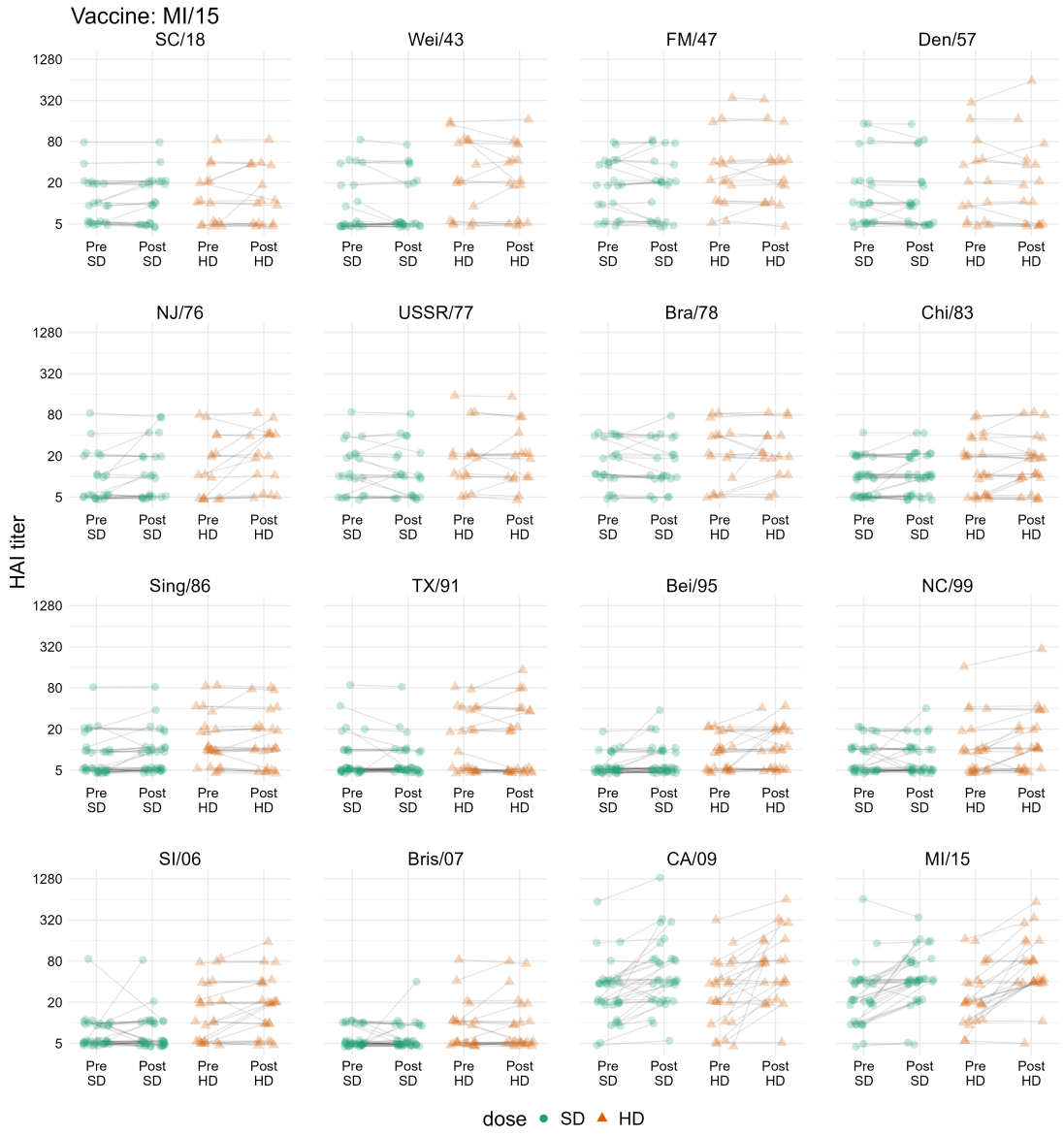

Figure A.3: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing MI/15-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against MI/15 vaccinated individual samples are shown.

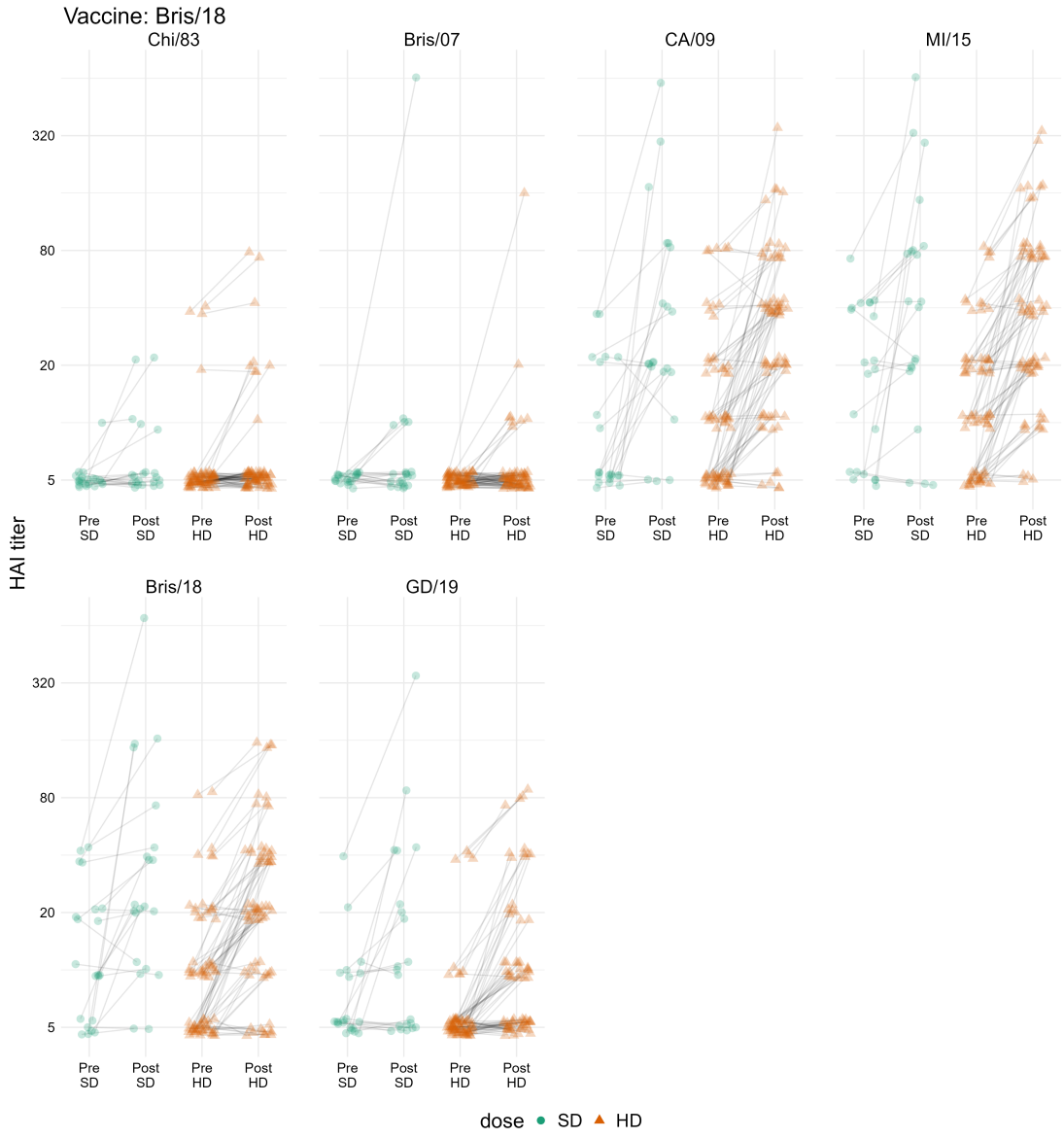

Figure A.4: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing Bris/18-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Bris/18 vaccinated individual samples are shown.

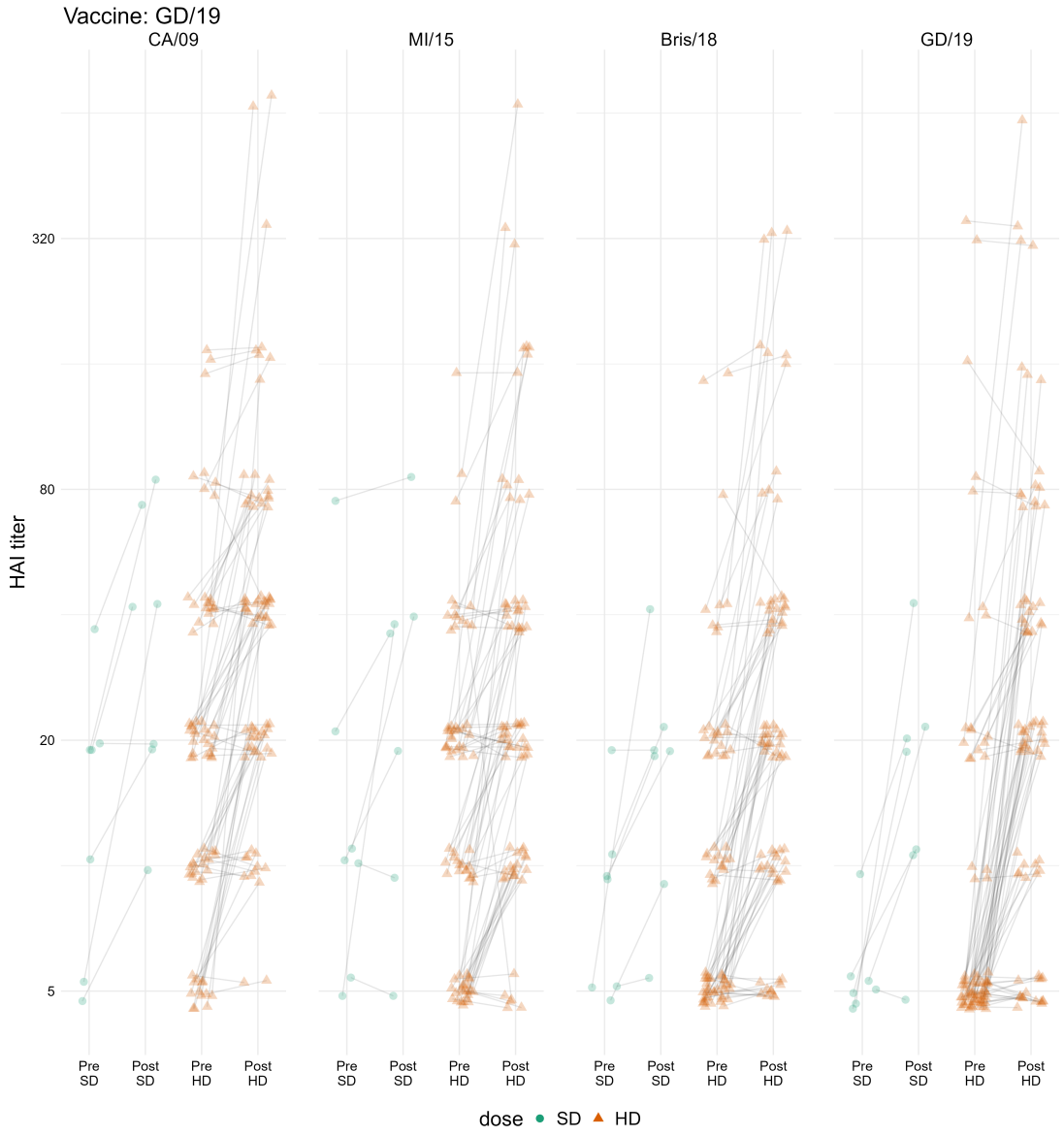

Figure A.5: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing GD/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against GD/19 vaccinated individual samples are shown.

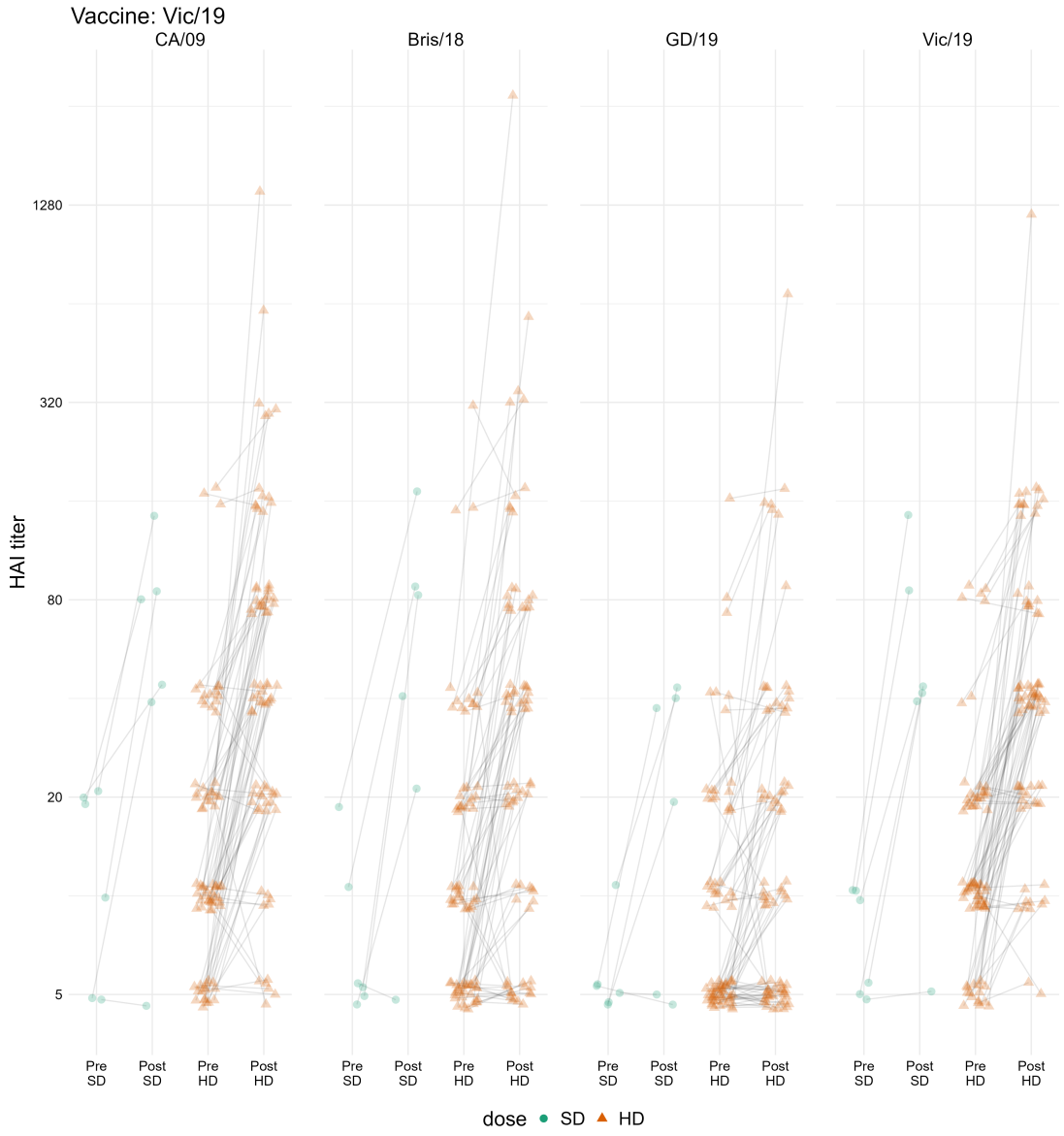

Figure A.6: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing Vic/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Vic/19 vaccinated individual samples are shown.

Figure A.7: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing TX/12-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against TX/12 vaccinated individual samples are shown.

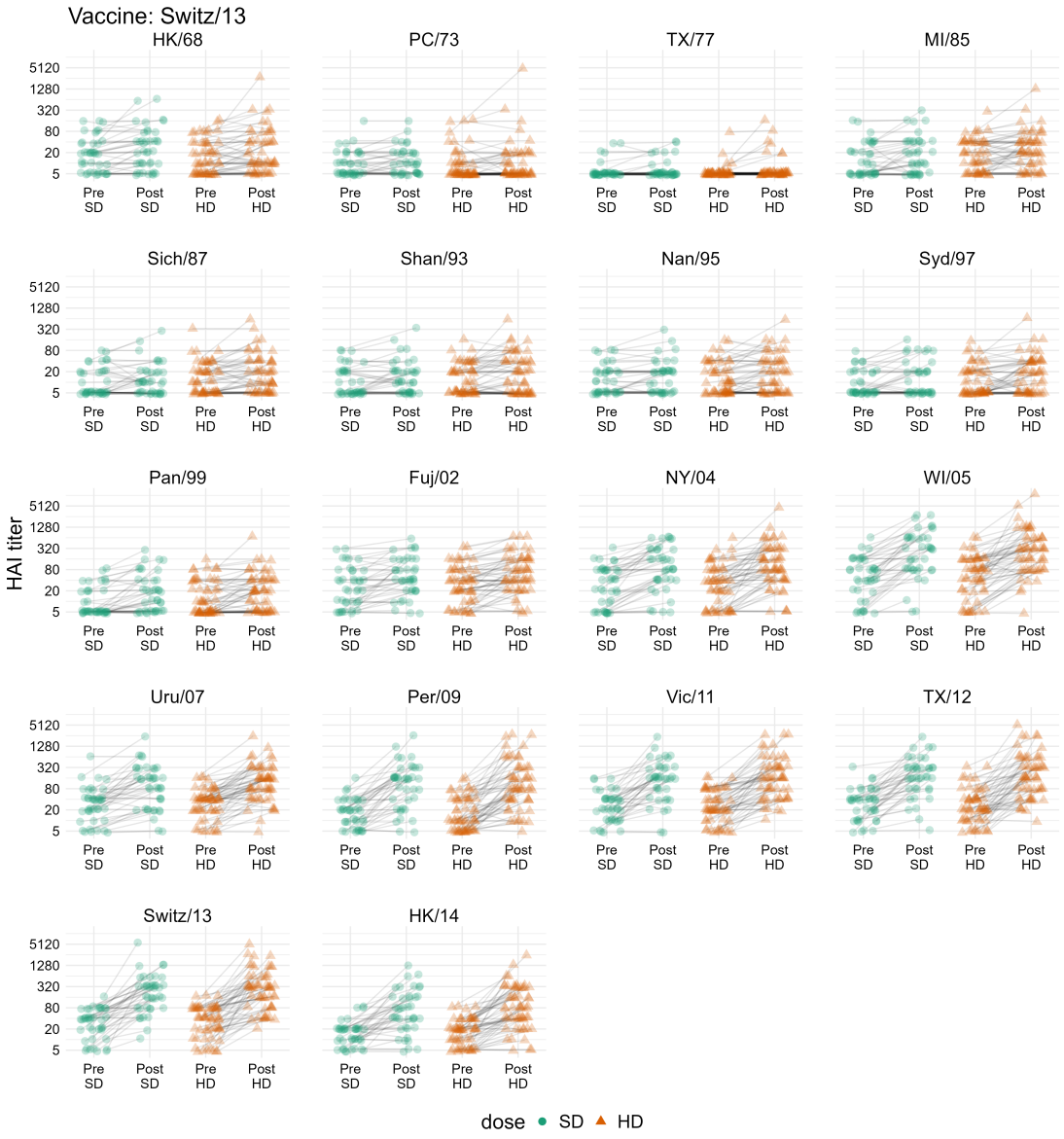

Figure A.8: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing Switz/13-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Switz/13 vaccinated individual samples are shown.

Figure A.9: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing HK/14-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against HK/14 vaccinated individual samples are shown.

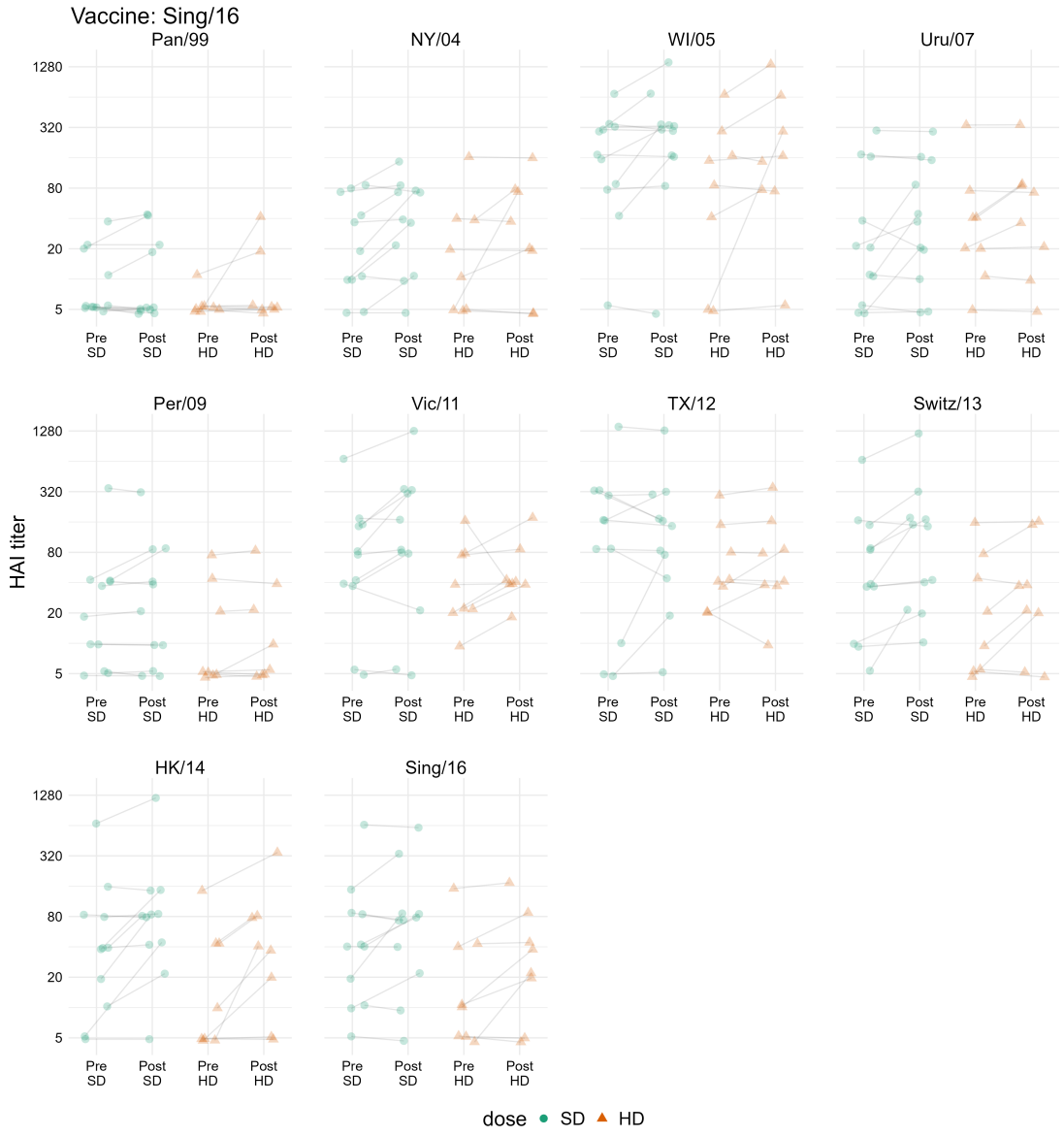

Figure A.10: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing Sing/16-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Sing/16 vaccinated individual samples are shown.

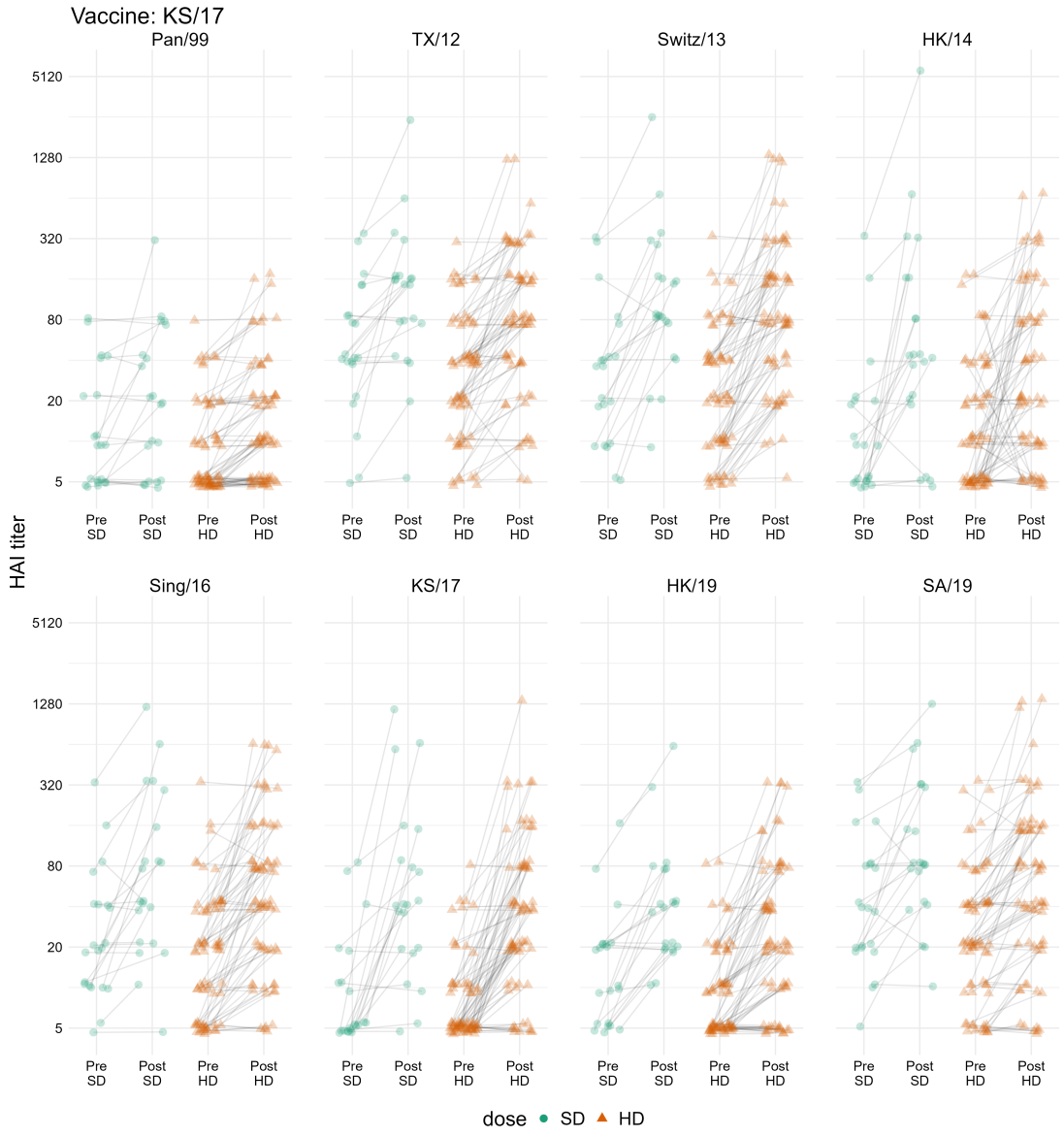

Figure A.11: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing KS/17-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against KS/17 vaccinated individual samples are shown.

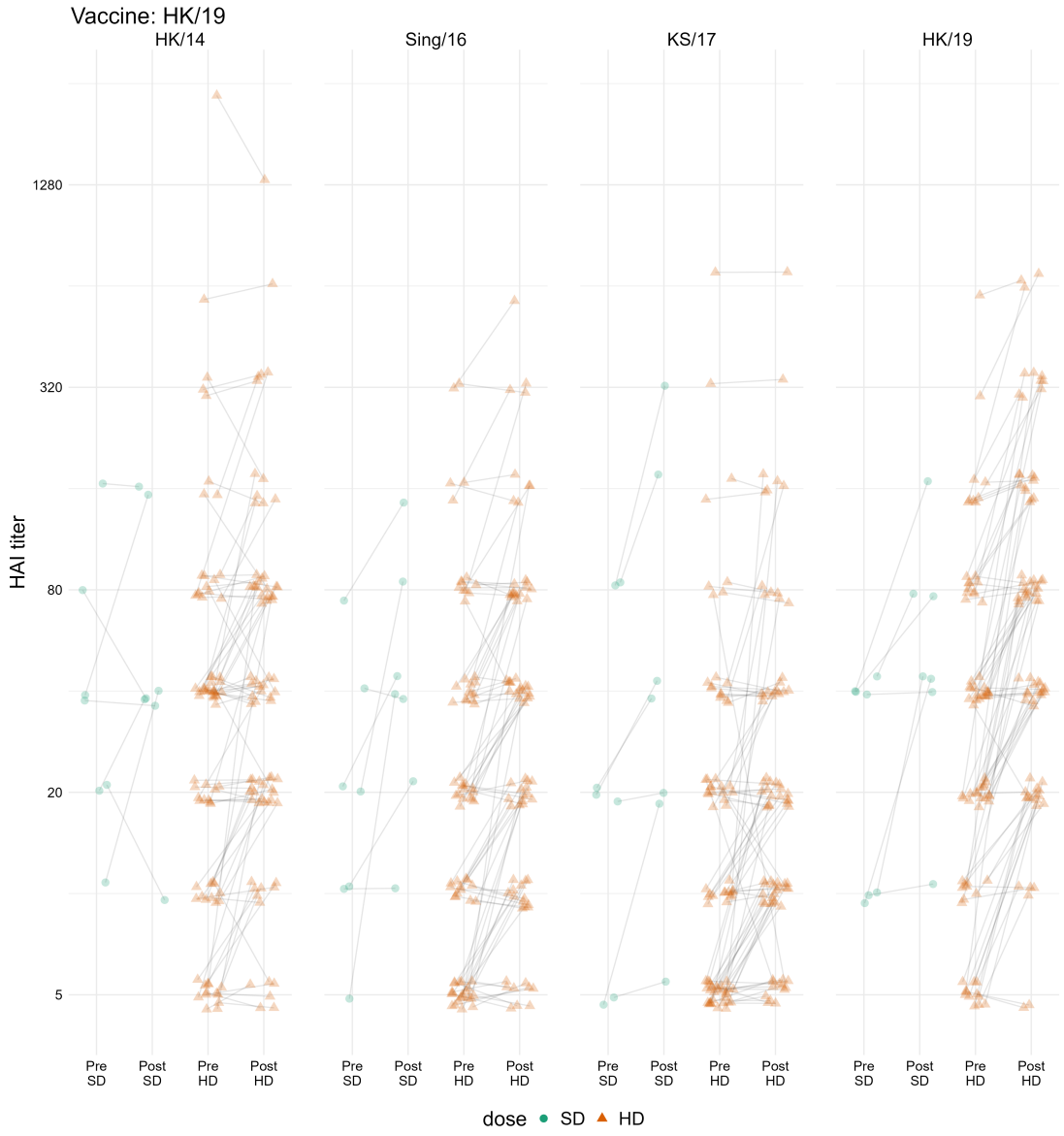

Figure A.12: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing HK/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against HK/19 vaccinated individual samples are shown.

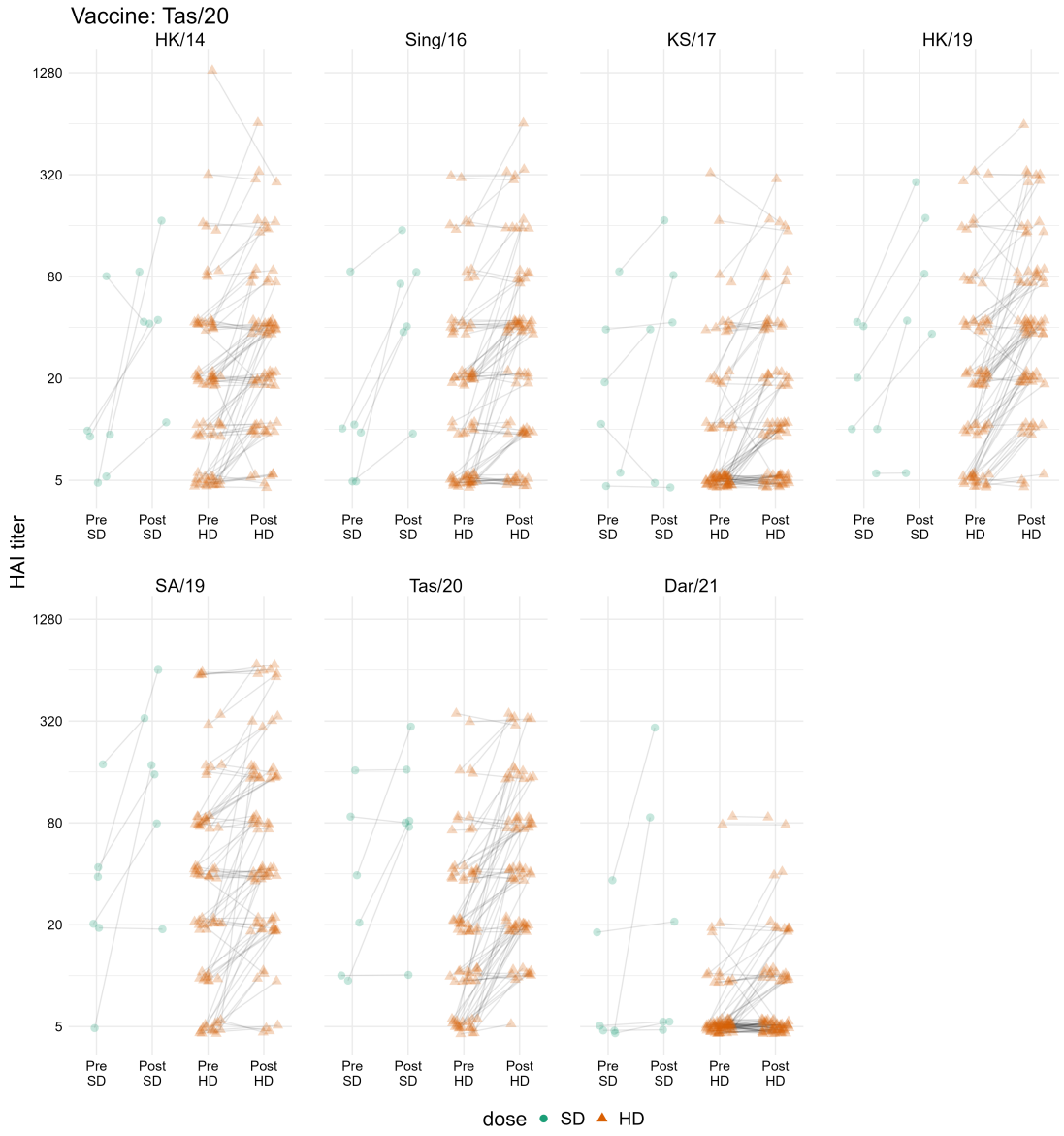

Figure A.13: Pre and post vaccination titers for all person-years where a participant was administered a vaccine containing Tas/20-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Tas/20 vaccinated individual samples are shown.

Fold change figures

In addition to the previous figures, we also show the distribution of fold changes to each assay strain that were induced by each vaccine component.

Similar to the previous section, in each of the following figures, every person-year of data is represented as one point, representing the fold change (i.e., post-vaccination titer divided by pre-vaccination titer) for that person in a particular season.

The following figure numbers correspond to the listed vaccine strains. Please reference Table A.6 and Table A.7 for the complete strain names.

- Figure A.14: Ca/09 (H1N1) vaccine strain;
- Figure A.15: MI/15 (H1N1) vaccine strain;
- Figure A.16: Bris/18 (H1N1) vaccine strain;
- Figure A.17: GD/19 (H1N1) vaccine strain;
- Figure A.18: Vic/19 (H1N1) vaccine strain;
- Figure A.19: TX/12 (H3N2) vaccine strain;
- Figure A.20: Switz/13 (H3N2) vaccine strain;
- Figure A.21: HK/14 (H3N2) vaccine strain;
- Figure A.22: Sing/16 (H3N2) vaccine strain;
- Figure A.23: KS/17 (H3N2) vaccine strain;
- Figure A.24: HK/19 (H3N2) vaccine strain; and
- Figure A.25: Tas/20 (H3N2) vaccine strain.

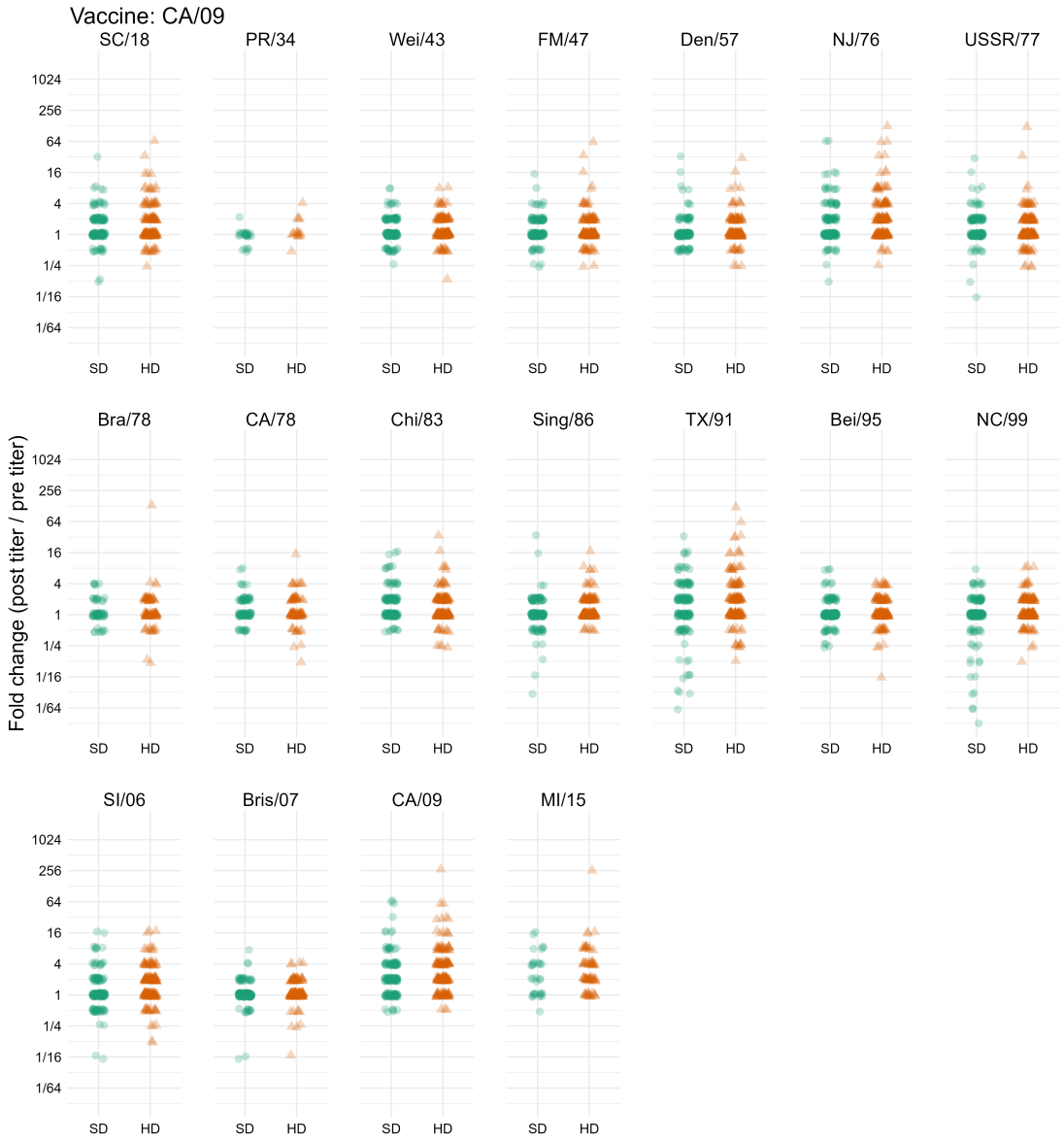

Figure A.14: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing CA/09-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against CA/09 vaccinated individual samples are shown.

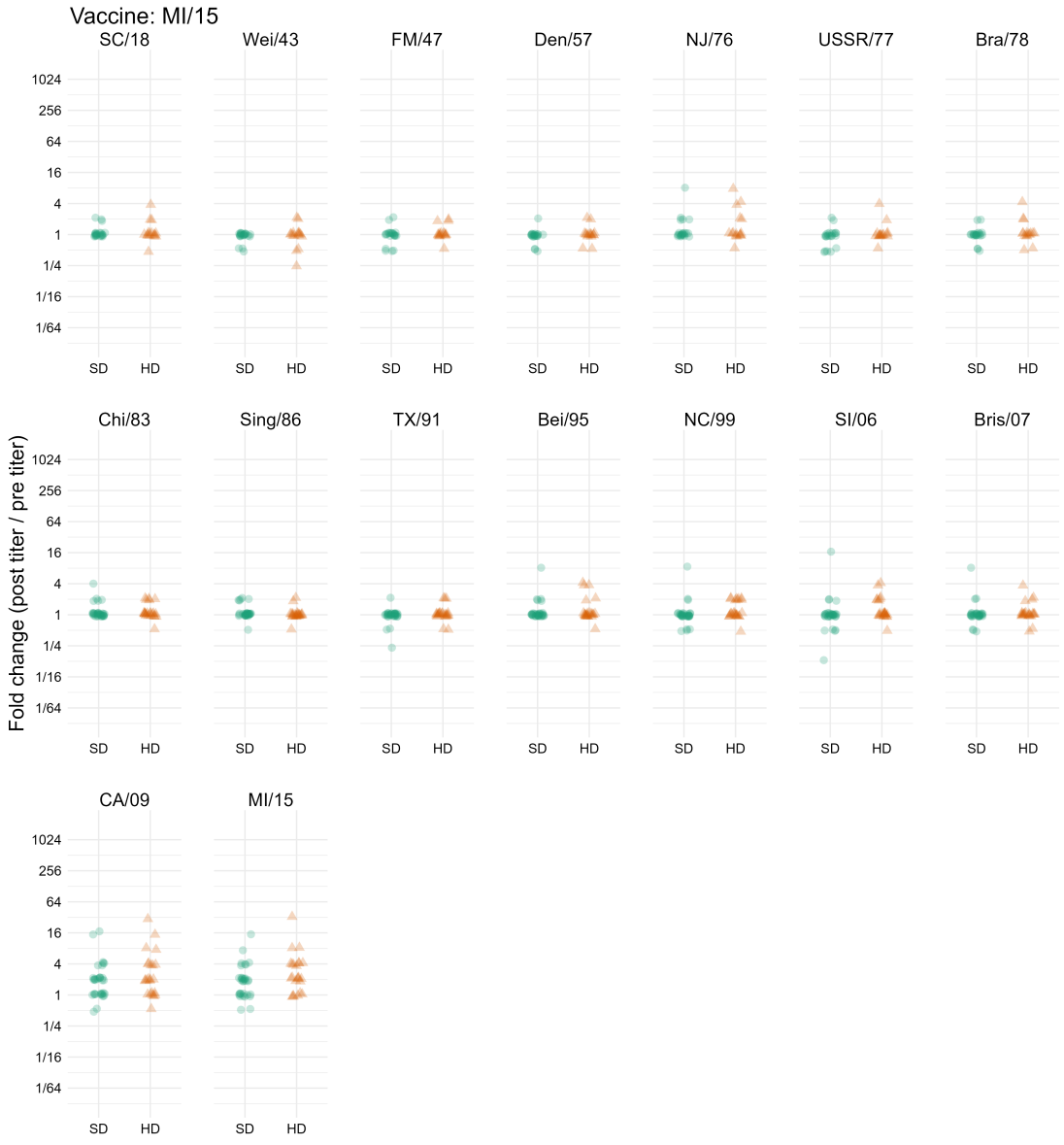

Figure A.15: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing MI/15-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against MI/15 vaccinated individual samples are shown.

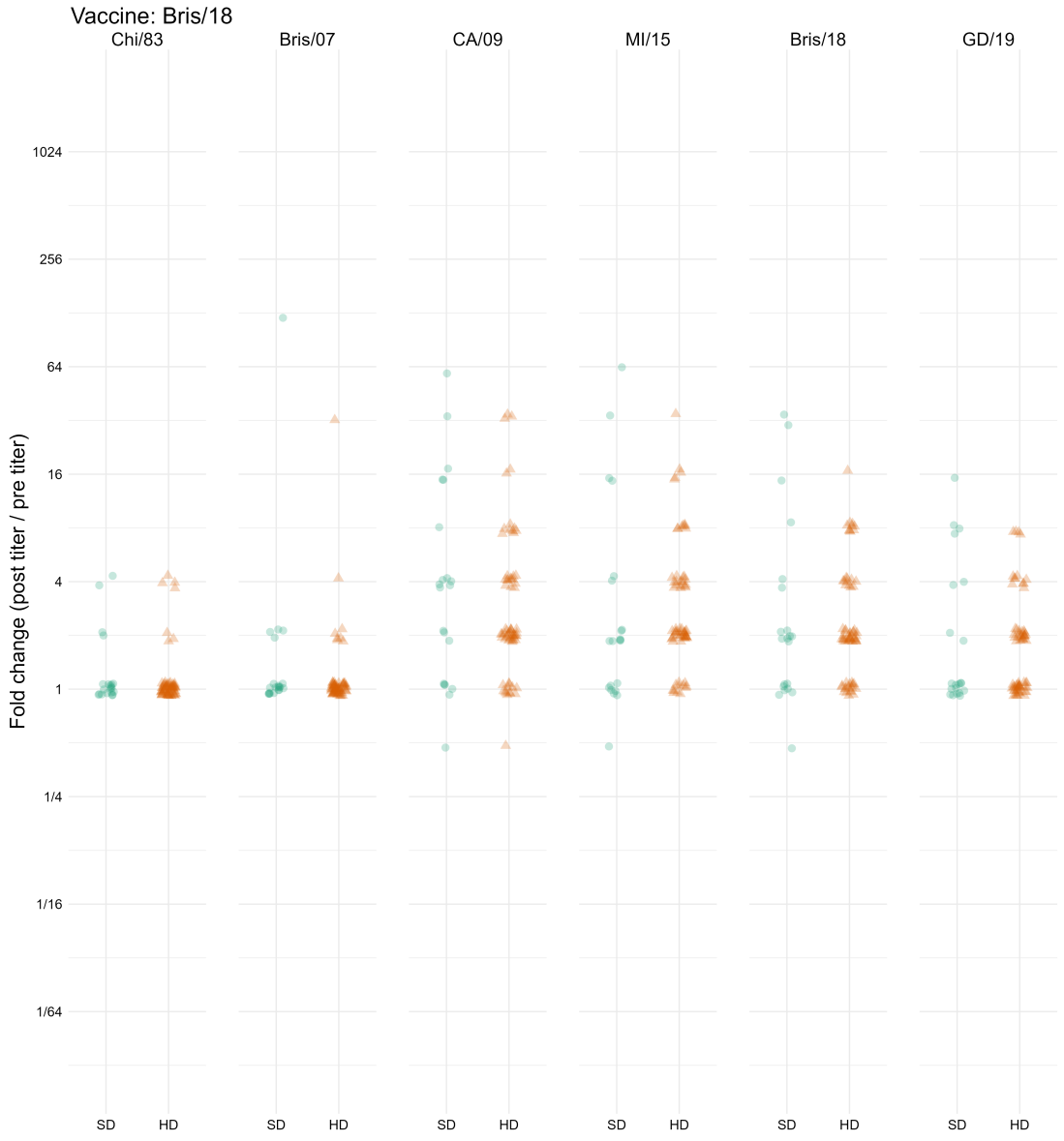

Figure A.16: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing Bris/18-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Bris/18 vaccinated individual samples are shown.

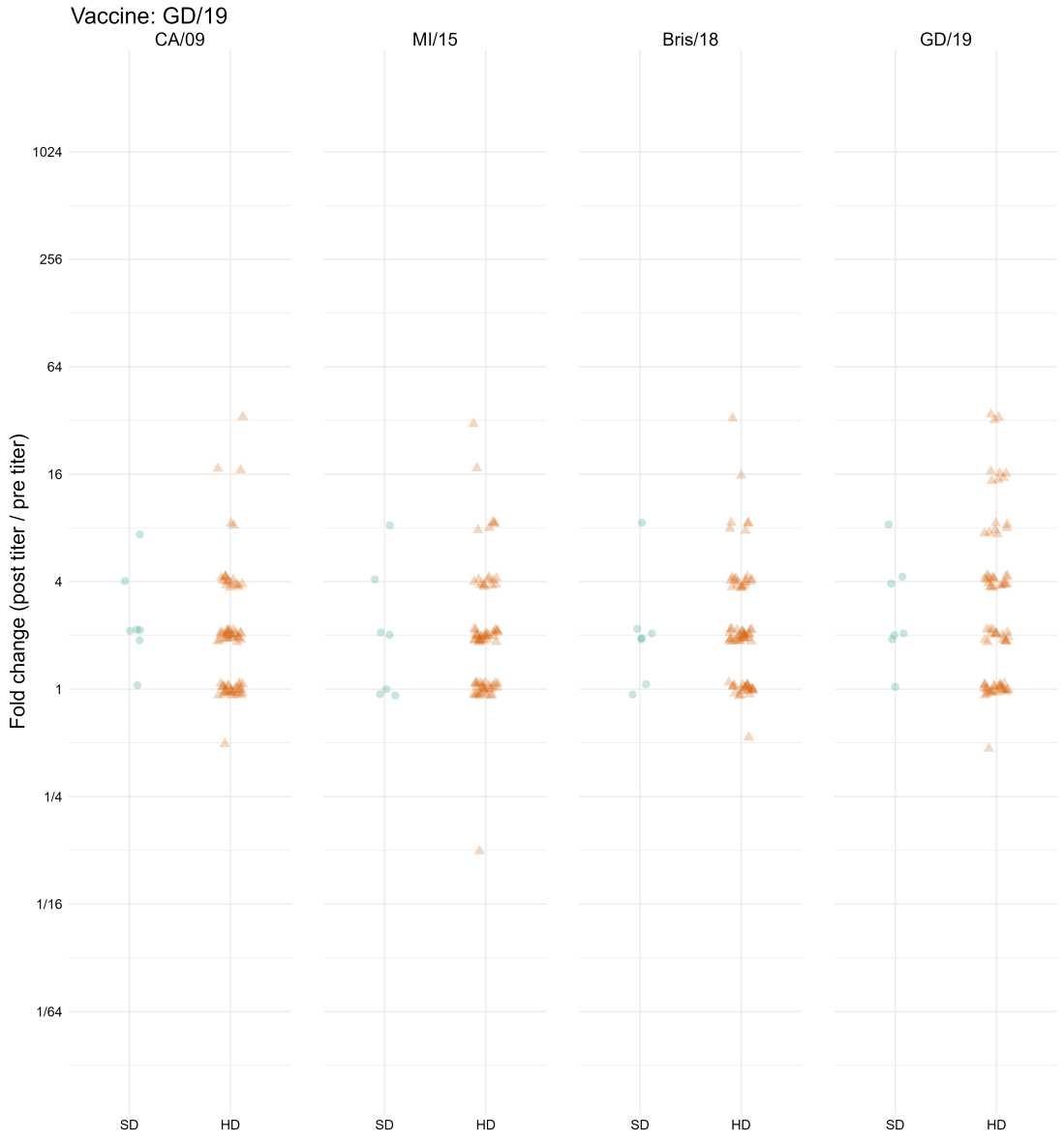

Figure A.17: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing GD/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against GD/19 vaccinated individual samples are shown.

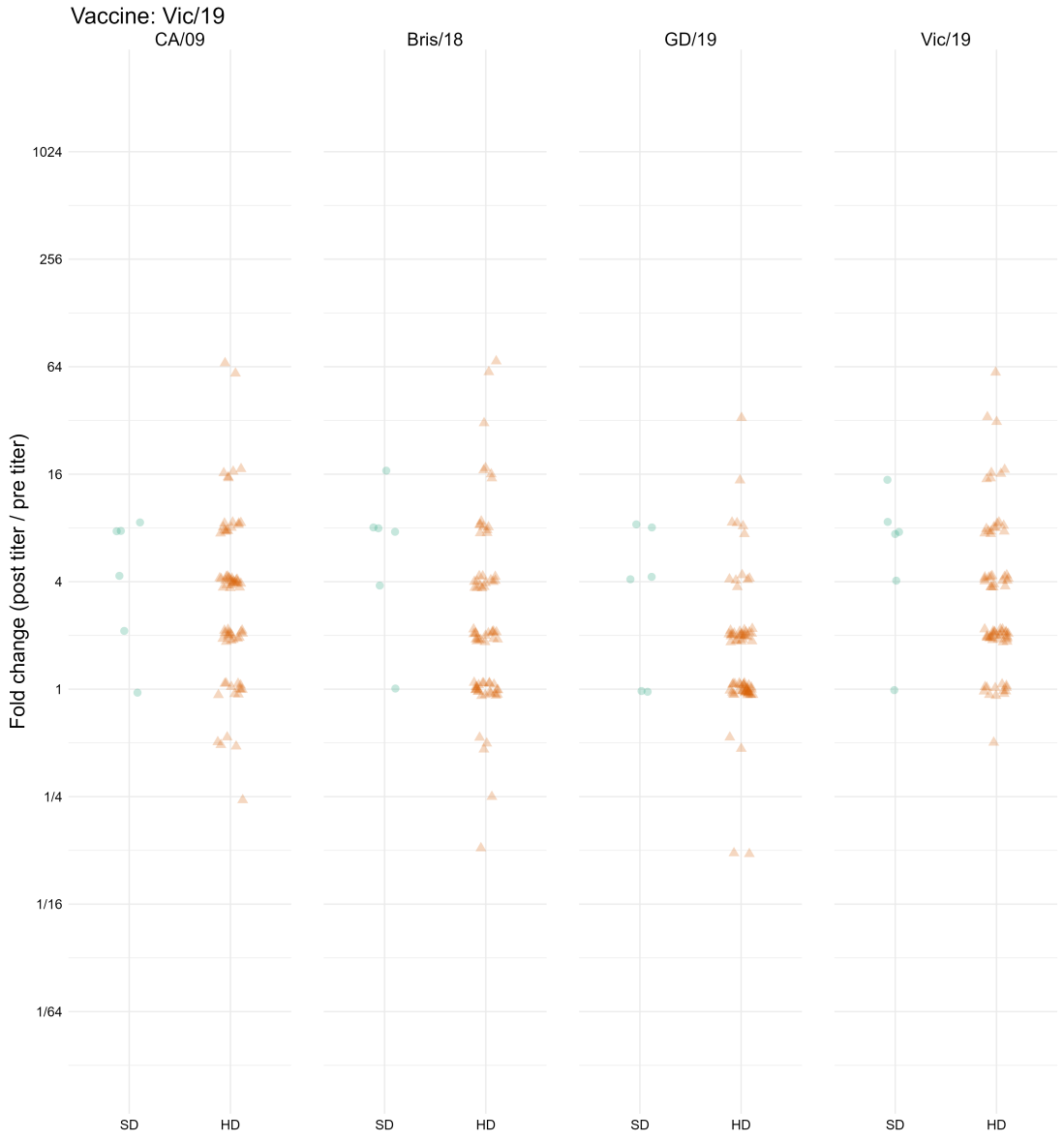

Figure A.18: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing Vic/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Vic/19 vaccinated individual samples are shown.

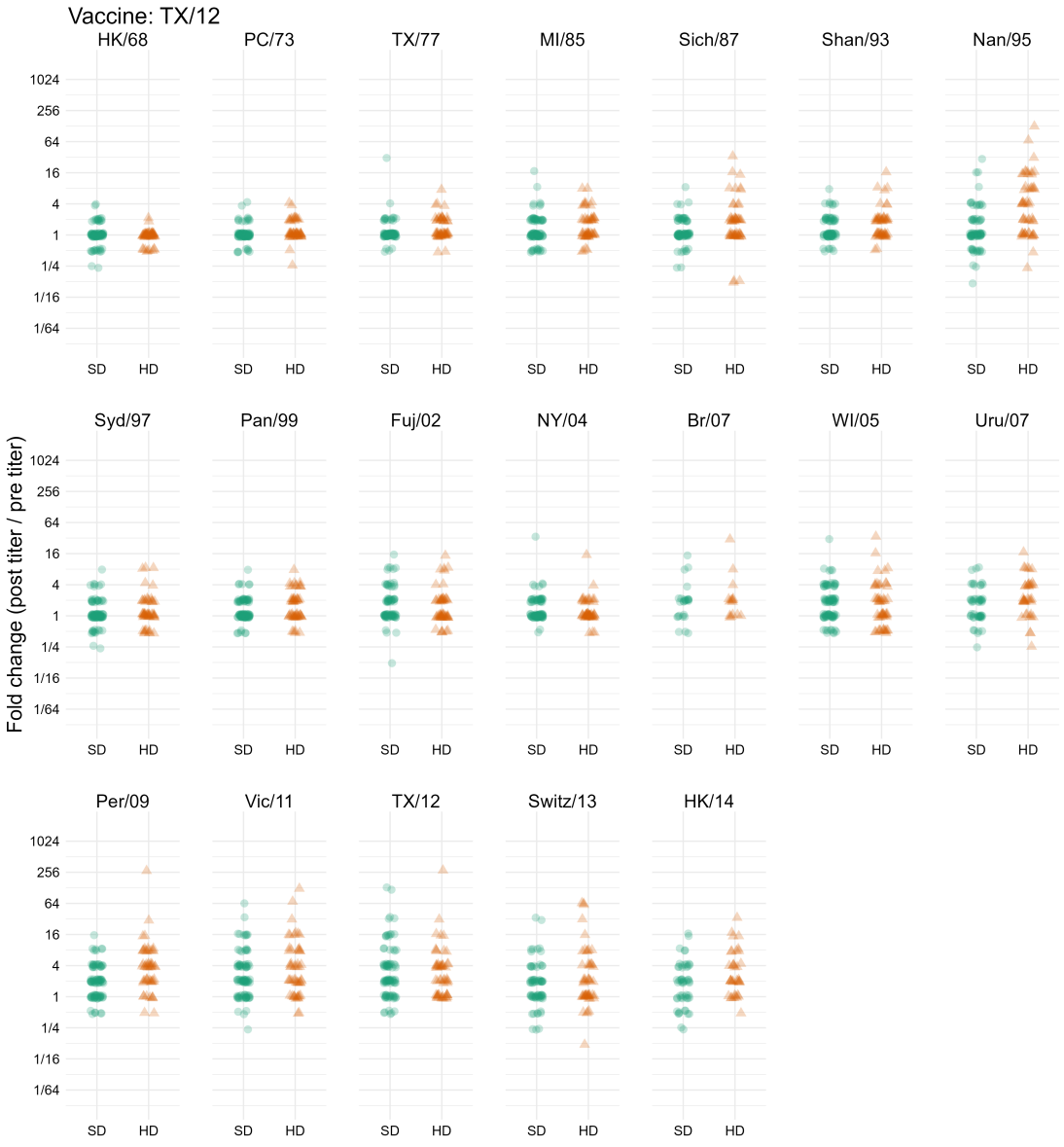

Figure A.19: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing TX/12-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against TX/12 vaccinated individual samples are shown.

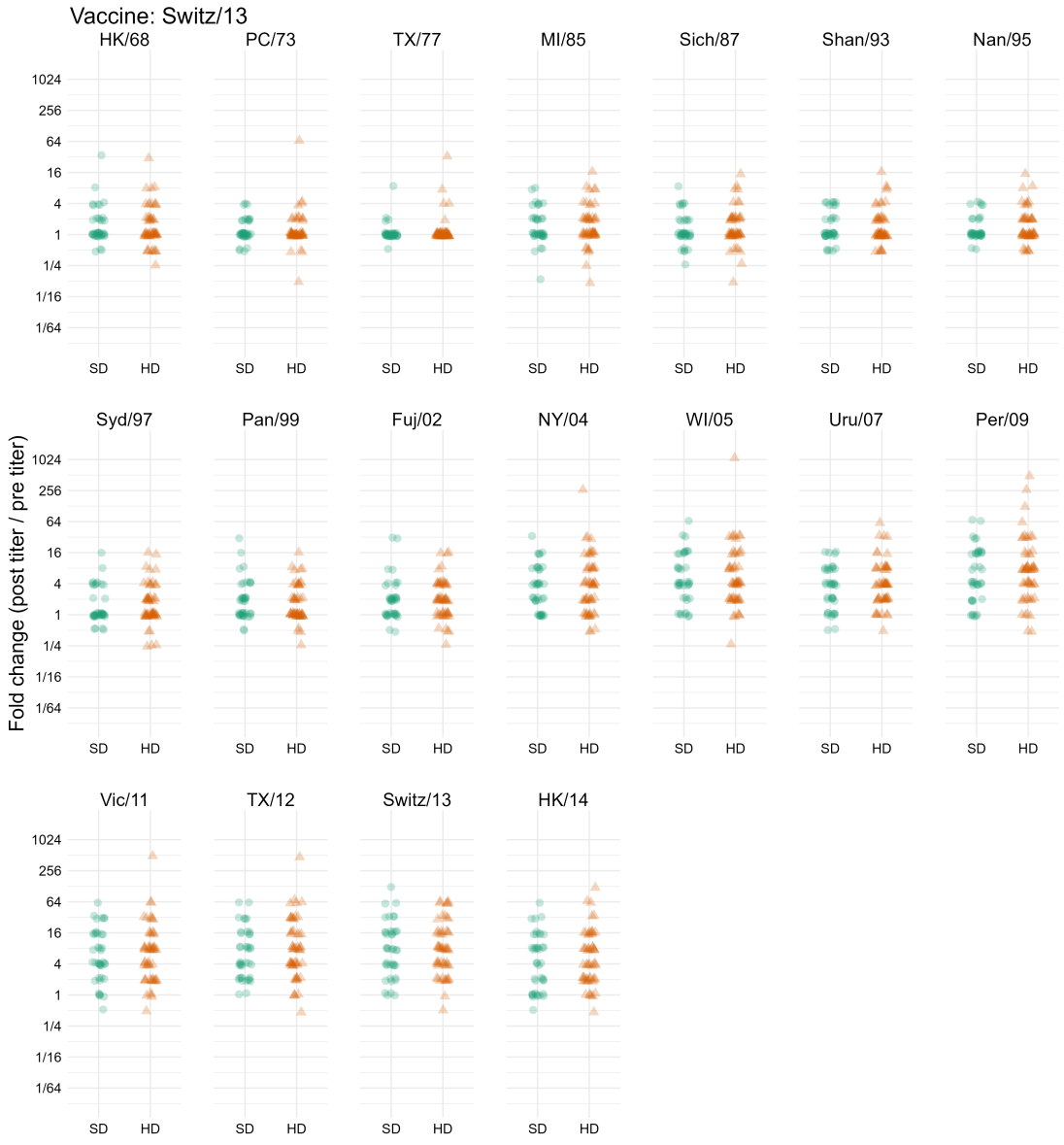

Figure A.20: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing Switz/13-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Switz/13 vaccinated individual samples are shown.

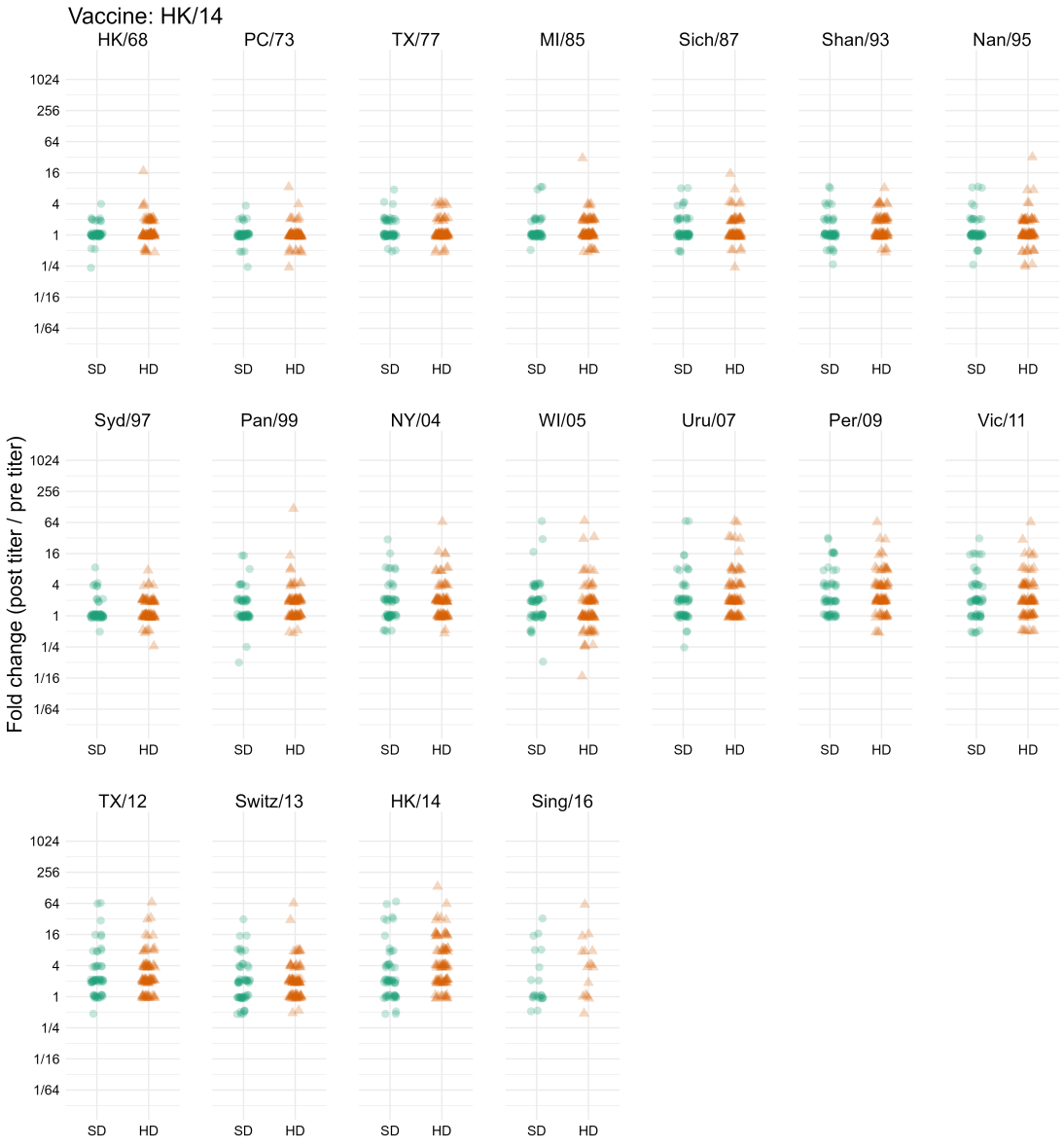

Figure A.21: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing HK/14-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against HK/14 vaccinated individual samples are shown.

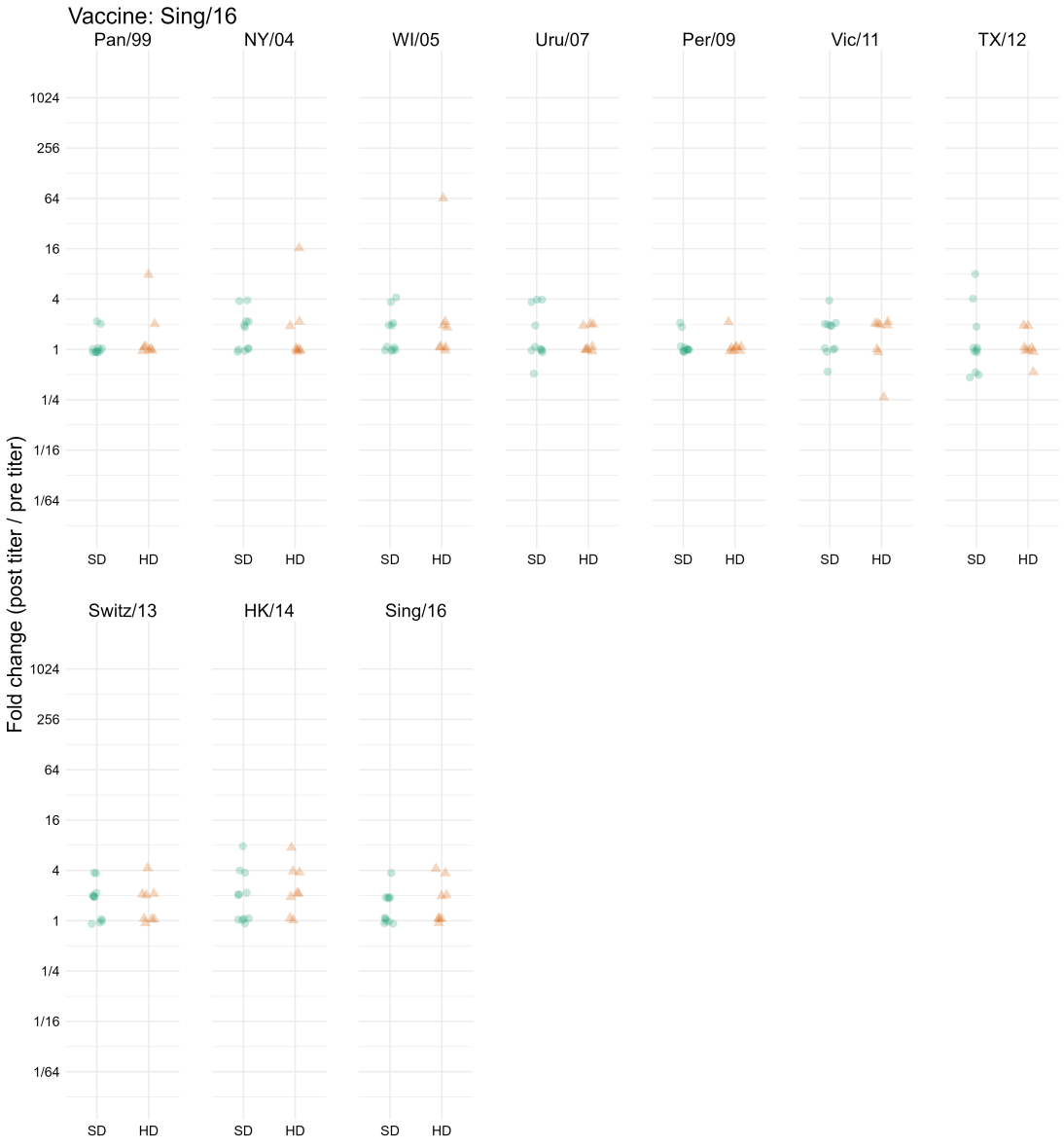

Figure A.22: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing Sing/16-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Sing/16 vaccinated individual samples are shown.

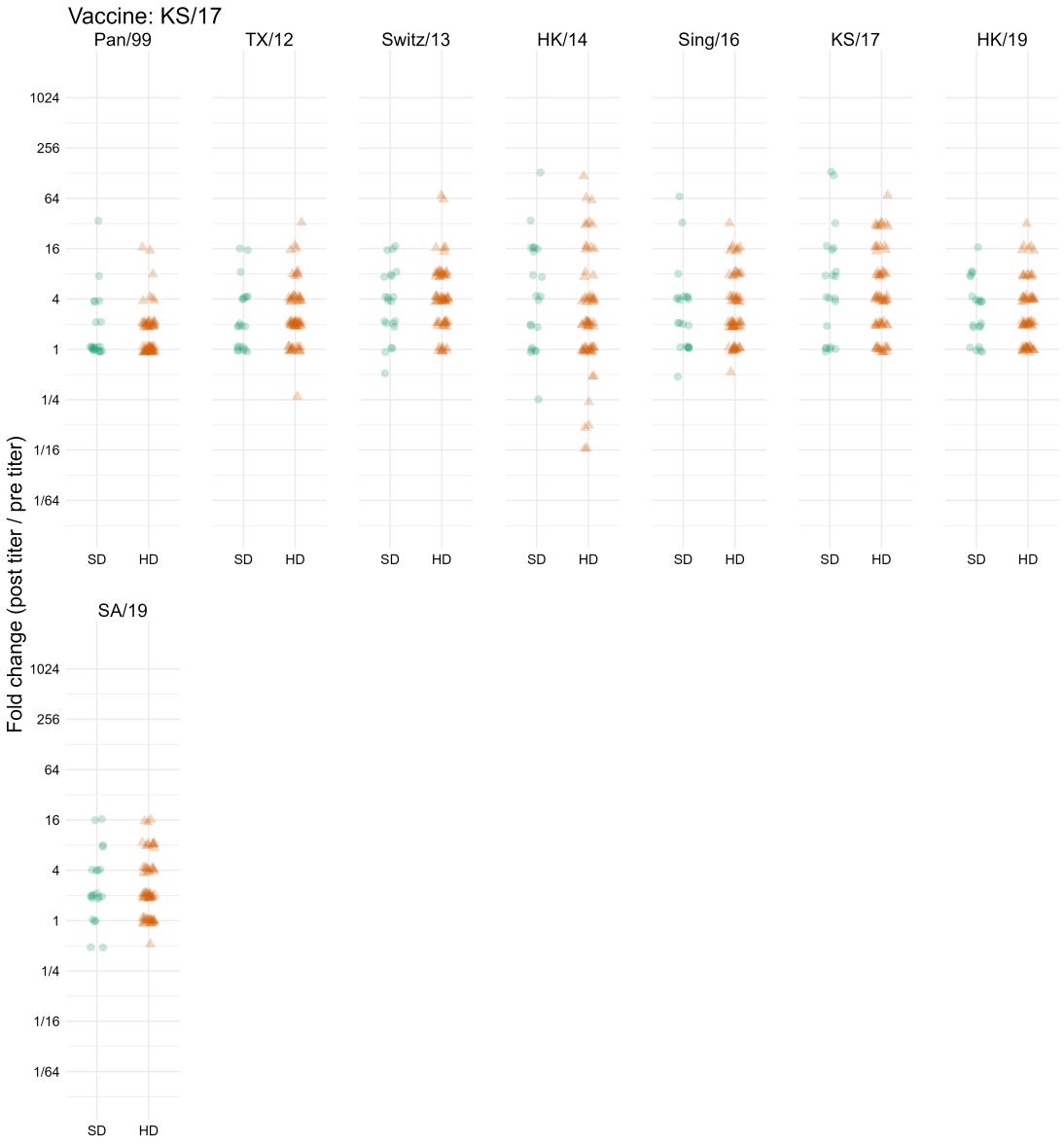

Figure A.23: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing KS/17-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against KS/17 vaccinated individual samples are shown.

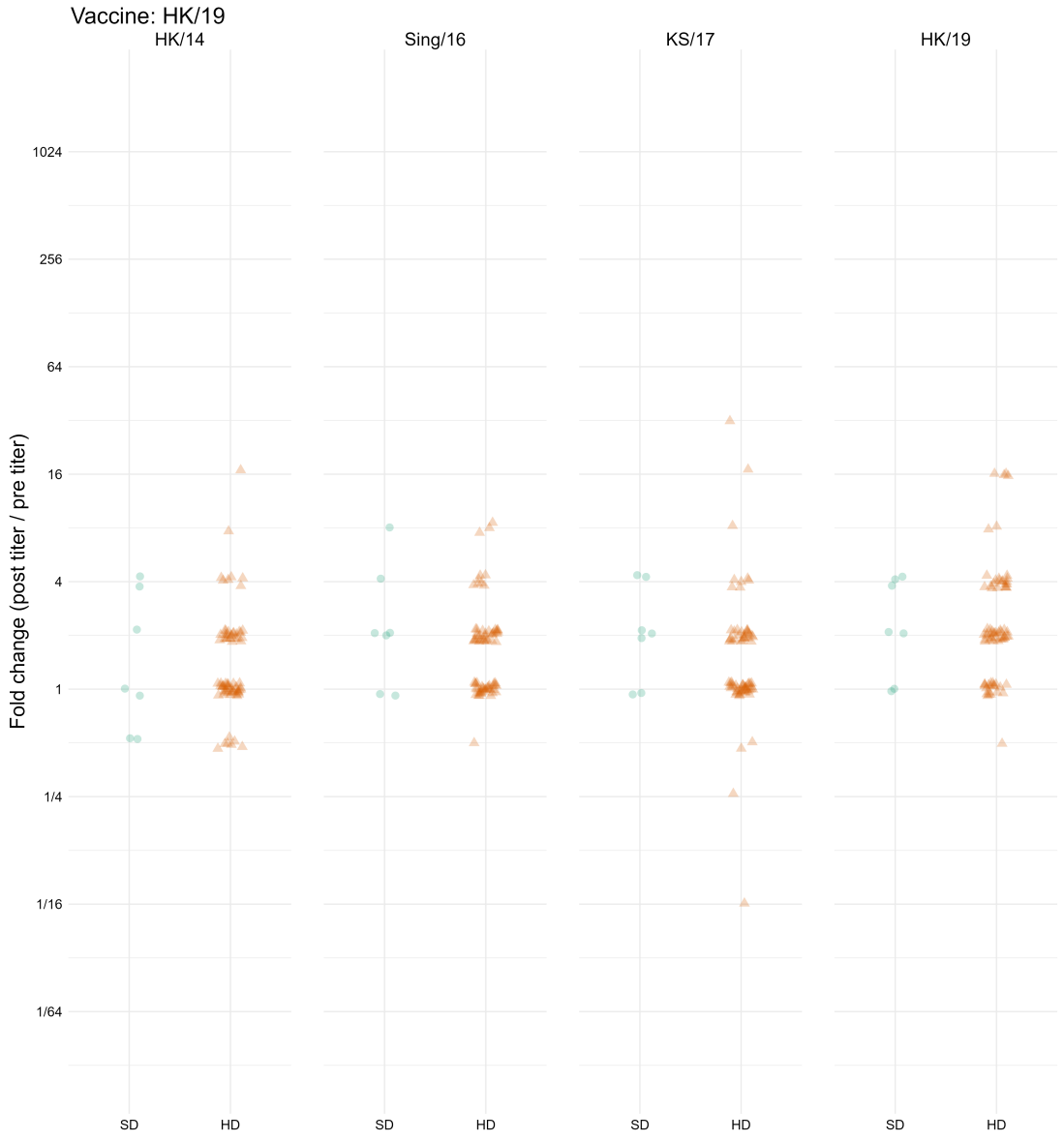

Figure A.24: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing HK/19-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against HK/19 vaccinated individual samples are shown.

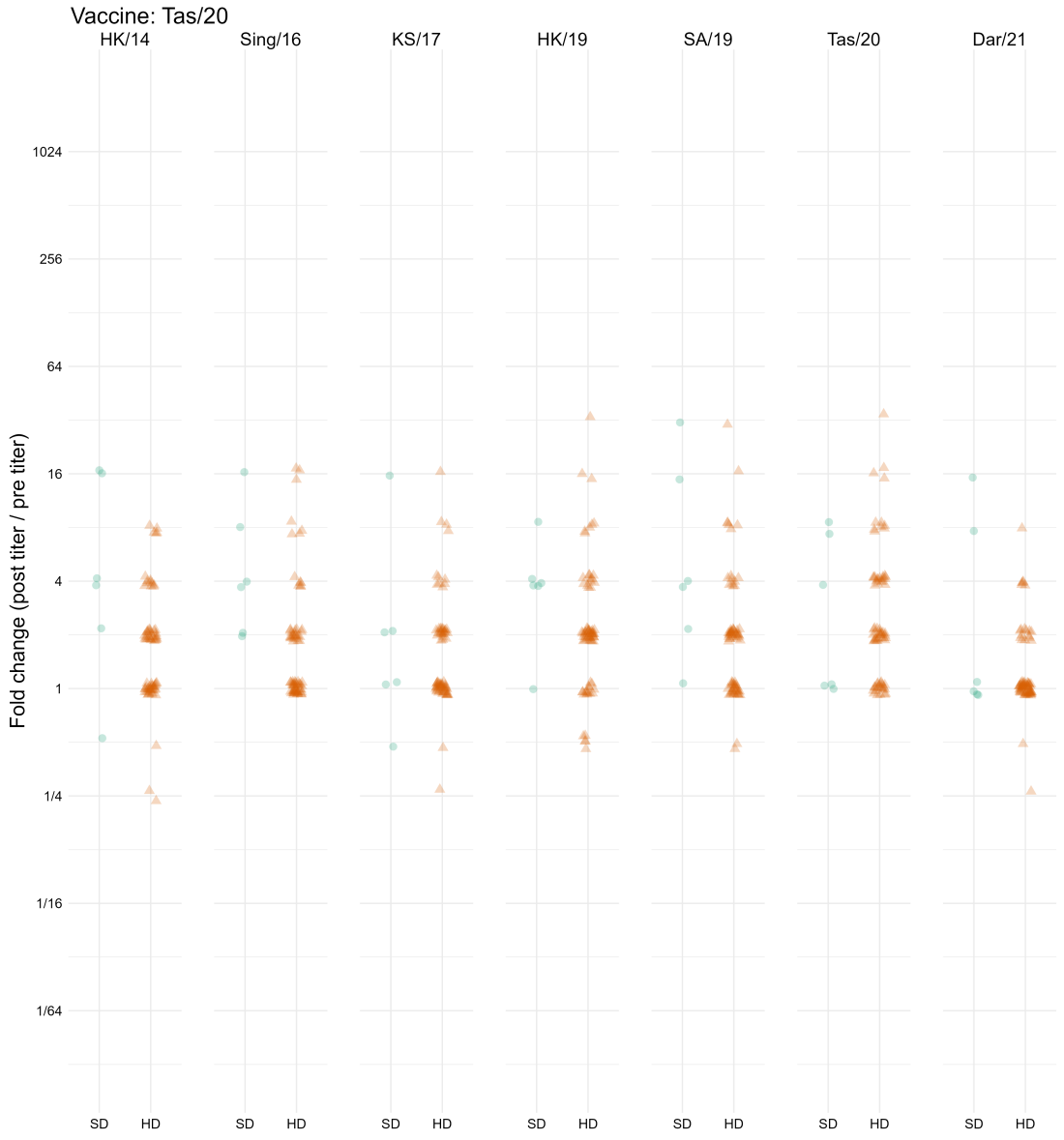

Figure A.25: Fold-change (post-vaccination titer divided by pre-vaccination titer) for all person-years where a participant was administered a vaccine containing Tas/20-like split inactivated virus. Each panel shows a historical strain which was used for HAI assays, and all H1N1 subtype historical strains that were used for running assays against Tas/20 vaccinated individual samples are shown.

Outcome summaries

For each outcome (and additionally the pre-vaccination titer), we computed crude summary statistics for the SD and HD groups in order to obtain a measure of the crude effect size. For the pre-vaccination titer, post-vaccination titer, and fold change, we computed the geometric mean and geometric SD, while for the seroprotection and serconversion, we computed the number and percentage of individuals for which each event occurred. We also computed standardized mean differences (SMDs) to compare the groups using the method of Yang and Dalton [214] via the R package `smd` [215].

We did not further stratify by each assay strain during the crude analysis, because the low sample size and number of comparisons would greatly inflate the amount of noise in the analysis, and understanding the stratified results would be very difficult. Notably, SMDs can be roughly interpreted by the guidelines shown in Table A.11, although these should not be strictly or decisively used to make decisions based on the qualitative guidelines alone [216, 217].

Table A.11: Suggested qualitative interpretations of the Cohen's d effect sizes represented by our SMD calculations. Note that these are only rough guidelines.

Cohen's d	**Interpretation**
0.01	Very small
0.20	Small
0.50	Medium
0.80	Large
1.20	Very large
2.0	Huge

Pre-vaccination titer

Table A.12 shows the crude analysis of the pre-vaccination titer. In contrast to the results shown in the main paper, the overall effect and subtype overall effects are all near zero. For some vaccine strains, the titer was clearly higher for the HD group, potentially due to the effect of receiving HD vaccines in multiple years. We using a flexible smoothing spline to control for the effect of pre-vaccination titer in our main model to reduce confounding by pre-vaccination titer in our main results.

Table A.12: Crude analysis of the pre-vaccination titer, stratified by dose.

		n[1]		**Pre-vaccination titer**[2]		
Subtype	**Vaccine strain**	SD	HD	SD	HD	**SMD**[3]
Overall	Overall	6,668	9,360	16.27 (±3.31)	16.29 (±3.21)	0.00 (-0.03, 0.03)
H1N1	Overall	3,082	4,288	13.40 (±3.04)	12.94 (±2.78)	-0.03 (-0.08, 0.01)
	CA/09	2,453	2,942	14.53 (±3.17)	13.96 (±2.88)	-0.04 (-0.09, 0.02)
	MI/15	451	328	10.46 (±2.43)	15.07 (±2.84)	0.38 (0.23, 0.52)
	Bris/18	126	414	8.16 (±2.14)	7.66 (±2.06)	-0.08 (-0.28, 0.11)
	GD/19	28	304	9.28 (±2.09)	12.03 (±2.57)	0.31 (-0.08, 0.70)
	Vic/19	24	300	7.49 (±1.71)	11.57 (±2.34)	0.61 (0.20, 1.03)
H3N2	Overall	3,586	5,072	19.22 (±3.46)	19.79 (±3.46)	0.02 (-0.02, 0.07)
	TX/12	1,544	972	19.42 (±3.43)	22.34 (±3.56)	0.11 (0.03, 0.19)
	Switz/13	720	1,042	15.90 (±3.11)	17.04 (±3.22)	0.06 (-0.04, 0.15)
	HK/14	974	1,597	20.16 (±3.62)	24.06 (±3.66)	0.14 (0.06, 0.22)
	Sing/16	110	80	36.39 (±4.62)	21.07 (±3.93)	-0.38 (-0.67, -0.09)
	KS/17	168	552	19.92 (±3.39)	14.80 (±2.94)	-0.26 (-0.43, -0.09)
	HK/19	28	304	23.20 (±2.54)	21.46 (±3.34)	-0.07 (-0.46, 0.31)
	Tas/20	42	525	15.87 (±2.88)	15.08 (±3.27)	-0.05 (-0.36, 0.27)

[1]Total number of HAI assays across all assay strains and seasons.
[2]Pre-vaccination HAI titer. Geometric mean (± geometric standard deviation).
[3]Standardized mean difference (HD - SD); SMD (95% CI).

Post-vaccination titer

Table A.13 shows the crude dose-stratified analysis of the post-vaccination titer outcome. We saw a weakly positive effect of the HD vaccine overall, which matched what we saw in our primary adjusted analysis. Notably, some strains showed a negative effect of the HD vaccine, which tended to correspond with the strains where the HD group had higher pre-vaccination titers (Table A.12) in the crude analysis.

Table A.13: Crude analysis of the post-vaccination titer, stratified by dose.

		n[1]		**Post-vaccination titer**[2]		
Subtype	**Vaccine strain**	SD	HD	SD	HD	**SMD**[3]
Overall	Overall	6,668	9,360	24.68 (±4.10)	28.99 (±4.01)	0.12 (0.08, 0.15)
H1N1	Overall	3,082	4,288	17.29 (±3.35)	20.09 (±3.24)	0.13 (0.08, 0.17)
	CA/09	2,453	2,942	18.44 (±3.40)	19.92 (±3.22)	0.06 (0.01, 0.12)
	MI/15	451	328	11.68 (±2.74)	19.01 (±3.19)	0.45 (0.30, 0.59)
	Bris/18	126	414	16.86 (±3.82)	14.07 (±3.01)	-0.15 (-0.35, 0.05)
	GD/19	28	304	21.02 (±2.22)	26.29 (±2.96)	0.24 (-0.15, 0.62)
	Vic/19	24	300	33.64 (±3.13)	28.95 (±3.49)	-0.13 (-0.54, 0.29)
H3N2	Overall	3,586	5,072	33.52 (±4.47)	39.53 (±4.39)	0.11 (0.07, 0.15)
	TX/12	1,544	972	28.48 (±3.96)	40.86 (±4.00)	0.26 (0.18, 0.34)
	Switz/13	720	1,042	39.09 (±5.33)	45.21 (±5.54)	0.09 (-0.01, 0.18)
	HK/14	974	1,597	32.43 (±4.55)	41.23 (±4.41)	0.16 (0.08, 0.24)
	Sing/16	110	80	52.78 (±4.80)	31.93 (±4.01)	-0.34 (-0.63, -0.05)
	KS/17	168	552	60.43 (±4.21)	40.81 (±4.09)	-0.28 (-0.45, -0.10)
	HK/19	28	304	44.16 (±2.70)	35.37 (±3.37)	-0.20 (-0.59, 0.19)
	Tas/20	42	525	50.40 (±3.82)	26.71 (±3.61)	-0.49 (-0.80, -0.17)

[1]Total number of HAI assays across all assay strains and seasons.
[2]Post-vaccination HAI titer. Geometric mean (± geometric standard deviation).
[3]Standardized mean difference (HD - SD); SMD (95% CI).

Fold change

The crude analysis of the fold change (Table A.14) was similar to our main adjusted analysis. The only qualitatively different result was the significant overall SMD, and the significant SMDs for all H1N1 and all H3N2 strains, indicating a small positive effect of the HD vaccine. These results are consistent with our main adjusted analysis.

Table A.14: Crude analysis of the fold change, stratified by dose.

		n[1]		**Fold change**[2]		
Subtype	**Vaccine strain**	SD	HD	SD	HD	**SMD**[3]
Overall	Overall	6,668	9,360	1.52 (±2.42)	1.78 (±2.51)	0.18 (0.15, 0.21)
H1N1	Overall	3,082	4,288	1.29 (±2.19)	1.55 (±2.24)	0.23 (0.19, 0.28)
	CA/09	2,453	2,942	1.27 (±2.19)	1.43 (±2.15)	0.15 (0.10, 0.21)
	MI/15	451	328	1.12 (±1.68)	1.26 (±1.83)	0.22 (0.07, 0.36)
	Bris/18	126	414	2.07 (±3.05)	1.84 (±2.15)	-0.12 (-0.32, 0.08)
	GD/19	28	304	2.26 (±1.98)	2.19 (±2.33)	-0.05 (-0.43, 0.34)
	Vic/19	24	300	4.49 (±2.48)	2.50 (±2.84)	-0.60 (-1.02, -0.19)
H3N2	Overall	3,586	5,072	1.74 (±2.56)	2.00 (±2.69)	0.14 (0.10, 0.18)
	TX/12	1,544	972	1.47 (±2.19)	1.83 (±2.62)	0.25 (0.17, 0.33)
	Switz/13	720	1,042	2.46 (±3.12)	2.65 (±3.44)	0.06 (-0.03, 0.16)
	HK/14	974	1,597	1.61 (±2.40)	1.71 (±2.36)	0.07 (-0.01, 0.15)
	Sing/16	110	80	1.45 (±1.79)	1.52 (±2.11)	0.07 (-0.22, 0.35)
	KS/17	168	552	3.03 (±3.19)	2.76 (±2.83)	-0.09 (-0.26, 0.09)
	HK/19	28	304	1.90 (±2.02)	1.65 (±2.04)	-0.21 (-0.59, 0.18)
	Tas/20	42	525	3.17 (±2.97)	1.77 (±2.11)	-0.63 (-0.95, -0.31)

[1]Total number of HAI assays across all assay strains and seasons.
[2]Fold change (Geometric mean ± geometric standard deviation).
[3]Standardized mean difference (HD - SD); SMD (95% CI).

Seroprotection

Table A.15 shows the crude analysis for the seroprotection outcome. The results were consistent with our main analysis, as well as the crude analyses of the other outcomes.

Table A.15: Crude analysis of the seroprotection rate, stratified by dose.

		n[1]		**Seroprotection**[2]		
Subtype	**Vaccine strain**	SD	HD	SD	HD	**SMD**[3]
Overall	Overall	6,668	9,360	2822 (42%)	4533 (48%)	0.12 (0.09, 0.15)
H1N1	Overall	3,082	4,288	998 (32%)	1638 (38%)	0.12 (0.08, 0.17)
	CA/09	2,453	2,942	847 (35%)	1116 (38%)	0.07 (0.02, 0.12)
	MI/15	451	328	88 (20%)	117 (36%)	0.37 (0.22, 0.51)
	Bris/18	126	414	36 (29%)	117 (28%)	-0.01 (-0.21, 0.19)
	GD/19	28	304	10 (36%)	133 (44%)	0.16 (-0.22, 0.55)
	Vic/19	24	300	17 (71%)	155 (52%)	-0.40 (-0.82, 0.02)
H3N2	Overall	3,586	5,072	1824 (51%)	2895 (57%)	0.12 (0.08, 0.17)
	TX/12	1,544	972	733 (47%)	599 (62%)	0.29 (0.21, 0.37)
	Switz/13	720	1,042	378 (52%)	616 (59%)	0.13 (0.04, 0.23)
	HK/14	974	1,597	471 (48%)	904 (57%)	0.17 (0.09, 0.25)
	Sing/16	110	80	72 (65%)	46 (57%)	-0.16 (-0.45, 0.12)
	KS/17	168	552	118 (70%)	318 (58%)	-0.27 (-0.44, -0.09)
	HK/19	28	304	21 (75%)	166 (55%)	-0.44 (-0.83, -0.05)
	Tas/20	42	525	31 (74%)	246 (47%)	-0.57 (-0.89, -0.26)

[1]Total number of HAI assays across all assay strains and seasons.
[2]Seroprotection events (indicator for post-titer ≥ 40); n (%).
[3]Standardized mean difference (HD - SD); SMD (95% CI).

Seroconversion

Table A.16 shows the crude analysis for the seroconversion outcome. The results were consistent with our main analysis, as well as the crude analyses of the other outcomes. Notably, some of the vaccines have a much lower rate of seroconversion than seroprotection, and the SMDs for seroconversion between the two groups are quite negative, which is affected by both the smaller sample size in those groups as well as the higher pre-vaccination titers in the HD group.

Table A.16: Crude analysis of the seroconversion rate, stratified by dose.

		n[1]		**Seroconversion**[2]		
Subtype	**Vaccine strain**	SD	HD	SD	HD	**SMD**[3]
Overall	Overall	6,668	9,360	981 (15%)	1710 (18%)	0.10 (0.06, 0.13)
H1N1	Overall	3,082	4,288	267 (09%)	535 (12%)	0.12 (0.08, 0.17)
	CA/09	2,453	2,942	203 (08%)	282 (10%)	0.05 (-0.01, 0.10)
	MI/15	451	328	19 (04%)	25 (08%)	0.14 (0.00, 0.29)
	Bris/18	126	414	23 (18%)	65 (16%)	-0.07 (-0.27, 0.13)
	GD/19	28	304	6 (21%)	62 (20%)	-0.03 (-0.41, 0.36)
	Vic/19	24	300	16 (67%)	101 (34%)	-0.70 (-1.12, -0.28)
H3N2	Overall	3,586	5,072	714 (20%)	1175 (23%)	0.08 (0.04, 0.12)
	TX/12	1,544	972	206 (13%)	211 (22%)	0.22 (0.14, 0.30)
	Switz/13	720	1,042	240 (33%)	378 (36%)	0.06 (-0.03, 0.16)
	HK/14	974	1,597	155 (16%)	276 (17%)	0.04 (-0.04, 0.12)
	Sing/16	110	80	13 (12%)	6 (08%)	-0.15 (-0.43, 0.14)
	KS/17	168	552	69 (41%)	191 (35%)	-0.13 (-0.31, 0.04)
	HK/19	28	304	8 (29%)	36 (12%)	-0.43 (-0.81, -0.04)
	Tas/20	42	525	23 (55%)	77 (15%)	-0.93 (-1.25, -0.61)

[1]Total number of HAI assays across all assay strains and seasons.
[2]Seroconversion events (indicator for post-titer ≥ 40 and fold change ≥ 4); n (%).
[3]Standardized mean difference (HD - SD); SMD (95% CI).

Model diagnostics

We assessed model convergence and sampling using the $\hat{R}$ statistic and the effective sample size (ESS) of the parameters. Detailed explainations of these metrics can be found in other sources [218]. Briefly, $\hat{R}$ assesses the mixing of the chains, and a large value indicates that chains have explored separate regions of the posterior or do not agree about the posterior density. The bulk ESS and tail ESS are two measures of the ESS, which provide information about how many draws of the parameters we would have if all of our draws were completely uncorrelated (typically draws from the posterior are correlated and thus have less information than independent draws)

Since each of the models contains hundreds of parameters, it is not feasible to display every diagnostic statistic in the summary. Table A.17 contains an abbreviated summary of the most important diagnostic criteria, which were within acceptable bounds ($\hat{R} \lesssim 1.01$ and both ESS $\gtrsim 1000$ for all parameters). While there were a handful of divergent transitions, they were negligible compared to the total amount of samples.

The main exception is the model for the seroprotection outcome, which had a high number of divergent transitions. While a few divergences are no cause for concern, and the number of post-warmup divergences for this model was around 1%, this may indicate that the seroprotection outcome is more difficult to model than the others. The lower ESS values for this model also support this assertion. Since the seroconversion model sampled normally, we attribute this problem to the data collected rather than the parametrization of the model. Since we encourage the use of the titer increase and post-vaccination titer models instead, rather than the binary outcome models, we did not attempt any reparametrizations or extended sampling of the seroprotection model.

Our models have many parameters (451 for the Gaussian models and 450 for the logistic models due to the lack of a residual variance), so we cannot show parameter-level diagnostics for all parameters. However, we selected a representative parameter of each major class to include representative diagnostics for each model. The parameters we selected are:

- `b_doseHD`, the population-level slope term for the effect of HD vaccination;

Table A.17: Summaries of $\hat{R}$, effective sample size, E-BFMI, and number of divergent transitions for each of the Bayesian models we fit.

Model	**Divergences**	**min E-BFMI**	**min tail ESS**	**min bulk ESS**	**max $\hat{R}$**
Post titer	2 / 20000	0.615	3682	1982	1.013
Seroconversion	2 / 20000	0.636	5956	2718	1.006
Seroprotection	249 / 20000	0.636	1118	1660	1.012
Titer increase	4 / 20000	0.609	2779	1550	1.013

- `bs_sbirth_year_c_1`, the first population-level slope term for the smoothing spline effect of birth year;
- `cor_strain_type__Intercept__doseHD`, the correlation between the varying intercept effect and the varying effect of HD vaccination for the strain type variable;
- `Intercept`, the population-level global intercept parameter;
- `sd_id__Intercept`, the variance parameter for the distribution of intercept effects which vary by subject ID;
- `sd_strain_type__doseHD`, the variance parameter for the distribution of HD slope effects which vary by strain type;
- `sd_strain_type__Intercept`, the variance parameter for the distribution of intercept effects which vary by strain type; and
- `sds_sage_c_1`, the variance parameter for the distribution of slope effects for the smoothing spline parameters for the effect of age.

Note that diagnostic statistics for every parameter can be generated from our provided code, and our code can easily be modified to create diagnostic plots for any other parameters.

Trace rank plots

We show trace rank plots (also called rank histograms) for each of these parameters for the titer increase model in Figure A.26, for the post-vaccination titer model in Figure A.27, for the seroprotection model in Figure A.28, and for the seroconversion model in Figure A.29. A rank histogram is created by ranking the relative value of each draw of a parameter across all chains, binning the ranks (we used a bin size of 500), and creating a separate histogram for each chains, which are superimposed in each plot. A parameter with chains that have

mixed will show histograms which constantly swap positions [219]. These plots indicate good mixing for each of our models, and there is no indication from these plots that hints at the higher number of divergent transitions for the seroprotection model.

Prior and posterior density comparison

In order to understand which parameters are affected the most by our priors, we compared the prior and posterior densities visually for the same set of representative parameters. Figure A.30 shows the density comparison for the model with titer increase as the outcome, Figure A.31 for the model with post-vaccination titer as the outcome, Figure A.32 for the model with seroprotection as the outcome, and Figure A.33 for the model with seroconversion as the outcome. In order to construct the densities, we sampled from the priors (without updating them with our data) using the same sampling algorithm parameters as we did for the posterior samples—that is, we obtained $20,000$ prior samples. For both the prior and posterior samples, we computed the Gaussian kernel density estimate across all samples for each of the representative parameters, which is the density curve shown in each of the figures.

In Figure A.30, we can see that many of the parameters behaved as expected. All of the parameters except for the correlation parameter have clearly moved towards a peak, indicating strong convergence to a value away from the prior distribution. The posterior distribution for the correlation parameter, `cor_strain_type__Intercept__doseHD`, has some differences from the prior but is largely the same. This is not surprising, because these correlation parameters can often be difficult to infer. There is no real alternative prior for the correlation parameters to use as a sensitivity analysis (shrinking the LKJ parameter would bias our model towards more extreme correlations, whereas 2 is a sensible default), and the posterior distribution of the correlation parameter merely indicates uncertainty in the posterior distribution of the correlation, since it is already constrained in the values which it can take. We do not expect changing the correlation parameter to strongly affect the model results.

While some of the variance parameters, represented here by `sd_strain_type__doseHD`

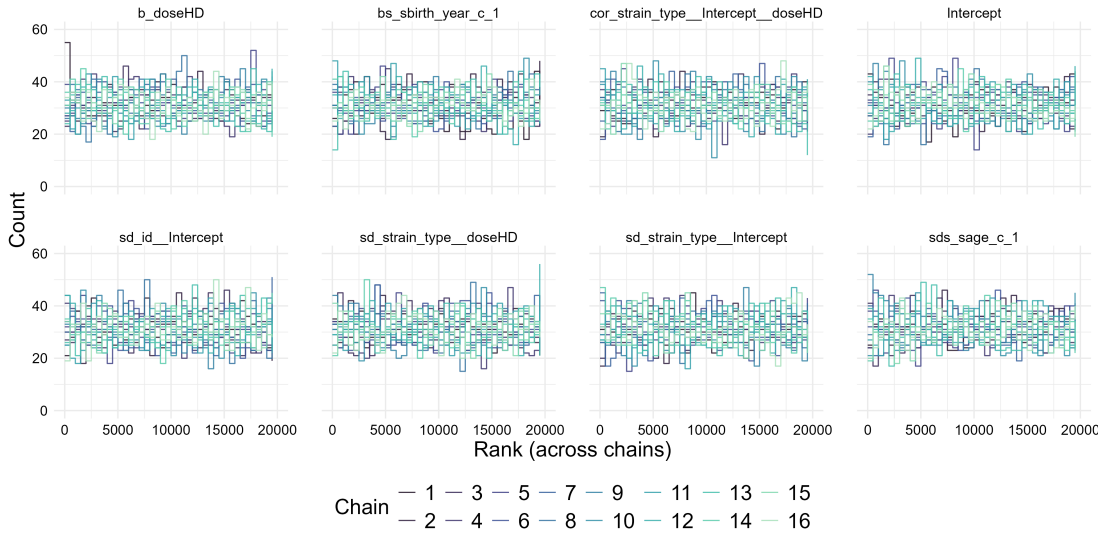

Figure A.26: Trace rank plot of the representative parameters from the model with titer increase as the outcome. The histograms oscillate randomly, and constantly swap positions, indicating good chain mixing.

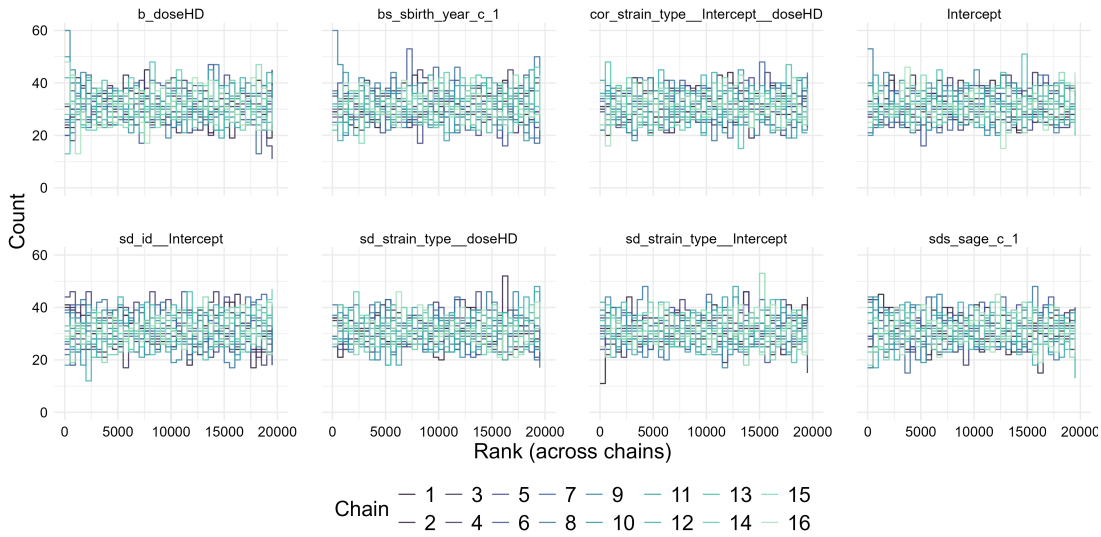

Figure A.27: Trace rank plot of the representative parameters from the model with post-vaccination as the outcome. The histograms oscillate randomly, and constantly swap positions, indicating good chain mixing.

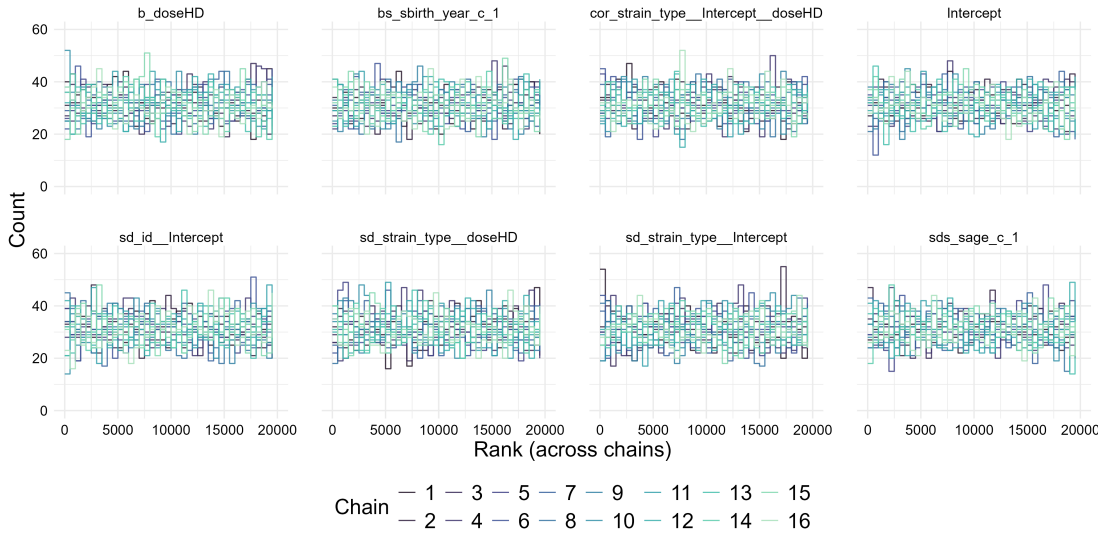

Figure A.28: Trace rank plot of the representative parameters from the model with seroprotection as the outcome. The histograms oscillate randomly, and constantly swap positions, indicating good chain mixing.

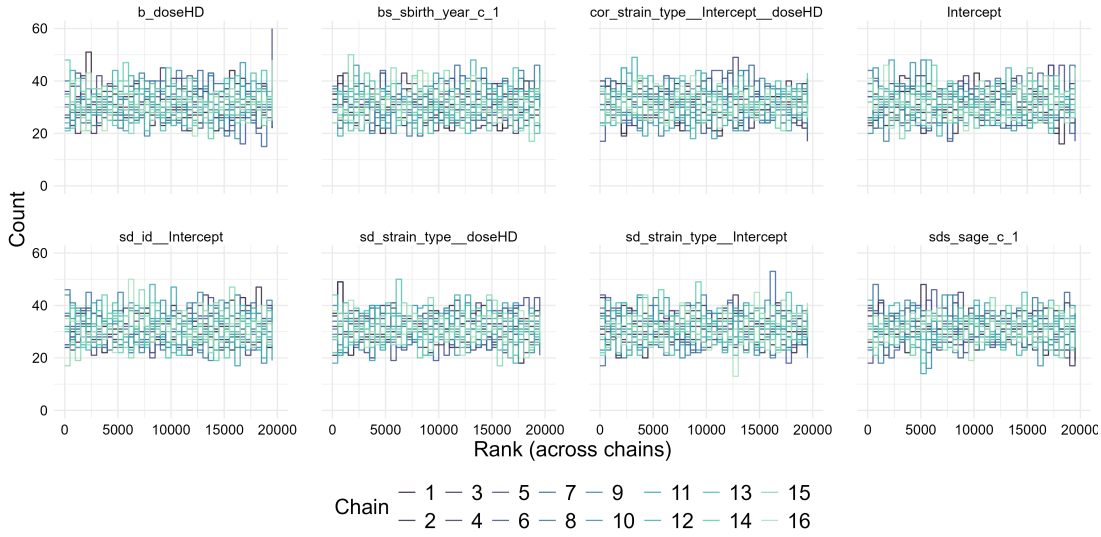

Figure A.29: Trace rank plot of the representative parameters from the model seroconversion as the outcome. The histograms oscillate randomly, and constantly swap positions, indicating good chain mixing.

and `sd_strain_type__Intercept` stay relatively similar in shape to the priors, they have more pronounced peaks. These terms are often very small (near zero) and would induce boundary convergence errors in a frequentist ML formulation of our model, so it is unsurprising that they stay relatively similar to our priors, which are close to zero, but allow for the parameter to move away from zero if the data allow this—in our case, the data do not appear to support a random effects variance far away from zero. This boundary variance also compounds the difficulties in estimating the random effects correlation parameters.

The conclusions from Figure A.31 are nearly identical to the conclusions from Figure A.30. We expect this, because the models are mathematically very similar.

In contrast to the conclusions from Figures A.30 and A.31, some of the conclusions in Figure A.32 are quite different. In a frequentist context, we would say that this model, with a dichotomous outcome, has much lower statistical power than the models with continuous outcomes. The majority of titers in our sample data were below 40, leading to a lack of events when we analyze seroprotection as a dichotomous outcome, especially for strains which are very antigenically different from the vaccine. We can see that the `Intercept` and `bs_sbirth_year_c_1` parameters have moved slightly away from the priors, but not much, indicating that we cannot learn much from the data about these parameters under this model. The random effects variance parameters (with the exception of the variance for the random effect of subject ID) are all very close to zero, near the priors, indicating that these parameters are not informative for learning about the seroprotection outcome. Since our priors were chosen as regularizing, skeptical priors (that allow the effects to become large easily if large effects are suggested by the data), the difficulty of predicting our underpowered dichotomous outcome is the cause of posterior densities similar to the prior densities—changing our priors to differently shaped skeptical priors would not make a difference. This analysis is not suggestive of any issues which would cause the higher number of divergent transitions for the seroprotection outcome.

The conclusions for the seroconversion model, shown in Figure A.33, are the same as those shown in Figure A.32. The seroconversion outcome contains even less events than the seroprotection outcome, leading to angrou even larger issue with statistical power. Because

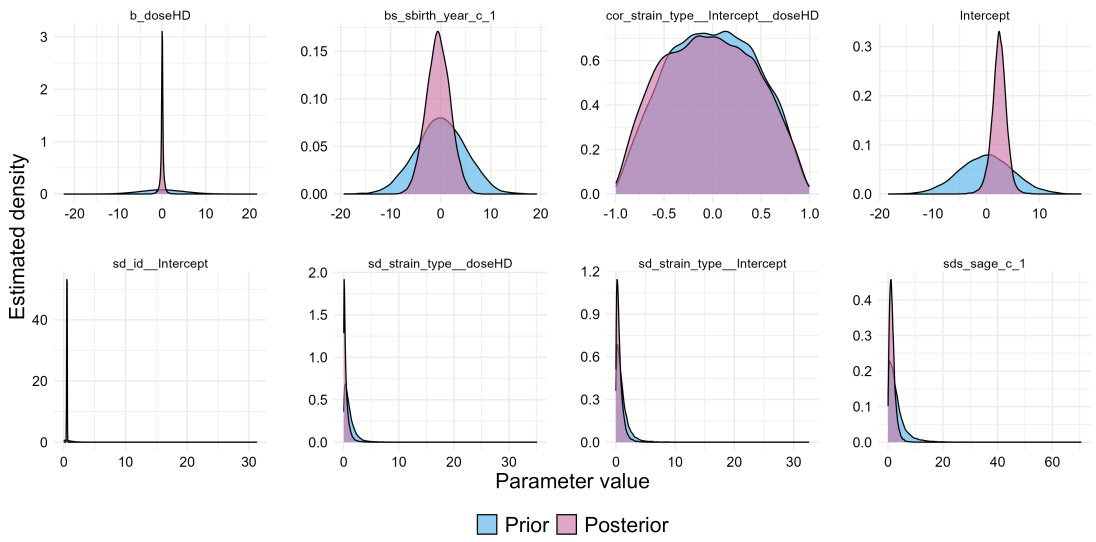

Figure A.30: Prior and posterior density plots for each of the representative parameters for the model where titer increase was the outcome.

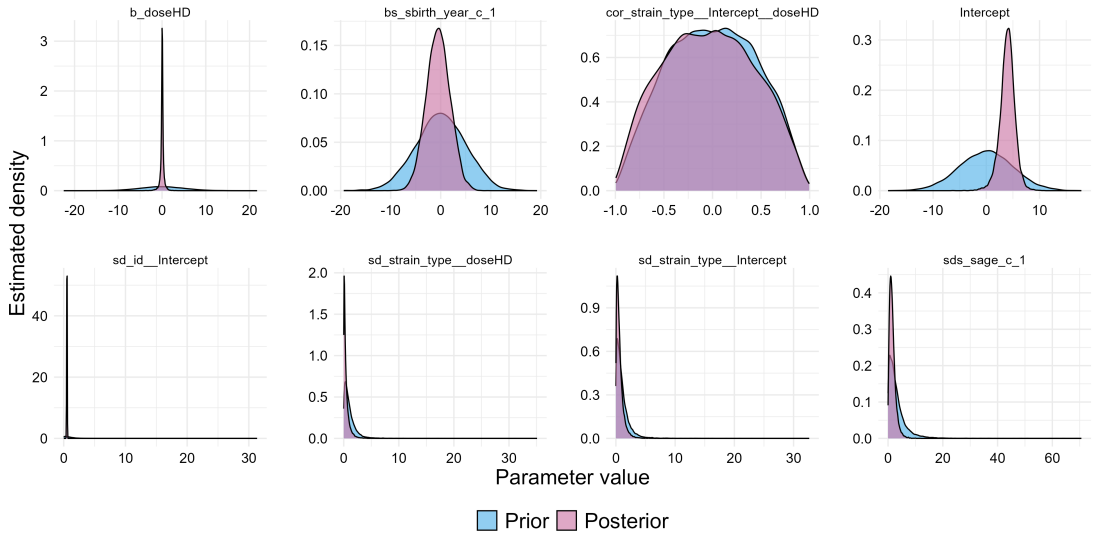

Figure A.31: Prior and posterior density plots for each of the representative parameters for the model where post-vaccination titer was the outcome.

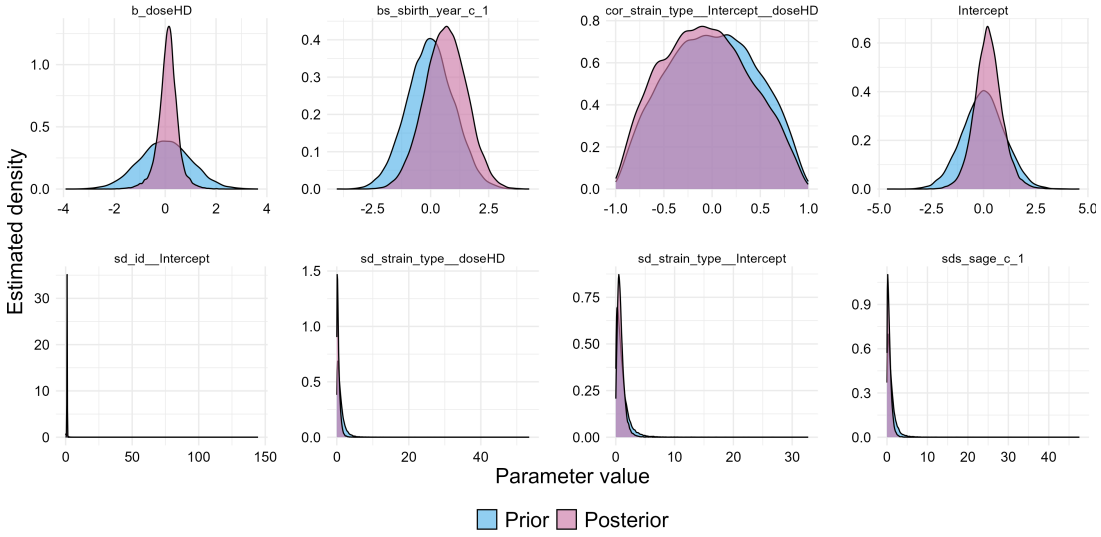

Figure A.32: Prior and posterior density plots for each of the representative parameters for the model where seroprotection was the outcome.

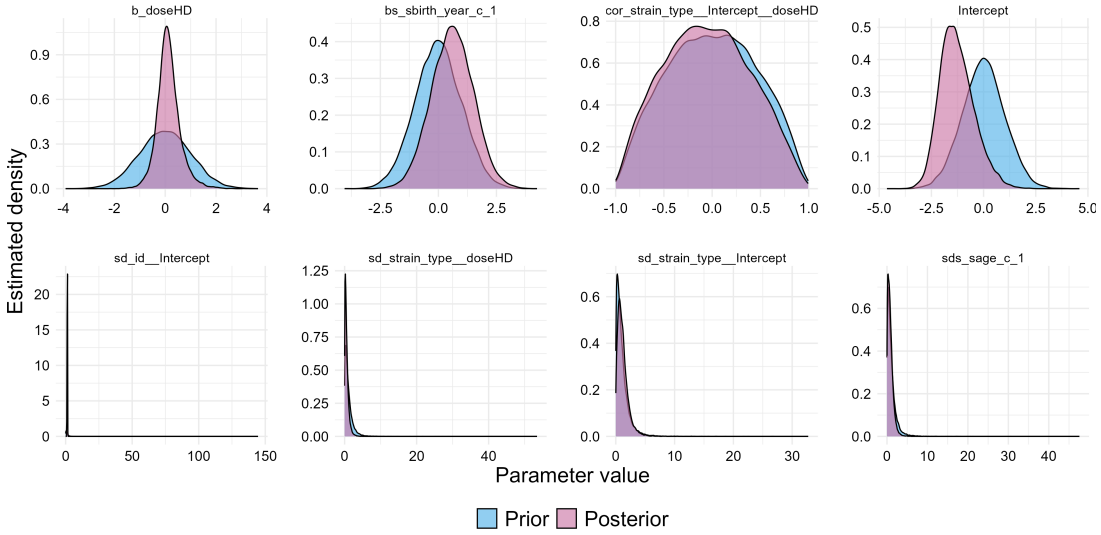

Figure A.33: Prior and posterior density plots for each of the representative parameters for the model where seroconversion was the outcome.

the logistic regression models have so little power and many parameters did not move away from the skeptical prior distributions, we do not recommend interpreting the binary outcome results directly, which is one reason we focus on the models for titer increase as our main results.

Homologous model results for titer increase outcome

In the main text, we briefly mentioned that results which compared only the homologous vaccine response supported a positive effect of the HD vaccine compared to the SD vaccine, as shown in previous literature. Figure A.34 shows our results when considering only the homologous response to each vaccine.

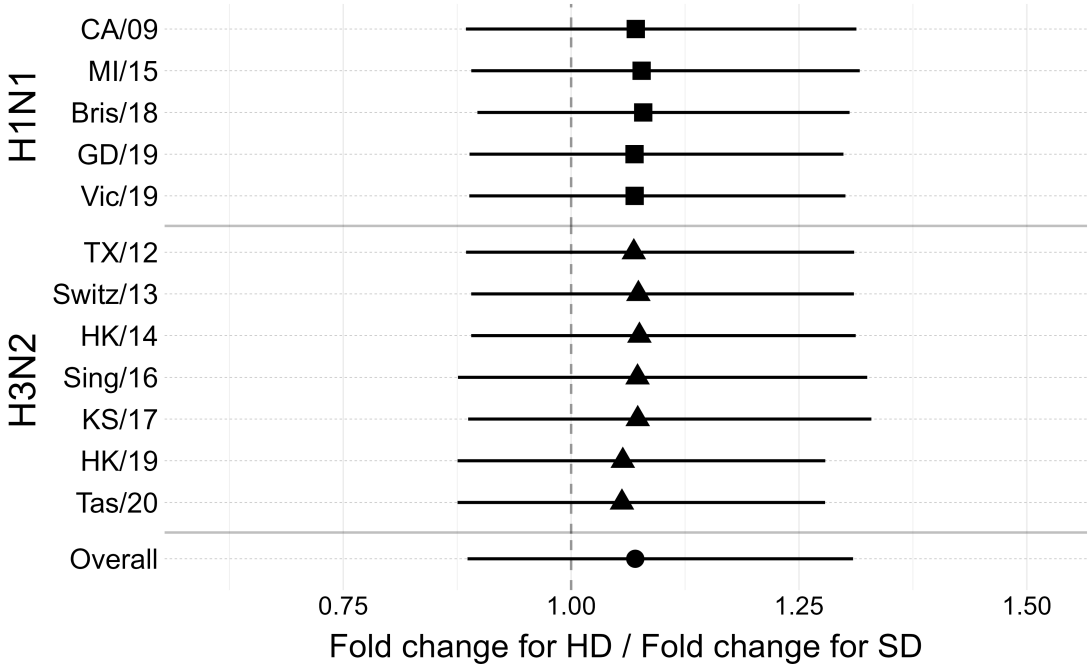

Figure A.34: Exponentiated cACE estimates for each vaccine strain and overall. Only homologous responses to each vaccine were considered.

The credible intervals are wide, consistent with our other findings and in general with this type of complex observational data. However, all of the point estimates are positive, which matches previous literature on the effect of the HD vaccine on the homologous response.

Model results for other outcomes

As a sensitivity analysis, we show the three figures from our main results, but using the alternative model outcomes post-vaccination titer, seroprotection, and seroconversion.

Post-vaccination titer

All of the figures in this section show the results for post-vaccination titer as the model outcome. All of our results agreed with the results in the main text.

Figure A.35 shows the cACEs for each vaccine strain when only the heterologous strains were included. Figure A.36 shows the cACEs for all assay strains. Figure A.37 shows the cACEs for all vaccine strains, pooling assay strains together within each vaccine strain. Figure A.38 shows the cACEs for each season, with the vaccine strain and assay strains for that season all pooled together.

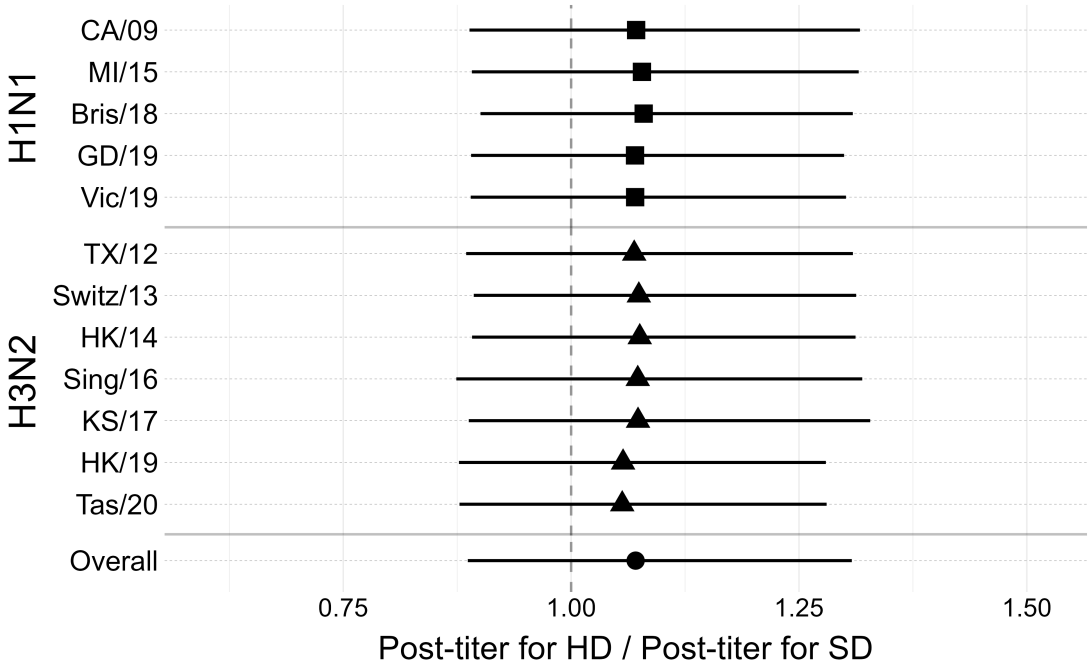

Figure A.35: Exponentiated cACE estimates from the post-vaccination titer model for each vaccine strain. Only homologous responses to each vaccine were considered.

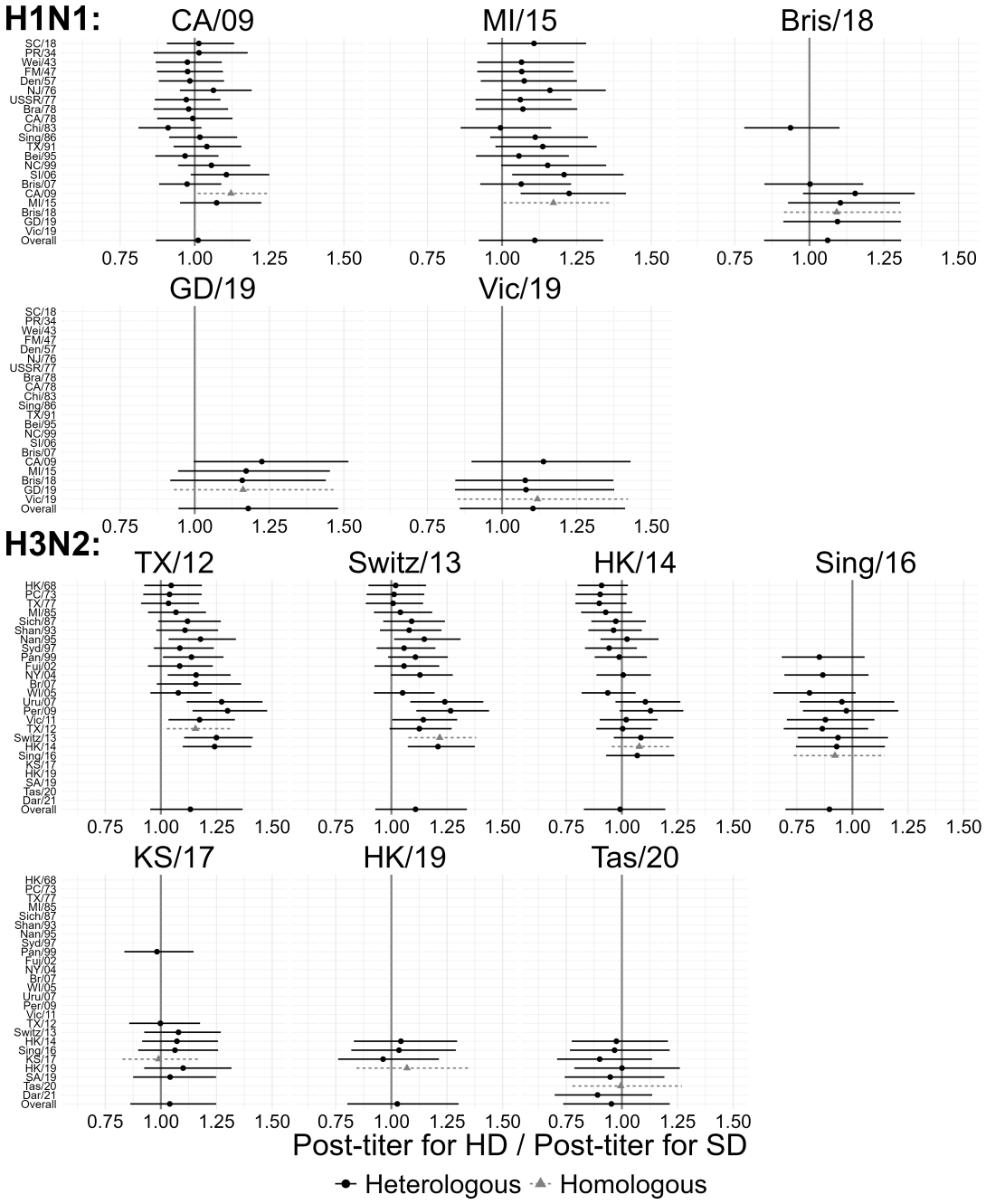

Figure A.36: Heterologous post-vaccination cACE estimates (exponentiated) for each assay strain and vaccine.

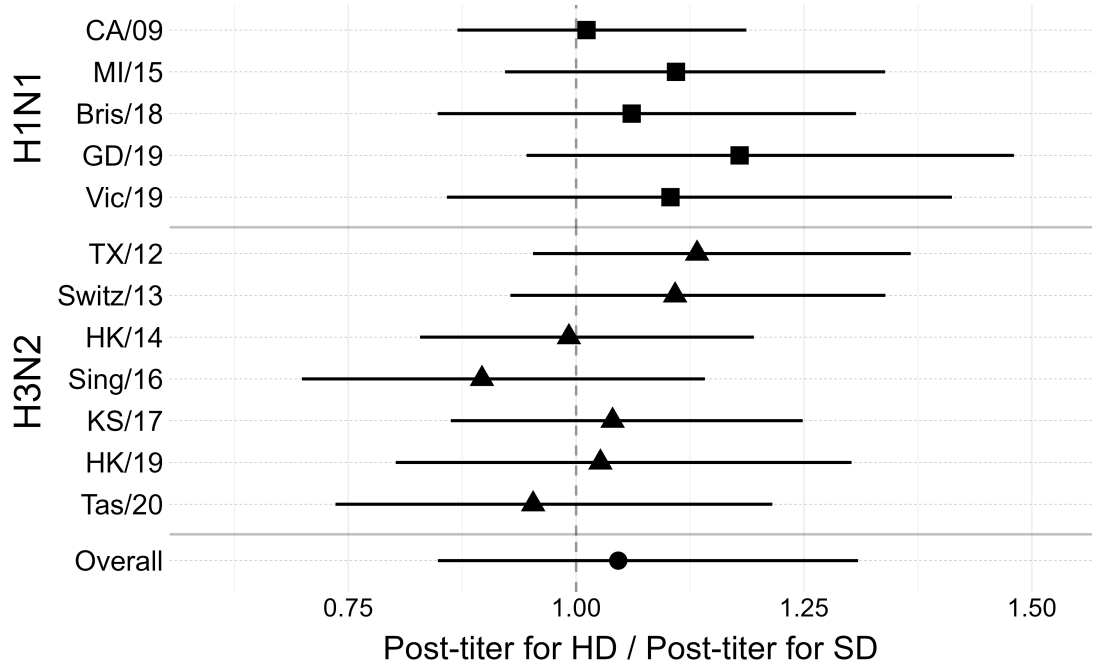

Figure A.37: Heterologous post-vaccination cACE estimates (exponentiated) for each vaccine.

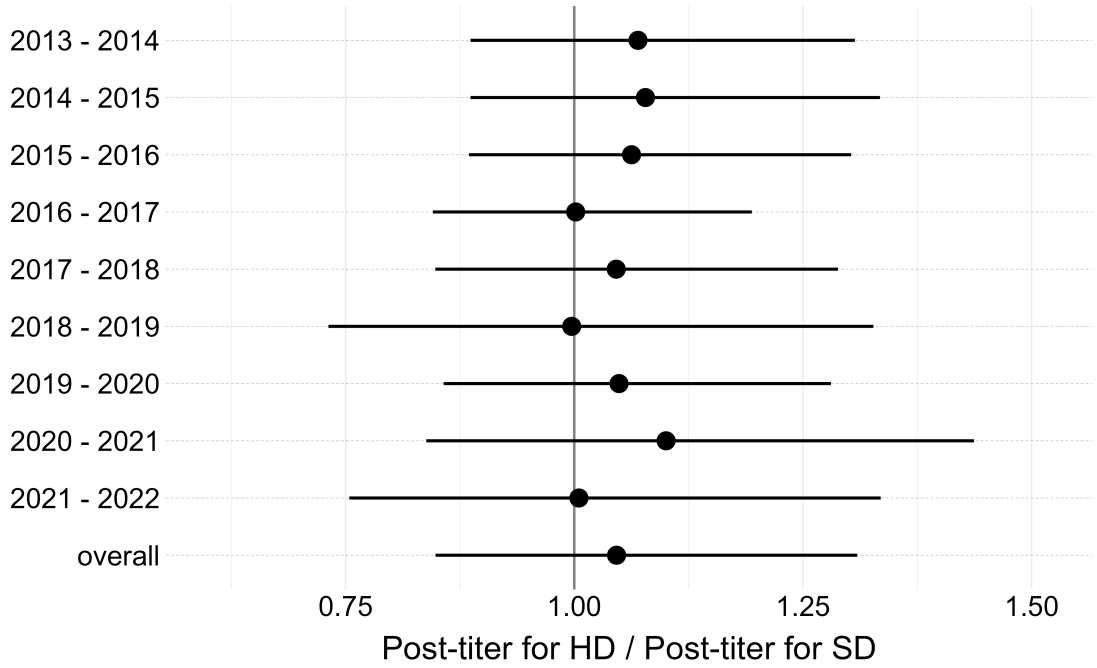

Figure A.38: Heterologous post-vaccination cACE estimates (exponentiated) for each season.

Seroprotection

All of the figures in this section show the results for seroprotection as the model outcome. All of our results agreed with the results in the main text and there were no major qualitative differences.

Figure A.39 shows the cACEs for each vaccine strain when only the heterologous strains were included.

Figure A.40 shows the cACEs for all assay strains.

Figure A.41 shows the cACEs for all vaccine strains, pooling assay strains together within each vaccine strain.

Figure A.42 shows the cACEs for each season, with the vaccine strain and assay strains for that season all pooled together.

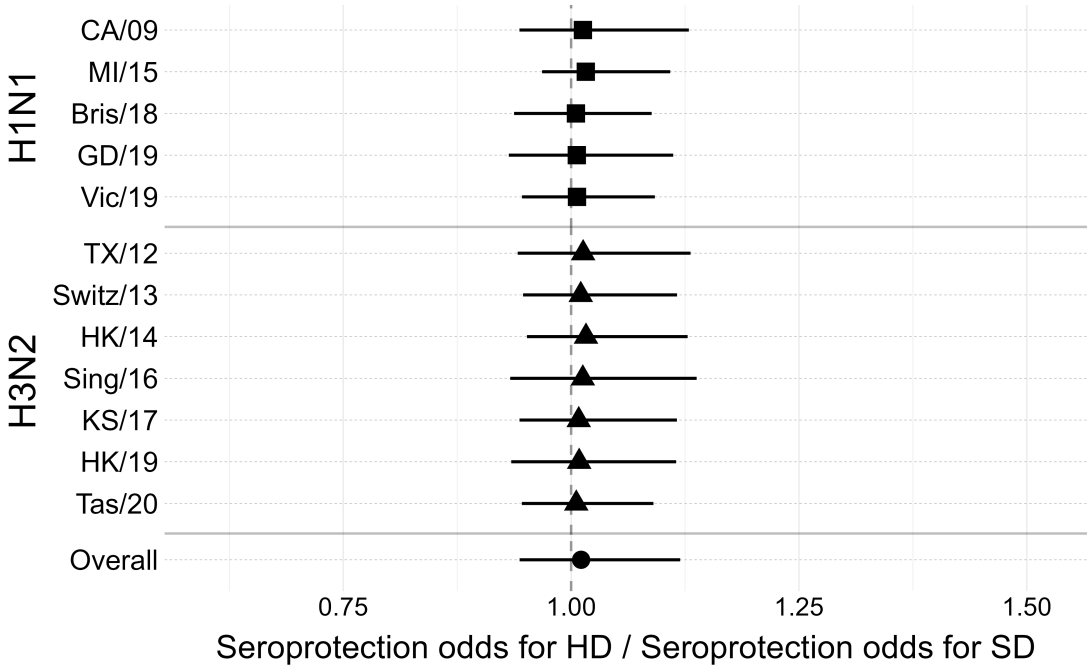

Figure A.39: Exponentiated cACE estimates from the seroprotection model for each vaccine strain and overall. Only homologous responses to each vaccine were considered.

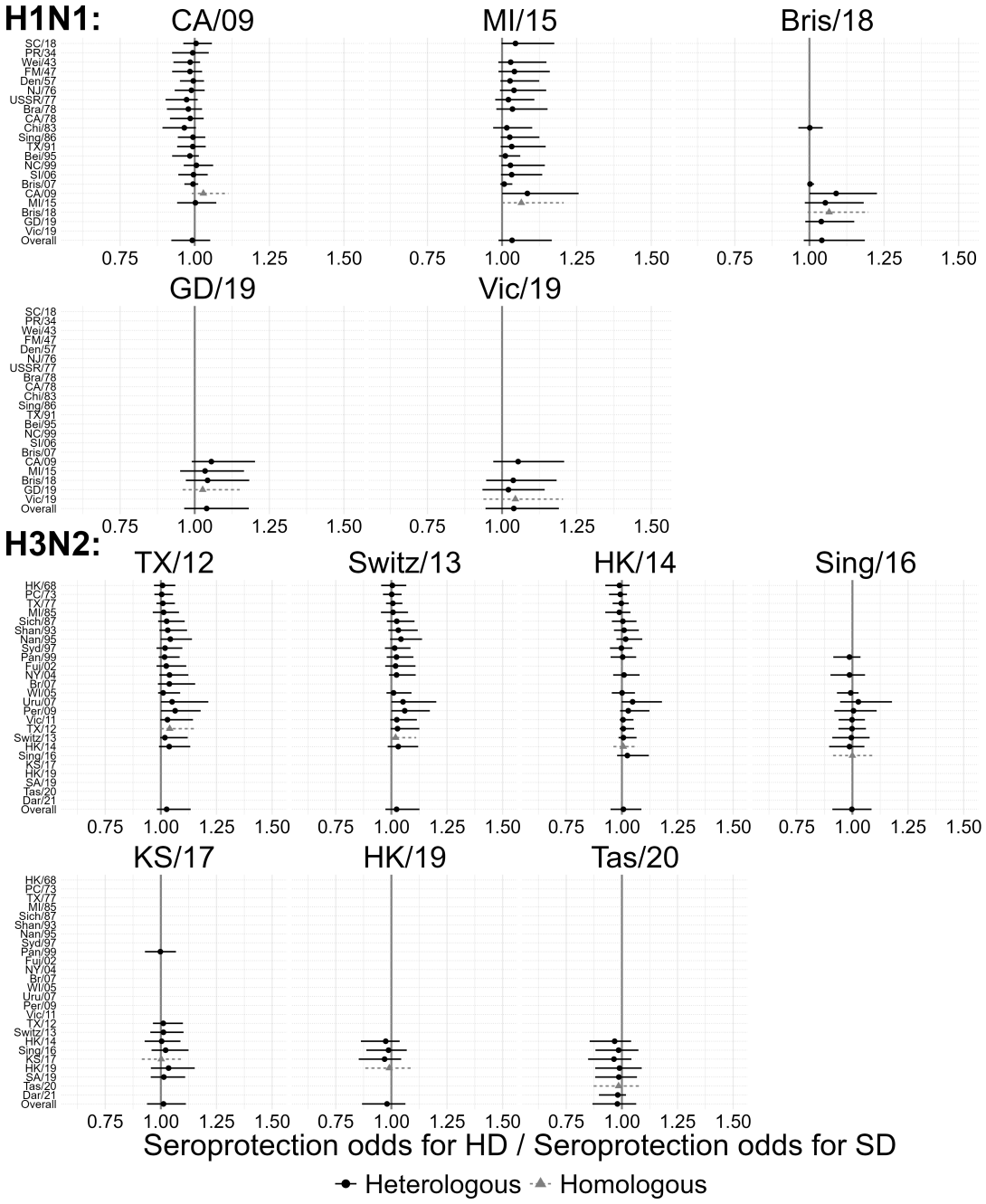

Figure A.40: Heterologous seroprotection cACE estimates (exponentiated) for each assay strain and vaccine.

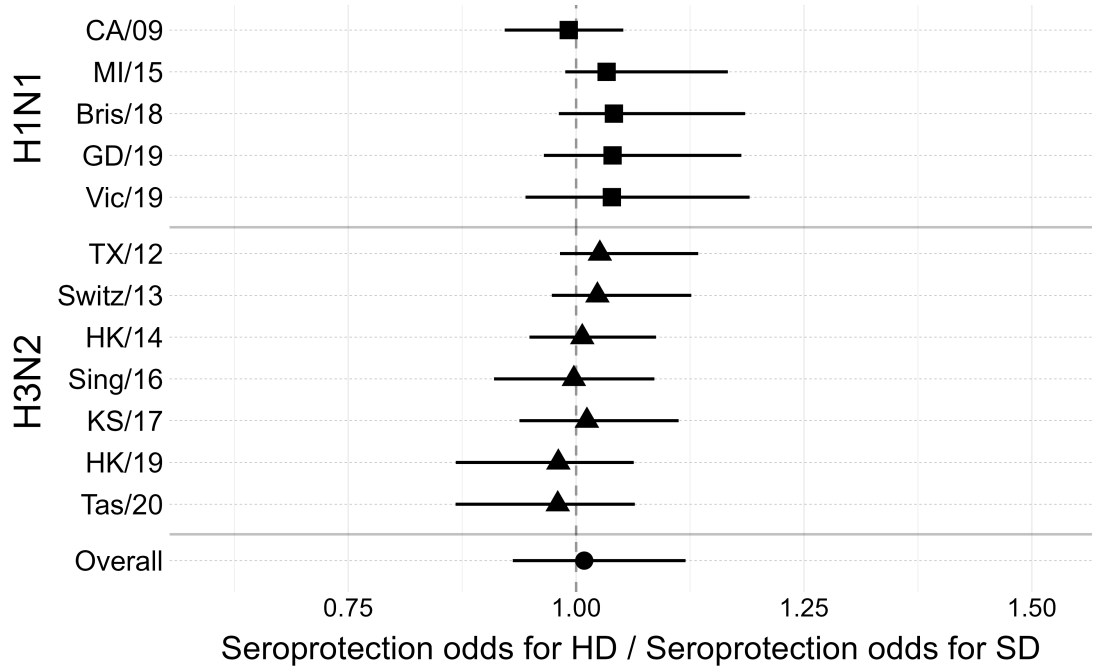

Figure A.41: Heterologous seroprotection cACE estimates (exponentiated) for each vaccine.

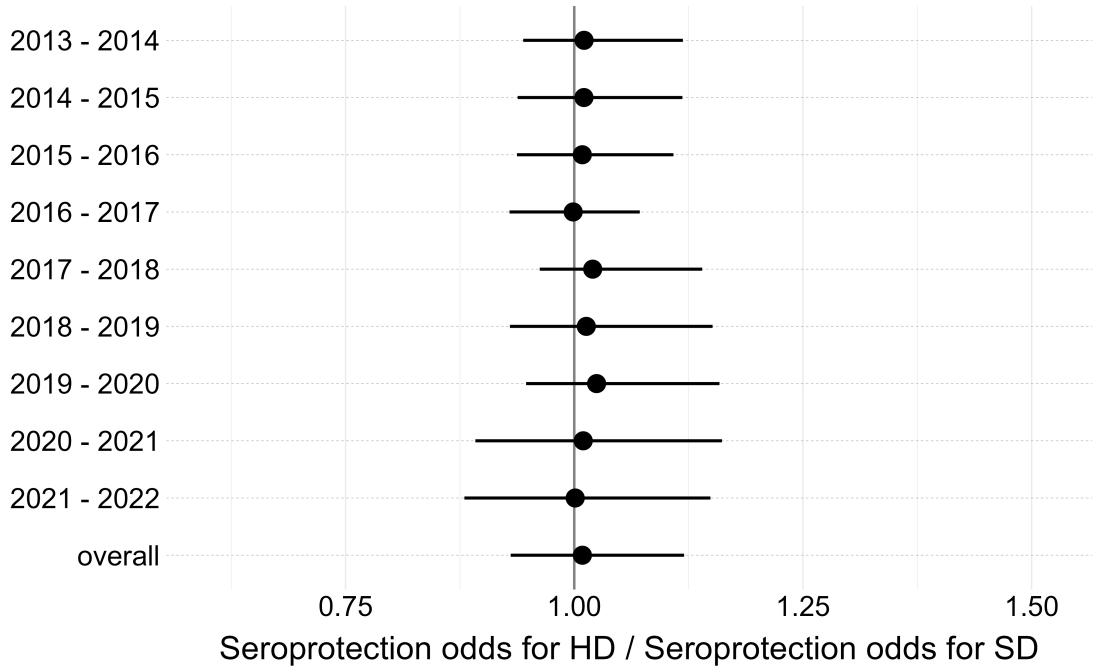

Figure A.42: Heterologous seroprotection cACE estimates (exponentiated) for each season.

Seroconversion

All of the figures in this section show the results for seroconversion as the model outcome. All of our results agreed with the results in the main text and there were no major qualitative differences.

Figure A.43 shows the cACEs for each vaccine strain when only the heterologous strains were included.

Figure A.44 shows the cACEs for all assay strains.

Figure A.45 shows the cACEs for all vaccine strains, pooling assay strains together within each vaccine strain.

Figure A.46 shows the cACEs for each season, with the vaccine strain and assay strains for that season all pooled together.

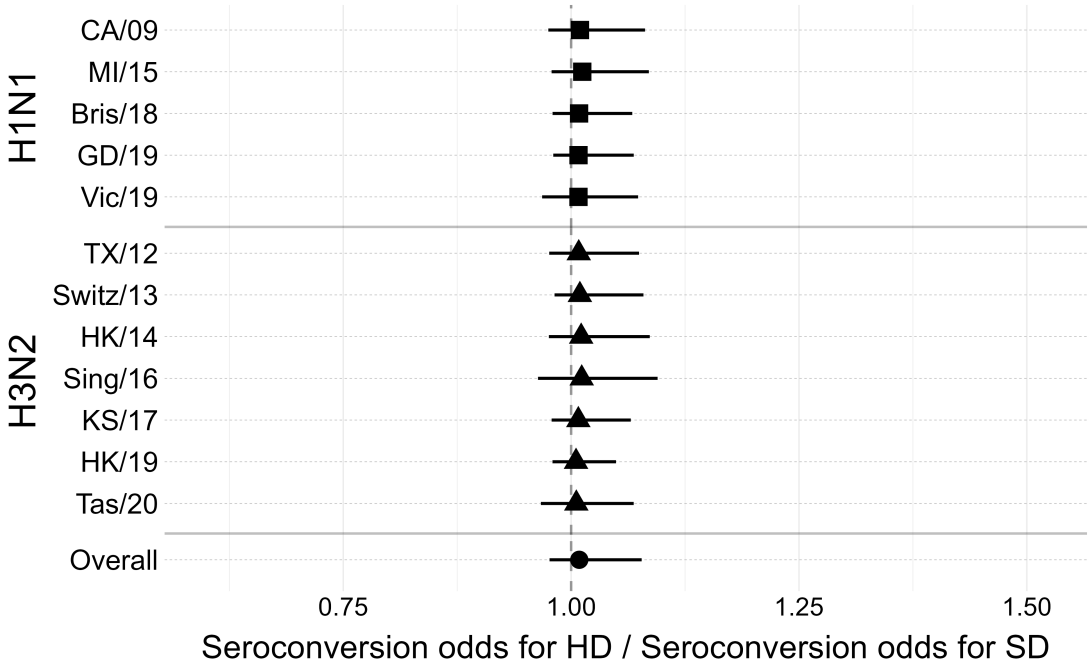

Figure A.43: Exponentiated cACE estimates from the seroconversion model for each vaccine strain and overall. Only homologous resconses to each vaccine were considered.

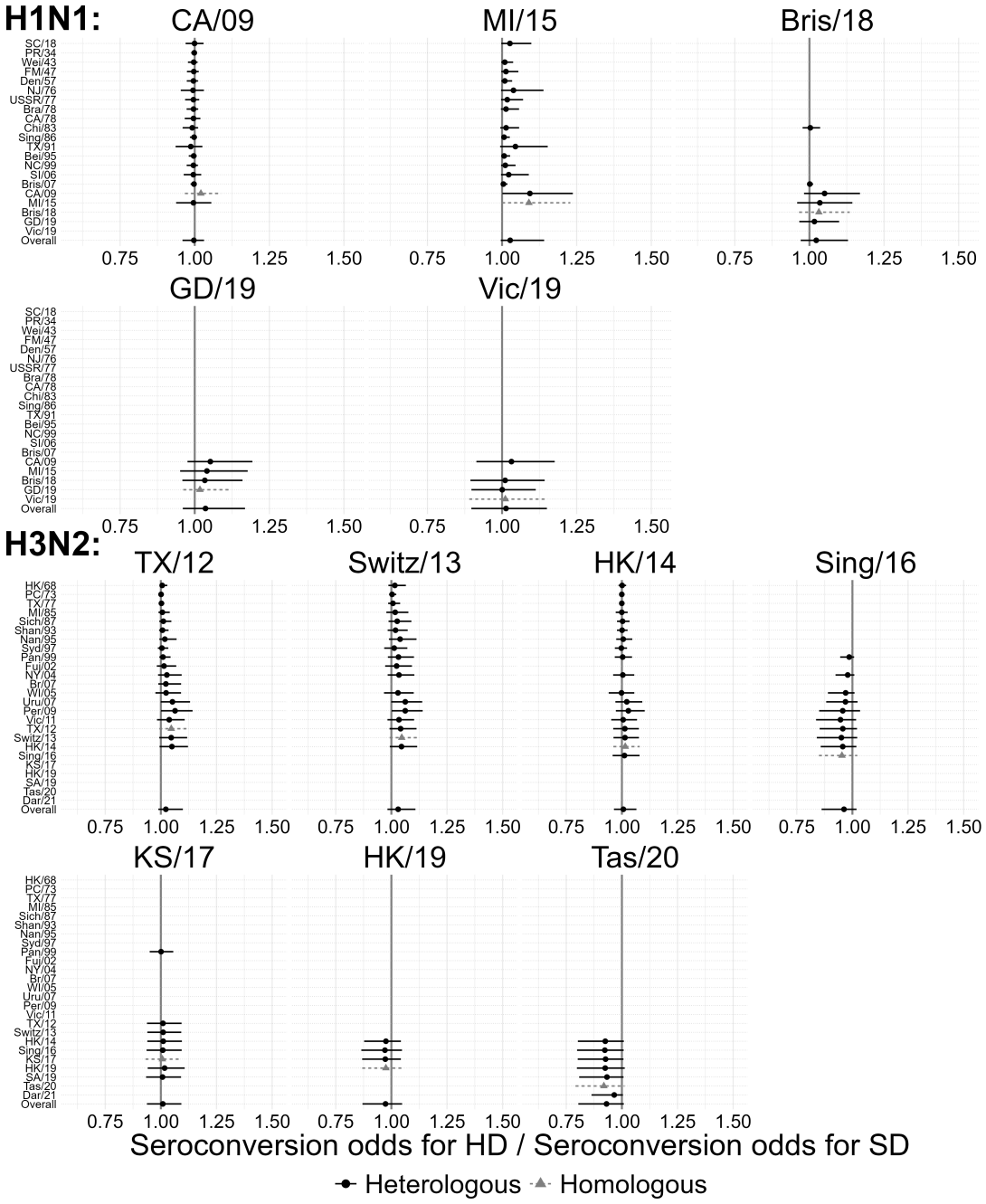

Figure A.44: Heterologous seroconversion cACE estimates (exponentiated) for each assay strain and vaccine.

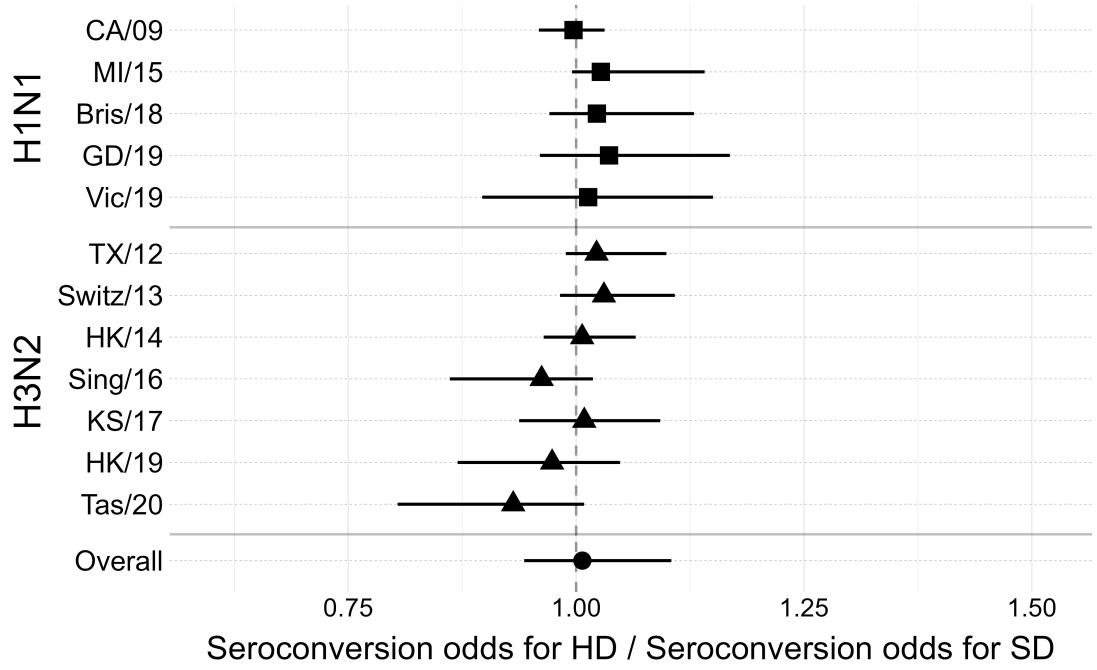

Figure A.45: Heterologous seroconversion cACE estimates (exponentiated) for each vaccine.

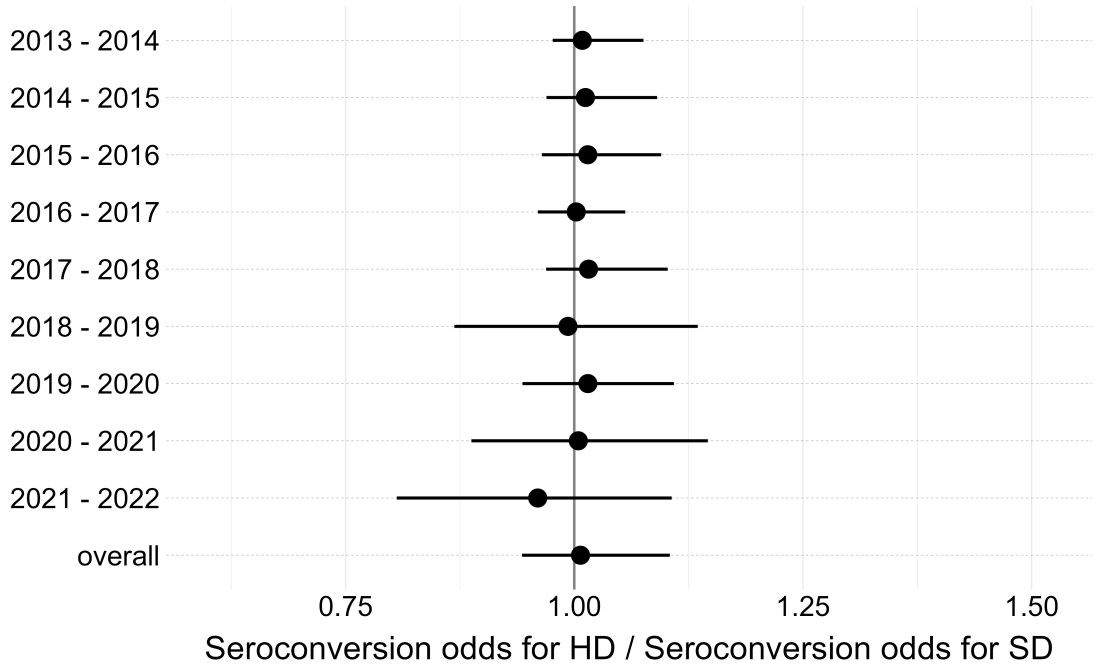

Figure A.46: Heterologous seroconversion cACE estimates (exponentiated) for each season.

Appendix B

Supplementary material for Chapter 3

Reproducibility Instructions

In order to reproduce our results, you should first download the archived repository from either Zenodo (DOI: 10.5281/zenodo.15522148) or clone/download the Git repository (hosted on GitHub here: `https://github.com/ahgroup/billings-comp-agdist-public`). Note that if you use different software, software versions, or run the results in a way that differs from these instructions, you may encounter errors or differences between your results and ours.

We ran our analysis pipeline on the University of Georgia's `sapelo2` computing cluster, which is a distributed computing cluster running CentOS Linux release 7.5 and uses Slurm to schedule jobs [160, 161]. Our code is written as a `targets` pipeline [158] and can detect whether you are running an HPC job in a Slurm environment or not. Notably, if you use a Slurm cluster computing environment which is configured differently from UGA's `sapelo2` environment, you may need to make changes to the Slurm submission script (`job.sh`) or to the portion of the `_targets.R` script that defines the Slurm jobs. We based our `targets` code to schedule HPC jobs on a pre-existing template project [159]. Our code will run on a local interactive R session as well and will automatically detect the number of cores available to use. **Each Bayesian model currently requests 32 cores and we therefore highly suggest running the main analysis on a cluster computing setup.**

Once you've downloaded the code, you should open the `.Rproj` file in RStudio. Using the R project file and RStudio is not mandatory, but if you don't, we assume you know what you are doing. You can then run the entire analysis pipeline by running `targets::tar_make()` in the console. If you are in an extremely limited computing environment, you can add the option `use_crew = FALSE` to force all targets to execute sequentially (although in this case you probably do not want to run the Bayesian models anyways). You can run the command `targets::tar_visnetwork(TRUE)` to see an interactive graph of our analysis pipeline, and you can pass a vector of target names to `tar_make()` (potentially using `tidyselect` to define the vector) to only run those targets. Note that due to differences in OS and file systems, targets will likely appear outdated for you even though they are up-to-date. We also do not provide all of our model result files in the GitHub repository because they are extremely large (over 100 GB) and infeasible to distribute, so if you want to edit or examine the Bayesian models, you will need to rerun the code.

Software Requirements

You will need the following software to run our code:

- R version 4.4.1, available from `https://cran.r-project.org`.
- A working C++ compiler:
  - On Linux clusters, this is probably already set up.
  - On Windows, you will need RTools 4.4, available from CRAN.
  - On macOS, you will need the latest version of the XCode command line tools.
- The RStudio IDE, available from `https://posit.co/download/rstudio-desktop`.
- Quarto version 16.40, available from `https://quarto.org`.
- The `renv` R package, version 1.1.4, available from `https://cran.r-project.org/web/packages/renv/index.html`. It will also attempt to install itself the first time you open our R project.
- Multiple system dependencies, including `CMake`. On Windows/macOS these are provided by RTools or XCode respectively. On Linux, there may be additional system requirements you need to download. Your system should prompt you.

- The packages specified in the file `renv.lock`, which can be installed as explained in the next section.

Running Our Code

With the software installed, follow these instructions to reproduce our results:

1. Open the `billings-comp-agdist-public.Rproj` file in RStudio.
2. Once `renv` initializes, run the command `renv::restore()` in the console to begin installing the required packages. If you have issues at this stage, you can also install the dependencies manually, but if you do not use `renv` or use different package versions than we did, our code might not work for you.
3. If you want to run any steps that involve Bayesian models, you need to install `cmdstan` following the `cmdstanr` quick start guide at this location: `https://mc-stan.org/cmdstanr/articles/cmdstanr.html`. We used cmdstan version 2.36.0 for this project.
4. Now you can run our pipeline by running `targets::tar_make()` in the console. If you are new to `targets` and want to learn more about how the pipeline works, we recommend reading the `targets` manual available at: `https://books.ropensci.org/targets/`.

Again, we note that our code is computationally intensive and we ran it on a distributed computing cluster. It still took multiple days to run, even running many operations in parallel with many cores each.

Extended Methods

Antigenic Distance Calculation

We calculated four different antigenic distance metrics for our study. In this section, we describe how each method is calculated. Note that we only considered pairwise distances between strains of the same subtype. That is, we only computed distances between two A(H1N1) strains, between two A(H3N2) strains, or between two influenza B strains; we did not compute distances between A(H1N1) and A(H3N2) strains or between any A and B strains. However, since the two B lineages are quite similar and our panel included pre-divergence influenza B strains, we performed pairwise comparisons of all influenza B strains. The methods we included in this section are abbreviated, for a full discussion on antigenic distance metric calculation see Appendix D.

Temporal distance is the absolute value of the difference in the years of isolation between the two strains. For example, the difference between `A/H1N1/California/09` and `A/H1N1/Michigan/15` would be $|2015 - 2009| = 6$.

Dominant p-Epitope distance is the maximum length-normalized Hamming distance across the five major epitope sites on the HA head. After aligning the HA amino acid sequences for all of the strains, we removed the signal peptides from the sequences and used the previously identified epitope site locations for influenza A [30] and influenza B [32]. Working pairwise with the sequences, we concatenated the residues for each epitope and calculated the Hamming distance between each epitope. We then divided the Hamming distance for a given epitope by the number of residues in that epitope. The p-Epitope distance for that pair of strains was the maximum of those epitope-wise distances.

Grantham's distance is a weighted distance based on biochemical properties that considers how different two differing residues at the same position are. We used Grantham's substitution matrix [151] to assign a value to each residue site between two sequences, based on the transition between amino acids. More different transitions are given higher weights. Then, for each pair of sequences, we summed the weights and divided by the length of the

sequence.

Finally, **cartographic distance** is the Euclidean distance between strains on an antigenic cartography map. We built our cartographic maps from the combined table of post-vaccination titer data in our study, treating all person-years as independent observations (as there is no clear meaning for repeat measurements in a dimension reduction analysis). We used `Racmacs`, which implements metric multidimensional scaling, to create and optimize the cartographic map [175]. All of our maps were two-dimensional, and we selected the best fitting map from 25 distinct `Racmacs` runs with random initializations, each of which was allowed to perform up to 100 L-BFGS optimization steps to refine the initial MDS cartography. Multiple optimization runs are necessary because different initial conditions can lead to different maps [125]. Although combining multiple runs using methods like generalized Procrustes analysis is theoretically possible (simple averaging is insufficient due to rotation and scaling), such approaches have not yet been published. Therefore, we selected the single best-fitting run.

For our models, we only considered the antigenic distance between the assay strain and the vaccine strain of the same subtype for a given HAI assay. Some of the assay strains used were influenza B strains isolated before the Victoria–Yamagata lineage divergence. Because our primary question was about antigenic distance, we compared pre-divergence B strains to both the Yamagata and Victoria vaccine strains in our analyses. To facilitate fair comparisons across subtypes and antigenic distance metrics, we min-max normalized the antigenic distance measurements within each combination of influenza season, subtype, and metric. After normalization, the antigenic distance for homologous measurements was set to 0, and the antigenic distance for the most different assay strain used in a given season was set to 1, with all other values falling in this interval.

Causal Modeling and Model Formulation

While we do not claim that our estimates are causal, we employed a graphical causal model to formulate our statistical models. While all statistical models are compromises between practicality and idealism, we hope that by formalizing our assumptions, our models are

robust and answer our research questions appropriately.

Our original dataset contained one record per HAI assay, indicating the individual, season, study site, time point (pre- or post-vaccination), vaccine dose, and assay strain for each record. The dataset also included the following demographic variables: age, birth year, sex assigned at birth, and reported race/ethnicity. The study provided a list of vaccine strains for each formulation (see the section on vaccine formulation for details). Note that we only analyzed standard-dose vaccine recipients, so we do not discuss vaccine dose further.

We built a causal model for the effect of antigenic distance as a directed acyclic graph (DAG). We include the following variables in our causal model: U, unobserved confounders that could be partially explained by nuisance variation but are not directly modeled; r, the self-reported race/ethnicity; s, the sex assigned at birth; p, the pre-vaccination titer; a, the individual's age; b, the individual's year of birth; sv, the vaccine strain (for a given subtype); and sa, the assay strain for a particular HAI assay. The causal model we selected is shown in Figure B.1.

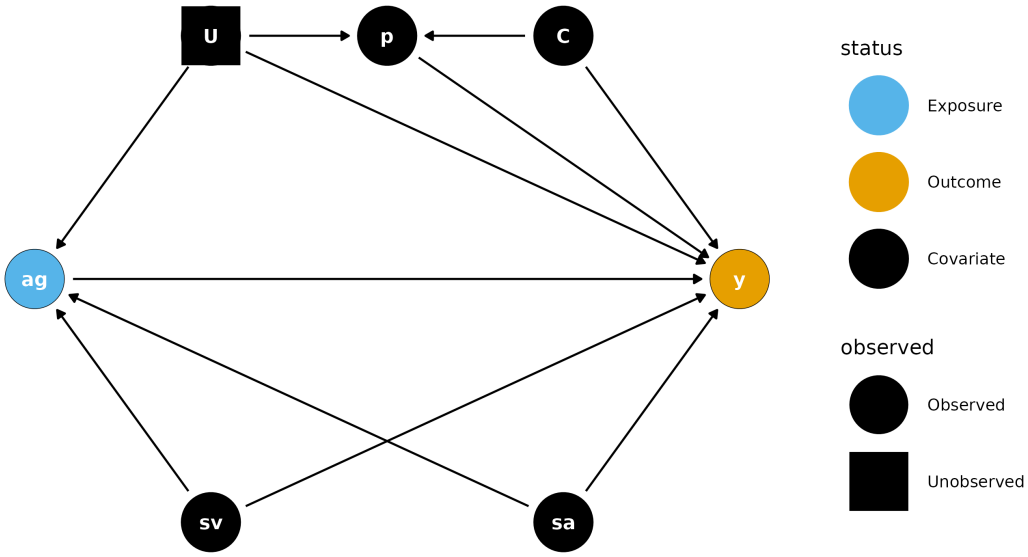

Figure B.1: The graphical causal model for our research question represented as a DAG.

Under this causal model, the only confounders are the vaccine strain and assay strain, and any unobserved confounders. If we assume no unmeasured confounding, then the

minimal sufficient adjustment set is only the vaccine strain and assay strain. However, our goal in this project was to analyze the effect of antigenic distance as a predictor without incorporating strain-specific effects. So, we stratified our models by vaccine strain (i.e., we fit all models separately for each vaccine component) and deliberately did not include a strain-specific effect.

To adjust for nuisance variation (potentially a source of unmeasured confounding), we included random effects to control for measurements at the same study site and on the same individual. Finally, we included specific ancestors of the outcome variable, which is not necessary to close backdoor paths, but does not improve bias and can improve the efficiency of the estimators of interest. We included pre-vaccination titer and age specifically in our model. In our previous work, we found that sex and race/ethnicity have minimal association with the observed HAI titers. Finally, our study was not specifically designed as an age-period-cohort analysis, and the age and birth year variables are highly correlated in our study (as one would expect). While our study has multiple longitudinal participants with different birth years, we elected to add only age to our model for simplicity. We included pre-vaccination titer in the model as-is, but since the age has a large range (from 11 to 65), we minmax scaled the age before using it in a model. Minmax scaling variables with large ranges can improve numerical stability of the model, but the model can still make predictions for any age.

Finally, we note that in some models it is also possible for cross-season differences to exist when the same vaccine strain was used for multiple years in a row. I.e., we might expect post-vaccination titers to change due to repeated usage of the same vaccine. However, since some of the vaccine strains were only used for one year before being replaced, this seasonal effect is not estimable in all of our models. Therefore, we decided not to include a seasonal effect in any of the models, especially since the effect of repeated usage of the same vaccine strain was not our primary research question.

Model implementation

We fit two models using 'brms', a generalized additive mixed model (GAMM) and a linear mixed model (LMM). The models were identical other than the specification for the effect of antigenic distance, so we will first describe the general parts of the model. Note that in the following mathematical descriptions, we adopt bracket notation rather than subscript notation following the convention of McElreath [68] due to the large number of subscripts in our model. That is, we use the notation $y[i]$ in place of the conventional y_i. We use subscripts to instead identify unique parameters. We also used the centered dot symbol ($\cdot$) to avoid repetition when there are many valid arguments that would have the same right-hand side in a formula. For example, $\zeta[\cdot]$ indicates that all subscripts for ζ use the same equation.

We modeled our outcome (post-vaccination titer) as a Gaussian random variable, but due to the censored nature of our data we applied a censoring correction in the likelihood. Letting the outcome for a specific vaccine component be y, we assumed that

$$f\left(y_i \mid \mu[i], \sigma^2\right) = \int_{L[i]}^{U[i]} \mathcal{N}\left(y[i] \mid \mu[i], \sigma^2\right) \, \mathrm{d}y[i]$$

$$\sigma \sim t^{+}\left(3, 0, 1\right)$$

$$i = 1, \ldots, n$$

where $L[i]$ and $U[i]$ are the lower and upper censoring bounds respectively (see the section on censoring bounds for details), $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian (Normal) probability density function with mean μ and variance σ^2, $t^{+}(\nu, \mu, \sigma)$ is the location-scale half Student's t distribution with degrees of freedom ν, location parameter μ, and scale parameter σ. We chose a Student's t prior with $\nu = 3$ degrees of freedom because the distribution has fat tails, which allows the variance to be large if supported by the data, but we assume *a priori* that the distribution of the variance has a finite location and scale parameter (which is only the case when $\nu > 2$). Here, i is the index for the current data record, representing one HAI assay, and n is the total number of HAI assays (records) in the dataset.

The model for the mean is shown below, including the priors for each parameter. For now, we represent the effect of antigenic distance as a function g, which we detail with its

priors in the next section.

$$
\begin{aligned}
\mu[i] = \beta_0 &+ u[1, \text{id}[i]] + u[2, \text{study}[i]] + u[3, \text{subtype}[i]] + \\
&u[4, \text{subtype}[i] \times \text{vaccine strain}[i]] + u[5, \text{subtype}[i] \times \text{assay strain}[i]] + \\
&g(\text{antigenic distance}[i]) + \\
&\beta_p(\text{log pre-vaccination titer[i]}) + \beta_a(\text{scaled age}[i]) + \\
&\beta_y(\text{scaled birth year}[i]) + \beta_r(\text{race/ethnicity[i]}) + \beta_s(\text{sex}[i]) \\
\beta_{(\cdot)} &\sim \mathcal{N}(0, 1) \\
u[r, \cdot] &\sim \mathcal{N}(0, \omega[r]) \quad r = 1, 2, \ldots, 5 \\
\omega[r] &\sim t^+(3, 0, 1)
\end{aligned}
$$

The priors follow the same formulation as before, but we chose Gaussian priors for the beta effects. Gaussian priors have flatter tails than Student's t priors, which provides a more regularizing effect for the beta parameters – that is, we presuppose that they are more likely to be close to zero, and our data needs to be strong enough to move the posterior distributions away from zero before we can make any conclusions.

The functional form of g is the only difference between the GAMM and the LMM. In the LMM, g takes a simple linear form:

$$
\begin{aligned}
g(\text{antigenic distance}[i]) &= \beta_d \, (\text{antigenic distance}[i]) \\
\beta_d &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

where the antigenic distance is minmax normalized for each model as described in the antigenic distance calculation section. For the GAMM, the function form of g is more complex. We modeled the antigenic distance effect using a thin-plate basis spline, which allows for the relationship to be curved in an arbitrary pattern, but constrains the fit so that rapid changes in the pattern are penalized and must be supported by data [201–204,

220] . The specific form is

$$g(\cdot) = \sum_{k=1}^{5} \gamma[k] \cdot \phi[k](\cdot)$$

$$\gamma[k] \sim \mathcal{N}(0, \tau)$$

$$\tau \sim t^{+}(3, 0, 0.25)$$

where the $\gamma[k]$ are coefficients which are regularized to be similar via an adaptive prior and the $\phi[k]$ are thin-plate spline basis functions. Thin-plate splines use a low-rank approximation of the spline basis for computational efficiency, which can be tuned to balance between accuracy and efficiency. The maximal k (or size of the spline basis) we can choose is equal to the number of unique values of the predictor, so we chose $k = 5$, which was estimable across all of our antigenic distance metrics. We used Student's t priors for the adaptive prior on the variance of the spline coefficients so that the spline can be wiggly if supported by the data, but we chose a conservative hyperprior variance (0.25, based on a prior predictive simulation) to constrain the spline towards being flat if the signal from the data is not strong.

The random effects we included in the model represented sources of nuisance variation which we were interested in controlling for, but not specifically estimating. We included random effects to capture interindividual variation, variation across study sites, and direct effects of the influenza strains not explained by the antigenic distance. We included random intercepts for individuals ($u[1, \cdot]$) and study sites ($u[2, \cdot]$) in a typical way with regularizing priors. To control for the direct effects of influenza assay and vaccine strains, we noted that each strain was nested within a subtype, but the assay strains and vaccine strains were not themselves crossed or nested (each assay strain could appear with an arbitrary combination of different vaccine strains, although all assay strains and all vaccine strains are only ever associated with a single influenza subtype). Including the subtype effect as $u[3, \cdot]$ and then including random effects which consider both the subtype and the vaccine strain ($u[4, \cdot]$) or the assay strain ($u[5, \cdot]$) allows for assay/vaccine strains within the same subtype to have a correlated effect, while assay/vaccine strains of different subtypes do not have correlated effects. Again, we assigne skeptical, regularizing priors to all of these random effects.

Censoring bounds

HAI titer assays, like all serial dilution assays, produce censored data values. In fact, all values produced by an HAI assay are censored. We take this censoring into account in the likelihood of our model by integrating over the censoring bounds for a given data point y_i.

All serial dilution assays are censored – for the case of HAI, we assume that there is some latent, true dilution y_i^* which is the minimal dilution where hemagglutination is not observed. This is likely some decimal number, and we will never observe this true value. Instead, we chose a starting dilution, $y_{\min}$, which is 10 in our dataset. If we observe agglutination at this starting dilution, we say the value is below the limit of detection and it is recorded as 5 in our dataset. These values are left censored. In reality, we know that the latent agglutination dilution for an assay can be any value less than 10, i.e., our censoring bounds for these assays are $(0, 10)$.

There is also a maximal dilution for the assay, $y_{\max}$, which was 20480 in our dataset. In practice, if researchers don't observe hemagglutination at any dilution, they can simply continue diluting the assay until they observe agglutination. However, a standard 96-well plate only has 12 columns, so most studies will report 20480 (the 12th serial dilution for an HAI assay starting at 10 and doubling each dilution). So these values are right censored, and the censoring bounds are $[20480, \infty)$. Note that the lower bound of the interval is included because the value *could* be exactly 20480 (though this occurs with probability zero for a continuous latent variable).

Finally, any other assay with a result between the limits of detection will also be interval censored, because we only observe certain dilutions. For example, if we observe inhibited hemagglutination at a dilution of 10, but agglutination occurs at a dilution of 20, we record the result as 10. However, we don't know that a dilution of 1:15 wouldn't cause inhibition, so we only know that the latent dilution is in the interval $[10, 20)$. Similarly for any value $y_{\min} < y < y_{\max}$, the latent dilution is in the interval $[y, 2y)$.

Converting to the log scale, the censoring bounds L and U that we refer to in the previous equations are as follows (here we omit subscripts to avoid confusion with interval notation,

but L, U, and y all vary by individual while $y_{\min}$ and $y_{\max}$ are constant):

$$(L, U) = \begin{cases} (-\infty, y_{\min}) & y = y_{\min} \\ [y, y+1), & y_{\min} < y < y_{\max} \\ [y_{\max}, \infty) & y = y_{\max} \end{cases}.$$

For our study, $y_{\min} = \log_2(10/5) = 1$ and $y_{\max} = \log_2(20480/5) = 12$.

Stan Implementation

We obtained posterior samples of the model parameters using the No-U-Turn Sampler (NUTS) algorithm implemented by Stan [71, 92], via the `brms` [87–89] and `cmdstanr` [90] packages for `R` [72]. In `brms`, we specified our model formulas as:

```
y | cens(c, y2) ~ 1 +
  birth_year_c + age_c + sex_i + race_i +
  log_pretiter + s(d_norm, k = 5, by = strain_type) +
  (1 | strain_type) +
  (1 | study) + (1 | subject_id) +
  (1 | strain_type:vaccine_name) + (1 | strain_type:strain_name)
```

for the GAMMs, and

```
y | cens(c, y2) ~ 1 +
  birth_year_c + age_c + sex_i + race_i +
  log_pretiter + d_norm + (1 + d_norm | strain_type) +
  (1 | study) + (1 | subject_id) +
  (1 | strain_type:vaccine_name) + (1 | strain_type:strain_name)
```

for the LMMs. We specified our prior distributions as:

```
brms::prior(normal(0,1), class = "Intercept"),
brms::prior(normal(0,1), class = "b"),
brms::prior(student_t(3, 0, 1), class = "sd", lb = 0),
brms::prior(student_t(3, 0, 1), class = "sigma", lb = 0),
brms::prior(student_t(3, 0, 0.25), class = "sds", lb = 0)
```

for the GAMMs, and

```
brms::prior(normal(0,1), class = "Intercept"),
brms::prior(normal(0,1), class = "b"),
brms::prior(student_t(3, 0, 1), class = "sd", lb = 0),
brms::prior(student_t(3, 0, 1), class = "sigma", lb = 0)
```

for the LMMs. Note that the LMMs do not include a prior for parameters of class `sds`, which represent the adaptive smoothing priors for spline coefficients.

We sampled the models using 32 chains with 200 warmup iterations and 625 post-warmup iterations per chain, resulting in 20,000 total post-warmup samples per parameter. The effective number of samples is reported in the model diagnostics table in a later section. We also specified an `adapt_delta` of 0.99 and a maximum `treedepth` of 12, using recommended Stan and `brms` defaults for all other algorithmic parameters. Notably, each chain was initialized with a random value—although using smarter initial values could potentially speed up sampling, we found this did not yield a meaningful benefit for our models. The primary cause of slower sampling was the large dataset size, although the hierarchical nature of our model also contributed.

Posterior Marginal Effects

To summarize our models, Figure 2 presents a posterior contrast we refer to as a marginal effect, though it is technically a marginal conditional effect that marginalizes over some variables while conditioning on others. The posterior effect of interest is the effect of antigenic distance on post-vaccination titer, conditional on the subtype, for each antigenic distance metric (each metric was modeled separately). To compute this effect, we constructed counterfactual predictions over an interpolated grid of antigenic distance values using the

`marginaleffects` package [154].

The effects reported in Figure 3.2 represent marginal effects at the mean (MEM), conditional on the subtype. All four models (corresponding to each antigenic distance metric) shared the same set of effects, except for the substitution of the respective distance variable. The marginal effects are conditional on the random effects structure, but the credible intervals do not incorporate random effect variances, except for the subtype.

We generated predictions on an interpolated grid of normalized antigenic distance values ranging from 0 to 1 in increments of 0.01, excluded random effects from the prediction, set categorical fixed effects to their mode, and continuous fixed effects (other than antigenic distance) to their mean. In this way, the marginal effects approximate the expected post-vaccination titer for a typical individual similar to those in our sample.

While using average marginal effects (AMEs) and integrating out random effects might allow more generalizable inference, these approaches significantly increase computational burden without substantially altering the predicted outcomes. Given the already wide credible intervals—appropriate for a non-causal observational analysis—we did not find it necessary to include additional nuisance variation in the predictions. The main goal of the analysis was to compare predictions across the four antigenic distance models, rather than isolate a causal effect of antigenic distance.

Specifically, we obtained predictions on a grid using the following `marginaleffects` package syntax (`model_i` refers to each of the individual models).

```
marginaleffects::datagrid(
  model = model_i,
  d_norm = seq(0, 1, 0.01),
  strain_type = c("H1N1", "H3N2", "B-Vic", "B-Yam")
)
```

Vaccine and heterologous strain information

Annual Fluzone vaccine formulation

Table B.1 shows the strains which were included in each season's formulation of the Fluzone vaccine. We only show the formulation for the standard dose (SD) vaccine (the HD vaccine was trivalent throughout the study years we selected, while the quadrivalent formulation of the SD vaccine became available in 2015/16).

Table B.1: Strains used in the Fluzone standard dose vaccine formulation during each influenza season for the seasons of data that comprise the study sample in Chapter 3. In 2013/14 and 2014/15, the vaccine formulation was trivalent and did not contain a B/Victoria component.

	Vaccine component			
Season	*A(H1N1)*	*A(H3N2)*	*B(Victoria)*	*B(Yamagata)*
2013/14	CA/09	TX/12	—	MA/12
2014/15	CA/09	TX/12	—	MA/12
2015/16	CA/09	Switz/13	Bris/08	Phu/13
2016/17	CA/09	HK/14	Bris/08	Phu/13
2017/18	MI/15	HK/14	Bris/08	Phu/13

Annual heterologous strain panel

The strains used in each panel are shown in Table B.2 (influenza A strains) and Table B.3. An X indicates that the strain indicated by the current row was used as part of the HAI panel in the season indicated by the current column.

Strain name abbreviations

Throughout the manuscript, we use abbreviated names for each strain. Table B.4 shows the corresponding abbreviation for the full name of each strain for influenza A strains and Table B.5 shows the same information for the influenza B strains.

Table B.2: Influenza A strains used in the heterologous panel during each influenza season for the seasons of data that comprise the study sample in Chapter 3.

		Season				
Subtype	**Strain**	*13/14*	*14/15*	*15/16*	*16/17*	*17/18*
A(H1N1)	SC/18	X	X	X	X	X
	PR/34	X				
	Wei/43	X	X	X	X	X
	FM/47	X	X	X	X	X
	Den/57	X	X	X	X	X
	NJ/76	X	X	X	X	X
	USSR/77	X	X	X	X	X
	Bra/78	X			X	X
	CA/78		X	X		
	Chi/83	X	X	X	X	X
	Sing/86	X	X	X	X	X
	TX/91	X	X	X	X	X
	Bei/95	X	X	X	X	X
	NC/99	X	X	X	X	X
	SI/06	X	X	X	X	X
	Bris/07	X	X	X	X	X
	CA/09	X	X	X	X	X
	MI/15				X	X
A(H3N2)	HK/68	X	X	X	X	X
	PC/73	X	X	X	X	X
	TX/77	X	X	X	X	X
	MI/85	X	X	X	X	X
	Sich/87	X	X	X	X	X
	Shan/93	X	X	X	X	X
	Nan/95	X	X	X	X	X
	Syd/97	X	X	X	X	X
	Pan/99	X	X	X	X	X
	Fuj/02	X	X	X		
	NY/04	X	X	X	X	X
	Br/07	X				
	WI/05	X	X	X	X	X
	Uru/07		X	X	X	X
	Per/09	X	X	X	X	X
	Vic/11	X	X	X	X	X
	TX/12	X	X	X	X	X
	Switz/13	X	X	X	X	X
	HK/14		X	X	X	X
	Sing/16					X

Table B.3: Full strain names and associated abbreviations for each influenza Astrain used in the Chapter 3 study sample.

		Season				
Lineage	**Strain**	*13/14*	*14/15*	*15/16*	*16/17*	*17/18*
B(Presplit)	Lee/40	X	X	X	X	
	MD/59		X	X	X	
	Sing/64		X	X	X	
B(Victoria)	Vic/87				X	X
	HK/01			X	X	X
	Mal/04			X	X	X
	Vic/06			X	X	X
	Bris/08			X	X	X
	CO/17			X	X	X
B(Yamagata)	Yam/88	X	X	X	X	X
	Harb/94	X	X	X	X	X
	Sich/99	X	X	X	X	X
	FL/06	X	X	X	X	X
	WI/10	X	X	X	X	X
	TX/11	X	X	X	X	X
	MA/12	X	X	X	X	X
	Phu/13	X	X	X	X	X

Table B.4: Influenza A strains used in the heterologous panel during each influenza season for the seasons of data that comprise the study sample in Chapter 3.

Subtype	**Strain name**	**Short name**
A(H1N1)	A/H1N1/South Carolina/1/1918	SC/18
	A/H1N1/Puerto Rico/8/1934	PR/34
	A/H1N1/Weiss/1943	Wei/43
	A/H1N1/Fort Monmouth/1/1947	FM/47
	A/H1N1/Denver/1957	Den/57
	A/H1N1/New Jersey/8/1976	NJ/76
	A/H1N1/Ussr/90/1977	USSR/77
	A/H1N1/Brazil/11/1978	Bra/78
	A/H1N1/California/10/1978	CA/78
	A/H1N1/Chile/1/1983	Chi/83
	A/H1N1/Singapore/6/1986	Sing/86
	A/H1N1/Texas/36/1991	TX/91
	A/H1N1/Beijing/262/1995	Bei/95
	A/H1N1/New Caledonia/20/1999	NC/99
	A/H1N1/Solomon Islands/3/2006	SI/06
	A/H1N1/Brisbane/59/2007	Bris/07
	A/H1N1/California/07/2009	CA/09
	A/H1N1/Michigan 45/2015	MI/15
A(H3N2)	A/H3N2/Hong Kong/8/1968	HK/68
	A/H3N2/Port Chalmers/1/1973	PC/73
	A/H3N2/Texas/1/1977	TX/77
	A/H3N2/Mississippi/1/1985	MI/85
	A/H3N2/Sichuan/2/1987	Sich/87
	A/H3N2/Shandong/9/1993	Shan/93
	A/H3N2/Nanchang/933/1995	Nan/95
	A/H3N2/Sydney/5/1997	Syd/97
	A/H3N2/Panama/2007/1999	Pan/99
	A/H3N2/Fujian/411/2002	Fuj/02
	A/H3N2/New York/55/2004	NY/04
	A/H3N2/Brisbane/10/2007	Br/07
	A/H3N2/Wisconsin/67/2005	WI/05
	A/H3N2/Uruguay/716/2007	Uru/07
	A/H3N2/Perth/16/2009	Per/09
	A/H3N2/Victoria/361/2011	Vic/11
	A/H3N2/Texas/50/2012	TX/12
	A/H3N2/Switzerland/9715293/2013	Switz/13
	A/H3N2/Hong Kong/4801/2014	HK/14
	A/H3N2/Singapore/infimh-16-0019/2016	Sing/16

Table B.5: Influenza B strains used in the heterologous panel during each influenza season for the seasons of data that comprise the study sample in Chapter 3.

Lineage	**Strain name**	**Short name**
B(Presplit)	B/Lee/1940	Lee/40
	B/Maryland/1959	MD/59
	B/Singapore/1964	Sing/64
B(Victoria)	B/Victoria/02/1987	Vic/87
	B/Hong Kong/330/2001	HK/01
	B/Malaysia/27127/2004	Mal/04
	B/Victoria/326/2006	Vic/06
	B/Brisbane/60/2008	Bris/08
	B/Colorado/06/2017	CO/17
B(Yamagata)	B/Yamagata/16/1988	Yam/88
	B/Harbin/7/1994	Harb/94
	B/Sichuan/379/1999	Sich/99
	B/Florida/4/2006	FL/06
	B/Wisconsin/01/2010	WI/10
	B/Texas/06/2011	TX/11
	B/Massachusetts/02/2012	MA/12
	B/Phuket/3073/2013	Phu/13

Descriptive analyses

A summary of the demographic information for the individuals included in our analysis is shown in Table B.6, and includes information about their reported race/ethnicity, sex assigned at birth, age at first enrollment, and year of birth. The majority of participants identified their race as White or Caucasian, and were assigned female at birth. All participants from the PA and FL study sites were adults, but the UGA study site also recruited teenagers, and all three study sites included elderly people over 65 years of age. Most participants returned to the study site in at least one subsequent year, contributing more than one person-year of data to the study.

Figure B.2 shows a visualization of the collected pre-vaccination titers, and Figure B.3 shows a visualization of the collected post-vaccination titers, ignoring all variables except for the assay strain.

Qualitatively summarizing the distribution of titers to all of the assay strains from plots alone is difficult, and the models in the main text are very helpful for understanding the variation in post-vaccination titers. However, we can make a few observations. Most people had some prior immunity (Figure B.2) to the A(H3N2) strains which have circulated since the 80's or 90's, with protective (40 or greater) titers to the strains from the 2000's and onward. However, most people only had protective titers to the two most recent A(H1N1) strains, CA/09 and MI/15 which represent the 2009 pandemic lineage. Some people had immunity to older strains, but the difference was much more stark than for H3N2. Many people had prior immunity to all of the B strains we examined, and the median was 40 or greater for all of the B strains except MD/59.

Post-titers were, in general, higher (Figure B.3). The two pandemic-like H1N1 strains showed a boost on average in the population, and there was noticeable back-boosting to some of the older H1N1 strains. Many of the H3N2 strains showed backboosting as well, although there was not much of a response to the oldest H3N2 strains which also had low pretiters. The median post-titers were above 40 for all of the B strains in our data, with Yamagata having the highest average titers, followed by Victoria and then the older lineages.

Table B.6: Demographic characteristics of the study participants. Summary statistics shown are count and column percent for sex, race, and contributed person-years; and median with range for age at first enrollment, birth year, and contributed HAI assays. Demographic variables were collected by a questionnaire from participants on the date they enrolled in a study season and received a vaccine.

Characteristic	**FL**, N = 2411	**PA**, N = 1331	**UGA**, N = 3031	**Overall**, N = 6771
Sex Assigned at Birth				
Female	184 (76%)	93 (70%)	168 (55%)	445 (66%)
Male	57 (24%)	40 (30%)	135 (45%)	232 (34%)
Race/Ethnicity				
White	190 (79%)	70 (53%)	233 (77%)	493 (73%)
Black or American American	14 (6%)	52 (39%)	24 (8%)	90 (13%)
Other	12 (5%)	8 (6%)	33 (11%)	53 (8%)
Hispanic or Latino	24 (10%)	3 (2%)	13 (4%)	40 (6%)
Unknown	1 (0%)	0 (0%)	0 (0%)	1 (0%)
Age at First Enrollment	42 (20 - 80)	60 (26 - 81)	25 (12 - 83)	40 (12 - 83)
Year of Birth	1972 (1933 - 1996)	1954 (1932 - 1987)	1991 (1934 - 2006)	1975 (1932 - 2006)
Contributed HAI assays	85 (40 - 189)	94 (8 - 185)	48 (47 - 95)	52 (8 - 189)
Contributed person-years				
1	114 (47%)	44 (33%)	206 (68%)	364 (54%)
2	52 (22%)	31 (23%)	97 (32%)	180 (27%)
3	61 (25%)	32 (24%)	0 (0%)	93 (14%)
4	14 (6%)	26 (20%)	0 (0%)	40 (6%)

1n (%); Median (Min - Max)

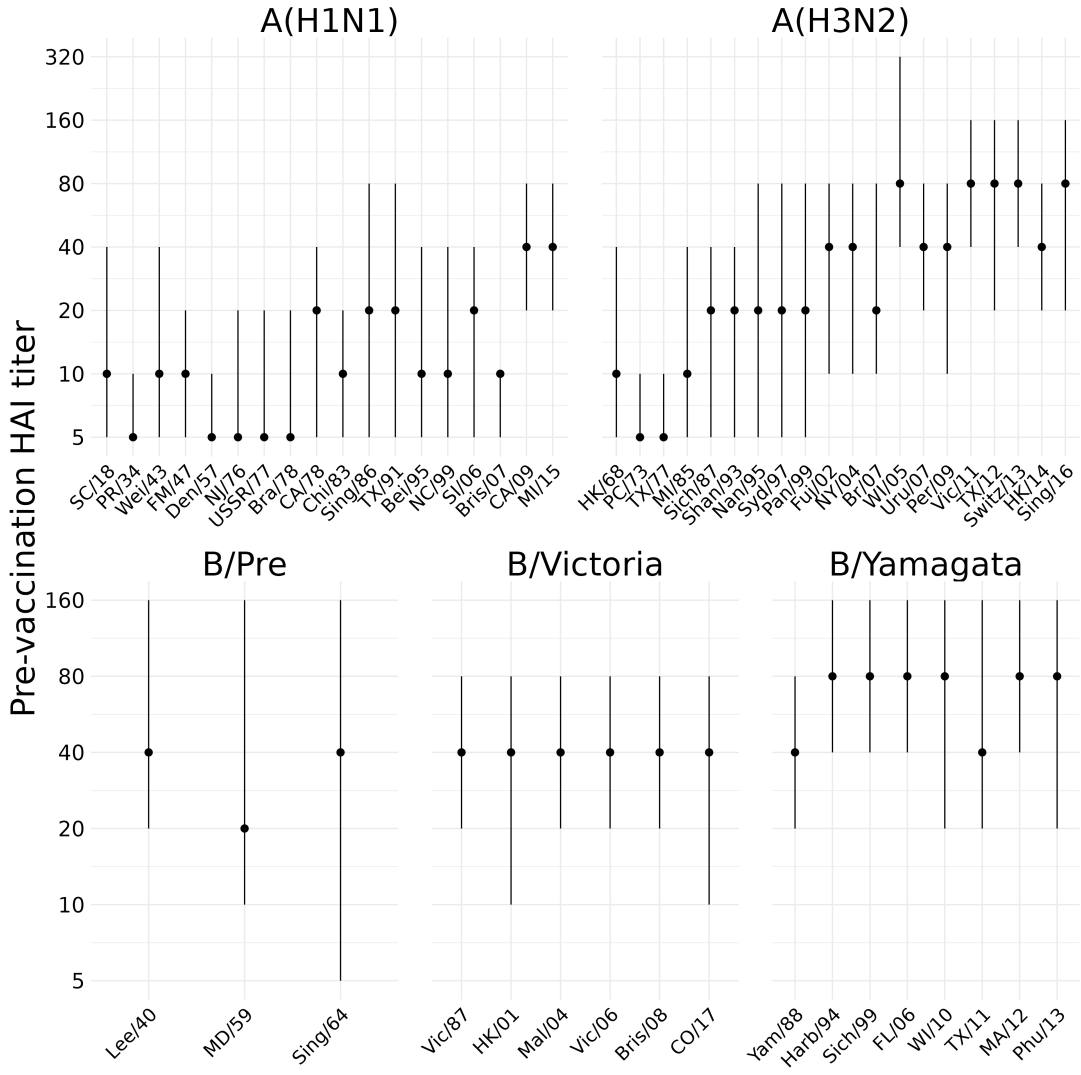

Figure B.2: Pre-vaccination titers in our study to each of the assay strains. The point shows the median and the line shows the IQR.

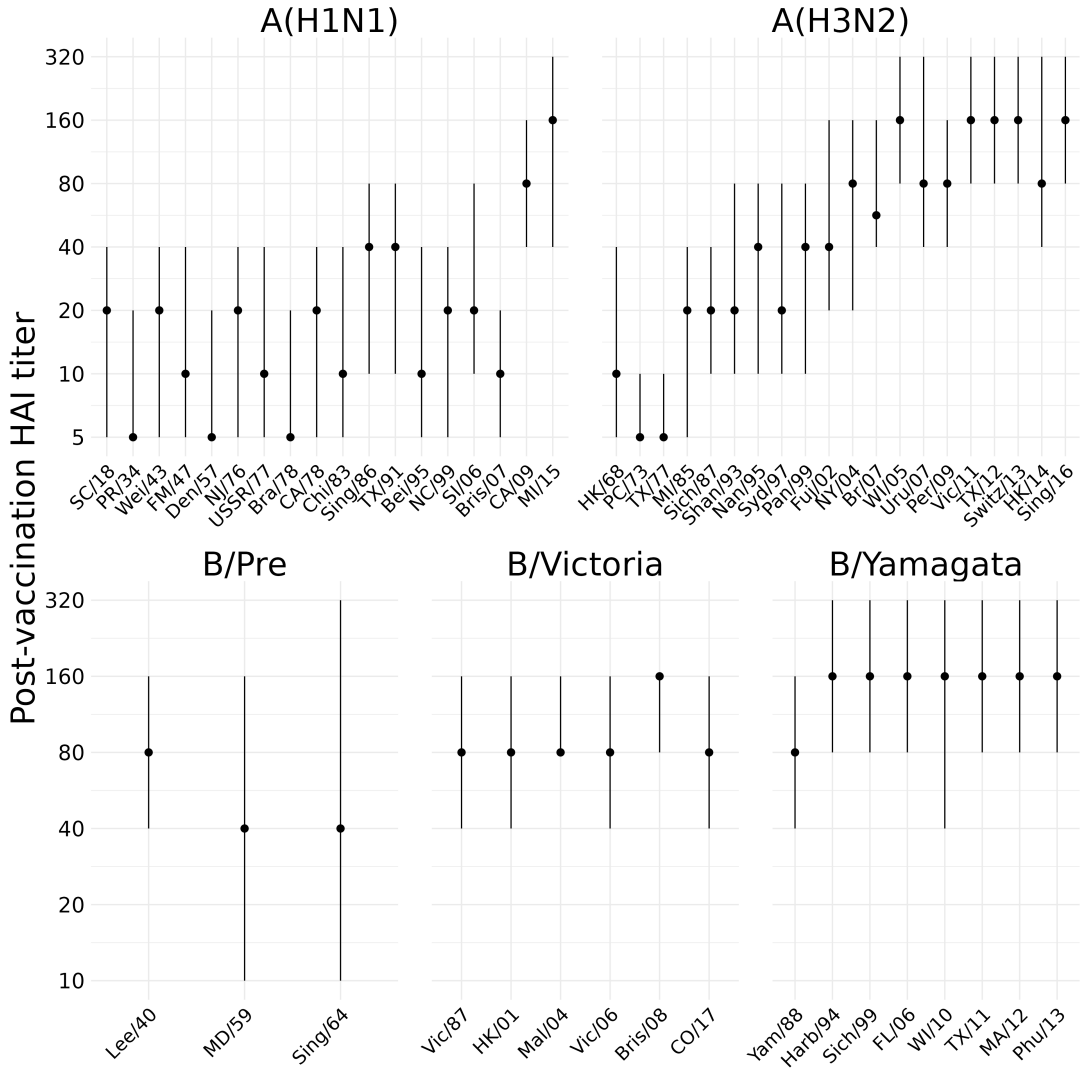

Figure B.3: Post-vaccination titers in our study to each of the assay strains. The point shows the median and the line shows the IQR.

Metric agreement analysis

Before we built statistical models for the post-vaccination titer, we first performed a simple unadjusted analysis of the consistency (or agreement) between the antigenic distance measurements. As an omnibus test of agreement, we calculated the intraclass correlation (ICC) across the four antigenic distance measurements (Table 3.2), separately for each strain type. We used a Bayesian model with a fixed effect for antigenic distance metric and random intercepts for both assay strain and vaccine strain, and calculated the ICC as the ratio of variance explained by the assay and vaccine strain variance components to the total variation. The Spearman rank correlations show in the main text can be viewed as a post-hoc analysis of the ICC which provide more information about specific comparisons.

Specifically, the model we fit for each subtype can be written as follows.

$$
\begin{aligned}
d[i] &\sim \mathcal{N}(\mu[i], \sigma^2) \\
\mu[i] &= \alpha_1 \cdot I(\text{method}[i] = \text{temporal}) + \alpha_2 \cdot I(\text{method}[i] = \text{p-Epitope}) + \\
&\quad \alpha_3 \cdot I(\text{method}[i] = \text{Grantham}) + \alpha_4 \cdot I(\text{method}[i] = \text{cartographic}) + \\
&\quad u[1, \text{assay strain}[i]] + u[2, \text{vaccine strain}[i]] \\
\alpha[k] &\sim t\,(3, 0, 5)\,; \quad k = 1, 2, 3, 4 \\
u[r, \cdot] &\sim \mathcal{N}(0, \zeta[r]); \quad r = 1, 2 \\
\zeta[r] &\sim t^{+}(3, 0, 1) \\
\sigma &\sim t^{+}(3, 0, 1)
\end{aligned}
$$

We fit the model using Stan's NUTS sampler using 12 chains, each with 1000 warmup iterations and 1000 post-warmup sampling iterations and an adaptive delta of 0.99. Model diagnostics were all sufficient (data not shown, the model is easy to sample from and samples quickly). We then calculate the ICC as

$$
\text{ICC} = \frac{\zeta[1]^2 + \zeta[2]^2}{\zeta[1]^2 + \zeta[2]^2 + \sigma^2},
$$

over the posterior samples of all parameters. That is, the ICC represents the ratio of variance due to strain effects only to the total variance after controlling for fixed effects. In the psychometric literature, this is referred to as a one-way ICC for consistency – if the ICC is close to one, it means the variance from the random effects dominates the model. We summarized the ICC as the mean and 95% HDCI across the posterior samples.

As a sensitivity analysis, we considered an alternative agreement statistic based on a different variance decomposition. We fit the same models as before, but then computed the variance of the posterior predictions for every point in the dataset without taking the random effects into account (the "fixed effects" predictions), i.e.

$$\sigma^2_{\text{FE}} = \text{Var}_{i=1}^{n} \left(\alpha[\text{method}[i]] \right),$$

where we choose the correct α parameter based on the method for dataset entry i (we omit writing all four alpha parameters and indicator functions for readability). Then, we compute the variance of the posterior predictions for each entry in the dataset taking the random effects and fixed effects into account:

$$\sigma^2_{\text{ME}} = \text{Var}_{i=1}^{n} \left(\alpha[\text{method}[i]] + u[1, \text{assay strain}[i]] + u[2, \text{vaccine strain}[i]] \right).$$

We can then compute an alternative agreement statistic as the variance ratio

$$1 - \sigma^2_{\text{FE}} / \sigma^2_{\text{ME}},$$

which will be close to one if the random effects dominate the prediction variance, or close to zero if the random effects have only a small contribution to the prediction variance. Table B.7 shows our results using this metric. All of the results indicate low agreement but with a much higher uncertainty, and this metric is less charitable to the A(H3N2) consistency, although we observed strong pairwise correlations between all of the A(H3N2) metrics as shown in the main text.

Table B.7: Prediction variance ratio across all antigenic distance measurements, calculated separately for each subtype or lineage (strain type). The posterior distribution for each ratio was calculated as one minus the ratio of the prediction variance ignoring random effects to the prediction variance including random effects, estimated with a Bayesian model. Numbers shown are the mean and 95% highest density credible interval (HDCI) of the posterior distribution of variance ratios.

Strain Type	**PPD Ratio**
H1N1	0.03 (-0.28, 0.30)
H3N2	0.21 (0.01, 0.39)
B-Yam	0.14 (-0.25, 0.48)
B-Vic	-0.05 (-0.75, 0.57)

Correlation coefficients and CIs

Table B.8 shows the Spearman correlation coefficients and 95% HDCIs for the correlations shown in Figure 1 of the manuscript. The estimates and CIs are arranged by subtype in the table in the same order in which they are shown in the plot. The coefficients may be slightly different from the point estimates we presented in the main text due to rounding error. We calculated the estimates and HDCIs shown here as the mean and 95% HDCI of 4000 samples from a posterior distribution created by Bayesian bootstrapping, which we performed independently on each subtype using 4000 resamples of the observed data points.

Notably, the credible intervals are quite wide for all subtypes other than A(H3N2), which showed strong agreement in both the reliability and correlation assessments. For influenza B subtypes, the width of the credible intervals is almost certainly due to the low number of vaccine/assay strain pairs we observed in our dataset. For A(H1N1), we suspect that the wide credible intervals are due to the multiple clusters in the data, which could violate the assumptions of calculating a correlation coefficient (specifically, Spearman's rank correlation assumes that the rank distributions are bivariate normal between the two variables of interest, which is unlikely to hold in a variable which represents information from multiple heterogeneous clusters). However, our point estimates which reflect low overall agreement are A(H1N1) are similar to the previous results obtained by Bedford et al. [33], and large credible intervals can indicate the lack of a strong signal in the data, so we feel that

the conclusions presented in the main text (a lack of agreement across metrics for A(H1N1) and B subtype, and a paucity of influenza B data) are not affected by the presence of wide credible intervals for the correlation coefficients.

Table B.8: Spearman correlation coefficients and 95% HDCIs estimated by Bayesian bootstrap for each influenza subtype. Each pairwise comparison is shown only once to prevent confusion.

Subtype		Cartographic	p-Epitope	Grantham
A(H1N1)	Temporal	0.45 (0.13, 0.74)	0.11 (-0.26, 0.51)	-0.25 (-0.64, 0.17)
	Cartographic		0.60 (0.34, 0.83)	0.35 (-0.01, 0.68)
	p-Epitope			0.74 (0.52, 0.93)
A(H3N2)	Temporal	0.93 (0.88, 0.97)	0.93 (0.89, 0.97)	0.96 (0.93, 0.98)
	Cartographic		0.93 (0.87, 0.97)	0.93 (0.88, 0.97)
	p-Epitope			0.97 (0.95, 0.99)
B/Yamagata	Temporal	0.66 (0.35, 0.89)	0.82 (0.69, 0.95)	0.83 (0.67, 0.97)
	Cartographic		0.64 (0.39, 0.86)	0.67 (0.36, 0.91)
	p-Epitope			0.91 (0.80, 0.99)
B/Victoria	Temporal	0.55 (0.01, 0.97)	0.78 (0.47, 0.99)	0.71 (0.31, 0.99)
	Cartographic		0.52 (-0.04, 0.92)	0.37 (-0.24, 0.88)
	p-Epitope			0.96 (0.85, 1.00)
Overall	Temporal	0.77 (0.67, 0.86)	0.65 (0.51, 0.77)	0.55 (0.39, 0.70)
	Cartographic		0.83 (0.77, 0.89)	0.77 (0.68, 0.86)
	p-Epitope			0.95 (0.92, 0.97)

Antigenic distance evenness and dispersion analysis

Rather than a uniform distribution of data points across distance space, each antigenic distancemetric had gaps in the distribution of observed distances, which varied by metric and subtype (Figure B.4A). The two B lineages had much larger gaps due to the sparser historical panels. For influenza A, all metrics were more uniform for A(H3N2) than for A(H1N1), suggesting their different evolutionary patterns across the time spanned by the historical panel. Notably, while the temporal metric was the most uniform for all strains (an artifact of how the historical panel was chosen), the Grantham and *p*-Epitope metrics tend to discretize the number of potential distances and result in less uniformly distributed values for the historical panel used in our study.

We quantified the uniform spread of points for each antigenic distance metric and subtype

using the gap standard deviation, where a higher gap standard deviation indicates more irregularity in the spacing of data points. The gap standard deviation is calculated as the standard deviation of the consecutive differences in the sorted antigenic distance values for a given metric. That is, assume x, a vector of measurements from $i = 1, \ldots, n$ is already sorted in increasing order so that $x_1 \leq x_2 \leq \ldots \leq x_n$.

Then, the gap standard deviation is computed as

$$\begin{aligned} d_k &= x_{k+1} - x_k; \quad k = 1, \ldots, i-1 \\ \bar{d} &= \frac{1}{n} \sum_{k=1}^{i-1} d_i \\ \sigma_{\text{gap}} &= \sqrt{\frac{1}{n-2} \sum_{k=1}^{i-1} \left(d_k - \bar{d}\right)^2}. \end{aligned}$$

For a random variable with a uniform distribution,

$$\lim_{n \to \infty} \sigma_{\text{gap}} = 0.$$

Figure B.4B shows the estimated gap standard deviations. Both B lineages had higher gap standard deviations for all methods than either influenza A subtype. For A(H3N2), the gap standard deviations were similar across antigenic distance methods, and for A(H1N1) the differences were still small but larger than A(H3N2), representing the diversity of strains in the historical panel for type A influenza strains. The differences were much more noticeable for both B lineages, with Grantham distance having notably higher gap standard deviation than the other metrics for both influenza B lineages, indicating lower diversity in the normalized distance values.

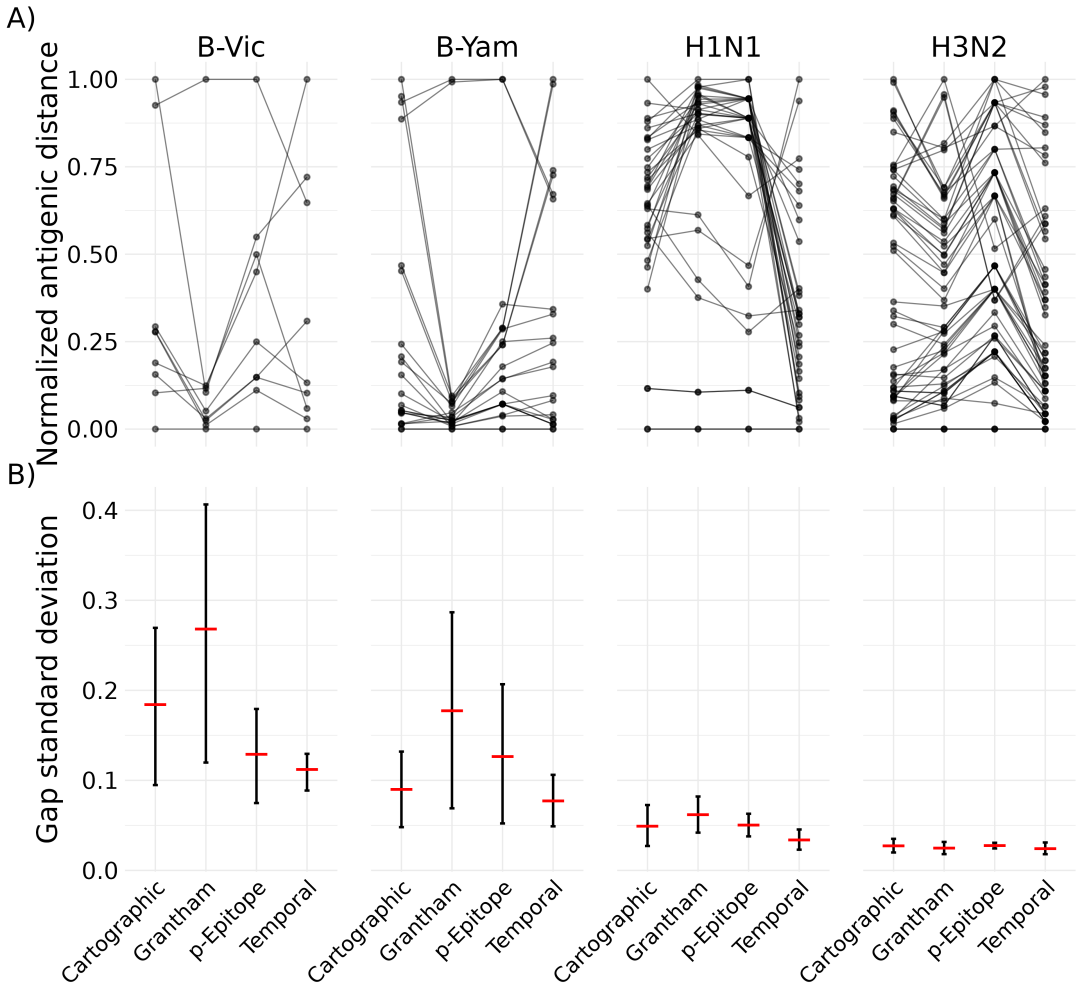

Figure B.4: A) parallel coordinates plot showing how the estimated pairwise antigenic distances change for each of the antigenic distance metrics. Each line in the plot represents one vaccine strain and assay strain pair, and the connected points are the pairwise distance measured under each metric shown on the x-axis. When two lines cross, this indicates that two metrics assigned a different relative order to the pairwise combination. Note that Grantham and especially p-Epitope distances are integer-valued and concentrate measurements to specific points which potentially overlap (temporal distance is also integer valued but has enough spread to avoid a similar issue). B) The gap standard deviation (gap SD) for each subtype and antigenic distance metric. The posterior distribution of gap SDs was calculated using the bayesian bootstrap with reweighting. The red horizontal bar shows the mean of the bootstrap posterior and the error bars show the 95% highest density credible interval (HDCI).

Model diagnostics

We examined the key model diagnostics for all of our models to ensure they converged. The main diagnostics with target criteria identified by the Stan development team [179] are:

- $\hat{R}$, which measures chain mixing, should be < 1.01 for all parameters;
- Bulk and tail ESS, measures of the number of samples drawn if all of the samples were independent, should be greater than 100 times the number of chains;
- Number of divergent transitions should be less than 1% of samples;
- Number of treedepth exceedences should be less than 1% of samples;
- E-BFMI should be greater than 0.3 for all chains.

These diagnostics are presented in Table B.9. We did not achieve all of the target diagnostic criteria for the most difficult to sample parameters, but we achieved sufficient diagnostic criteria for preliminary evaluation of our results.

We also examined trace plots of the parameters to ensure there were no obvious errors (and, in general, errors in the trace plots will be noticeable in the $\hat{R}$ statistic). We also examined the prior/posterior shrinkage and visually inspected prior/posterior plots. Since we have many models, each with thousands of parameters, we did not include the plots here. We observed good values of shrinkage (far from 1, indicating a divergence away from the prior) for most parameters, with the exception of some highly constrained parameters, typically correlations and GAMM regularizing variance parameters. Some of the random effects for individuals had poor shrinkage as well, but overall the shrinkage for random effects and for the random effects variances was far from 1. Since the GAMM was not supported by ELPD anyways, we did not investigate prior sensitivity analysis further since all of the LMM parameters had good shrinkage. Therefore, we feel safe about our choice of regularizing priors and a prior sensitivity analysis would require extensive computational time without being useful.

While less important for our purposes, we also sampled from the priors in order to examine the prior/posterior shrinkage and to visualize our prior predictive simulations. Such an analysis requires substantially less computational power than sampling from the posterior

distribution, but we still need to ensure that we have sampled from the priors enough to get good estimates of the prior distributions of some highly constrained parameters. So, our prior sampling diagnostics are shown in Table B.10.

Since there are thousands of parameters per model, we do not show the shrinkage parameters or prior distributions of all parameters here, but they are easy to produce from the code and results we provide.

ELPD Diagnostics

Similar to frequentist AIC/BIC and Bayesian WAIC, ELPD relies on a computationally efficient approximation to leave-one-out cross validation that allows estimation of a goodness-of-fit metric without having to refit a computationally impossible number of models. However, unlike other informatic criteria, the LOO-IC based on the leave-one-out expected log pointwise predictive density provides diagnostics to determine if the approximation is trustworthy [155, 156]. Table B.11 shows the diagnostic measures for each of our models. The maximum Pareto k diagnostic is the primary value indicating whether the LOO-ELPD approximation is accurate – all Pareto k values (one per observation) should be below 0.7. The N_{eff} value is the effective sample size for the approximation, and the ratio of the effective sample size to the actual sample size should be greater than 0.5 to ensure that the threshold of 0.7 is reliable. If the number of effective samples is greater than 2200 however, the threshold of 0.7 is useful regardless of the ratio.

Pointwise prediction comparisons

To examine the difference in predictions across each of the antigenic distance metrics, we computed the fold change in predicted post-vaccination HAI titer conditional on normalized antigenic distance and strain type for each unique pair of antigenic distance metrics. We visually inspected the conditional fold changes between metrics using a limit of agreement approach with a clinically defined threshold for whether the difference between predictions should matter, which is commonly defined as a 4-fold change for HAI measurements.

Table B.9: Model diagnostics for the posterior samples from the GAMMs and LMMs fit with each of the antigenic distance metrics. We show the total number of divergences out of the number of samples, percent treedepth exceedences, minimum ESS across all parameters, the minimum E-BFMI across chains, and the maximum R hat across all parameters.

Model	Pct. Divergences	Pct. TD Exceeded	min ESS (tail)	min ESS (bulk)	min E-BFMI	max R_hat
GAMM	0.0%	100.0%	2858	1595	0.57	1.02
	0.2%	99.9%	4182	2407	0.53	1.01
	0.0%	100.0%	477	464	0.57	1.05
	0.1%	100.0%	567	372	0.51	1.06
	0.0%	100.0%	2066	956	0.53	1.03
		100.0%	1062	529	0.54	1.05
LMM		100.0%	2054	970	0.49	1.03
	0.2%	99.8%	3068	2109	0.50	1.01
	0.0%	100.0%	2969	1188	0.54	1.03
	0.2%	100.0%	2239	1077	0.52	1.02
	0.0%	100.0%	1793	894	0.57	1.03
	0.2%	99.8%	2400	1108	0.56	1.03

Table B.10: Model diagnostics for the prior samples from the GAMMs and LMMs fit with each of the antigenic distance metrics. We show the total number of divergences out of the number of samples, percent treedepth exceedences, minimum ESS across all parameters, the minimum E-BFMI across chains, and the maximum R hat across all parameters.

Model	Pct. Divergences	Pct. TD Exceeded	min ESS (tail)	min ESS (bulk)	min E-BFMI	max R_hat
GAMM	0.0%	0.0%	9051	6896	0.89	1.00
			8915	7012	0.88	1.00
			8830	7469	0.90	1.01
			8644	7494	0.88	1.01
			8312	7353	0.94	1.01
			8475	7486	0.91	1.00
LMM			7363	6808	0.89	1.00
			7849	6856	0.82	1.00
			7719	6873	0.83	1.01
			7327	6789	0.87	1.01
			6999	6599	0.82	1.01
			8269	6823	0.87	1.00

Table B.11: Diagnostics for the LOO-IC ELPD approximation. Pareto k is the primary diagnostic indicating whether the approximation is trustworthy and all Pareto k values should be below 0.7. The N_{eff} is the effective sample size and R_{eff} is the ratio of the effective sample size to the true sample size – if there are too few effective samples relative to actual samples, we can get an optimistic evaluation of the approximation quality, but in general this matters less if the ESS is sufficiently high.

Metric	**Model**	**Max. Pareto k**	**Min. N_eff**	**Max. R_eff**
Cartographic	GAMM	0.39	4368.8	1.00
	LMM	0.44	3704.8	1.00
p-Epitope	GAMM	0.40	3738.7	1.00
	LMM	0.44	4289.5	1.00
Grantham	GAMM	0.36	3803.0	1.00
	LMM	0.35	3936.8	1.00
Temporal	GAMM	0.49	3365.6	1.00
	LMM	0.43	3850.4	1.00

We performed the fold change between predictions calculations for both the GAMM and LMM with each antigenic distance metric. Figure B.5 shows the prediction comparisons across antigenic distance metrics for each subtype using the LMMs. In contrast to our agreement analysis, where the H3N2 metrics showed the strongest agreement across metrics (and the highest pairwise correlations), H3N2 was the only strain with noticeable trends in the contrasts between metrics. In particular, all of the comparisons with p-Epitope for H3N2 had a noticeable trend – even though the mean fold change in predictions always stayed within the measurement error boundaries we set *a priori*, sometimes the credible interval did not fully cover the measurement error boundaries and there was a noticeable slope. These trends suggested that p-Epitope measurements underestimated the expected change in post-vaccination titer compared to Grantham and cartographic distance, while p-Epitope overestimated the difference compared to temporal methods. These results suggest that perhaps biochemical features like glycosylation sites or changes to the virus outside of the immunodominant epitope region are important, because these features are detected by cartographic and Grantham distance, but not by p-Epitope distance.

Figure B.6 shows the prediction comparisons across antigenic distance metrics for each subtype using the GAMMs. Even though the GAMM was not supported by our ELPD analysis, we used the GAMM for analyzing pairwise differences in predictions in case the

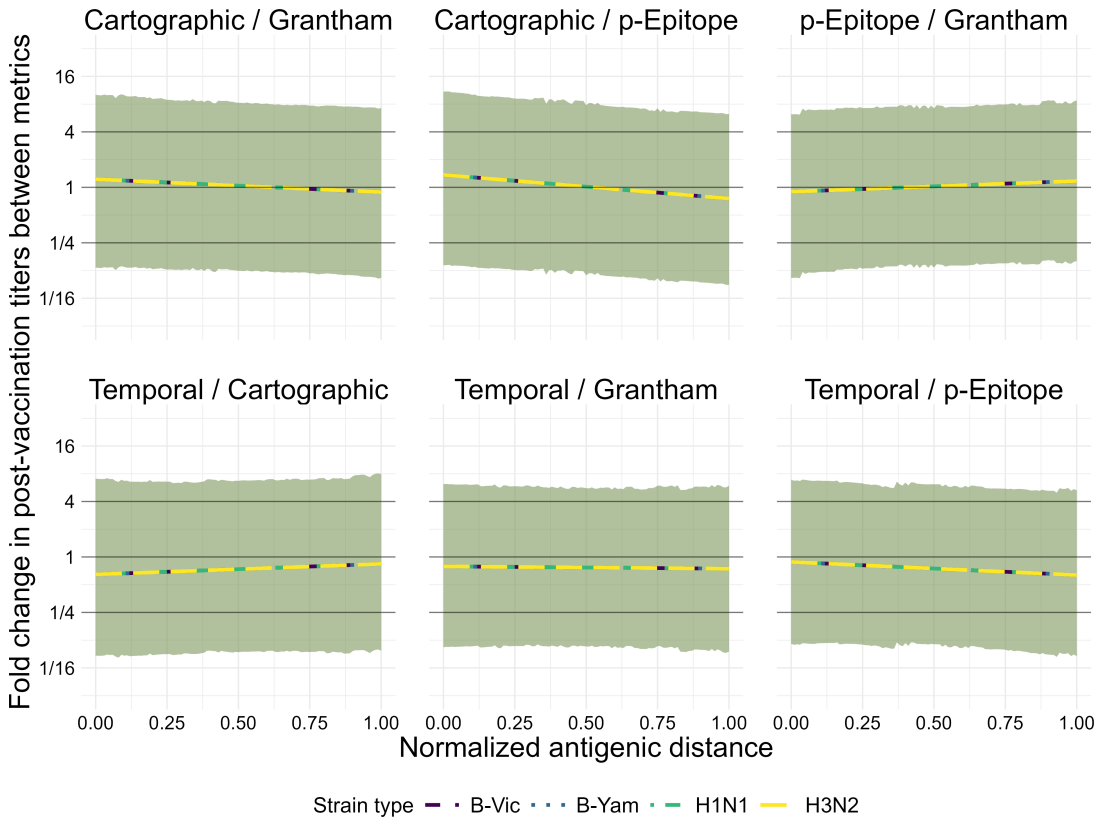

Figure B.5: Pairwise comparisons of predictions (from the LMMs) between each unique set of two metrics. The y-axis shows the fold change in predictive titers between metrics, and the two metrics being compared in each subplot are shown as the subplot labels. Each line represents the predictions for the first metric in the pair at a given antigenic distance value divided by the predictions for the second metric in the pair. Color and linetype correspond to different strain types. The solid black lines on the plot are reference lines at a value of 1 for no effect, and at 4 and 1/4, effect values which would represent a clinically notable deviation in HAI predictions beyond what is expected from measurement error. Lines represent the mean of the posterior distribution of the contrast and the colored ribbons represent the 95% highest density credible interval (HDCI) for each strain type in each subplot.

nonlinear signal was biologically important with a weak signal. Unlike our simple correlation analysis, this analysis examines the predicted protection for an average individual exposed to an antigenically distant strain after vaccination, rather than only taking antigenic distance into account. We saw that the fold change in predicted HAI titers was almost always less than four for every pairwise comparison between two metrics. A four-fold change in HAI titer is considered a clinically relevant difference between two measurements, so in almost every case we saw that changing the antigenic distance metric would not lead to a clinically relevant difference in predicted post-vaccination HAI titer. The primary exception was strain type A(H1N1), which exceed 40 at a few antigenic distance values for some of the pairwise comparisons (around a normalized antigenic distance of 0.25 for the cartographic/Grantham and Cartographic/p-Epitope comparisons, and around a normalized antigenic distance of 0.75 for the Grantham/temporal distance comparisons). Due to the large standard errors and the number of comparisons we make, we are comfortable attributing these fluctuations to measurement error, although the large variability across antigenic clusters for A(H1N1) strains (pdm-like vs. non-pdm-like) could contribute as well.

However, the differences in comparisons for A(H3N2) was not completely trivial either. Figure B.6 shows that for A(H3N2), the temporal distance overwhelming underestimates the fold change in predictions for the largest antigenic distances compared to both Grantham and p-Epitope measurements, with some interesting trends in the comparisons between cartographic distance as well. These results support our conclusion that further research into which of these metrics actually captures useful and interesting features is warranted, because it is difficult to tell whether we are capturing noise from our study or actual patterns that suggest different metrics are identifying different relevant characteristics of the viruses.

In both models, nearly all contrast predictions fall within the clinically irrelevant reference bounds, although the credible intervals for all predictions are wide because our bayesian models fairly account for many sources of uncertainty in the data. However, our results for the GAMM model suggest some interesting exceptions for the A(H1N1) strains that are likely related to the pandemic-like and non-pandemic-like cluster differences. Our results for the GAMM and LMM model for A(H3N2) seem to suggest that perhaps different metrics

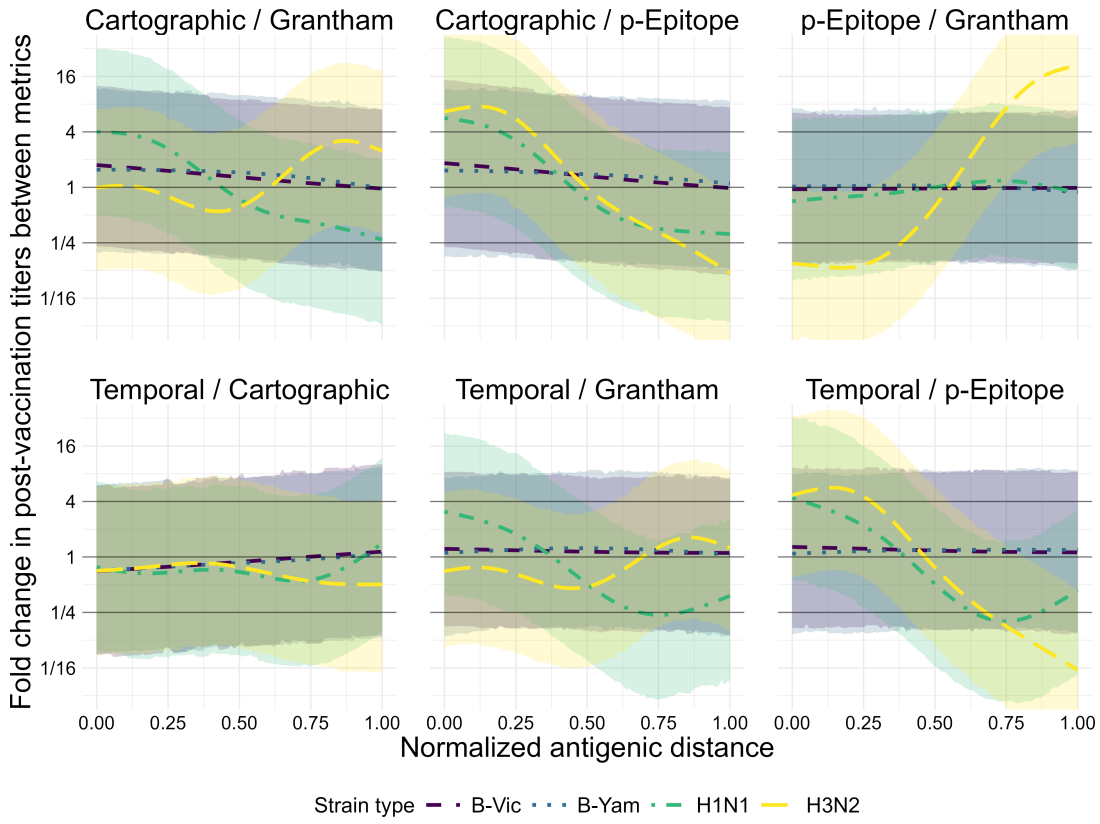

Figure B.6: Pairwise comparisons of predictions (from the GAMMs.) between each unique set of two metrics. The y-axis shows the fold change in predictive titers between metrics, and the two metrics being compared in each subplot are shown as the subplot labels. Each line represents the predictions for the first metric in the pair at a given antigenic distance value divided by the predictions for the second metric in the pair. Color and linetype correspond to different strain types. The solid black lines on the plot are reference lines at a value of 1 for no effect, and at 4 and 1/4, effect values which would represent a clinically notable deviation in HAI predictions beyond what is expected from measurement error. Lines represent the mean of the posterior distribution of the contrast and the colored ribbons represent the 95% highest density credible interval (HDCI) for each strain type in each subplot.

are picking up different relevant features, as we noted in the main text discussion.

Vaccine-specific predictions

In order to analyze the differences between vaccine strains, we also examined the results conditional on the specific strain used in a vaccine (for a given subtype). Figure B.7 shows the fitted models. Within a given subtype, there were no striking results across the different vaccine components. In the main text, we show that direct causal effects of vaccine and assay strain contribute very little to the variance in the outcomes after controlling for antigenic distance.

Alternative distance metrics

Because we chose to include the dominant p-Epitope distance and Grantham's distance in our final manuscript, we also analyzed other sequence-based and biochemical distances to determine if our arbitrary choice was misleading and we should consider further antigenic distance measures. So, we compared the dominant p-Epitope distance [30, 31] with the p-all-Epitope distance [31, 35] and the Hamming distance [221]. Furthermore, we compared Grantham's distance [151] with the Hamming distance and with the FLU substitution model, an evolutionary amino acid substitution matrix model derived specifically for influenza sequences [173] (there are other indices like Grantham's index, but we felt that the comparison to a model specifically for influenza amino acid substitutions was sufficient).

When we compared the FLU substitution model to Grantham's distance and the simple Hamming distance, we found that all three metrics were highly correlated for all subtypes except A(H1N1), with relatively small credible intervals from Bayesian bootstrapping (Table B.12). For A(H1N1), Grantham and Hamming distances were also highly correlated, but the correlations between Hamming and FLU substitution and Grantham and FLU substitution distances were moderate at best, with credible intervals that covered quite low values. The original study which developed the FLU substitution matrix used a mix of influenza virus sequences across multiple proteins and types/subtypes [173], so it is unclear

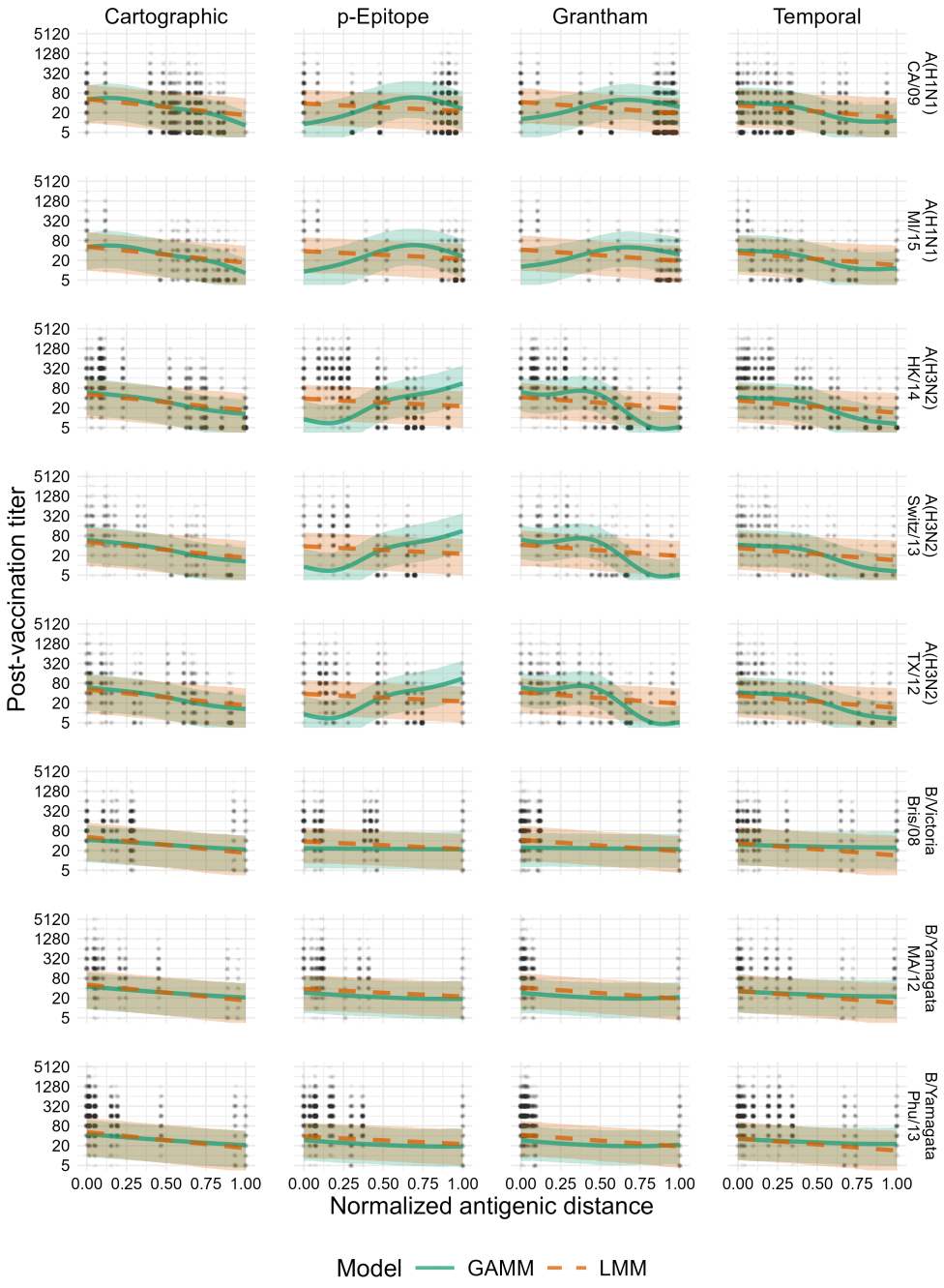

Figure B.7: Model predictions for both the GAMM and LMM, conditional on the vaccine strain rather than only on the subtype (shown in the main text). Solid green lines and green ribbons show the mean and 95% highest density continuous interval (HDCI) for GAMM predictions. Dashed orange lines and orange ribbons show the mean and 95% HDCI for LMM predictions. Circular points show the data values. Each subplot shows the model predictions for a particular subtype (changes by row) and distance metric (changes by column). Outcomes shown on the plot are predicted post-vaccination titers for an average individual to an average strain.

why the difference would be so stark for A(H1N1). Regardless, because the difference was only noticeable for A(H1N1) we decided to use the Grantham distance in our main analysis. Despite high similarity to the Hamming distance across all subtypes, Grantham distance contains more information by design and better antigenic coverage of Influenza B strains in a future study might reveal further differences between Grantham distance and Hamming distance.

Table B.12: Pairwise Spearman rank correlations between antigenic distance values using the Grantham, FLU Substitution, and Hamming distance metrics. We calculated correlations between two distances using the normalized distance values between every vaccine/assay strain pair for the given subtype. Numbers shown are the mean and 95% highest density continuous interval (HDCI) calculated by Bayesian bootstrapping.

Subtype		Grantham	FLU Substitution
A(H1N1)	Hamming	0.91 (0.82, 0.98)	0.62 (0.35, 0.85)
	Grantham		0.41 (0.05, 0.75)
A(H3N2)	Hamming	0.99 (0.98, 1.00)	0.97 (0.94, 0.99)
	Grantham		0.96 (0.93, 0.98)
B/Yamagata	Hamming	0.98 (0.93, 1.00)	0.94 (0.86, 1.00)
	Grantham		0.91 (0.80, 0.98)
B/Victoria	Hamming	0.99 (0.97, 1.00)	0.97 (0.89, 1.00)
	Grantham		0.97 (0.88, 1.00)
Overall	Hamming	0.99 (0.97, 1.00)	0.94 (0.91, 0.96)
	Grantham		0.91 (0.88, 0.95)

We also examined the pairwise Spearman correlations between the Hamming distance, the (dominant) p-Epitope method which we present in the main analysis, and the p-all-Epitope distance, which is calculated by averaging the Hamming distance across all 5 of the immunodominant HA epitope sites. Again, the correlations were overall high Table B.13, with A(H1N1) displaying a notably lower correlation across differences. These supplementary results suggest that different antigenic distance metrics may have the strongest effect on understanding the immune response to A(H1N1), probably in accounting for noteable differences across clusters. The multiple clusters in A(H1N1) antigens are the primary differentiating factor from the ladder-like continuously evolutionary pattern in A(H3N2) and might explain the differences, although we lack the ability to analyze this further. The high correlation between sequence-based metrics for A(H3N2) provides strong support against the

idea that specific epitopes or mutations outside of the epitope region drive the unexpected results we saw with H3N2 and the *p*-Epitope distance in our model. In particular, there may be few specific sites [133] or specific changes in glycosylation sites [165] that are lost in the noise from other mutations when we use these metrics.

Table B.13: Pairwise Spearman rank correlations between antigenic distance values using the Grantham, FLU Substitution, and Hamming distance metrics. We calculated correlations between two distances using the normalized distance values between every vaccine/assay strain pair for the given subtype. Numbers shown are the mean and 95% highest density continuous interval (HDCI) calculated by Bayesian bootstrapping.

Subtype		p-Epitope	p-All-Epitope
A(H1N1)	Hamming	0.81 (0.64, 0.95)	0.89 (0.78, 0.97)
	p-Epitope		0.83 (0.68, 0.95)
A(H3N2)	Hamming	0.97 (0.93, 0.99)	0.99 (0.97, 0.99)
	p-Epitope		0.97 (0.94, 0.99)
B/Yamagata	Hamming	0.93 (0.84, 0.99)	0.94 (0.86, 1.00)
	p-Epitope		0.98 (0.95, 1.00)
B/Victoria	Hamming	0.95 (0.81, 1.00)	0.97 (0.88, 1.00)
	p-Epitope		0.93 (0.78, 1.00)
Overall	Hamming	0.95 (0.92, 0.97)	0.98 (0.96, 0.99)
	p-Epitope		0.97 (0.95, 0.99)

Due to the relative consistency across these other antigenic distance metrics, we did not fit further models to other antigenic distance metrics. The models require a great deal of computational time and power, and since we found overall good agreement between these additional metrics (and a large amount of disagreement within A(H1N1), as we found for the metrics in the main analysis), we felt that this did not justify a further investigation. However, a future study with an expansive panel of serological data to A(H1N1) and A(H3N2) strains to further explore why A(H1N1) metrics have lower agreement would be useful for further understanding these results.

Appendix C

Supplementary material for Chapter 4

Reproducibility instructions

To produce our results, you can access our code and data on the Zenodo archive (`https://doi.org/10.5281/zenodo.15578876`) or the GitHub page (`https://github.com/ahgroup/billings-breadth-quantification-public`). You will need to install the following software requirements.

- R version 4.4.2, available from `https://cran.r-project.org`.
- A C++ compilation toolchain that is compatible with R – on Windows you should install RTools 4.4 (also available from CRAN). On MacOS you need the XCode command line tools. On a Linux computing cluster, this has probably already been set up for you.
- (Optional) We used the RStudio IDE which provides many convenience features, especially when working with `targets`, which is available from `https://posit.co/download/rstudio-desktop`.
- If you want to reproduce the manuscript or supplement documents you will need Quarto version 1.6.40, available from `https://quarto.org`.
- The R package `renv`, version 1.1.4, which can be installed from CRAN and will attempt to bootstrap itself the first time you open our R project.
- The packages specified in the `renv.lock` file, which can be installed with

`renv::restore()`.

- Once you've installed the `cmdstanr` package, you also need to install the `cmdstan` utility by following the `cmdstan` quick start guide (`https://mc-stan.org/cmdstanr/articles/cmdstanr.html`). We used `cmdstan` version 2.36.0.
- Many of those packages have system utilities, which you will need to install yourself on most Linux systems.

We ran our pipeline on the `sapelo2` computing cluster at the University of Georgia, which runs CentOS Linux release 7.5. and uses Slurm to schedule jobs [160, 161]. We developed our pipeline using the R workflow management package `targets` [158], and configured the pipeline to run automatically on both our local Windows environment and on sapelo2. If you use a different HPC environment or a different local environment, you may need to make edits to `_targets.R`, which configures the distributed computing setup for the pipeline, or `job.sh`, the Slurm submission script, for the pipeline to run the way you intended (this is no different from running any other scripting language program on multiple platforms, we cannot guarantee cross-platform compatibility).

You can run our entire pipeline in one step by opening the `Rproj` file in RStudio, and running `targets::tar_make()` in the console. Notably, to run all code sequentially and avoid any issues with distributed computing setup, you can run `targets::tar_make(use_crew = FALSE)` but this will be much slower if you are on a multicore computer. You do not need to use RStudio or `targets` to reproduce our results, but if you run our code in any other way we assume you know enough about R programming to know what you are doing.

Finally, note that our pipeline is computationally intensive and requires a large amount of storage — you should have at least 100 GB of available storage space on the drive where you store this project. Regardless, the pipeline will take a long time to execute, typically at a full day. Depending on your HPC setup (with potential alterations to the setup scripts) or local machine, it might be faster but this analysis will take multiple hours at a minimum to run. We do not recommend running this pipeline on a machine with less than 16 cores and 32 GB of RAM. Note that differences between local systems and program setups can make even robust code difficult to reproduce.

Expanded methods

The methods included in the main text are a summary of the most important methods we used to conduct our analysis. Here, we include a more detailed explanation of many aspects of the methods used.

Antigenic distance calculation

Here, we briefly summarize our methods for calculating antigenic distances. We give a complete description of the methods in Appendix D.

We calculated the antigenic distance for all of the influenza strains that were used in the cohort data. We calculated the distances pairwise between each set of two strains, and we calculated all distances separately between H1N1 and H3N2 influenza strains. That is, we did not calculate the distance between any H1N1 and H3N2 strains. We used three different distance metrics to measure antigenic distance. All three have been previously described.

The first metric we used was the temporal antigenic distance. Many existing papers in the literature either order strains in year order and then count the distances between them as equal, regardless of the number of years elapsed [38, 140]; or assign distance based on the year of isolation [45, 142]. The temporal distance is based on the year of isolation for each strain, and measures how much chronological time has passed since the strains were first observed and sequenced. Letting s_1 and s_2 be two influenza strains from the same subtype, and $t(s_{(\cdot)})$ be the year of isolation for an arbitrary strain, the temporal distance is given by

$$d_{\text{temporal}}(s_1, s_2) = |t(s_1) - t(s_2)| .$$

That is, the temporal distance is equal to the absolute difference between the year of isolation for the strains. For our study, the absolute value is not necessary because we do not have any assay strains that were isolated after the vaccine strain to which they were compared. However, in such a study, the absolute temporal distance should be compared with the directional temporal distance as future years might have a greater impact on evolution than

past years.

The second metric we used was a sequence-based metric called the dominant p-epitope metric [30]. First, the p-epitope distance is calculated for each of the five antigenic sites on the hemaaglutinin head domain. The five antigenic sites are composed of residues with known positions using a standard numbering scheme for HA proteins which varies by subtype [222]. Notably, the numbering for the residues refers to the sequence after the sequence for the signal peptide has been removed. For H1 HA proteins, the signal peptide is typically 18 residues in length. For a given epitope site, the p-epitope distance is calculated as

$$d_{\text{epitope}_i}(s_1, s_2) = \frac{d_H(s_1, s_2)}{\text{length of sequences}}$$

where $d_H(\cdot, \cdot)$ is the standard Hamming distance [221] which counts the number of differences between two strings. The dominant p-epitope distance is then the maximum of the p-epitope distance for each set:

$$d_{p-\text{epitope}}(s_1, s_2) = \max \left\{ d_{\text{epitope}_i}(s_1, s_2) \right\}_{i \in I} \ ; \quad I = \{A, B, C, D, E\}.$$

Here the index set I contains the letters which are typically used as names for each of the antigenic sites. A table of the site residues can be found in Appendix D.

Appendix D also gives details on how we obtained the sequences used in our study. In order to calculate the p-epitope distance, we first preprocessed the sequences by removing the signal peptide sequence from the beginning of each sequence. After preprocessing, we computed the p-epitope distance for each of the epitopes, and then we took the maximum across all epitopes for each strain pair to obtain the dominant p-epitope distance.

The third method we used to calculate antigenic distance was antigenic cartography [29, 44]. The goal of antigenic cartography is to estimate distances between strains from immunological data by a dimension reduction algorithm. We used the HAI titers against the panel of strains from our cohort data for antigenic cartography. First we had to create a titer matrix, which has one row for each individual in the dataset and one row for each strain observed in the dataset.

We used the methods from the Racmacs [175] R package for antigenic cartography, which accepted our titer matrix as the input. Racmacs uses a variant of classical multidimensional scaling (MDS) for the dimension reduction algorithm, which finds a two-dimensional representation of the data that minimizes a particular loss function. In simpler terms, Racmacs finds a way to represent our data in two dimensions that preserves the most amount of information possible. Classical MDS does not account for missing or censored data, which are both present in our titer matrices. However, Racmacs implements an optimization routine called dimensional annealing to numerically minimize the classical MDS loss function.

First, the censored and missing data points are randomly imputed, taking into account only the range of the immunological values provided. We pass the imputed data to a standard implementation of classical MDS and calculate the resulting stress (another term for the loss function in this scenario). Then, the data are passed to a numerical optimization routine (L-BGFS) which relaxes the coordinates in the reduced dimensional space in order to reduce the stress. We performed 100 optimization rounds for our cartographic maps in order to account for the randomness in the algorithm. We repeated the entire map creation process 25 times to allow for different initial conditions which could lead the optimizer to local minima, and chose the best overall map after all optimizations.

Once we obtained the optimized antigenic map, we found the positions of each strain in the resulting two-dimensional space and calculated the pairwise Euclidian distances, giving us the pairwise cartographic distances.

Likelihood adjustment for censoring

In the following sections about estimating our various metrics, we will employ the same correction to the regression model likelihood to adjust for censoring [153]. Letting y represent an HAI titer measurement, and the outcome for a regression model, if we say that

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$$

we can adjust for censoring in y by writing the likelihood function as

$$f(y_i \mid \mu_i, \sigma^2) = \int_L^U \phi\left(y_i \mid \mu_i, \sigma^2\right) \, \mathrm{d}y_i$$

where f is the probability density function of Y_i and ϕ is the normal density function. Here, L is the lower censoring bound and U is the upper censoring bound.

The censoring bounds are easy to determine as long as we know how our assay of interest was conducted. HAI is a serial dilution assay, and so produces interval censored values, and has both a lower and an upper limit of detection (LoD). For an observed HAI value y_i, we assume there is some latent "true" dilution y_i^* which will be a positive real number that represents the minimal dilution where hemagglutination is not observed. The latent dilution can never be observed because of how we conduct the assay. First, we choose a starting dilution, which is constrained by the proportion of reagents we need to add to our assay. We call the starting dilution $y_{\min}$ and it is 10 in our dataset. That is, the HAI assays in our study were performed with an initial dilution of 1:10 serum to other assay ingredients.

If we observe hemagglutination at this starting titer, we record the assay value as below the limit of detection – by convention, this is often recorded as half the limit of detection, as it is in our study data. Although it is important to note that as long as we keep track of which assay measurements are below the limit of detection, these values are really meaningless and such not be treated as numbers (but they usually are).

Regardless, we will dilute the assay solution 2-fold and look for agglutination again, continuing to dilute the amount of serum in solution. The maximal dilution we perform for the assay, $y_{\max}$ is the upper limit of detection. While 20480 was the theoretical upper LoD in our study, we did not observe any values at this level.

Any result that is within the limits of detection is still interval censored, i.e., we only know that it lies within a particular interval. For example, if we observe agglutination at a dilution of 1:40, but not at 1:20, we know that the true minimal inhibition dilution is somewhere between 20 and 40. We know it is higher than 20, and less than 40, but not where it falls in that range.

For an HAI titer assay conducted in this manner, we can write the censoring bounds L and U for a given titer y (leaving out subscripts from this formula for simplicity) on the log scale as

$$(L, U) = \begin{cases} (-\infty, y_{\min}), & y = y_{\min} \\ [y, y+1), & y_{\min} < y < y_{\max} \\ [y_{max}, \infty), & y = y_{\max} \end{cases}.$$

We use this same censoring correction in all of the corrected models.

Methods for calculating current immunogenicity metrics

Our methods for calculating the metrics are laid out without much technical detail in Appendix E, so we refer readers looking for a tutorial to that section. In this section we will briefly list the relevant technical details for our estimation models.

We will use the following notation to define our formulas for the current immunogenicity metrics.

- $i = 1, \ldots, n$ indexes study subjects, where n is the sample size.
- The variable $s = 0, 1, \ldots, S$ indexes different strains, and $s = 0$ is the homologous strain.
- We define seroprotection as a post-vaccination HAI titer greater than 40, and seroconversion as a post-vaccination fold-change of 4-fold or higher along with seroprotection.
- We use GMT as an abbreviation for geometric mean titer.

Homologous GMT (magnitude): The simple formula for the homologous GMT without considering censoring is

$$GMT_0 = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log \text{ titer}_{i,s=0}\right).$$

The regression model we fit to estimate the homologous GMT in the same framework as our novel metrics is given as

$$\log \text{ titer}_{i,s=0} \sim \mathcal{N}\left(\alpha, \sigma^2\right).$$

In this model, the parameter α estimates the mean of the outcome values.

Seroconversion rate (breadth): The simple formula for the seroconversion rate without considering censoring or the overall estimation framework is

$$\text{SCR} = \frac{1}{n}\frac{1}{S}\sum_{i=1}^{n}\sum_{s=0}^{S} I(\text{patient } i \text{ seroconverted to strain } s)$$

where $I()$ is the indicator function.

The regression model we fit to estimate the seroconversion rate is

$$\text{Seroconverted}_{i,s} \sim \text{Bernoulli}(p)$$

$$\text{logit}(p) = \alpha$$

where α estimates the overall seroconversion rate.

Overall GMT (magnitude): The simple formula for the overall GMT is

$$GMT = \exp\left(\frac{1}{n}\frac{1}{S}\sum_{i=1}^{n}\sum_{s=0}^{S}\log \text{ titer}_{i,s}\right).$$

and the regression model is

$$\log \text{ titer}_{i,s} \sim \mathcal{N}\left(\alpha, \sigma^2\right).$$

This regression model is exactly the same as the model for GMT_0, but for the GMT_0 model we only include homologous strains in the fitting, and for the GMT, which is estimated by α as well, we include all strains in the model fitting to get an overall estimate.

Methods for calculating novel immunogenicity metrics

In order to calculate our novel immunogenicity metrics, we first need to fit a summary antibody landscape model to the titer data. Letting d_s be the antigenic distance from the current season's vaccine to assay strain s (recall that we treat all seasons independently and fit a separate model to each season) we define a multilevel linear regression model to

construct the summary antibody landscape. The multilevel model is defined as

$$
\begin{aligned}
y_{i,s} &\sim \mathcal{N}\left(\mu_{i,s}, \sigma^2\right) \\
\mu_{i,s} &= (\beta_0 + b_{0,i}) + (\beta_1 + b_{1,i}) \cdot d_s \\
\begin{pmatrix} b_{0,i} \\ b_{1,i} \end{pmatrix} &\sim \text{MVN}\left(\vec{0}, \Sigma_b\right)
\end{aligned}
$$

where we have a random intercept and random antigenic distance slope for each individual i, and the random effects for each individual are allowed to be correlated with a correlation matrix shared by all individuals. We will discuss priors in the model implementation section, but for all other models we have chosen weakly informative independent priors. However, allowing the random effects to be correlated as we do here is often beneficial for model fitting.

In order to calculate our metrics from this model we need to calculate what `brms` calls expectation predictions or epreds. This refers to the predicted values $\hat{\mu}$ for some values of the predictors. Specifically, we use the population average epreds which are estimated conditionally on the random effects but do not include random effects deviations in the predictions, i.e.,

$$\hat{\mu} \mid d_{new} = (\hat{\beta}_0 + 0) + (\hat{\beta}_1 + 0) \cdot d_{new}.$$

We calculated these epreds on an interpolated grid of normalized antigenic distance values, $\vec{d}_{new} = 0, 0.01, 0.02, \ldots, 0.99, 1.00$. The granularity of the predictions can be increased if desired, but we found this granularity to give a good balance between resolution and memory usage.

Our novel metrics are then derived from the epreds.

Intercept (magnitude): The intercept is simply defined as

$$E(\hat{\mu} \mid d_{new} = 0).$$

That is, we average over the posterior samples of the epreds for an antigenic distance value

of 0.

Proportion above threshold (breadth): The proportion above the threshold is defined as the distance value where the horizontal line $y = 3$ (on the log scale) intersects the summary antibody landscape. That is,

$$E\left(\underset{d^*}{\operatorname{argmin}} \left| (\hat{\mu} \mid d^*) - 3 \right| \right),$$

which for the linear regression model has a simple closed form of

$$E\left(\frac{3 - \hat{\beta}_0}{\hat{\beta}_1}\right)$$

but might not have a simple closed form for more complicated models, so we find the interpolated value of d_{new} in our grid which minimizes $g(d_{new}) = \operatorname{abs}(\hat{\mu} \mid d_{new} - 3)$.

Area under the curve (total strength): The formula for the area under the curve (AUC; between the normalized antigenic distance values of 0 and 1) is

$$E\left(\int_0^1 \hat{\mu} \mid x \ \mathrm{d}x\right).$$

For each posterior sample, we approximate this integral numerically using the trapezoid method implemented by `pracma::trapz()` with default iteration settings.

General details for model implementation

We estimated all of these metrics using bayesian models, which produce a distribution of posterior samples for all model parameters. Whenever we summarize posterior samples to obtain a point estimate and credible interval, we calculate the mean and 95% highest density continuous interval (HDCI), implemented as `tidybayes::mean_hdci()`.

For all of our regression models, we specified general weakly informative priors. Because the models converged and moved away from the priors, the specific priors are not very important. That being said, we used $t(3, 0, 3)$ priors for regression coefficients and $t^+(3, 0, 1)$ priors for variance parameters, where $t(\nu, \lambda, \tau)$ is the location-scale Student's t distribution

with ν degrees of freedom, location parameter λ and scale parameter τ; and $t^+(\cdot)$ is the half-Student's t distribution constrained to be strictly positive with the same arguments. We use 3 degrees of freedom for all of our Student's t priors because the variance is infinite for smaller degrees of freedom. Using $\nu = 3$ allows for the prior distribution to have fat tails, so if the data support a large parameter value the model likelihood will allow that, but this tends to be much less pathological than trying to sample from priors with $\nu = 1$ or 2. For our multilevel summary landscape models, we specified the priors on the covariance matrix using a Cholesky factor decomposition with an LKJ-corr-cholesky(2) prior [207] on the Cholesky factor and independent $t^+(3, 0, 1)$ priors on the vector of variance parameters.

Intraclass correlation (ICC) calculation

To estimate the intraclass correlation coefficients, we employed a second Bayesian model. We fit a one-way random effects model where the outcome was the posterior samples of a given metric (from a given season) and the predictors were a global intercept parameter and a random intercept for the subsample. We can write the regression model for calculating the ICC as

$$\begin{aligned} \text{Metric sample}_{i,k} &\sim \mathcal{N}\left(\mu_{i,k}, \sigma^2\right) \\ \mu_i &= \alpha + b_k \\ b_k &\sim \mathcal{N}(0, \sigma_k^2) \end{aligned}$$

where k indexes the subsamples from the same cohort. We also fit the models separately for each of the six different metrics, using the p-epitope and cartographic distances for the novel metrics, and for metrics calculated with and without the censoring correction.

Once we fit the regression model, we calculated the posterior samples of the ICC as

$$\text{ICC} = \frac{\sigma_k^2}{\sigma_k^2 + \sigma^2}$$

and summarized this as the mean and 95% HDCI over the posterior samples, which is what we report as our ICC results.

Supplementary results

We conducted a number of additional analyses which are tangential to those presented in the main text. However, they provide more context to our results so we present them here. In the main text and in any of these results, we use abbreviated names for the influenza strains we discussed. The abbreviated strain names along with the full strain names used on GenBank are shown in Table C.1.

Table C.1: Full strain names as they appear on Genbank along with our abbreviated names for each strain included in the study sample for Chapter 4.

Short Name	**Full Name**
SC/18	A/H1N1/South Carolina/1/1918
PR/34	A/H1N1/Puerto Rico/8/1934
Wei/43	A/H1N1/Weiss/1943
FM/47	A/H1N1/Fort Monmouth/1/1947
Den/57	A/H1N1/Denver/1957
NJ/76	A/H1N1/New Jersey/8/1976
USSR/77	A/H1N1/Ussr/90/1977
Bra/78	A/H1N1/Brazil/11/1978
CA/78	A/H1N1/California/10/1978
Chi/83	A/H1N1/Chile/1/1983
Sing/86	A/H1N1/Singapore/6/1986
TX/91	A/H1N1/Texas/36/1991
Bei/95	A/H1N1/Beijing/262/1995
NC/99	A/H1N1/New Caledonia/20/1999
SI/06	A/H1N1/Solomon Islands/3/2006
Bris/07	A/H1N1/Brisbane/59/2007
CA/09	A/H1N1/California/07/2009
MI/15	A/H1N1/Michigan 45/2015

For the duration of our study, the A(H1N1) vaccine component was only updated once. The Fluzone standard dose vaccine included the CA/09 strain from 2013/14 through 2016/17, and was updated to contain the MI/15 strain in 2017/18.

The panel of historical strains for A(H1N1) stayed fairly consistent for the duration of our study, and was the same at all study sites within a given season. The 2013/14 season used a panel of 16 strains, the 2014/15 and 2015/16 seasons used a panel of 15 seasons, and the 2016/17 and 2017/18 seasons used a panel of 16 strains. The additional strain used

in 2013/14 was PR/34, which was only used during that season. In 2016/2017 after the MI/2015 virus was isolated and sequenced, that virus was added to the panel of viruses for all following study seasons.

Participant demographics for each seasonal cohort are shown in Table C.2, including sex assigned at birth, self-reported race/ethnicity, age in years, birth year, and the study site an individual enrolled at. The majority of participants in our study were white and female. Demographics were fairly consistent across the seasonal cohorts with the exception of age. The UGA study site which began recruiting patients in 2016/2017 recruited younger patients, which was not a primary focus of the other two study sites, so the 2016/17 and 2017/18 cohorts have more young people than the preceding cohorts.

Table C.2: Participant demographics for each of the seasonal cohorts in our study sample.

Characteristic	2013/14, N = 128	2014/15, N = 236	2015/16, N = 209	2016/17, N = 328	2017/18, N = 255	Overall, N = 1156
Sex assigned at birth, n (%)						
Female	97 (76)	170 (72)	156 (75)	221 (67)	142 (56)	786 (68)
Male	31 (24)	66 (28)	53 (25)	107 (33)	113 (44)	370 (32)
Race/Ethnicity, n (%)						
White	89 (70)	158 (67)	147 (70)	235 (72)	203 (80)	832 (72)
Black or African American	24 (19)	39 (17)	35 (17)	43 (13)	17 (7)	158 (14)
Other	13 (10)	17 (7)	13 (6)	31 (9)	25 (10)	99 (9)
Hispanic or Latino	2 (2)	21 (9)	13 (6)	18 (5)	10 (4)	64 (6)
Unknown	0 (0)	1 (0)	1 (0)	1 (0)	0 (0)	3 (0)
Age (years), Median (Min - Max)	58 (23 - 81)	57 (20 - 83)	56 (21 - 84)	44 (18 - 82)	25 (12 - 83)	48 (12 - 84)
Birth year, Median (Min - Max)	1956 (1932 - 1990)	1957 (1932 - 1994)	1960 (1932 - 1994)	1972 (1934 - 1998)	1991 (1934 - 2006)	1968 (1932 - 2006)
Study site, n (%)						
FL	60 (47)	150 (64)	128 (61)	119 (36)	0 (0)	457 (40)
PA	68 (53)	86 (36)	81 (39)	64 (20)	0 (0)	299 (26)
UGA	0 (0)	0 (0)	0 (0)	145 (44)	255 (100)	400 (35)

Summary antibody landscapes for all seasons

Here, we include the summary antibody landscapes for all of the seasons in our study. The general trends were the same across all landscapes, so we chose to report the 2016/2017 landscape in the main text due to the large number of participants.

- Figure C.1 shows the summary landscape for the 2013/14 seasonal cohort.
- Figure C.2 shows the summary landscape for the 2014/15 seasonal cohort.
- Figure C.3 shows the summary landscape for the 2015/16 seasonal cohort.
- Figure C.4 shows the summary landscape for the 2016/17 seasonal cohort. This figure is also included in the main text but is reproduced here for easier comparison.
- Figure C.5 shows the summary landscape for the 2017/18 seasonal cohort.

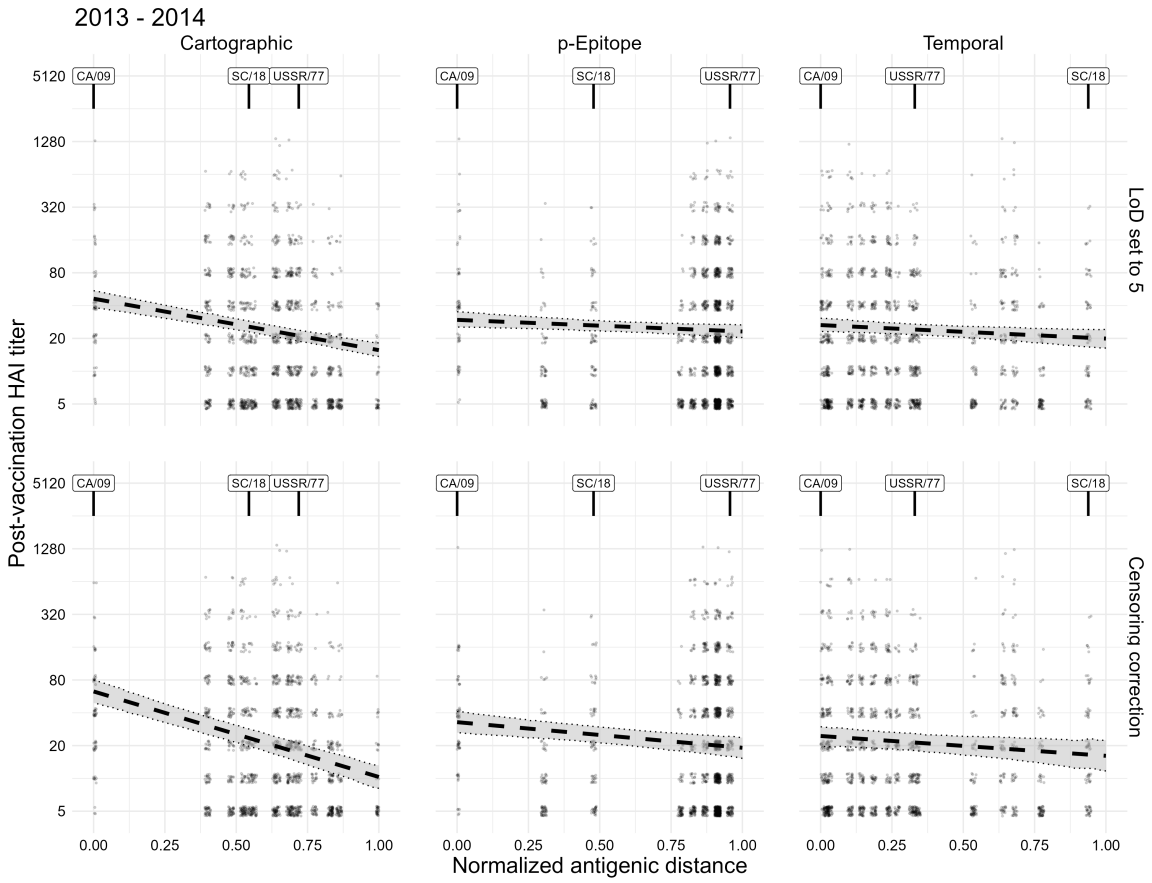

Figure C.1: Raw data and summary antibody landscapes for the 2013 - 2014 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (CA/09 in 2013/14). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape.

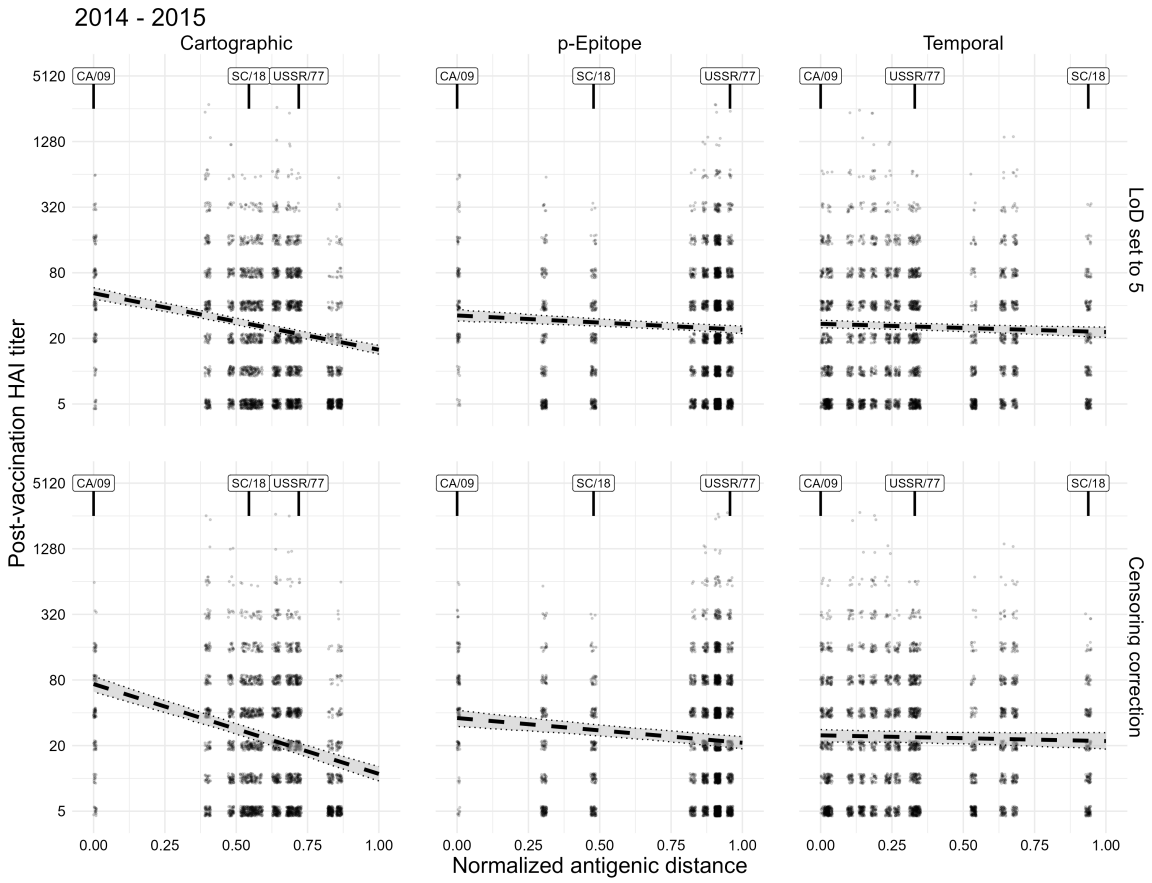

Figure C.2: Raw data and summary antibody landscapes for the 2014 - 2015 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (CA/09 in 2014/15). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape.

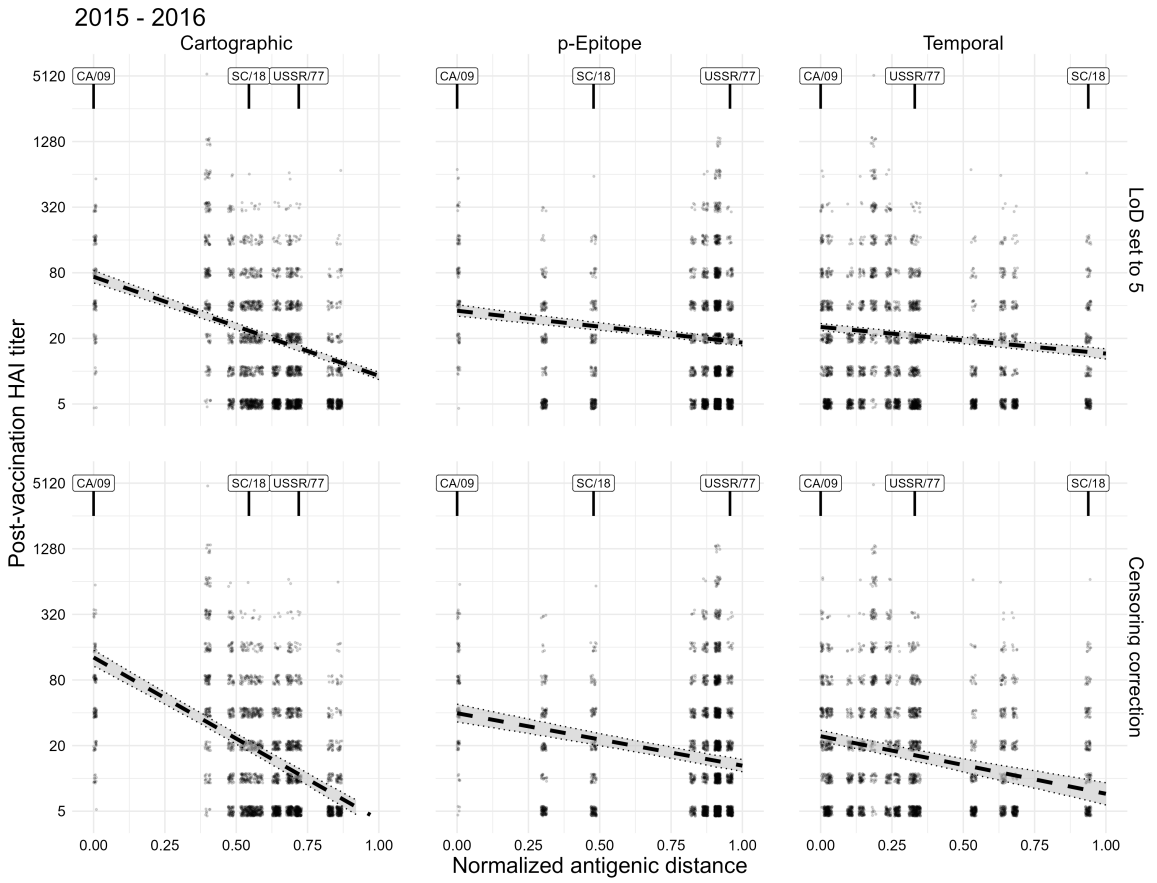

Figure C.3: Raw data and summary antibody landscapes for the 2014 - 2015 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (CA/09 in 2014/15). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape.

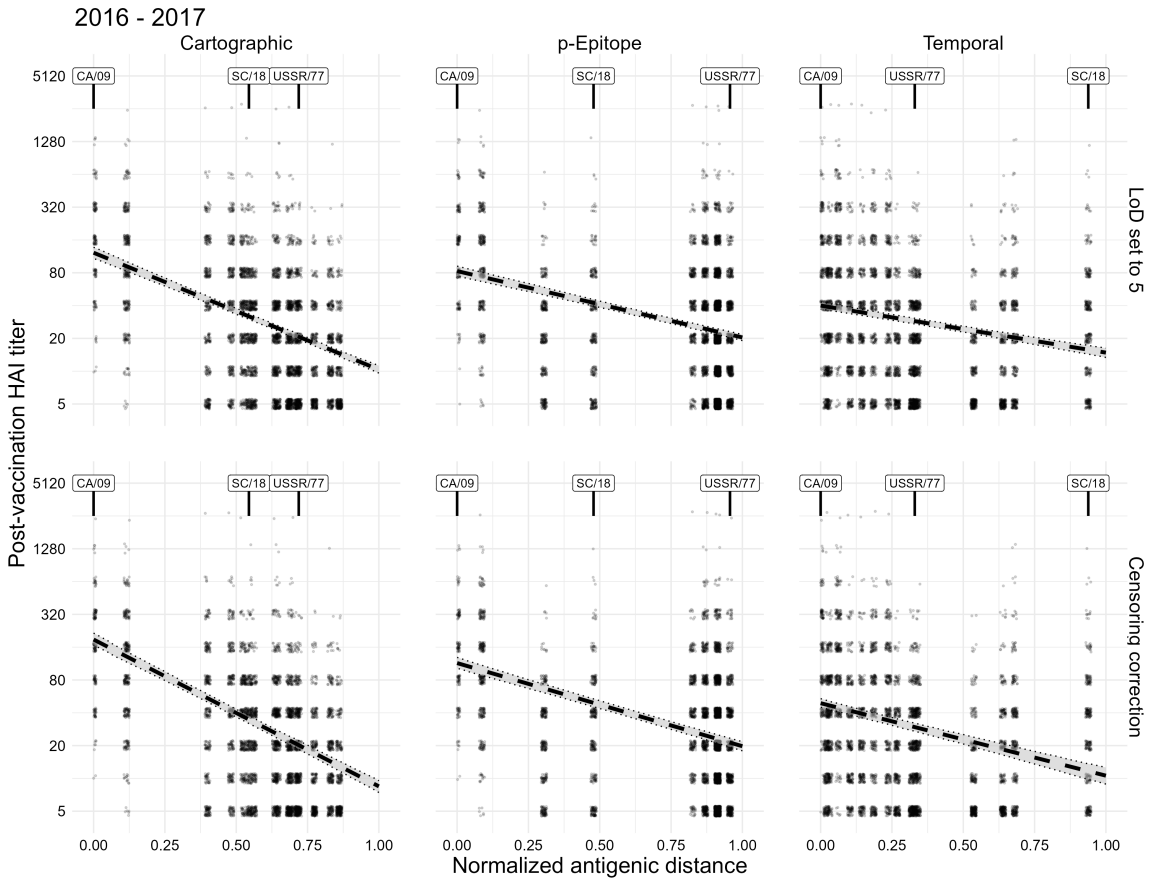

Figure C.4: Raw data and summary antibody landscapes for the 2016 - 2017 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (CA/09 in 2016/17). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape.

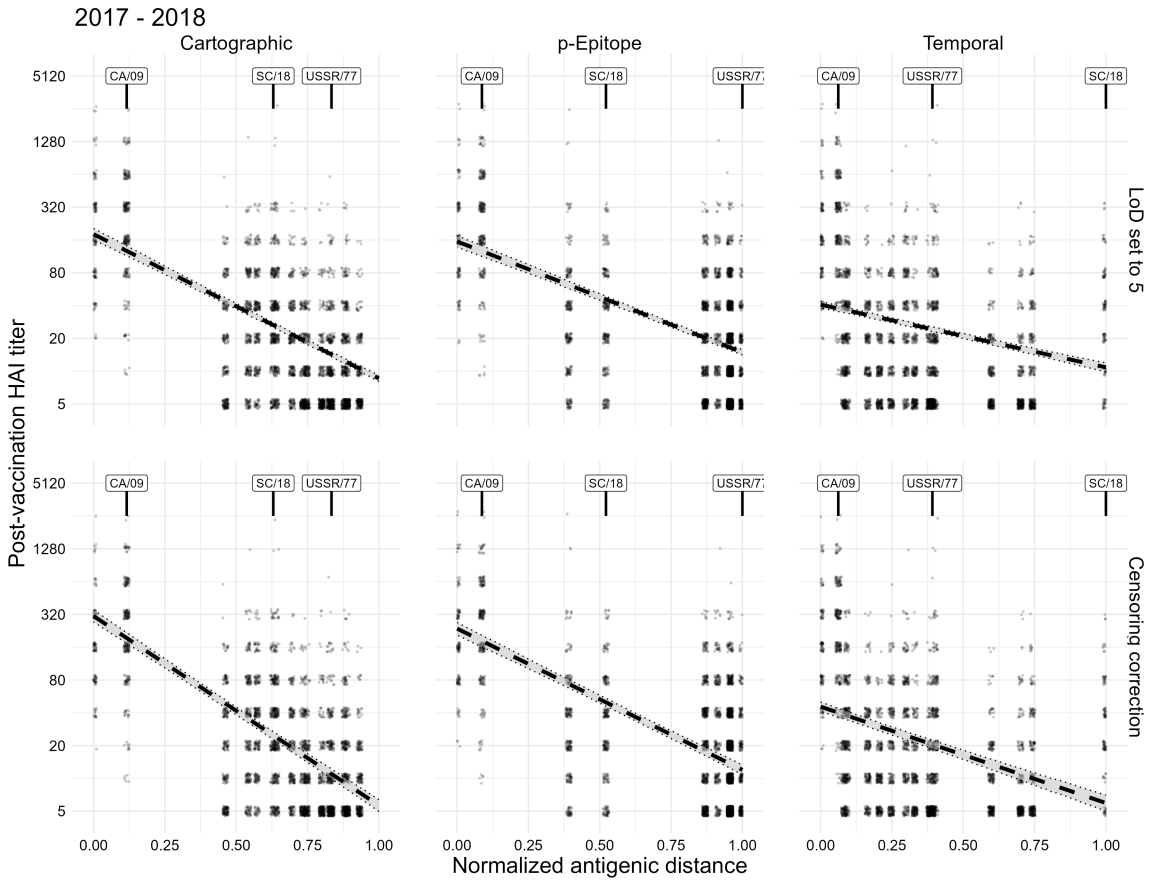

Figure C.5: Raw data and summary antibody landscapes for the 2017 - 2018 influenza season. Each point shows the post-vaccination HAI titer to a specific strain with a specified normalized antigenic distance from the vaccine strain (MI/15 in 2017/18). The dashed line and envelope show the mean and 95% credible interval (CrI) of the posterior summary antibody landscape.

Summary landscape metrics for all seasons

Table C.3 shows the summary metrics for all of the seasonal cohorts. The trends for the other seasons are similar to the 2016/17 results, but we include the other seasons here for completeness. In general the metrics are overall more optimistic in the 2016/2017 and 2017/18 cohorts, but this is because those cohorts had much more young people relative to the total sample size for that season. We elected to present these results in the main text because the signal is consistent with the other seasons, but easier to see. We chose the 2016/17 cohort for the main results because it used the same vaccine as the older cohorts and contains a mix of individuals from both, so it is easier to compare to the older cohorts directly. We only show the censoring-corrected metrics here.

Table C.3: Current and novel vaccine immunogenicity metrics for each season, shown both with and without the censoring correction. The 'Current' metric set uses the homologous GMT for magnitude, the seroconversion rate across the historical panel for breadth, and the GMT across all strains for the total strength. The 'Novel' metric sets are based on the corresponding summary landscape. The novel metrics are regression line intercept for magnitude, proportion of the line above a titer of 40 for breadth, and area under the curve for total strength. All metrics were derived from bayesian regression models and numbers shown are the posterior mean and 95% CrI.

Season	**Metric Set**	**Magnitude**	**Breadth**	**Total Strength**
2013 - 2014	Current	3.45 (3.15, 3.75)	0.09 (0.07, 0.10)	2.09 (1.96, 2.21)
	Novel (Cartographic)	3.65 (3.30, 4.00)	0.25 (0.15, 0.37)	2.35 (2.09, 2.62)
	Novel (p-Epitope)	2.71 (2.39, 3.04)	0.00 (0.00, 0.00)	2.32 (2.08, 2.57)
	Novel (Temporal)	2.29 (1.98, 2.58)	0.00 (0.00, 0.00)	1.99 (1.72, 2.29)
2014 - 2015	Current	3.61 (3.40, 3.82)	0.08 (0.07, 0.08)	2.26 (2.18, 2.34)
	Novel (Cartographic)	3.88 (3.65, 4.12)	0.32 (0.26, 0.39)	2.51 (2.35, 2.67)
	Novel (p-Epitope)	2.84 (2.58, 3.08)	0.01 (0.00, 0.06)	2.46 (2.29, 2.63)
	Novel (Temporal)	2.31 (2.11, 2.49)	0.00 (0.00, 0.00)	2.23 (2.06, 2.40)
2015 - 2016	Current	3.93 (3.73, 4.13)	0.07 (0.06, 0.08)	1.73 (1.64, 1.82)
	Novel (Cartographic)	4.69 (4.42, 4.90)	0.34 (0.31, 0.38)	2.21 (2.05, 2.37)
	Novel (p-Epitope)	2.99 (2.72, 3.27)	0.03 (0.00, 0.14)	2.19 (2.00, 2.37)
	Novel (Temporal)	2.29 (2.12, 2.47)	0.00 (0.00, 0.00)	1.41 (1.20, 1.60)
2016 - 2017	Current	5.22 (5.07, 5.39)	0.10 (0.09, 0.11)	2.64 (2.58, 2.70)
	Novel (Cartographic)	5.23 (5.06, 5.43)	0.50 (0.48, 0.52)	3.00 (2.87, 3.11)
	Novel (p-Epitope)	4.52 (4.36, 4.69)	0.60 (0.55, 0.64)	3.25 (3.13, 3.37)
	Novel (Temporal)	3.29 (3.17, 3.43)	0.14 (0.08, 0.19)	2.19 (2.05, 2.34)
2017 - 2018	Current	5.57 (5.40, 5.75)	0.10 (0.09, 0.11)	2.14 (2.06, 2.23)
	Novel (Cartographic)	5.95 (5.77, 6.14)	0.51 (0.50, 0.53)	3.06 (2.94, 3.18)
	Novel (p-Epitope)	5.57 (5.37, 5.75)	0.59 (0.57, 0.62)	3.41 (3.30, 3.54)
	Novel (Temporal)	3.19 (3.04, 3.33)	0.07 (0.02, 0.11)	1.72 (1.58, 1.86)

Subsample metrics plot for each season

We also repeated the subsampling analysis using data from the other cohorts. Again, because the signal is stronger in the 2016/17 and 2017/18 cohorts we presented those results in the main text. The results for the other seasons showed a similar pattern, and we include them here for completeness.

- Figure C.6 shows the posterior distribution of the vaccine metrics for all of the subsamples from the 2013/14 cohort data.
- Figure C.7 shows the posterior distribution of the vaccine metrics for all of the subsamples from the 2014/15 cohort data.
- Figure C.8 shows the posterior distribution of the vaccine metrics for all of the subsamples from the 2015/16 cohort data.
- Figure C.9 shows the posterior distribution of the vaccine metrics for all of the subsamples from the 2016/17 cohort data. This figure is included in the main text but reproduced here for easier comparisons.
- Figure C.10 shows the posterior distribution of the vaccine metrics for all of the subsamples from the 2017/18 cohort data.

All of these figures show the current metrics for magnitude, breadth, and total strength, and the novel metrics using both the cartographic distance and the p-Epitope distance for each subsample. The black circles show samples from the posterior distribution of each metric. The red dotted line shows the overall mean metric estimate across the subsample, and the red x for each subsample shows the mean metric estimate for that subsample. In general, metrics with lower ICCs (less variation explained by subsample grouping) will have group means that are more similar to the overall mean. We show only 1000 posterior samples for each subsample/metric to avoid unnecessary overplotting.

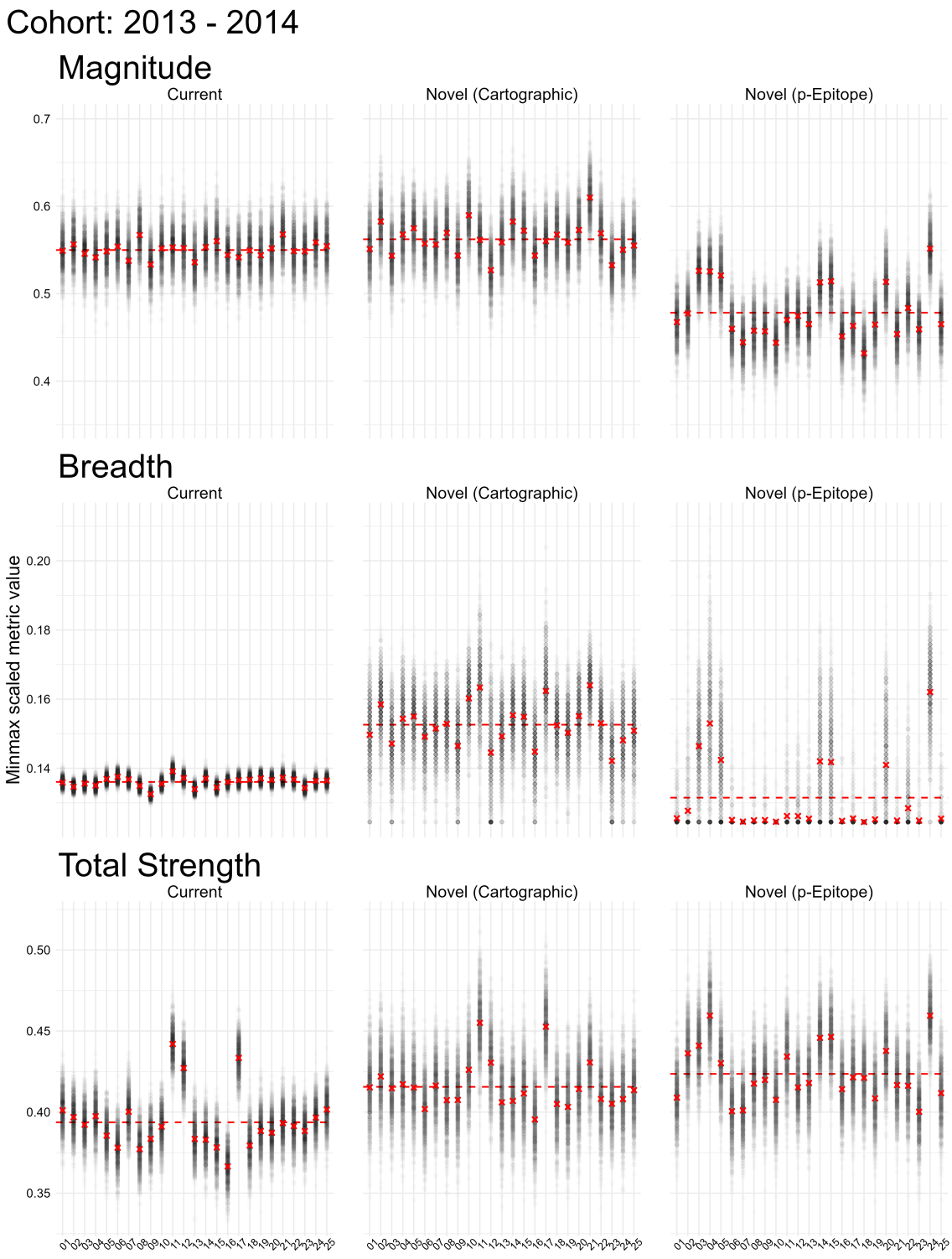

Figure C.6: Estimated immunogenicity metrics for each simulated lab drawn from the 2013/14 subcohort data.

Figure C.7: Estimated immunogenicity metrics for each simulated lab drawn from the 2014/15 subcohort data.

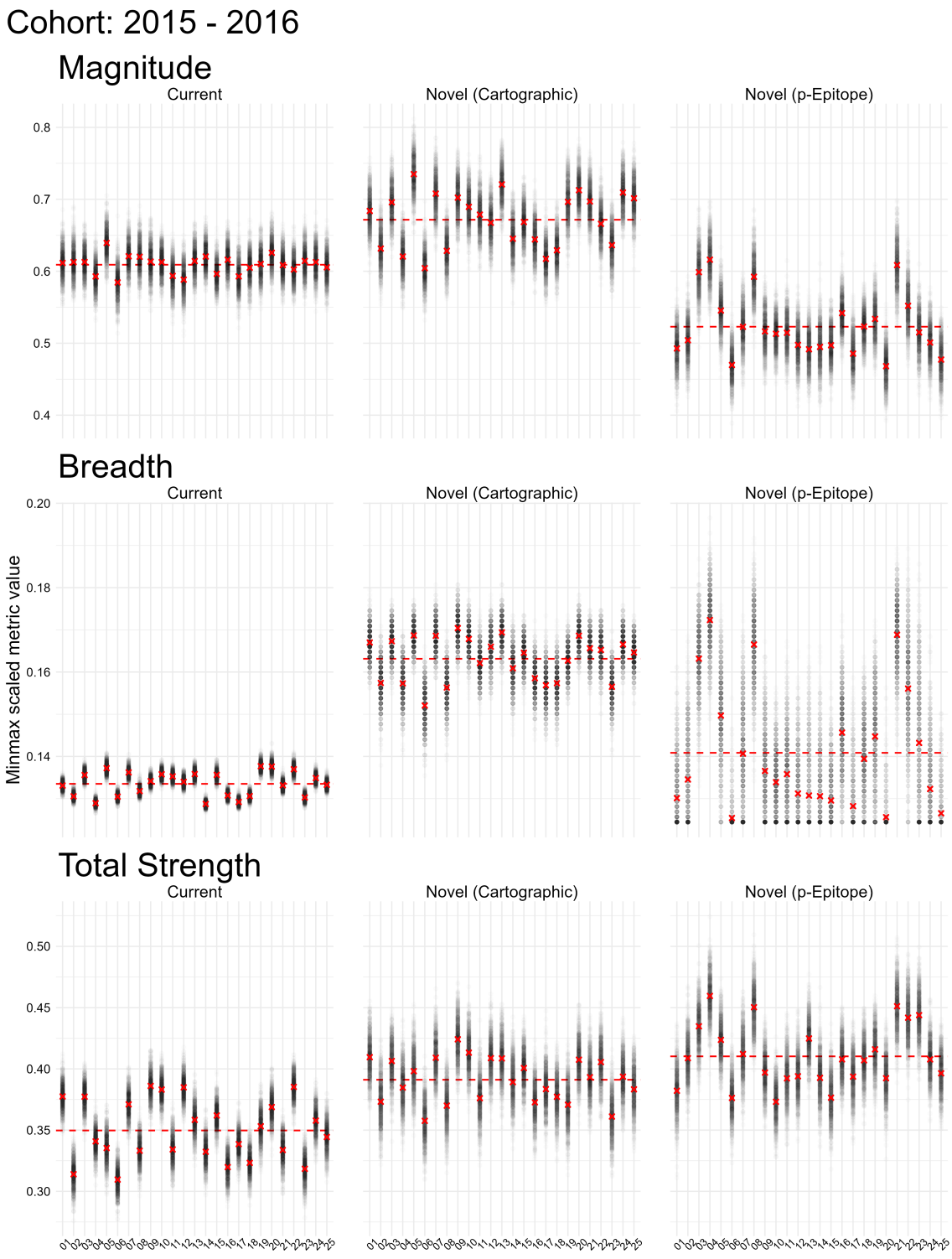

Figure C.8: Estimated immunogenicity metrics for each simulated lab drawn from the 2015/16 subcohort data.

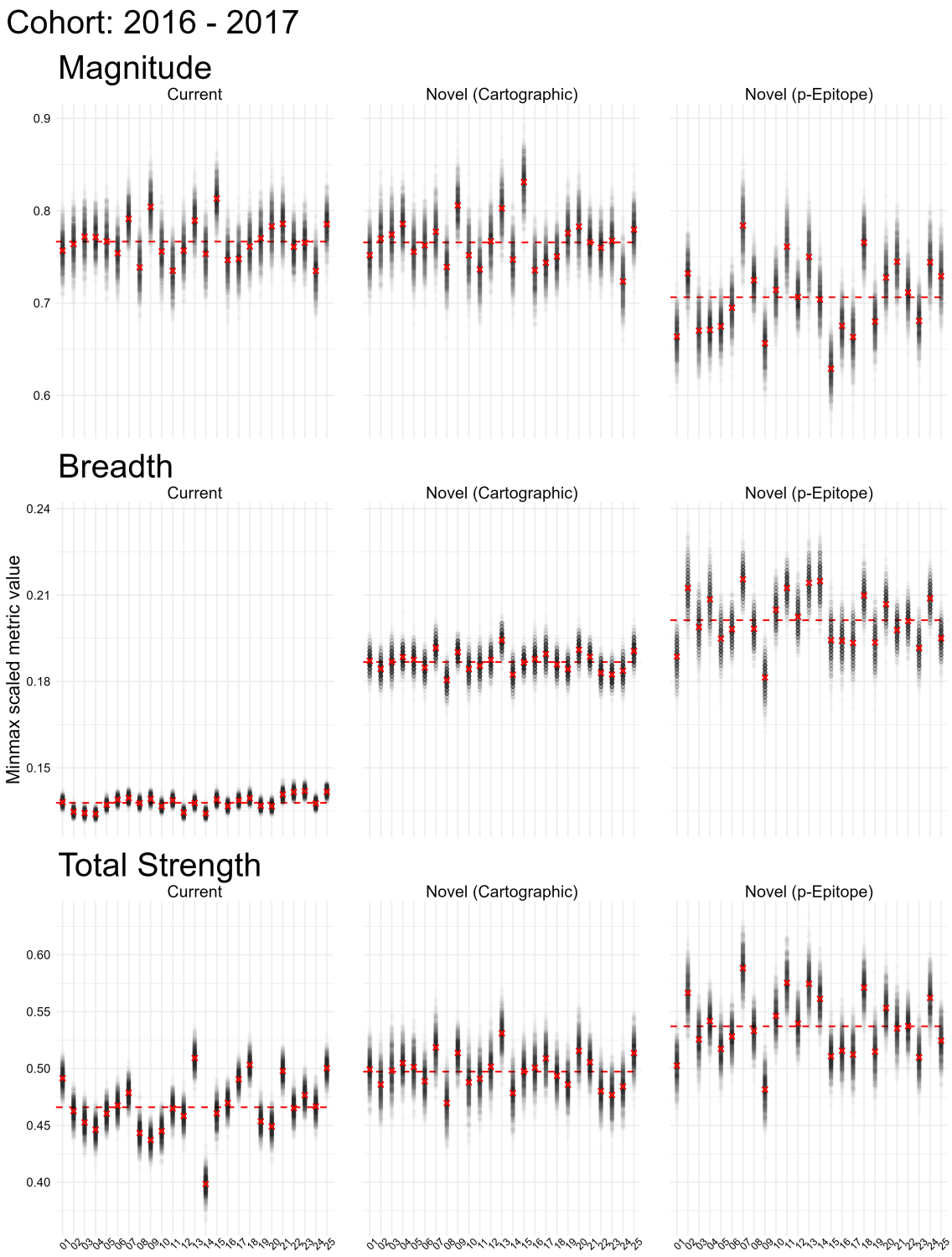

Figure C.9: Estimated immunogenicity metrics for each simulated lab drawn from the 2013/14 subcohort data. This figure is included in the main text but reproduced here for easier comparisons.

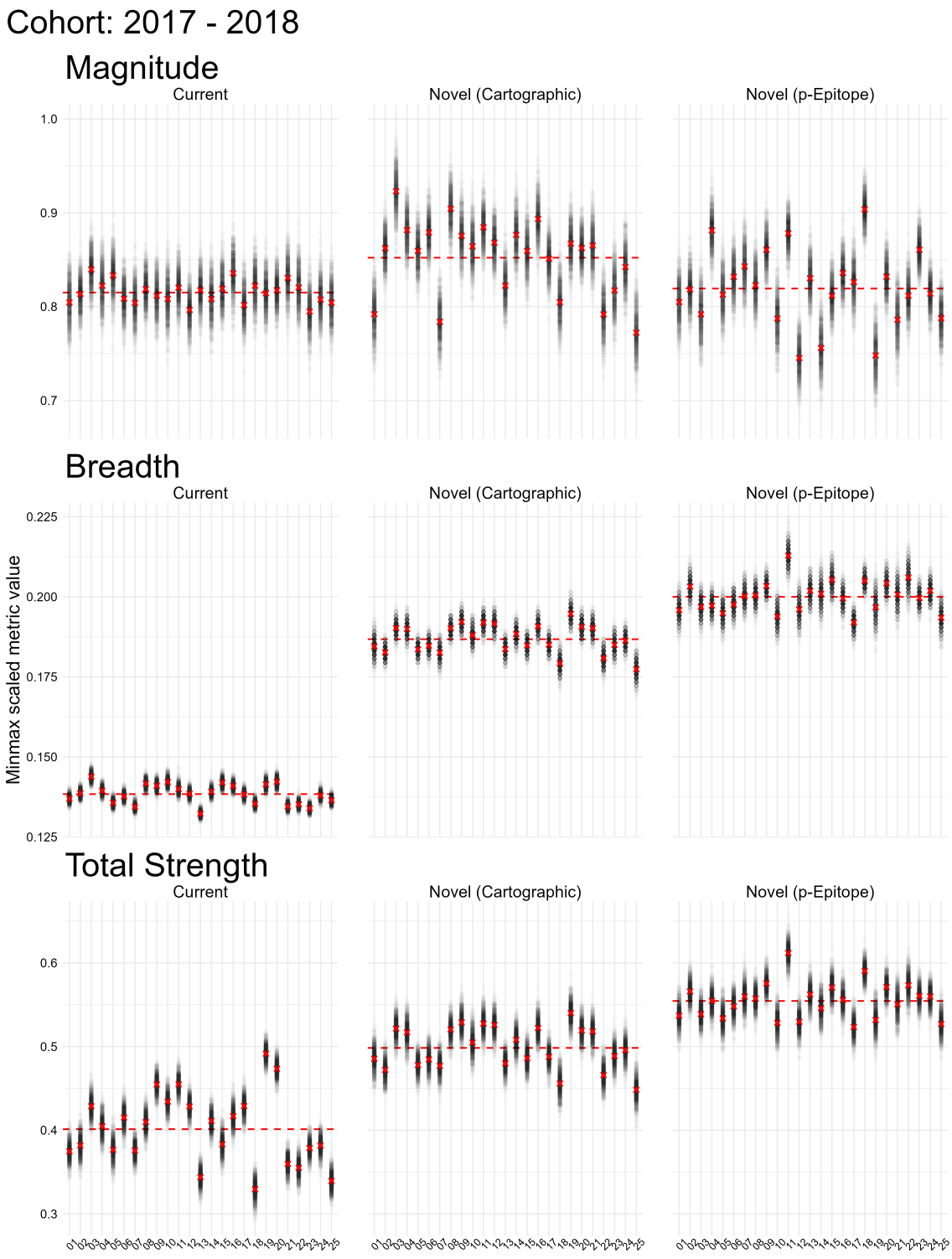

Figure C.10: Estimated immunogenicity metrics for each simulated lab drawn from the 2017/18 subcohort data.

Intraclass correlation (ICC) analysis for all seasons

To quantify the results in the subsampling figures, we also include the ICC estimates for each of the seasonal cohorts. Yet again, the results were similar to the 2016/17 cohort with some outliers and expected sampling variation, but we include them here for completeness in Table C.4.

Table C.4: Intraclass correlation coefficients (ICCs) for consistency across the subsampled studies. Each number shown is the posterior mean and 95% CrI for the ICC, which we calculated as the between-groups variance for subgroups divided by the between-groups variance for subgroups plus the residual error variance. An ICC closer to zero indicates that little of the variance in metric estimates is due to variability across subsamples, while an ICC closer to one indicates that variability across subsamples makes up the majority of the variation.

Season	**Metric Set**	**Magnitude**	**Breadth**	**Total Strength**
2013 - 2014	Current	0.14 (0.08, 0.22)	0.54 (0.36, 0.66)	0.71 (0.54, 0.80)
	Novel (Cartographic)	0.35 (0.24, 0.48)	0.30 (0.19, 0.44)	0.39 (0.24, 0.47)
	Novel (p-Epitope)	0.66 (0.49, 0.73)	0.63 (0.50, 0.74)	0.52 (0.36, 0.63)
2014 - 2015	Current	0.42 (0.29, 0.54)	0.51 (0.35, 0.62)	0.90 (0.82, 0.93)
	Novel (Cartographic)	0.51 (0.38, 0.60)	0.42 (0.28, 0.55)	0.64 (0.46, 0.74)
	Novel (p-Epitope)	0.63 (0.47, 0.73)	0.59 (0.40, 0.70)	0.69 (0.51, 0.76)
2015 - 2016	Current	0.30 (0.18, 0.44)	0.89 (0.80, 0.93)	0.85 (0.75, 0.90)
	Novel (Cartographic)	0.72 (0.60, 0.80)	0.65 (0.48, 0.74)	0.57 (0.44, 0.67)
	Novel (p-Epitope)	0.74 (0.58, 0.82)	0.72 (0.59, 0.80)	0.71 (0.55, 0.80)
2016 - 2017	Current	0.55 (0.36, 0.65)	0.79 (0.68, 0.85)	0.88 (0.82, 0.93)
	Novel (Cartographic)	0.59 (0.43, 0.70)	0.53 (0.38, 0.67)	0.54 (0.39, 0.66)
	Novel (p-Epitope)	0.79 (0.69, 0.85)	0.75 (0.62, 0.82)	0.78 (0.66, 0.86)
2017 - 2018	Current	0.30 (0.21, 0.40)	0.86 (0.75, 0.91)	0.94 (0.87, 0.96)
	Novel (Cartographic)	0.81 (0.71, 0.87)	0.80 (0.65, 0.87)	0.80 (0.70, 0.87)
	Novel (p-Epitope)	0.82 (0.73, 0.88)	0.72 (0.61, 0.80)	0.74 (0.61, 0.83)

Appendix D

Extended methods for antigenic distance calculation

We calculated pairwise antigenic distances using multiple methods for all of the influenza strains used in the UGAFluVac cohort from the 2013/14 season through the 2017/18 season. The methods we support (either here or in the extended R package at `https://github.com/ahgroup/agdist`) include the following methods: temporal, Hamming (and any other string-based distances supported by the `stringdist` R package), p-epitope, cophenetic distances from trees, cartographic distances from antigenic maps, and substitution matrix distances such as Grantham's distance. Here, we will explain in detail the methods we use for calculating antigenic distances.

Sequence retrieval

We manually obtained hemagglutinin (HA) amino acid sequences for each of the strains used in UGAFluVac from either the U.S. National Center for Biotechnology Information (NCBI)'s GenBank database [223, 224], the UniProt dataset [225], or GISAID's EpiFlu database [226, 227]. The attribution and accession numbers for each strain are listed in Table D.1.

GenBank and UniProt information

Most of the sequences we used from GenBank and UniProt are not associated with particular publications and are only able to be referenced via their accession numbers. The following

sequences have formal references: AAD17229.1 [228]; AAA67338.1 [229]; AAP34324.1 [230]; ADE28750.1 [231]; ACP41953.1 [232]; ABQ97200.1 [233]; AAA62338.1 [234]; AIW60702.1 [235]; P03460 and P03461 [236]; and P12443 [237].

GISAID information

The sequences we used from GISAID are accessible via GISAID Identifier EPI_SET_250609vz and DOI 10.55876/gis8.250609vz. EPI_SET_250609vz is composed of 9 individual genome segments. The collection dates range from 1999-12-31 to 2023-04-25. Data were collected in 7 countries and territories.

All genome sequences and associated metadata in this dataset are published in GISAID's EpiFlu database. To view the contributors of each individual sequence with details such as accession number, virus name, collection date, originating lab and submitting lab, and the list of authors, visit 10.55876/gis8.250609vz.

Table D.1: Strain names, source locations, and accession numbers for each of the sequences we used in our dataset

Strain Name	**Source**	**Accession #**
A/H1N1/South Carolina/1/1918	GenBank	AAD17229.1
A/H1N1/Puerto Rico/8/1934	GenBank	AGU93019.1
A/H1N1/Weiss/1943	GenBank	ABD79101.1
A/H1N1/Fort Monmouth/1/1947	GenBank	AAA67338.1
A/H1N1/Denver/1957	GenBank	ABD15258.1
A/H1N1/New Jersey/8/1976	GenBank	AGB51356.1
A/H1N1/Ussr/90/1977	GenBank	ABD95350.1
A/H1N1/Brazil/11/1978	GenBank	ABO38065.1
A/H1N1/California/10/1978	GenBank	ABP49338.1
A/H1N1/Chile/1/1983	GenBank	ABO38340.1
A/H1N1/Singapore/6/1986	GenBank	ABO38395.1
A/H1N1/Texas/36/1991	GenBank	ACF41933.1

Table D.1: Strain names, source locations, and accession numbers for each of the sequences we used in our dataset (continued)

Strain Name	**Source**	**Accession #**
A/H1N1/Beijing/262/1995	GenBank	ACF41867.1
A/H1N1/New Caledonia/20/1999	GenBank	AAP34324.1
A/H1N1/Solomon Islands/3/2006	GenBank	ABU99109.1
A/H1N1/Brisbane/59/2007	GenBank	ADE28750.1
A/H1N1/California/07/2009	GenBank	ACP41953.1
A/H1N1/Michigan 45/2015	GenBank	AMV49034.1
A/H1N1/Brisbane/02/2018	GISAID	EPI1415369
A/H1N1/Guangdong-Maonan/SWL1536/2019	GISAID	EPI3133357
A/H1N1/Victoria/2570/2019	GenBank	WEY08940.1
A/H3N2/Hong Kong/8/1968	GenBank	ABQ97200.1
A/H3N2/Port Chalmers/1/1973	GenBank	ABE12532.1
A/H3N2/Texas/1/1977	GenBank	AFM68965.1
A/H3N2/Mississippi/1/1985	GenBank	AAA62338.1
A/H3N2/Sichuan/2/1987	GenBank	AFG72085.1
A/H3N2/Shandong/9/1993	GenBank	AFH00285.1
A/H3N2/Nanchang/933/1995	GenBank	AFG72625.1
A/H3N2/Sydney/5/1997	GenBank	ACO95259.1
A/H3N2/Panama/2007/1999	GenBank	ABF21273.1
A/H3N2/Fujian/411/2002	GenBank	AFG72823.1
A/H3N2/New York/55/2004	GenBank	ACF41900.1
A/H3N2/Wisconsin/67/2005	GenBank	AHG96791.1
A/H3N2/Brisbane/10/2007	GenBank	AIW60702.1
A/H3N2/Uruguay/716/2007	GenBank	ACD47213.1
A/H3N2/Perth/16/2009	GenBank	ACS71642.1
A/H3N2/Victoria/361/2011	GenBank	AIU46088.1

Table D.1: Strain names, source locations, and accession numbers for each of the sequences we used in our dataset (continued)

Strain Name	**Source**	**Accession #**
A/H3N2/Texas/50/2012	GenBank	AGL07159.1
A/H3N2/Switzerland/9715293/2013	GISAID	EPI530687
A/H3N2/Hong Kong/4801/2014	GISAID	EPI834581
A/H3N2/Singapore/infimh-16-0019/2016	GISAID	EPI780183
A/H3N2/Kansas/14/2017	GenBank	AVG71503.1
A/H3N2/South Australia/34/2019	GISAID	EPI1387331
A/H3N2/Hong Kong/2671/2019	GenBank	WMW30924.1
A/H3N2/Tasmania/503/2020	GenBank	WMW30850.1
A/H3N2/Darwin/9/2021	GenBank	WND60806.1
B/Lee/1940	UniProt	P03460
B/Maryland/1959	UniProt	P03461
B/Singapore/1964	UniProt	P12443
B/Victoria/02/1987	UniProt	A4D5N9
B/Hong Kong/330/2001	GenBank	ABL77178.1
B/Malaysia/27127/2004	GenBank	AFJ80733.1
B/Victoria/326/2006	GenBank	AGX18732.1
B/Brisbane/60/2008	GenBank	AFH57909.1
B/Colorado/06/2017	GenBank	ASK81305.1
B/Washington/02/2019	GenBank	WIM08940.1
B/Michigan/01/2021	GenBank	WMW30908.1
B/Austria/1359417/2021	GISAID	EPI1868375
B/Yamagata/16/1988	GenBank	ABL77255.1
B/Harbin/7/1994	GenBank	ACR15721.1
B/Sichuan/379/1999	GISAID	EPI2085837
B/Florida/4/2006	GenBank	ACA33493.1

Table D.1: Strain names, source locations, and accession numbers for each of the sequences we used in our dataset (continued)

Strain Name	**Source**	**Accession #**
B/Wisconsin/01/2010	GenBank	AET22057.1
B/Texas/06/2011	GenBank	AGI64713.1
B/Massachusetts/02/2012	GenBank	AGL06036.1
B/Phuket/3073/2013	GISAID	EPI3555941

Sequence processing

We saved all of our sequences in the file `data/raw/raw-sequences.xlsx`, which we imported into `R` using the `readxl` package [73]. We standardized sequence formatting by converting all sequences to lowercase and removing all white space characters from the sequence. We aligned the sequences separately for each subtype (A/H1N1, A/H3N2, and all B sequences pooled together) using a two-stage alignment method. Since some sequences were not full length, we first performed a multiple sequence alignment with the `msa` package [238] using the MUSCLE algorithm [239]. From a brief examination, we noticed several inconsistencies across the alignment generated using MUSCLE, so we only used the multiple alignment to find the start and end position of the non-full length sequences. The only sequences which were not full length were A/H3N2/Mississippi/1/1985, B/Singapore/1964, and B/Sichuan/379/1999.

Once we estimated the start and end positions of the non-full length sequences relative to the full length sequences (which made up the majority of the dataset), we manually aligned the sequences for each subtype based on the known deletions for different clades of influenza viruses. Since all of the HA sequences were very similar in length, with known deletions that explained the differences, the manual alignment was easier than dealing with the errors introduced by an MSA algorithm.

All of the A(H3N2) sequences were aligned with a final length of 566 amino acids (starting from the beginning of the signal peptide, which we included in our sequences) after we found

the starting position for the MI/85 sequence. In order to align the A(H1N1) sequences, we added a gap character at position 148 (position 130 after the 17 residue signal peptide is accounted for) [240] if the sequence was length 565. All A(H1N1) sequences were full length and were either 565 or 566 residues long. The type B sequences were the most complicated to align because we needed to account for the 1DEL, 2DEL, and 3DEL lineages. We based our alignment off the gaps in the Lee40 standard numbering: if a sequence was length 585, we did not need to add gaps; if a sequence was length 584 it was a 1DEL strain and we added a gap at position 163; if a sequence was length 583 it was a 2DEL strain and we added gaps at positions 163 and 165; and finally if a sequence was length 582 it was a 3DEL strain and we added gaps at positions 163, 164, and 164 [241, 242]. After alignment, all type B sequences were 585 residues long, including the signal peptide.

Distance calculation

Once the sequences are aligned (within each subtype), we can calculate the different antigenic distances. All of the antigenic distance metrics we used are based on the sequences and metadata, except for the cartographic distance which relies on actual titer data. We'll discuss the data that is necessary for cartography in that subsection.

As a brief note, we applied our sequence-based distances to the entire HA sequence (except for p-Epitope which only uses specific sites and is discussed in detail). Some people have suggested applying sequence differences to NA as well, or to the entire genome, or to specific parts of HA like HA1 or HA2. Using the entire uncleaved HA sequence is the easiest and appears to work well enough, although a detailed comparison might be necessary to determine the best way to calculate these distances.

Temporal distance

The simplest distance metric we used, what we call the temporal distance, does not use the actual sequence or titer information about any strains. Instead, the temporal distance is

based only on the years of isolation between two strains:

$$d_{\text{temporal}}(\text{strain } i, \text{strain } j) = \mid \text{isolation year}_{\text{strain } i} - \text{isolation year}_{\text{strain } j} \mid.$$

Some studies have proposed using a directional temporal difference (i.e. without the absolute value) that allows for positive and negative temporal distances [243]. This might be useful in some cases, but we caution using this definition with the word *distance* because a mathematical distance cannot take on negative values. Some authors have also used a much simpler version of the temporal distance where they place strains in order of isolation year, and treat the distance between any two consecutive ordered strains as 1, regardless of the difference in isolation years between those strains. This approach should not be used in practice, as the difference in years is simple to calculate and more accurate.

However, we encourage authors **not to use any temporal distances**. Sequences for all well-characterized influenza strains are available, and even the simple Hamming distance on the HA sequence tends to perform better than the temporal distance. We know that influenza evolution does not occur at a constant rate, particularly for lineages like A(H1N1) which have multiple clades that are better defined by genetic similarity than time – indeed, using the temporal metric for H1N1 strains assigns a large distance value to 1918 pandemic-like strains and 2009-pandemic like strains, when these lineages are more similar to each other than they are to the 1999 seasonal virus-like circulating strains with an acquired deletion.

Temporal distances might work okay for lineages like A(H3N2), but they are less accurate than any distances based on actual metrics of genetic or antigenic change, and so we believe they should not be used. We include them in our comparisons to show how different they are from other distance metrics, not because they should be interpreted as a valid antigenic distance metric.

Hamming distance and string distances

The simplest metric for calculating the antigenic distance from sequence data is the Hamming distance [221]. The Hamming distance between two strings is simply the number of spots between those two (aligned) strings that are different, so in our case, it is the number of mutant residues between two HA sequences. We can also compute the normalized Hamming distance by dividing the Hamming distance for two strings by the length of those strings.

In general, we use the fast implementation of the Hamming distance in the `stringdist` package [244]. The `stringdist` package also supports other information-theoretic distance metrics between two strings such as the (restricted) Damerau-Levenshtein distance and multiple distances based on q-grams. There is no current evidence showing that any of these other distance metrics are preferable to the Hamming distance for applications to influenza evolution, but we refer the interested reader to the `stringdist` package for more information.

p-Epitope distance

The p-Epitope distance is based on computing normalized Hamming distances across five immunodominant epitope sites on the influenza HA, and has been shown to correlate with vaccine effectiveness in the literature [30–32]. Defining the residue sites is the difficult part; once the residue sites are defined, the normalized Hamming distance at each site between two sequences is easy to calculate.

The original papers using p-Epitope for A(H3N2) distance calculation [30] and A(H1N1) distance calculation [31] do not include the list of residues included in each epitope, and the citations they refer to for these epitope sites appear to be permanently broken. Thus, we used previously compiled tables of residue sites based on combining computational modeling strategies with earlier structural biochemistry results for A(H1N1) [245] with numbering following the A/California/04/2009; and A(H3N2) [246]. Influenza A numbering is standardized and begins after the signal peptide which is 17 residues long for A(H1N1) and 16 residues long for A(H3N2) [222]. We reproduce the residue sites from those sources

in Table D.2.

Table D.2: Subtype-specific epitope sites used for p-epitope distance for influenza A.

Model	**Epitope**	**Residues**
A(H1N1)	A	118 120 121 122 126 127 128 129 132 133 134 135 137 139 140 141 142 143 146 147 149 165 252 253
	B	124 125 152 153 154 155 156 157 160 162 183 184 185 186 187 189 190 191 193 194 195 196
	C	34 35 36 37 38 40 41 43 44 45 269 270 271 272 273 274 276 277 278 283 288 292 295 297 298 302 303 305 306 307 308 309 310
	D	89 94 95 96 113 117 163 164 166 167 168 169 170 171 172 173 174 176 179 198 200 202 204 205 206 207 208 209 210 211 212 213 214 215 216 222 223 224 225 226 227 235 237 241 243 244 245
	E	47 48 50 51 53 54 56 57 58 66 68 69 70 71 72 73 74 75 78 79 80 82 83 84 85 86 102 257 258 259 260 261 263 267
A(H3N2)	A	122 124 126 130 131 132 133 135 137 138 140 142 143 144 145 146 150 152 168
	B	128 129 155 156 157 158 159 160 163 165 186 187 188 189 190 192 193 194 196 197 198
	C	44 45 46 47 48 50 51 53 54 273 275 276 278 279 280 294 297 299 300 304 305 307 308 309 310 311 312
	D	96 102 103 117 121 167 170 171 172 173 174 175 176 177 179 182 201 203 207 208 209 212 213 214 215 216 217 218 219 226 227 228 229 230 238 240 242 244 246 247 248
	E	57 59 62 63 67 75 78 80 81 82 83 86 87 88 91 92 94 109 260 261 262 265

Calculating the p-Epitope distance for influenza B is a bit trickier. Influenza B has different major immunodominant epitopes than influenza A [247], but previous studies still found a correlation between p-epitope distance and VE for influenza B (likely because the Hamming distance would also be informative and the epitope sites captured by p-Epitope partially overlap with the real immunodominant sites) [32]. We cannot ensure that our p-Epitope definition is consistent with the previous literature, because the previous literature does not provide an example of the non-standard numbering they use to define the residue sites used for flu B strains. We used numbering based on our B/Lee40 numbering alignment [241], and since Pan and Deem report using only HA1 in their alignment, we removed the first 15 residues, which comprise the signal peptide for influenza B [248, 249]. We then used the residues they identified in their table, which is reproduced below in Table D.3 [32].

We also found that, in general, the two B lineages were much more cross-reactive than the influenza A subtypes, and we preferred to use trees and cartographic maps built on all B strains simultaneously (data not shown, a formal comparison is likely necessary to make

Table D.3: Lineage-specific epitope sites used for p-epitope distance for influenza B.

Model	Epitope	Residues
Victoria	A	121 122 123 125 126 134 135 136 137 139 141 142 144 146 147 148 149 150 151 155 157 177
Victoria	B	127 129 133 160 161 162 163 164 165 166 168 172 174 196 197 198 199 200 202 203 204 206 207 208 209
Victoria	C	34 35 36 37 38 39 40 289 291 292 293 294 309 315 317 318 320 321 323 324 325 326 327
Victoria	D	93 101 102 116 120 176 179 180 182 183 184 185 186 187 188 190 212 214 218 219 220 223 224 225 226 227 228 229 230 233 242 243 244 245 246 254 255 256 257 258
Victoria	E	42 44 48 56 58 59 63 71 73 75 77 78 79 80 83 84 85 88 89 91 108 273 276 277 280
Yamagata	A	121 122 123 125 126 134 135 136 137 139 141 142 144 146 147 148 149 150 151 155 157 176
Yamagata	B	127 129 133 160 161 162 163 164 165 167 171 173 195 196 197 198 199 201 202 203 205 206 207 208
Yamagata	C	34 35 36 37 38 39 40 288 290 291 292 293 308 314 316 317 319 320 322 323 324 325 326
Yamagata	D	93 101 102 116 120 175 178 179 181 182 183 184 185 186 187 189 211 213 217 218 219 222 223 224 225 226 227 228 229 232 241 242 243 244 245 253 254 255 256 257
Yamagata	E	42 44 48 56 58 59 63 71 73 75 77 78 79 80 83 84 85 88 89 91 108 272 275 276 279

further conclusions). So instead of using the lineage-specific epitope sites, we combined them together to create a "unified" list of sites for influenza B – for each epitope site, we combined together all of the B/Yamagata and B/Victoria residues. The combined residues are shown together in Table D.4.

Table D.4: Unified epitope sites used for p-epitope distance for influenza B.

Epitope	Residues
A	121 122 123 125 126 134 135 136 137 139 141 142 144 146 147 148 149 150 151 155 157 176 177
B	127 129 133 160 161 162 163 164 165 166 167 168 171 172 173 174 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209
C	34 35 36 37 38 39 40 288 289 290 291 292 293 294 308 309 314 315 316 317 318 319 320 321 322 323 324 325 326 327
D	93 101 102 116 120 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 211 212 213 214 217 218 219 220 222 223 224 225 226 227 228 229 230 232 233 241 242 243 244 245 246 253 254 255 256 257 258
E	42 44 48 56 58 59 63 71 73 75 77 78 79 80 83 84 85 88 89 91 108 272 273 275 276 277 279 280

Once the residue sites are defined, we compute the normalized Hamming distance between all five epitopes for the two sequences (that is, the proportion of amino acids that are different

at the same site within each epitope only). This gives us five epitope distances for each pair of sequences. Multiple methods have been suggested in the literature for reducing these five distances into one measure of antigenic distance between each pair of strains.

- The **dominant p-Epitope** method was the original proposed method [30] and uses the maximum normalized Hamming distance across the five epitope sites.
- The **p-all-Epitope** method uses the normalized Hamming distance across all epitope sites (the total number of changes divided by the total number of epitope residues in all sites) [31].
- An alternate definition of **p-all-Epitope** uses the mean of the epitope-specific normalized Hamming distances [35]. We refer to this definition as **p-epitope-Anderson** to avoid confusion.
- The **p-2-Epitope** method uses the fraction of substitutions in the two epitopes with the highest number of substitutions [32], and could easily be generalized to **p-k-Epitope** for $k \in \{2, 3, 4\}$.

When we say **p-Epitope** without an additional descriptor, we always refer to the dominant p-Epitope distance.

Substitution matrix distances

While the p-Epitope distance has been correlated with vaccine effectiveness in the literature, it uses a function of the normalized Hamming distance at each epitope. The Hamming distance treats all substitutions equally, and basic protein biochemistry tells us that this is, in general, untrue — some substitutions matter more than others because they can change protein structure or function. In general, a leucine to isoleucine mutation will affect the basic properties of a protein less than a leucine to glutamine mutation. Incorporating an amino acid substitution matrix into our antigenic distance calculations can help to account for the specific mutations between two sequences instead of treating all substitutions equally. This can, for example, assign higher distances to sequences where a glycosylation site has changed which might be an important factor in antigenic evolution [165].

To our knowledge, substitution matrix distances have not been validated against im-

munological or epidemiological data in the same way that p-Epitope has, but they offer a compelling alternative to the simple Hamming distance. However, the selection of an appropriate substitution matrix is a necessary component of calculating the antigenic distance in this way. We have primarily used Grantham's amino acid substitution index [151], but there are many other amino acid substitution matrices (`https://en.wikipedia.org/wiki/Amino_acid_replacement`) including those developed in a sequence alignment or phylogenetic context (`https://en.wikipedia.org/wiki/Substitution_matrix`). The FLU substitution model [173] was developed specifically for influenza evolution, but to our knowledge there is no comparison of these different substitution matrices in the context of sequence distances.

Once we have chosen a suitable substitution matrix, calculating the distance is simple. For every site in two aligned sequences, we look up the score for that substitution (0 if there is no substitution) in the substitution matrix. We can then sum those scores to get a substitution matrix difference, or take the mean to get a normalized substitution matrix difference. We typically refer to these by the name of the substitution matrix, e.g., Grantham's distance or the FLU substitution distance.

Cophenetic distance

The cophenetic distance is a generalized notion of distance based on dendrogram clustering. In the specific case of protein sequences, we can build a phylogenetic tree (a specific type of dendrogram or hierarchical clustering model) and extract the cophenetic distances from the tree, which estimate the amount of total genetic divergence that has occurred between two strains. Building a correct (or at least acceptable) phylogenetic tree is the difficult part; once the tree is built with branch lengths, the cophenetic distance is well-defined and easy to calculate.

In order to calculate the cophenetic distances for the UGAFluVac data, we built one maximum likelihood tree for each subtype (that is, A(H1N1), A(H3N2), and all B strains together). We used the R package `phangorn` [250, 251] to build unrooted trees with the following specifications. We used the `FLU + G(20) + I` evolutionary model which assumes substitution likelihoods follow the `FLU` substitution matrix [173], with a discrete gamma

model for rate variation across sites that has 20 categories. The FLU matrix is time-reversible and uses fixed substitution rates and amino acid frequencies estimated from a large dataset of influenza sequences. The evolutionary rate at each site is drawn from the discrete gamma distribution with a shape parameter estimated from the data and then discretized; the model also estimates the proportion of invariant sites that are conserved and never change [252, 253]. We ran a preliminary test of multiple models (data not shown) and found that this model performed well. Most models only use $k = 4$ or $k = 8$ for the discrete gamma model, but a high k ensures a robust fit if the discrete gamma model assumption is violated [254]. We used Laguerre gamma quadrature [255] to estimate the gamma likelihood accurately with a high number of categories, but we had a low number of sequences per tree so the computational time was sufficiently fast. Finally, we implemented 100 stochastic search rearrangements per tree similar to the method used by IQ-TREE [256].

We calculated the cophenetic distance matrix from each tree using the method from the `ape` package [257].

Cartographic distance

Antigenic cartography uses multidimensional scaling of immunological assay data to estimate distances in antigenic shape space between influenza strains [29]. Specifically, cartography for influenza uses a matrix of data where one dimension (say, rows, without loss of generality) indexes different serum samples, and the other dimension indicates influenza virus strains which were used to conduct immunological assays. For a simple setup, suppose a study recruits multiple individuals. Each of those individuals donates a serum sample, and each serum sample is used for multiple hemagglutination inhibition (HAI) assays, each to a specific strain of influenza. (In principle, any strain-specific assay can be used, but in practice, HAI is by far the most common assay at time of writing.) An HAI assay for a specific strain of influenza measures the amount of antibodies in an individual's serum that are capable of binding to the HA head domain of that influenza strain. We then arrange our data into a matrix with individuals (or serum samples if individuals donated multiple serum samples) indexing columns and influenza strains indexing rows. We use a statistical method like

multidimensional scaling (MDS) [258] to reduce the data into k columns, typically two, so we then have two dimensions associated with each influenza strain and we can plot the reduced coordinates on a two-dimensional figure. Some results suggest that two is too few dimensions [259], but it is the most common because plotting in two dimensions is easy.

We specifically used `Racmacs` [175] to build a two-dimensional cartographic map for each subtype (A(H1N1), A(H3N2), and all B strains together) using post-vaccination HAI titers from individuals who received standard dose influenza vaccines at a UGAFluVac study site from the 2013/14 through 2017/18 influenza seasons. Beginning in 2018/19, UGAFluVac decreased the size of the panel of viruses they were using for HAI assays, which limits our ability to create a stable cartographic map, so we only used data from the seasons with a large panel of historical strains. These maps created from post-vaccination human serum have some differences from previously published cartographic maps on ferrets, but a detailed investigation warrants its own publication (data not shown). `Racmacs` does not use the simple MDS implemented that we used as an example: instead, `Racmacs` performs a matrix expansion that allows us to place both serum samples and virus strains on the cartographic map simultaneously, and randomly assigns starting values to missing and censored (below the HAI limit of detection) titer values in the matrix. `Racmacs` then applies metric MDS to the full distance table before using a stochastic L-BFGS optimization routine to reduce the stress of the map. The stress of an MDS map measures how much pairwise distances have changed in the dimension reduction step, so the stochastic optimization step improves the quality of the map in terms of the original goal of MDS.

Since the stochastic optimization method used by `Racmacs` can be sensitive to initial conditions [125], we created 25 cartographic maps (per subtype), and ran 100 L-BFGS optimization rounds per map. Then, for each subtype, we selected the map with the overall minimum stress across all initial conditions and optimization rounds. Finally, we extracted the pairwise Euclidean distances between all of the viral strains in each of the fitted maps.

Implementation

We implemented our pipeline for calculating antigenic distances in `R` 4.4.2 [72] using RStudio 2024.12.1+563 "Kousa Dogwood" Release [177]. We developed our pipeline using `targets` [158] and ran it locally on a Windows 10 Enterprise 64-bit machine and on UGA's `sapelo2` computing cluster, which is a distributed computing cluster running CentOS Linux release 7.5 that uses Slurm [160, 161] to schedule jobs. We developed this writeup using Quarto version 1.6.40 [260] and the `flextable` [81] and `softbib` [86] packages.

Where possible, we have specifically acknowledged the `R` packages we used throughout our methods. However, for completeness we include a list of all `R` packages we explicitly used in our code here: `ape` [257], `BiocManager` [261], `crew` [262], `crew.cluster` [263], `dplyr` [264], `forcats` [265], `here` [77], `janitor` [266], `knitr` [83–85], `parallelly` [267], `phangorn` [250, 251], `purrr` [268], `Racmacs` [175], `readr` [269], `readxl` [270], `renv` [78], `rlang` [271], `rmarkdown` [272–274], `stringdist` [244], `stringr` [275], `tarchetypes` [276], `tibble` [277], `tidyr` [278], and `visNetwork` [279]. Many of these packages are part of the `tidyverse` suite of packages [73].

We have also implemented these methods in an `R` package called `agdist` (this link is to an archived version at time of writing), as well as a pipeline that computes all distances for the UGAFluVac study up to 2017/18 (also link to an archived version). At time of writing, the `agdist` package can compute Hamming (and other distances from `stringdist`), p-Epitope, Grantham, FLU substitution, and temporal distances; and can process tree models and antigenic map objects to extract the distances from those. However, package development is ongoing, and the latest package source or latest pipeline source may have different or altered functionality.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for

generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. Similarly, we gratefully acknowledge all data contributors for generating the genetic sequence and metadata and sharing via NCBI GenBank and UniProt, on which this research is based.

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

We thank Klaus Schliep for his assistance on GitHub with using the `phangorn` package, as well as William Michael Landau and Eric Scott for their assistance with developing our computational pipeline.

Appendix E

Extended methods for calculation of vaccine evaluation metrics

In this document, we will use an example to show how our summary statistics are calculated. Note that using our method requires adopting a Bayesian approach. **At this time, it is quite difficult to fit a frequentist model that simultaneously accounts for censoring and incorporates random effects.** Because the censoring correction seems to be incredibly important for our results, we *strongly* recommend using the Bayesian approach.

However, if you *must* use a frequentist method for some reason, you could apply frequentist mixed-effects models to our data and use bootstrapping procedures to get confidence intervals for the statistics of interest. If you know of an `R` package which can fit mixed-effects models with censored outcomes, please let us know so we can update these instructions.

Data description and format

The type of data that our methods apply to are panels of immunological assays to multiple different influenza strains. The following example data are a completely randomized, anonymized subset of the UGAFluVac HAI data from Ted Ross's lab group. So our results in this document might not make sense, but they are used for a calculation example, not to draw any conclusions. Table E.1 shows the example dataset before preprocessing. We do not include any pre-titer or demographic information in these calculations right now.

We store the data in "long form" so that each immunological assay represents one record in the dataset (i.e., one row of data). Before we can fit any models there are several preprocessing steps that should be performed first.

First, post-vaccination titer should be transformed to the log-scale. We always normalize the HAI titers based on the limit of detection (LoD) of 10 also, so that the values coded as 5 (below LoD) become 0 on the log scale. If x is the natural scale HAI titer, like the ones shown in Table E.1, you can get the log-scale values y using the transformation

$$y = \log_2\left(\frac{x}{5}\right).$$

The antigenic distance should be normalized. We use min-max normalization to ensure that the range of the datapoints is transformed to $[0, 1]$. If d is a raw distance value like you see in Table E.1, and d_{norm} is the normalized antigenic distance, the formula for min-max normalization is

$$d_{\text{norm}} = \frac{d - \min(d)}{\max(d) - \min(d)}.$$

Min-max normalization is a linear and monotonic transformation, so it preserves all of the statistical properties we care about. Using min-max normalization on the antigenic distance makes sure that AUC values using different antigenic distance methods and different assay panels can be fairly compared against each other (as long as the outcome is HAI). Other normalization methods like z-score standardization might be useful but we haven't compared them.

Table E.2 shows the same example data after processing.

We also recommend visualizing the data before proceeding to ensure there are no issues. Since the HAI titers are integer-valued, we apply a small amount of random jitter to the data before plotting. (If you use `ggplot2` like we do, we recommend adding jitter manually rather than using `ggplot2`'s built-in jitter functionality. As of the time of writing, there is no way to have `ggplot2`'s built-in jitter apply the same jitter to both ends of a line and to a point.)

As you can see in Figure E.1, the individual landscapes are extremely messy. This is due

Table E.1: Example uncleaned dataset for analysis.

ID	Assay Strain	Post-vac HAI	Ag. Distance
01	HK/68	5	5.37
01	PC/73	5	6.23
01	TX/77	5	6.29
01	MI/85	5	4.78
01	Sich/87	20	4.68
⋮			
83	Uru/07	1280	0.50
83	Per/09	2560	0.35
83	Vic/11	2560	0.70
83	Tx/12	2560	0.87
83	HK/14	1280	0.00

Table E.2: Example pre-processed dataset for analysis.

ID	Assay Strain	Post-vac HAI	Ag. Distance
01	HK/68	0	0.85
01	PC/73	0	0.99
01	TX/77	0	1.00
01	MI/85	0	0.76
01	Sich/87	2	0.74
⋮			
83	Uru/07	8	0.08
83	Per/09	9	0.06
83	Vic/11	9	0.11
83	Tx/12	9	0.14
83	HK/14	8	0.00

in part to the inherent measurement error in HAI titers, along with the variance induced by strain-specific effects (i.e. some strains tend to work better with HAI than others, regardless of how many antibodies are in the serum sample). The messy appearance of these individual landscapes **is to be expected**. That is why we need to fit a summary curve.

Coding censored data

We still have one more essential data cleaning step, however. We will fit all of our Bayesian models using the accessible and performant `brms` package. However, `brms` requires us to code our censored outcome in a specific way, which is documented in the `brmsformula` information (`https://paulbuerkner.com/brms/reference/brmsformula.html`).

We know that our HAI data has both left-censored and interval-censored values, and no

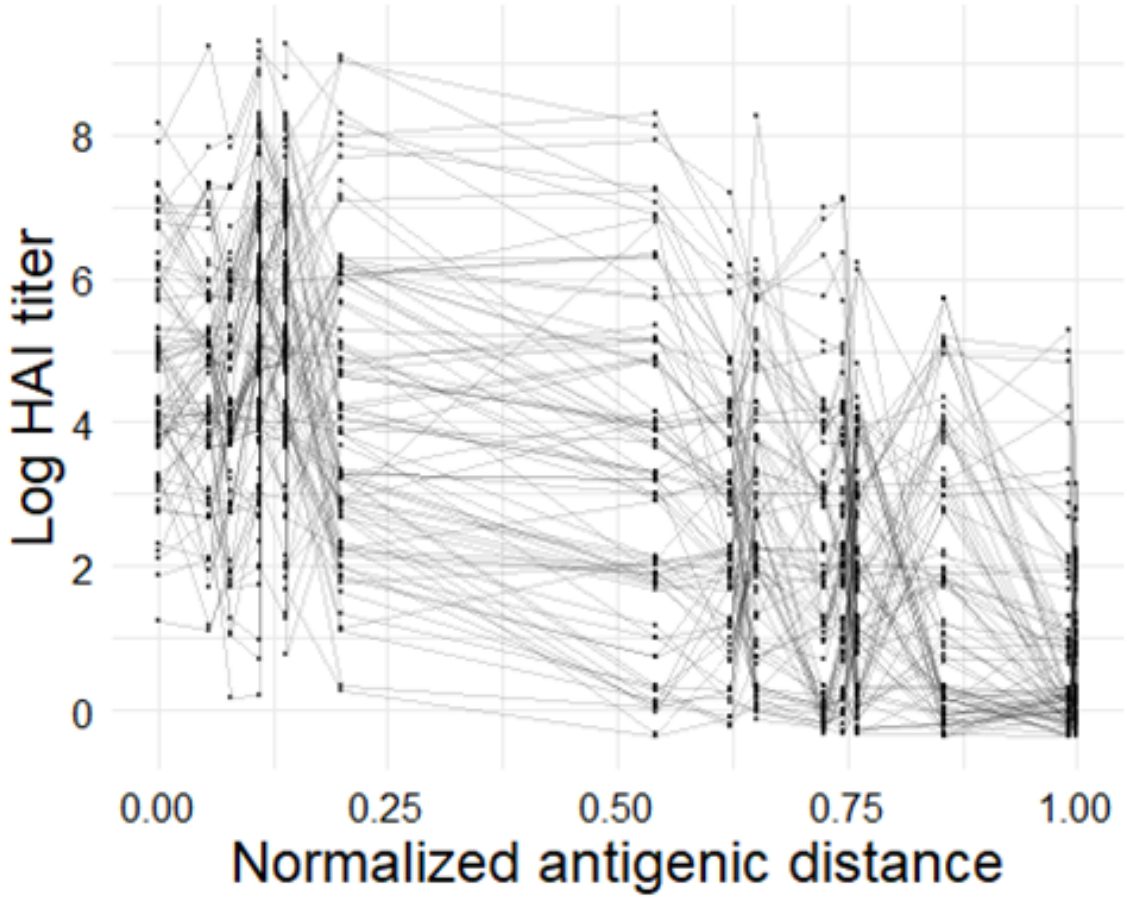

Figure E.1: Transformed titer vs. normalized antigenic distance. Each point represents one HAI titer measurement and each line connects all of the measurements for a particular subject.

completely observed (un-censored). Figure E.2 shows an example of the censoring mechanism for HAI. Other immunological assays will have different mechanisms for censoring, so this particular step only applies to HAI.

Coding censored data with mixed censoring types in `brms` format is a bit annoying. Instead of having one `posttiter` variable, we need to have three variables in the data which are called `y`, `c`, and `y2` in the `brms` documentation. Since the data below the LoD are coded as 0, but the actual LoD is 10 (or 1 in the transformed HAI data), we have to do some cleanup.

1. The variable `c` is an indicator for the censoring status. We set it to `"left"` if our value is left-censored, which occurs when the posttiter is 0, and we set it to `"interval"`

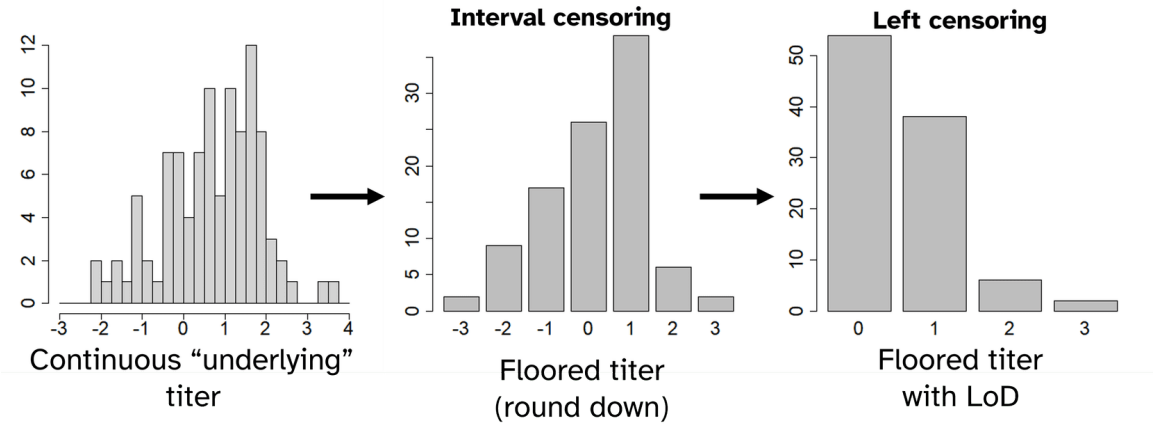

Figure E.2: Censoring mechanisms that apply to the latent or underlying continuous titers (left panel) during the measurement process to produce interval censored observations (middle panel), which are binned continuous values, and left censored observations (right panel), which are below the limit of detection.

otherwise, which is all of our other observations.

2. The variable `y` should be coded as the lower limit of detection (i.e., 1 on our transformed scale) for left-censored observations, which is easy, because we already did that in the HAI processing transformation. For interval-censored observations, `y` needs to be the lower bound of the interval. Since HAI titers are always rounded down as part of the data-generating mechanism, our outcome variable is already correctly coded, so we don't need to do anything else here.
3. The variable `y2` represents the upper bound for an interval-censored observation, so for left-censored observations it should be the same as `y`, but for interval-censored observations, it should be `y + 1`.

If you use `R` and `tidyverse`, you can use this code for cleaning up the censored data. If you prefer not to use `tidyverse`, the `ifelse()` statements in the code will still work, you just need to make sure to use a function like `transform()` or `with()`, or manually specify the dataset for each column every time, i.e. `titer_data$posttiter` if you aren't using a function that adds the column names to the scope. In this code, `titer_data` is the dataset shown in Table E.2, and `posttiter` is the name of the column containing the HAI data.

```
titer_data_cens <-
        titer_data |>
        dplyr::mutate(
                c = ifelse(posttiter == 0, "left", "interval"),
                y = ifelse(c == "left", 1, posttiter),
                y2 = ifelse(c == "left", 1, posttiter + 1)
        )
```

Table E.3 shows the dataset with the variables correctly encoded for the censoring correction. Now that the data are processed and ready to use we can calculate the statistics of interest.

Table E.3: The cleaned dataset with the new columns `c`, `y`, and `y2` that must be passed to `brms` to correctly apply the likelihood correction for censoring.

id	assay_strain	posttiter	distance	c	y	y2
1	HK/68	0	0.85	left	1	1
1	PC/73	0	0.99	left	1	1
1	TX/77	0	1	left	1	1
1	MI/85	0	0.76	left	1	1
1	Sich/87	2	0.74	interval	2	3
⋮						
83	Uru/07	8	0.08	interval	8	9
83	Per/09	9	0.06	interval	9	10
83	Vic/11	9	0.11	interval	9	10
83	TX/12	9	0.14	interval	9	10
83	HK/14	8	0	interval	8	9

Mathematical notation

We'll use the following notation in all of our math formulas.

- $i = 1, \ldots, n$ indexes study subjects, where $i = 1$ represents the first study subject and n represents the sample size.
- $s = 0, \ldots S$ indexes the different assay strains. Here, $s = 0$ is the homologous strain, and $s = 1, \ldots, S$ are the heterologous strains – the particular order of the strains doesn't matter.

- d_s is the antigenic distance between strain s and the vaccine strain. Thus, $d_0 = 0$ for the homologous strain and $0 < d_s \leq 1$ for every other $0 < s \leq S$.

We also define seroprotection and seroconversion as follows:

$$\text{Seroprotected}_{i,s} = I\left(\text{post-vaccination titer}_{i,s} \geq 40\right),$$

and

$$\text{Seroconverted}_{i,s} = I\left(\frac{\text{post-vaccination titer}_{i,s}}{\text{pre-vaccination titer}_{i,s}} \geq 4\right) \times I\left(\text{Seroprotected}_{i,s}\right).$$

Note that on the log-scale, we have

$$\text{Seroconverted}_{i,s} = I\left(\log \frac{\text{post-vaccination titer}_{i,s}}{\text{pre-vaccination titer}_{i,s}} \geq 2\right) \times I\left(\log \text{post-vaccination titer}_{i,s} \geq 3\right).$$

Current statistics

First, we want to calculate the statistics that do not take antigenic distance into account. The formulas **without taking censoring into account** are as follows. All are simple to calculate and can be done in a spreadsheet or with basic `R` code.

1. **Magnitude**: measured by the geometric mean titer to the homologous strain:

$$\text{GMT}_0 = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log \text{titer}_{i,s=0}\right).$$

2. **Breadth**: measured by the seroconversion rate across all strains. Note that since seroconversion status is binary, taking the mean gives us a proportion:

$$\text{SCR} = \frac{1}{n}\frac{1}{S}\sum_{i=1}^{n}\sum_{s=0}^{S} \text{Seroconverted}_{ij}.$$

3. **Strength**: measured by the geometric mean titer across all of the observed strains:

$$\text{GMT} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{S} \log \text{titer}_{i,s}\right).$$

However, when we take censoring into account we actually need to fit a model.

Homologous GMT

Note that when we fit a regression model that *only has an intercept term*, say

$$y_i = \alpha + \varepsilon_i,$$

the estimate of α actually estimates $\bar{y}$, the mean of the response variable. Since we can easily apply a censoring correction to a regression model, fitting an intercept-only regression model provides an easy way to estimate the mean while correcting for censoring.

To estimate GMT_0, we need to fit an intercept-only regression model to the post-vaccination titer data for only the homologous strain. We can do this in `brms` while applying a censoring correction like this (make sure to run `library(brms)` at the top of your R script!). Note that you can set the seed to any integer number, but if you use our example data, setting the seed will ensure our results can be reproduced without differences due to random number generation. A full tutorial on the `brms` package is beyond the scope of our instructions, but excellent tutorials can be found on the `brms` package website (`https://paulbuerkner.com/brms/`). We will provide a brief explanation of the arguments we chose.

```
titer_data_homologous <- dplyr::filter(titer_data_cens, distance == 0)

gmt_0_model <- brms::brm(
        formula = y | cens(c, y2) ~ 1,
        data = titer_data_homologous,
        family = gaussian,
        prior = c(
                brms::prior(student_t(3, 0, 3), class = "Intercept"),
                brms::prior(student_t(3, 0, 3), class = "sigma")
        ),
        cores = 4,
```

```
    chains = 4,
    warmup = 250,
    iter = 1250,
    backend = "cmdstanr",
    seed = 213948
)
```

- `formula`: follows the `brms` syntax for censoring, which is `y | cens(c, y2)` as explained earlier. The `~` separates the response terms from the predictor terms, and having `1` as the only predictor fits a regression model with only an intercept.
- `data`: the data we want to fit the model to goes here. All of the variables mentioned in `formula` have to match a column in `data`.
- `family`: `"gaussian"` indicates that we want to fit a standard regression model where the error term has a normal (or Gaussian) distribution.
- `prior`: since we are using a Bayesian model, we have to set priors. Setting priors can be contentious, and in general you can use whichever priors you prefer. We advocate for using generic `student_t(3, 0, 3)` priors for these models, because they allow for large parameter estimates if they are supported by the data, but they prefer to be skeptical and assume that parameters are small, similar to how we assume a null hypothesis is true and try to falsify it in frequentist hypothesis testing.
- `cores`: if your computer has multiple cores (most modern computers have at least 4), you can run multiple model chains in parallel to speed up the model fitting.
- `chains`: each "chain" of the model starts from a different initial condition and then tries to sample from the posterior distribution. Running multiple chains and making sure they agree provides evidence that we are sampling from the whole posterior distribution. Most Bayesian statisticians recommend running at least 4 chains per model.
- `warmup`: the number of samples the model should draw in the warmup step, when it is calibrating. These will not be used for inference, but if you do not run enough it will degrade the quality of your model fit.

- `iter`: the number of total samples the model should draw (per chain), so the number of post-warmup samples you will use for inference is `iter - warmup`.
- `backend`: we strongly recommend using the `cmdstanr` package here, because it is more modern and has more tools available for use than the default `rstan` package. However, if you have difficulties installing `cmdstanr`, you can use `rstan` instead, which is easier to set up.
- `seed`: you have to specify this to make sure your results are exactly reproducible. It can be any integer number, but different values will result in different random number draws and thus slightly different results.

We should also check the model diagnostics by looking at the summary. Again, we will not do an in-depth explanation of Bayesian model diagnostics here, but typically you want all of the `Rhat` estimates to be `1.00` (sometimes `1.01` can be acceptable for quick checks), and you want both `bulk ESS` and `tail ESS` to be above 500 (or above 1000 if you need precise estimates of quantiles or variances).

```
summary(gmt_0_model)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: y | cens(c, y2) ~ 1
   Data: titer_data_homologous (Number of observations: 83)
  Draws: 4 chains, each with iter = 1250; warmup = 250; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     5.19      0.16     4.87     5.51 1.00     2291     2578

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.49      0.13     1.27     1.77 1.00     2572     2569
```

```
Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

We see that our model passes those quick diagnostic guidelines. We'll skip these checks for the rest of these instructions because I already checked the models, but whenever you fit your bayesian models, you need to check these diagnostics. If any of the checks don't pass, you should try increasing `warmup` (above 500 typically is not useful, though) and `iter` (increase this one as much as needed until you reach the diagnostic targets). More troubleshooting advice is easy to find on the Stan discourse forum (`https://discourse.mc-stan.org/`).

Now with the arguments explain we can use this code to extract the posterior samples of the intercept, which correspond to the posterior samples of the `GMT`. Then we can summarize the posterior samples into a point estimate and credible interval for our metric. We typically use the mean as the point estimate along with a 95% highest density continuous interval (HDCI), implemented through `tidybayes::mean_hdci()`. You could also use the median or the mode (also called the MAP or Maximum *a posteriori* estimate; this is actually what we prefer but some reviewers find it too technically difficult) for the point estimate, and you could also use an equal-tailed credible interval. The choice depends on personal preference and interpretation.

```
gmt_0_samples <-
        gmt_0_model |>
        tidybayes::tidy_draws() |>
        dplyr::select(b_Intercept)


tidybayes::mean_hdci(gmt_0_samples) |>
        # This step back-transforms the log-scale estimates
        dplyr::mutate(
                dplyr::across(c("b_Intercept", ".lower", ".upper"), \(x) 5 * 2 ^ x
    )
        )

# A tibble: 1 x 6
```

```
  b_Intercept .lower .upper .width .point .interval
        <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>
1        183.   146.   227.   0.95 mean   hdci
```

So our censoring-corrected estimate of the GMT_0 metric is 183 with a 95% CI of $(146, 227)$.

Seroconversion rate

Now that we've estimated the GMT_0, the other parameters will be simpler because almost everything stays the same. Because all seroconversion measurements are binary (they are either 0 for not seroconverted or 1 for seroconverted), several things are different about this model.

First we need to calculate the seroconversion status, which is easy. *However, you must have the pre-vaccination titers.* We have them for our example data, but if you don't have pre-vaccination data, you can only look at seroprotection, not seroconversion. Note that we calculate the fold change (ratio of post-vaccination titer divided by pre-vaccination titer) using a difference instead of a division because the log of a fraction is equal to the difference of the logs. That is, $\log(a/b) = \log(a) - \log(b)$.

```
titer_data_sc <-
        titer_data_cens |>
        tibble::add_column(pretiter = log2(example_data_pretiters / 5)) |>
        dplyr::mutate(
                fold_change = posttiter - pretiter,
                seroprotection = posttiter >= 3,
                seroconversion = seroprotection * (fold_change >= 2)
        )
```

We don't have to correct for censoring, and we need to fit a logistic regression model instead of a linear regression model. Even though we don't apply censoring corrections here, we choose to fit this with a bayesian model anyways to ensure our CI's are comparable. In the code for this model, notice that we only have one prior – because logistic regression

models don't have a σ^2 residual variance parameter, we don't need a prior for it.

```
scr_model <- brms::brm(
        formula = seroconversion ~ 1,
        data = titer_data_sc,
        family = bernoulli("logit"),
        prior = c(
                brms::prior(student_t(3, 0, 3), class = "Intercept")
        ),
        cores = 4,
        chains = 4,
        warmup = 250,
        iter = 1250,
        backend = "cmdstanr",
        seed = 213948
)

scr_samples <-
        scr_model |>
        tidybayes::tidy_draws() |>
        dplyr::select(b_Intercept)

tidybayes::mean_hdci(scr_samples)

# A tibble: 1 x 6
  b_Intercept .lower .upper .width .point .interval
        <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>
1       -1.30  -1.43  -1.18   0.95 mean   hdci
```

So we fit the model, and summarize the posterior samples of the intercept, and our estimate of the seroconversion rate is 0.21 with a 95% CI of $(0.19, 0.24)$.

Overall GMT

Estimating the GMT is exactly the same as estimating the GMT_0, except we use the entire dataset instead of only the homologous strain measurements. We extract and summarize the posterior samples of the intercept, exactly the same as the GMT_0.

```
gmt_model <- brms::brm(
        formula = y | cens(c, y2) ~ 1,
        data = titer_data_cens,
        family = gaussian,
        prior = c(
                brms::prior(student_t(3, 0, 3), class = "Intercept"),
                brms::prior(student_t(3, 0, 3), class = "sigma")
        ),
        cores = 4,
        chains = 4,
        warmup = 250,
        iter = 1250,
        backend = "cmdstanr",
        seed = 213948
)

gmt_samples <-
        gmt_model |>
        tidybayes::tidy_draws() |>
        dplyr::select(b_Intercept)

tidybayes::mean_hdci(gmt_samples) |>
        dplyr::mutate(
                dplyr::across(c("b_Intercept", ".lower", ".upper"), \(x) 5 * 2 ^ x
    )
        )
```

```
# A tibble: 1 x 6
  b_Intercept .lower .upper .width .point .interval
        <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>
1        65.9   59.8   73.0   0.95 mean   hdci
```

So our censoring-corrected estimate of the GMT metric is 66 with a 95% CI of $(60, 73)$.

Novel metrics

Now, calculating the models with antigenic distance included is a bit more involved. The first step is to fit our antibody landscape model.

Summary antibody landscape

The first step in calculating the antigenic-distance based metrics is to compute the **summary antibody landscape**. This is a mathematical model that takes all of the individual antibody landscapes from Figure E.1 as the input, and produces a study-level summary curve that represents the average antibody landscape in the study sample.

We use a multilevel linear regression (AKA mixed-effects regression) model to construct the summary antibody landscape. Letting the log postvaccination titers by y, the model is as follows.

$$
\begin{aligned}
y_{i,s} &\sim \text{Normal}\left(\mu_{i,s}, \sigma^2\right) \\
\mu_{i,s} &= \beta_0 + b_{0,i} + (\beta_1 + b_{1,i}) \cdot d_s \\
\begin{pmatrix} b_{0,i} \\ b_{1,i} \end{pmatrix} &\sim \text{MVN}\left(\vec{0}, \Sigma_b\right)
\end{aligned}
$$

We have a population level intercept and slope for distance, and each individual i also gets their own intercept and slope, which are correlated and drawn from a distribution that we can estimate. Because we are using a Bayesian model, we also have to specify priors. Here we use priors that are sensible for most similar problems. The Student t priors are the same as those we used in the simpler models. However, the multivariate normal distribution can be problematic to parametrize, especially in terms of a prior for Σ_b, the covariance matrix

of the random effects. The best solution, and the one used by `brms`, is to use a Cholesky decomposition of the covariance matrix and apply a Lewandowski-Kurowicka-Joe prior to the Cholesky factor of Σ_b. We recommend $\eta = 2$ as a sensible default for the covariance Cholesky factor prior.

```
summary_landscape_model <-
        brms::brm(
                formula = y | cens(c, y2) ~ 1 + distance + (1 + distance | id),
                data = titer_data_cens,
                family = gaussian,
                prior = c(
                        brms::prior(student_t(3, 0, 3), class = "Intercept"),
                        brms::prior(student_t(3, 0, 3), class = "b"),
                        brms::prior(student_t(3, 0, 3), class = "sd"),
                        brms::prior(student_t(3, 0, 3), class = "sigma"),
                        brms::prior(lkj(2), class = "cor")
                ),
                cores = 4,
                chains = 4,
                warmup = 250,
                iter = 1250,
                backend = "cmdstanr",
                seed = 213948
        )
```

Note that in the above code, the formula part `1 + distance` adds the study-level (fixed effects) parameters for the intercept and the antigenic distance slope, and the part `(1 + distance | id)` adds the individual-level (random effects) parameters for each individual id.

Now that we fit the model (and I made sure to check that the diagnostics are sufficient as mentioned earlier) we want to extract the summary antibody landscape fit. Because we used a multilevel model, there are many ways to get predictions and uncertainty about the best antibody landscape that summarizes the population level. However, we chose to use

the global grand mean prediction and associated HDCI. When we talk about the grand mean prediction, there is unfortunately no standard language used to describe these different methods, other than getting lost in describing different parts of the model as marginal or conditional. We adopt the terminology of Andrew Heiss, who provides an excellent layman-friendly explanation of this issue in a blog post (`https://www.andrewheiss.com/blog/2021/11/10/ame-bayes-re-guide/`).

Make sure to keep the entire set of posterior predictions, because the AUC and other landscape metrics are calculated from the posterior samples. (That's `summary_landscape_preds` in our code below.) Also note the variable h that we assign below – this is called the step size for our interpolated predictions. Making h smaller will result in more fine-grained predictions and thus more accurate metric calculations, but the storage space and memory requirements for the posterior samples increase rapidly as h decreases. We have found that $h = 0.01$ offers a good balance between memory usage and prediction accuracy.

```
h <- 0.01

summary_landscapes_preds <-
        summary_landscape_model |>
        tidybayes::epred_draws(
                newdata = data.frame(distance = seq(0, 1, h)),
                re_formula = NA
        ) |>
        dplyr::ungroup()

summary_landscape_intervals <-
        summary_landscapes_preds |>
        dplyr::summarize(
                tidybayes::mean_hdci(.epred),
                .by = distance
        )
```

Now we'll take our previous plot of the individual antibody landscapes and add the

summary antibody landscape on top (Figure E.3). We're currently experimenting with nonlinear methods for fitting the summary antibody landscape, but right now this linear regression version appears to be doing okay. Now that we have the summary antibody landscape posterior samples, we can calculate the metrics based on the posterior samples. Now, all three of the landscape-based statistics can be expressed as functions of the posterior predictions of the summary antibody landscape, so we'll use that to calculate them.

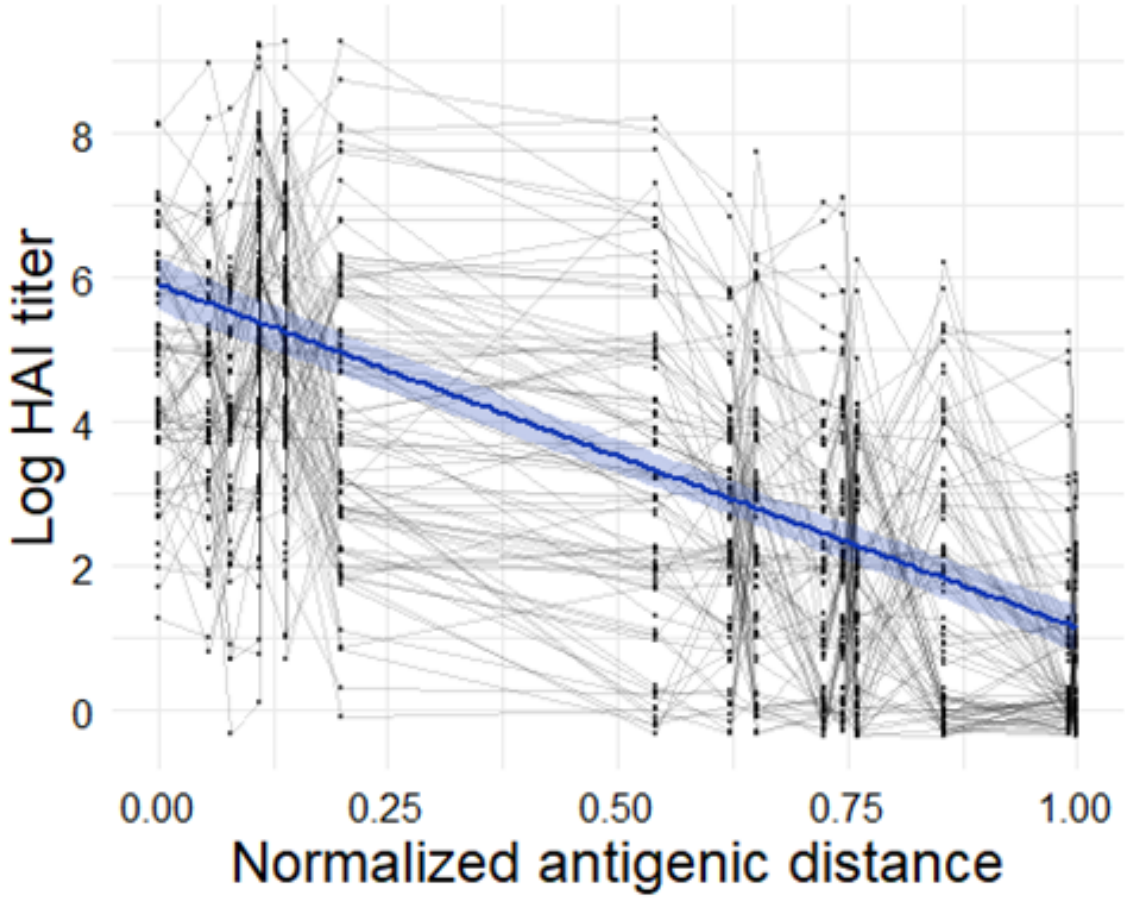

Figure E.3: Individual antibody landscapes (lines) and observed HAI data (points), with the study-level posterior mean summary landscape and 95% HDCI (blue line and ribbon).

Magnitude

We measure the magnitude of the summary landscape as the intercept of the line – recall that we only saved the posterior samples of the conditional predictions though, so we should

express this as the average prediction when $d = 0$. That is,

$$\text{Magnitude} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{0,j},$$

where $j = 1, \ldots, m$ indexes the posterior samples from the model. Estimating this from the posterior samples is easy because all we have to do is average the predictions where $d = 0$.

```
summary_landscapes_preds |>
        dplyr::filter(distance == 0) |>
        dplyr::summarize(tidybayes::mean_hdci(.epred))
```

```
# A tibble: 1 x 6
      y  ymin  ymax .width .point .interval
  <dbl> <dbl> <dbl>  <dbl> <chr>  <chr>
1  5.92  5.57  6.28   0.95 mean   hdci
```

So we see that our estimate for the magnitude is 5.92 with a credible interval of $(5.57, 6.28)$. We can also backtransform this one to HAI units to compare it to the GMT_0 estimate.

```
5 * 2 ^ c(5.92, 5.57, 6.28)
```

```
[1] 302.7384 237.5238 388.5424
```

As you can see, this is actually larger than the GMT_0 estimate – accounting for the antigenic distance allows us to get a better estimate of the average homologous response.

Breadth

Our estimate of the breadth is the intersection between the horizontal line $y = 3$ (or $y = 40$ on the natural scale of the HAI titers) and the summary antibody landscape. This estimates the antigenic distance where we expect post-vaccination HAI titers to drop below the clinical protection threshold. The first thing we need to do is for each sample find the predicted values that are closest to 3. Then we get the distance value associated with that prediction, and that is our measurement of the breadth.

```
summary_landscapes_preds |>
```

```
        # minimizing |prediction - 3| over the draws will tell us which x-value is
        # closest to a y-value of 3.
        dplyr::mutate(
                diff_from_threshold = abs(.epred - 3)
        ) |>
        dplyr::summarize(
                # For each posterior sample (aka posterior draw) of the summary
    landscape,
                # get the predicted y-value that is closest to 3.
                closest = min(diff_from_threshold),
                # Now get the x-value that's associated with that y-value -- this
    is the
                # estimate for that posterior draw.
                breadth = distance[which.min(diff_from_threshold)],
                .by = .draw
        ) |>
        # Summarize the posterior samples into a point estimate and CI
        dplyr::summarize(
                tidybayes::mean_hdci(breadth)
        )
```

```
# A tibble: 1 x 6
      y  ymin  ymax .width .point .interval
  <dbl> <dbl> <dbl>  <dbl> <chr>  <chr>
1 0.613  0.56  0.66   0.95 mean   hdci
```

So our estimate of the breadth for our example data is 0.61 with a 95% CI of $(0.56, 0.66)$.

Strength

The last metric we need to calculate is the AUC. Because we're using a linear regression model, we could actually write out a formula to easily calculate the AUC from the intercept and slope samples. But because we want to generalize to potentially more complicated models, we need to do some numeric calculations (this is true for the breadth estimate as

well).

Since we already have the interpolated posterior samples, we can calculate the numeric AUC using any of the existing methods for numerical integration (see, e.g., `https://en.wikipedia.org/wiki/Numerical_integration`). The computationally fastest and simplest to implement is the trapezoidal rule, which has a computationally efficient implementation in the function `pracma::trapz()`. All we have to do is put in the interpolated distance values and their corresponding posterior predictions and this function estimates the area under the curve for us — we repeat this for each posterior sample of the summary landscape to get a credible interval.

```
summary_landscapes_preds |>
        # Calculate the AUC for each posterior sample of the line
        dplyr::summarize(
                AUC = pracma::trapz(x = distance, y = .epred),
                .by = .draw
        ) |>
        # And summarize it into a point estimate and interval
        dplyr::summarise(
                tidybayes::mean_hdci(AUC)
        )
```

So our estimate for the AUC for the example data is 3.54 with a 95% CI of $(3.31, 3.78)$. Normally, the units for the AUC estimate are a product of the units for the x-variable and the units for the y-variable. But since we min-max scaled the x-variable, it is actually unitless — that means we can interpret the AUC as an average log-titer measurement over the antigenic distances. So, we can backtransform the AUC to the natural HAI scale to get a better understanding.

```
5 * 2 ^ c(3.54, 3.31, 3.78)
```

These numbers are on the same order of magnitude as the GMT estimates, but a bit smaller – it's still hard to interpret this number, but it can be easily used for comparison across multiple vaccines, panels of historical viruses, or subsamples.

Appendix F

Comparing neighbor-joining antigenic distance trees with phylogenetic trees

Project summary: Developing universal influenza vaccines will require improved understanding of how influenza variants differ from each other. We find that temporal distances perform poorly overall, but even sequence distances which match phylogenetic distances well do not match cartographic distances based on actual immune response data.

Introduction

Influenza A virus causes seasonal epidemics worldwide, primarily driven by continual evolution of the virus under selective pressure by host immunity [1]. Development of a universal influenza vaccine which can protect against novel strains of influenza has many challenges, including surveillance of new genomic variants and predicting which will be successful [8]. Statistical modeling and phylodynamic approaches are crucial tools in the development of a more broadly-protective influenza vaccine, but these methods rely on understanding how different each genomic variant of influenza actually is from its predecessors [280].

Many different metrics for assessing the antigenic difference between two influenza strains currently exist, including phylogenetic methods [281], sequence distances [30, 35], and antigenic cartography, which is based on observed immunological data [44]. To understand

the agreement in different distance measurements, we obtained data from a cohort study that has been previously described [61, 63]. Using data from this study, we compared antigenic cartography and sequence methods to phylogenetic methods.

Study methodology

Briefly, our study data [61, 63] consisted of volunteers enrolled at three different study sites from 2013 – 2019 who received a FluZone (Sanofi Pasteur) vaccine, and gave pre-vaccination and post-vaccination (21 or 28 day) serum samples. The serum samples were used for HAI assays against a panel of historical viruses. We computed the Hamming distance [35], p-Epitope distance [30], and the absolute difference in the year of isolation of strains [45] from the sequences of all influenza viruses used for HAI assays, and used Racmacs to compute antigenic cartography distances from the HAI data [175]. All of our analyses were conducted separately for H1N1 and H3N2 strains.

In order to compare with phylogenetic methods, we first computed a multiple sequence alignment (MSA) using the MUSCLE algorithm [239]. The Hamming and p-Epitope distances were computed based on this MSA. We had 18 H1N1 strains and 21 H3N2 strains in total. We then used both alignments to construct maximum likelihood (ML) unrooted phylogenetic trees using the FLU amino acid substitution model. We extracted the cophenetic distances between taxa from the ML trees, and compared these distances to our other distance metrics (temporal, Hamming, p-Epitope, and cartography) using Pearson's correlation.

For each of the four distance metrics, we also built distance-based trees using neighbor joining. To compare the methods, we calculated the likelihood of each of the distance-based trees, then estimated the Shimodaira-Hasegawa test statistic to compare each of the distance trees to the ML tree. Finally, we computed the Robinson-Foulds distance between each set of trees. Our analyses were implemented with R version 4.3.3 [72] using the packages `phangorn` [250] and `msa` [238].

Study Results

We found that all four distance metrics were strongly correlated with cophenetic tree distance for H3N2, but for H1N1, only the Hamming and p-Epitope distances had a strong correlation with the tree distance (Figure F.1). H1N1 has two clusters, 2009 pandemic-like (pdm) and non-pdm. The pdm-like strains are genetically more similar to the 1918 pandemic strain than to most strains which circulated from 1950 – 2009, so the temporal distance correlation is weak, as expected.

The cartographic distance correlation for H1N1 is also moderate, indicating that the evolutionary pattern of H1N1 strains does not necessarily explain variation in observed immune responses. For H3N2, the cartographic correlation was the lowest, and the two distances become less correlated as the distance values become larger. For closely related H3N2 strains the ability of the tree distance to predict differences in immune response appears to attenuate as strains drift further away.

The ML trees for both subtypes were able to reconstruct the patterns we expect for H1N1 and H3N2 influenza (Figure F.2). The H1N1 strains form two clades, one pdm-like clade which contains SC/18 (the 1918 pandemic strain), NJ/76 swine influenza, and the modern pdm-like strains. The other clade contains the H1N1 strains which circulated between the 1918 pandemic and the 2009 pandemic. The H3N2 strains tend to follow a similar ladder-like pattern, beginning with HK/68 and primarily separating by temporal distance, which corroborates the correlations between temporal and cophenetic distance (Figure F.1).

For brevity, we do not show all 8 of the distance-based neighbor joining phylogenies. However, we conducted SH tests and computed the RF distance between each of the distance-based trees and the ML tree for the same subtype (Table F.1). For the H1N1 strains, the temporal distance and cartographic distance trees were different from the maximum likelihood tree based on the SH test, and these trees also had a much higher RF distance from the ML tree than the Hamming and p-Epitope distance trees. For the H3N2 strains, the p-Epitope distance tree was different from the ML tree, and the cartographic tree was extremely different from the ML tree. The ML tree, temporal distance, and Hamming distance trees

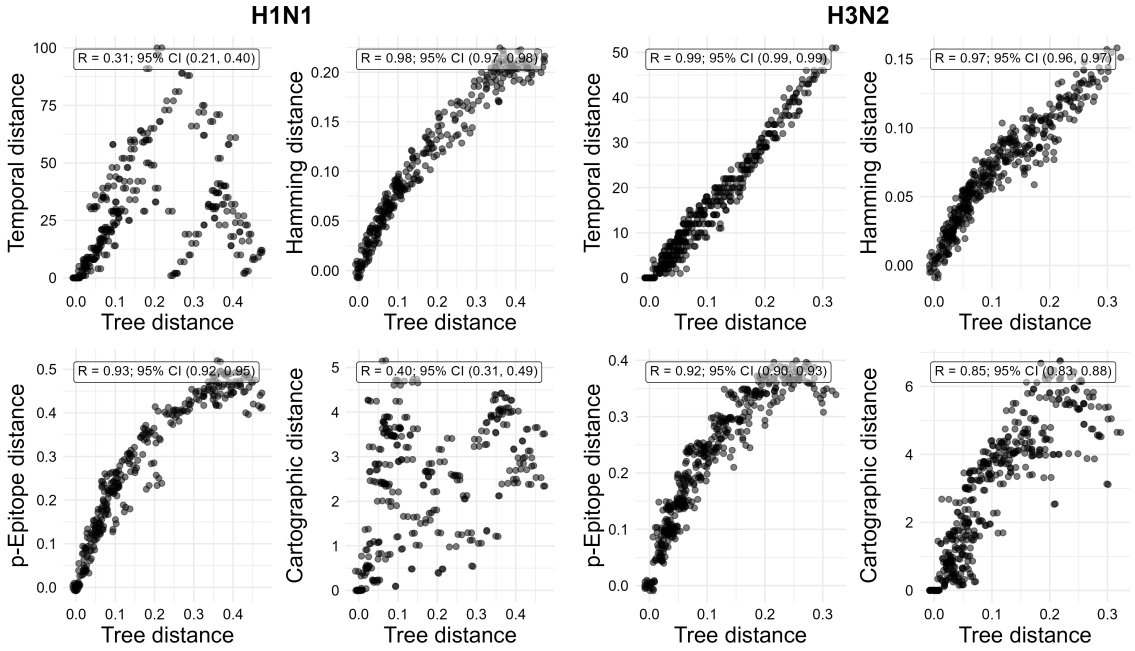

Figure F.1: Scatterplots showing the cophenetic tree distance on the x-axis and the other distance metrics we calculated on the y-axes. The plots on the left are for H1N1 strains and the plots on the right are for H3N2 strains. The box shows Pearson's correlation (R) along with a 95% Wald-type confidence interval.

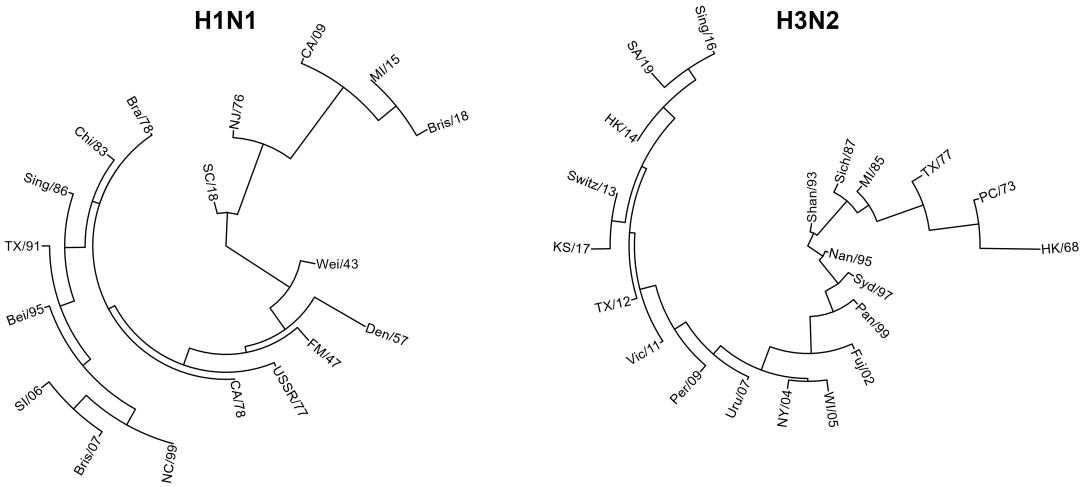

Figure F.2: Maximum likelihood phylogenetic trees for H1N1 strains (left) and H3N2 strains (right). Both trees are rooted at the midpoint for display purposes, but the root was not optimized during fitting.

were all similar. All of the changes in log likelihood for the H3N2 trees were smaller in magnitude than for H1N1. Notably, the temporal distance tree had a much lower likelihood than the ML model for H1N1.

Table F.1: Log likelihood of all constructed trees (ℓ), along with the decrease in log likelihood ($\Delta\ell$) from the ML model, the p-value of the Shimodaira-Hasegawa test (p_{SH}; evaluated on one million bootstrap resamples), and the Robinson-Foulds distance from the ML tree (d_{RF}).

	Tree	ℓ	$\Delta\ell$	p_{SH}	d_{RF}
H1N1	ML (baseline)	-3468.8	0.0	N/A	0
	Temporal	-5708.2	2239.4	<0.001	26
	Hamming	-3469.4	0.6	0.875	2
	p-Epitope	-3543.3	74.5	0.299	8
	Cartographic	-3980.1	511.3	<0.001	24
H3N2	ML (baseline)	-3065.6	0.0	N/A	0
	Temporal	-3102.7	37.2	0.270	8
	Hamming	-3110.2	44.6	0.214	4
	p-Epitope	-3171.3	105.7	0.014	12
	Cartographic	-3442.2	376.6	<0.001	30

Conclusion

Many papers still use the temporal method for calculating antigenic distance. However, for H1N1, the temporal distance completely fails to reconstruct any genetic changes. For H3N2, the temporal distance was similar to the ML distance. The Hamming and p-Epitope distances were similar for both subtypes.

The cartographic distance tree was substantially different from the ML tree for both H1N1 and H3N2. Since cartographic distance is based on observed immune response data, this implies that the hemagglutinin sequence is not the only factor in determining individual immune responses. Our sample is likely not representative, so similar analyses should be repeated on other cohorts. Performing similar analyses using neuraminidase sequence and inhibition data would complete our findings well.

Overall we find that temporal methods should be avoided and are not suitable for calculating evolutionary distance between influenza strains. Additionally, the genetic distance

between influenza strains does not match the cartographic difference from observed immune response data, indicating that genetic and antigenic evolution do not always agree.

Acknowledgments

We thank the following contributors for their contributions to this work: Justin Bahl, Lambodhar Damodaran, Yang Ge, Murphy John, Hayley Hemme, Savannah L. Miller, Andreas Handel, Ted Ross, Ye Shen, Amanda L. Skarlupka, and Meng-Hsuan Sung. Finally, we thank the University of Georgia graduate school for providing funding.

Data availability. The code and data for this project are avaible on GitHub (`https://github.com/wzbillings/flu-dist-v-phylo-trees`).