SeroTracker: Synthesising Seroprevalence Data to Map the True Extent of COVID-19 Infection and Immunity



Rahul K. Arora
Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2022

This thesis is dedicated to my grandparents.
Barrima, Barepapa, Nani, and Nana,
thank you for making my education possible.

Acknowledgements

Personal

I am deeply grateful to my DPhil supervisor, Professor David Clifton. I benefited tremendously from his academic mentorship, strategic advice, and personal guidance.

The SeroTracker team deserves particular thanks. It is an absolute privilege to have led such an incredible group of individuals. I am particularly grateful to my cofounder Tingting Yan, without whom this project would not have existed; to Niklas Bobrovitz, for his systematic review expertise and leading SeroTracker's data collection efforts; and to Mairéad Whelan and Harriet Ware, the next generation of SeroTracker's leadership. I also owe my thanks to the University of Calgary for hosting SeroTracker operationally, to McGill University for its tremendous academic partnership, and to Tim Evans and David Naylor for trusting Tingting and I to take our half-baked idea and make it reality.

Finally, this thesis would not have been possible without my family and friends. The close friends I made in Oxford – Matthew, Alex, Emily, Brian, Caro, and Dan, among many others – who made it from the place where I studied into the place where I made my life. My childhood friends – especially Arjun, Jarred, Graham, and Vlad – were my rocks through the tumultuous times. And many others who defy such neat categories, Anna and Nathan among them. To my family – Mama, Dada, and Rohit – thanks for dealing with me as I took constant Zoom meetings from the basement, entertaining me walking around the family home in a work-fueled haze, and always being there to support me and reinforce my values.

Institutional

SeroTracker would not have been possible without the generosity of its funders. These include the Public Health Agency of Canada through Canada's COVID-19 Immunity Task Force, the World Health Organization Health Emergencies Programme, the Robert Koch Institute, and the Canadian Medical Association Joule Innovation Fund. This manuscript does not necessarily reflect the views of the World Health Organization or any other funder.

I am grateful to the Rhodes Trust for funding my doctoral work, and to Rhodes House (particularly Mary Eaton) for tremendous support beyond this. I will never forget walking into the Rhodes House rotunda on my first day in Oxford – the sense of awe that, over the years, became deep comfort and a home base.

Abstract

Effective decision-making during pandemics requires understanding the true extent of infection and immunity. Seroprevalence studies measure the population prevalence of antibodies, providing key insight into these quantities. However, the results from these studies are dispersed, hindering their use. We developed SeroTracker, a dashboard and data platform for SARS-CoV-2 seroprevalence studies. Through its ongoing systematic review, SeroTracker identified over 3,800 seroprevalence studies and made their data available to global health, government, and academic stakeholders. Meta-analysing these studies, we show that global SARS-CoV-2 seroprevalence was 59.2% in September 2021 (95% confidence interval $[56.1 - 62.2\%]$), and rose steeply during 2021 due to infection in some regions and vaccination in others. We also identify limitations of seroprevalence studies: their timeliness, where the median study is published 154 days after being conducted; measurement bias, where assay manufacturers systematically overestimate sensitivity by 5.4% and specificity by 2.8%; and coverage, where seroprevalence studies are sparse in time and space. To address these challenges, we develop a Bayesian modelling framework that synthesises information from case and seroprevalence data to provide up-to-date estimates of true infections. We validate this model on simulated data and apply it to Canadian data, showing that only approximately 20% of infections were detected as cases at the start of the pandemic, and revealing substantial geographical heterogeneity in infection dynamics. Returning to SeroTracker's evidence synthesis process, we develop an automated tool for risk of bias assessment with excellent reliability compared to manual review (intraclass correlation 0.77), and implement a natural language processing-based process to screen abstracts for our systematic review that reduces screening time by approximately 70%. Finally, we reflect on the increased risk posed by pandemics and discuss future work applying engineering to public health to reduce these risks, including the development of a central data repository for respiratory virus surveillance data.

Contents

List of Figures **ix**

List of Tables **xi**

List of Abbreviations **xiii**

1 Introduction to SARS-CoV-2 serosurveillance **1**

1.1 Introduction and overview 1

1.2 Infectious disease surveillance 2

1.2.1 Types of surveillance 3

1.2.2 Characteristics of effective surveillance systems 4

1.3 SARS-CoV-2 virology, immunology, and diagnostics 4

1.3.1 The immune response to SARS-CoV-2 infection 4

1.3.2 Kinetics of the SARS-CoV-2 immune response 5

1.3.3 Correlates of protection against reinfection 6

1.3.4 Diagnostic testing for SARS-CoV-2 6

1.3.5 Antibody testing for SARS-CoV-2 7

1.4 Use and evolution of serosurveys during COVID-19 8

1.4.1 Serosurvey interpretation early in the pandemic 9

1.4.2 Serosurvey design after broad vaccine availability 9

1.4.3 Serosurvey interpretation after broad vaccine availability 11

1.4.4 Limitations and caveats 13

1.4.5 Reporting and synthesis 13

1.5 Evidence synthesis during health emergencies 14

1.5.1 Systematic reviews and meta-analyses 14

1.5.2 Dashboards 15

1.5.3 The future of pandemic monitoring 15

1.6 Aims of this thesis 16

2 Engineering the SeroTracker dataset and dashboard **18**

2.1 Motivation 18

2.2 Systematic review of SARS-CoV-2 serosurveys 19

2.2.1 Data sources and search strategy 20

2.2.2 Study inclusion criteria . . . 21
2.3 SeroTracker data extraction and database . . . 22
2.4 The SeroTracker dashboard . . . 26
2.4.1 Development process and philosophy . . . 26
2.4.2 Explore page and mapping . . . 26
2.4.3 Analyse page and charts . . . 30
2.4.4 Data page, process, and availability . . . 32
2.4.5 SeroTracker's users . . . 33
2.5 Discussion . . . 33
2.5.1 Embracing user-centered design in internal decisions . . . 34
2.5.2 Ensuring a focused scope . . . 34
2.5.3 Choosing the right tool for the right application . . . 35

3 Meta-analysing SARS-CoV-2 seroprevalence data 36
3.1 Motivation . . . 36
3.2 Jan 2020 - June 2020 systematic review . . . 37
3.3 Jan 2020 - Dec 2020 meta-analysis . . . 39
3.3.1 Characteristics of included studies . . . 39
3.3.2 Key limitations . . . 42
3.4 Jan 2020 - Apr 2022 meta-analysis . . . 42
3.4.1 Seroprevalence data . . . 44
3.4.2 Contextualising data . . . 45
3.4.3 Included studies . . . 46
3.4.4 Global and regional estimates of seroprevalence over time . . . 51
3.4.5 Determining the extent of case ascertainment . . . 57
3.4.6 Meta-analysis of seroprevalence differences between groups . . . 59
3.4.7 Meta-regression to examine study design and geographic factors . . . 61
3.4.8 SARS-CoV-2 seroprevalence in Africa . . . 64
3.5 Discussion . . . 66
3.5.1 Strengths and limitations . . . 69
3.5.2 Implications and next steps . . . 71

4 Delays and biases challenging SARS-CoV-2 seroprevalence estimation 73
4.1 Early methodological challenges and motivation . . . 73
4.2 Timeliness of seroprevalence study reporting . . . 76
4.2.1 Motivation . . . 76
4.2.2 Defining timeliness, representativeness, and data quality . . . 77
4.2.3 Dataset and descriptive analysis . . . 78
4.2.4 Timeliness by study publication venue . . . 79
4.2.5 Discussion . . . 84

4.3 Serological assays and measurement error . . . 87
4.3.1 Motivation . . . 87
4.3.2 Serological assays used in seroprevalence studies . . . 88
4.3.3 Comparison of assay performance evaluations . . . 91
4.3.4 Impact of misspecification on seroprevalence estimates . . . 96
4.3.5 Discussion . . . 98

5 Bayesian modelling to estimate the true extent of SARS-CoV-2 infection 100
5.1 Background . . . 100
5.2 Objective and workflow . . . 104
5.3 Bayesian inference and numerical Bayesian computation . . . 105
5.3.1 Numerical methods for Bayesian inference . . . 106
5.3.2 Monte Carlo Bayesian inference . . . 107
5.3.3 Markov Chain Monte Carlo . . . 108
5.3.4 Hamiltonian Monte Carlo . . . 110
5.4 Model architecture . . . 112
5.4.1 Infections process . . . 113
5.4.2 Indicator-based surveillance measures: reported cases . . . 115
5.4.3 Population-based surveillance measures: seroprevalence . . . 118
5.5 Model simulation . . . 121
5.5.1 Generate simulated data . . . 121
5.5.2 Simulated observations and prior predictive checks . . . 124
5.5.3 Simulated data recovery . . . 126
5.6 Application to real-world Canadian data . . . 129
5.6.1 Data . . . 129
5.6.2 Model fit and diagnostics . . . 130
5.6.3 Posterior predictive checks . . . 131
5.6.4 Interpreting model fit . . . 134
5.7 Discussion . . . 141

6 Expediting rigorous evidence synthesis during health emergencies 143
6.1 Motivation . . . 143
6.2 Automating reproducible risk of bias assessment for seroprevalence studies 144
6.2.1 Operationalising and automating the JBI checklist . . . 145
6.2.2 Developing decision rules for risk of bias assessment . . . 149
6.2.3 Evaluating coverage, reliability, and efficiency . . . 151
6.2.4 Discussion . . . 153
6.3 Real-world assessment of NLP-based screening . . . 154
6.3.1 Developing a NLP-enabled tool to expedite abstract screening . . . 156

6.3.2 Real-world evaluation of the NLP-enabled tool . . . 159
6.3.3 Process measures: screening efficiency and accuracy . . . 160
6.3.4 Outcome measures: review results and tool usage . . . 161
6.3.5 Structure measures: one-person one-tool voting model . . . 164
6.3.6 Discussion . . . 164

7 Discussion and future work **166**
7.1 Motivation . . . 166
7.2 SeroTracker in context . . . 167
7.2.1 Public health surveillance data sharing . . . 167
7.2.2 SARS-CoV-2 seroprevalence evidence syntheses . . . 169
7.2.3 Seroprevalence dashboards and computable evidence synthesis . . . 170
7.2.4 Enabling user-centered evidence synthesis . . . 171
7.3 Future work . . . 173
7.3.1 Extended Bayesian modelling framework for surveillance data . . . 173
7.3.2 PathoTracker: unified data platform for population-based surveillance . . . 175
7.3.3 Emerging detection technology assessment and prioritisation . . . 177
7.4 Conclusion . . . 178

Appendices

A SeroTracker full data schema and tables **180**
A.1 Sources of seroprevalence data . . . 180
A.2 Seroprevalence studies . . . 181
A.2.1 Country data for each study . . . 181
A.2.2 Risk of bias table . . . 184
A.3 Seroprevalence estimates . . . 184
A.3.1 Serological assays table . . . 187
A.3.2 Serological assay evaluations table . . . 188

B SeroTracker Analyse page visualisations **192**

C Natural language processing for abstract screening: additional results **194**
C.1 Summary of project measures, methods, and results . . . 194
C.2 Additional analyses . . . 194
C.2.1 Process measures: Inclusion recommendation performance on the pilot abstracts . . . 194

References **198**

List of Figures

1.1 Conceptual decision tree for distinguishing between infection-induced and vaccine-induced antibodies . . . 11

2.1 SeroTracker entity-relationship diagram . . . 25
2.2 May 2020 Explore page heatmap design . . . 27
2.3 August 2020 Explore page binary design . . . 28
2.4 Current SeroTracker Explore page . . . 28
2.5 Current SeroTracker Explore page modals . . . 29
2.6 SeroTracker data architecture . . . 30
2.7 May 2022 Analyse page . . . 30
2.8 Current SeroTracker Analyse page . . . 31
2.9 Current SeroTracker Data page . . . 32

3.1 Map of planned, ongoing, and completed seroprevalence studies as of 26 June 2020 . . . 38
3.2 Map of serosurvey distribution by *Global Burden of Disease* region as of 31 December 2020 . . . 40
3.3 Flow diagram of included studies . . . 47
3.4 WHO member states with seroprevalence data identified . . . 48
3.5 Risk of bias assessment for included studies . . . 50
3.6 Estimated seroprevalence globally and by WHO region from January 2020 to March 2022 . . . 56
3.7 Meta-analysis of seroprevalence ratios between demographic groups within studies . . . 61
3.8 Meta-regression to identify study design, country, and COVID-19 context factors associated with seroprevalence pre-vaccination . . . 64
3.9 Reported seroprevalence, variants of concern, and cumulative incidence by selected countries over time, March 2020 - December 2021 . . . 66

4.1 Characteristics of SARS-CoV-2 seroprevalence studies . . . 74
4.2 Kaplan-Meier curve and risk table for time to publication by publication venue . . . 80
4.3 Pareto chart of serological assay use . . . 90

4.4 Comparison of sensitivity and specificity from manufacturer, third-party, and independent group evaluations 92
4.5 Consequences of correcting seroprevalence estimates using biased estimates of sensitivity and specificity 97

5.1 Overall model architecture 113
5.2 Infections model diagram and specification 114
5.3 Model diagram for cases and indicator-based surveillance 117
5.4 Model diagram for seroprevalence and population-based surveillance . . . 122
5.5 Simulated infections and observed cases 125
5.6 Simulated infections, proportion vaccinated, and observed seroprevalence . 125
5.7 Trace plot and model diagnostics $\hat{R}$ and $\log_{10}(N_{eff})$ for parameter recovery 127
5.8 True infections curve and model estimated infections for parameter recovery 128
5.9 All parameters and region-specific intercepts for parameter recovery . . . 128
5.10 Trace plot and model diagnostics $\hat{R}$ and $\log_{10}(N_{eff})$ for fit to Canadian data 131
5.11 Posterior predictive distributions and observed values for cases C in Canadian data fit 132
5.12 Posterior predictive distributions and observed values for seroprevalence from infection SP_I in Canadian data fit 132
5.13 Posterior predictive distributions and observed values for seroprevalence from infection or vaccination SP_{IV} in Canadian data fit 133
5.14 Modelled infections I for Canadian data fit 134
5.15 Modelled proportion of infections detected ρ for Canadian model 136
5.16 All parameters and region-specific intercepts for Canadian model 138

6.1 Decision rules to determine risk of bias for seroprevalence studies 151
6.2 The NLP-enabled tool augmenting Covidence abstract screening 156
6.3 Calibration curve of the inclusion recommendation system 158
6.4 Receiver operating characteristic curve of the inclusion recommendation model 159
6.5 Operating characteristics of the tool compared to user votes 163

List of Tables

2.1 Inclusion criteria for our systematic review of SARS-CoV-2 serosurveys . 22

3.1 Summary of general population seroprevalence studies as of 31 December 2020 41
3.2 Characteristics of included studies, Jan 2020 - May 2022 48
3.3 Seroprevalence to cumulative incidence ratios by WHO region, income level, and quarter 58

4.1 Attributes and measures of effective public health surveillance used in this study 78
4.2 Cox regression table 83
4.3 Features of commercial and self-developed serology assays used by studies 89
4.4 Predictors of assay sensitivity and specificity, estimated with mixed-effects Beta regression 94

5.1 Summary of three models of true SARS-CoV-2 infections 103
5.2 Model indices and variables used in the infections model 115
5.3 Variables in model for cases and indicator-based surveillance 116
5.4 Variables in model for seroprevalence and population-based surveillance. . 123

6.1 Risk of bias table fields 146
6.2 Operating characteristics of the tool at different confidence thresholds on the test set of abstracts. Abbreviations as follows: TN: true negatives; FP: false positives; FN: false negatives; TP: true positives; Prec.: precision; Rec.: Recall; F1: F1 score 158
6.3 Precision, recall and F1-scores in all user-tool pairings for texts that were included in full-text screening 161
6.4 Summary statistics of the seroprevalence estimates from the included texts in the pre-project, without-tool and with-tool stages 162
6.5 Tool feature usefulness as rated by nine users 163

A.1 Source table fields 180
A.2 Country table fields 182
A.3 Study table fields 183

A.4 Estimates table fields 184
A.5 Serological assay table fields 188
A.6 Reference panels of five third-party lab evaluations 190
A.7 Serological assay evaluations table fields 191

C.1 Operating characteristics of the tool evaluated on the pilot abstracts at different inclusion thresholds 194
C.2 Key methods and results of process, outcome and structure measures evaluated 195

List of Abbreviations

CI . . . . . Confidence interval; reflects uncertainty in the estimate of a parameter. In meta-analysis, reflects uncertainty in summary effect size.

COVID-19 . . . Coronavirus disease 2019.

HDCrI . . . . . Highest density credible interval; reflects uncertainty in the Bayesian estimate of a parameter.

HR . . . . . Hazard ratio.

Ig . . . . . Immunoglobulin.

IQR . . . . . Interquartile range; i.e., the 25th percentile and 75th percentile.

JBI . . . . . The Joanna Briggs Institute; an internationally recognised evidence synthesis group, which has produced a checklist for critical appraisal of prevalence studies.

NLP . . . . . Natural language processing.

OPOT . . . . . One person, one tool; a model for systematic review screening where one vote is provided by a human reviewer and the second by a language model. The human reviewer may be reviewing with (OPOT-W) or without (OPOT-W/O) access to tool recommendations.

OR . . . . . Odds ratio.

PCR . . . . . Polymerase chain reaction.

PHSM . . . . . Public health and social measures; used to mitigate infectious disease spread. Sometimes referred to as non-pharmaceutical interventions.

PIO . . . . . Population, intervention, and outcome; three parts of an epidemiological research question used to assess article eligibility for inclusion in systematic reviews.

PR . . . . . Prevalence ratio; risk ratio for seroprevalence.

SARS-CoV-2 . Severe acute respiratory syndrome coronavirus 2; the virus that causes COVID-19.

SP . . . . . Seroprevalence.

RoB . . . . . Risk of bias.

TPR . . . . . Test positivity rate; proportion of conducted diagnostic tests which have positive results

WHO . . . . . World Health Organization.

1

Introduction to SARS-CoV-2 serosurveillance

1.1 Introduction and overview

The COVID-19 (coronavirus disease 2019) pandemic, caused by the coronavirus SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), represents one of the largest public health emergencies of our time [1, 2]. Effective infectious disease surveillance is crucial to monitor this pandemic, understand its nature, and inform decision-making.

COVID-19 surveillance often relies on diagnostic testing for active infection to identify confirmed cases [1]. As of 1 November 2022, there have been over 630 million confirmed cases of COVID-19 worldwide [3]. However, confirmed cases inevitably underestimate the true cumulative incidence of infection: many infected individuals go undiagnosed because of asymptomatic infection [4], limited diagnostic test supply [5], and disparities in test availability [6]. As a result, the 630 million confirmed cases of COVID-19 worldwide represent a substantial underestimate of the true number of infections.

In contrast, serological assays measure levels of antibodies against SARS-CoV-2, thereby providing evidence of previous SARS-CoV-2 infection or vaccination [7, 8]. Data from population-wide antibody testing studies, referred to as seroprevalence studies or serosurveys, therefore enable estimation of the cumulative incidence of infection and population levels of immunity. In this way, they complement diagnostic testing and can help shape public health responses to COVID-19 [1, 9].

Over 3,500 SARS-CoV-2 serosurveys have been done to date, but the results from these studies have been dispersed across published manuscripts, preprints, government reports, and media articles. Moreover, seroprevalence data is also delayed, subject to biases, and difficult to interpret.

This DPhil thesis focuses on effectively synthesising seroprevalence data to enable effective public health decision-making. The present chapter sets the stage for this work. Here, we review the principles of infectious disease surveillance, discuss COVID-19 background and surveillance technologies, critically examine the use and interpretation of serosurveys for SARS-CoV-2, discuss evidence synthesis during health emergencies, and describe the aims and contributions of this thesis.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Arora RK*, Cheng B*, . . . , Subissi L, Bergeri I. 2021. Operational brief on interpreting SARS-CoV-2 seroprevalence data. Internal World Health Organization brief. *Rahul's contributions:* outlining, writing, and revising policy brief, in equal proportion to Cheng B.

Duarte NI, Yanes-Lane M, **Arora RK**, . . . , Papenburg J**, Langlois M-A**. 2021. Adapting serosurveys for the SARS-CoV-2 vaccine era. *Open Forum Infectious Diseases. Rahul's contributions:* conceptualization of manuscript and its central points; substantial rewriting and revision of the drafted version.

Donnici C*, Ilincic N*, Cao C, . . . , Bobrovitz N, **Arora RK.** 2022. Timeliness of reporting of SARS-CoV-2 seroprevalence results and their utility for infectious disease surveillance. *Epidemics. Rahul's contributions:* conceptualization of manuscript; defining research question and objectives; methodology development, including conceptual framework and analytical methods; substantial rewriting and revision of the drafted version.

1.2 Infectious disease surveillance

Infectious disease surveillance systems are monitoring programs which aim to provide decision-relevant information on infectious disease emergence, spread, and immunity [10].

1.2.1 Types of surveillance

We discuss three major types of infectious disease surveillance: event-based, indicator-based, and population-based surveillance.

Event-based surveillance is a strategy for early detection of infectious disease outbreaks [11]. This strategy draws on a breadth of unstructured information, such as news reports, social media, community discussion, and others to identify early indicators of potential outbreaks. In relying on these informal sources of information, event-based surveillance can enable rapid identification of potential outbreaks. This was the case for COVID-19, where the event-based surveillance platform *ProMED* identified a cluster of pneumonia of unknown origin in Wuhan, China, on 31 December 2019 [12].

In contrast to event-based surveillance, indicator-based surveillance relies on structured indicators to measure the dynamics of a known infectious disease [11]. Examples of common indicators include confirmed cases of notifiable diseases [13], hospitalisations for those diseases, and deaths where the cause of death is linked to such a disease. This is invaluable standardised and quantitative data.

However, indicator-based surveillance relies on the identification of the indicator and its reporting back to a central system, and therefore tends to systematically underestimate the true extent of a disease's spread or consequence [14]. This underestimation results from two phenomena: underascertainment, where not all individuals seek diagnosis or care for an infection, and under-reporting, where this diagnosis or care is not centrally reported for all individuals receiving it.

Population-based surveillance aims to systematically surveil the population to measure the true extent of infection and immunity in the population [15]. In this way, population-based surveillance is not subject to the underestimation bias inherent to indicator-based surveillance [14]. This sort of surveillance strategy can be laborious and expensive to conduct, as it requires public health agencies to conduct systematic investigations that extend beyond the indicators routinely generated in patient care[15].

Three sorts of population-based surveillance tools have seen considerable use:

1. First, seroprevalence studies, which aim to sample a population to estimate the prevalence of antibodies against a pathogen [9]

2. Second, point prevalence studies, which involve the systematic testing of a random sample of the population for active infection, enabling the point prevalence of active infection to be estimated [16]

3. Finally, environmental studies, which examine wastewater or other environmental samples for indicators of infection – for example, in the case of COVID-19, SARS-CoV-2 genetic material shed by infected individuals into wastewater [17]

1.2.2 Characteristics of effective surveillance systems

For these surveillance systems to be useful for public health decision-making, it is crucial that they have certain key attributes [10]. Such attributes include timeliness, representativeness, and bias, among several others [18]. On timeliness, it is crucial that these systems provide information in a fashion rapid enough to inform decision-making. On representativeness, to make decisions that affect a population based on this data, it is essential that the information from these systems can be taken to be representative of that population. On bias, for the information generated by surveillance systems to be relevant to decision-making, that information must not be substantially biased away from its true value.

1.3 SARS-CoV-2 virology, immunology, and diagnostics

SARS-CoV-2 is a highly transmissible respiratory virus in the betacoronavirus family. The spike (S) protein on the surface of SARS-CoV-2 mediates its entry into infected cells, while its nucleocapsid (N) protein helps form its protein shell. Infection with SARS-CoV-2 can cause COVID-19, an infectious disease spread via aerosols and other respiratory modes of transmission.

1.3.1 The immune response to SARS-CoV-2 infection

Immune responses to a pathogen like SARS-CoV-2 have two components: the innate immune response and the adaptive immune response [19]. Innate immunity is rapid and not pathogen-specific, characterised by inflammation and endocytosis or phagocytosis (engulfment by immune cells) of foreign bodies [19]. In contrast, adaptive immunity is

pathogen-specific, and generates an immune memory which can potentiate rapid and robust immune responses upon subsequent exposure to the same pathogen [19].

In the adaptive immune response, antigen-presenting cells present protein fragments to T cells for recognition [19]. These T cells have antigen-specific receptors that can recognise proteins from a foreign pathogen. If a T cell receptor recognises a presented antigen, T cells amplify and differentiate into cytotoxic T cells, which destroy infected host cells, and helper T cells, which promote the amplification of B cells and their maturation into plasma cells [19].

Plasma cells produce antibodies that can neutralise pathogens and otherwise promote their destruction [19]. There are five isotypes of antibodies, or immunoglobulins (Ig), each with different roles: for example, IgM is the first Ig expressed during the adaptive immune response, and IgG is the primary antibody responsible for neutralising viruses. After active infection resolves, memory B cells remain to respond to subsequent infections [19]. In this way, the presence of antibodies indicates past infection or vaccination. Antibodies against SARS-CoV-2 usually bind specifically to the virus's spike protein, termed anti-spike antibodies, or its nucleocapsid protein, termed anti-nucleocapsid antibodies [20].

1.3.2 Kinetics of the SARS-CoV-2 immune response

Each component of the SARS-CoV-2 immune response has distinct kinetics. Individuals infected by or vaccinated against SARS-CoV-2 take one to three weeks to seroconvert (exhibit a detectable antibody response), with 90% of individuals ultimately seroconverting [20, 21]. Anti-spike and anti-nucleocapsid antibodies decay at different rates, with anti-spike protein antibodies slowly decaying but remaining detectable for several months after infection [20, 22–25], while anti-nucleocapsid antibodies decline more rapidly in the months following infection [20, 26].

Cellular immunity appears more stable. Both CD4+ and CD8+ T cells appear to last for at least 12 months [25, 27]. In contrast, memory B cells and resting plasma cell levels also appear stable for several months after infection or vaccination [25, 28, 29].

1.3.3 Correlates of protection against reinfection

A measurable correlate of protection against SARS-CoV-2 infection would help identify the individuals most vulnerable to reinfection [30]. Binding and neutralising antibody levels appear to correlate with vaccine efficacy [30–32], suggesting that measuring antibodies could be used to estimate an individual's level of protection from subsequent infection. Importantly, cellular immunity may protect previously infected or vaccinated individuals even after neutralising antibody levels fade [31].

The extent of protection depends on the SARS-CoV-2 variant in question. Several such variants have emerged since the beginning of the pandemic, with the dominant variants becoming progressively more transmissible. The currently dominant Omicron variant has a distinct spike protein antigen, making it extremely transmissible and much more adept at evading immunity from previous infection or vaccination [33, 34]. With the Omicron variant, infection or vaccination confers transient protection from infection but long-lasting protection from severe disease.

1.3.4 Diagnostic testing for SARS-CoV-2

Testing for SARS-CoV-2 can serve two purposes: diagnosing SARS-CoV-2 infections in individuals [1] and conducting population-level surveillance of SARS-CoV-2 spread [7].

COVID-19 diagnostic testing assesses patient samples to see whether the SARS-CoV-2 virus is present [1]. SARS-CoV-2 infection primarily affects the respiratory tract, meaning that testing usually involves taking respiratory samples. These diagnostic tests provide evidence of active SARS-CoV-2 infection, but do not indicate or rule out previous infection [1]. This individual-level diagnostic testing can ultimately contribute to indicator-based infectious disease surveillance.

Nucleic acid amplification tests, like polymerase chain reaction (PCR), detect any viral nucleic acids that may be present in patient samples by amplifying them [35, 36]. In PCR-based tests for SARS-CoV-2, the ribose nucleic acid that makes up the viral genetic material is first reverse transcribed into deoxyribose nucleic acid [36]. Virus-specific genetic material selectively amplified with virus-specific fragments of genetic material termed primers, a replication enzyme, and a thermal cycler. The amount of

amplified genetic material can then be quantified [36]. If viral genetic material was present in the initial sample, it will be amplified and detected, and the test considered positive. Collectively, this process is referred to as quantitative reverse transcription PCR [36], and is the gold standard for diagnosing SARS-CoV-2 infection [1].

The use of primers makes PCR highly specific to a specific segment of the viral genome. False positives are very uncommon in the absence of sample contamination, and these tests have nearly perfect specificity (approximately 100%)[37, 38]. However, patient samples may contain less genetic material than a test's limit of detection, so these tests sometimes produce false negatives and hence have a lower sensitivity (approximately 88%)[39].

Antigen tests can also be used to diagnose SARS-CoV-2 infection [1]. These tests detect viral proteins. Compared to PCR, antigen tests have worse sensitivity and specificity [40], but are simpler to carry out and more affordable [41]. These tests are often done at home or at the point of care, meaning that their results are often not centrally reported – limiting their use for indicator-based surveillance.

Diagnostic testing supplies and resources have been in short supply during the pandemic, severely limiting diagnostic testing capacity [42]. Available testing capacity is poorly distributed, with worse access to testing among non-White individuals [6], individuals with lower socioeconomic status [43], or individuals with limited access to health care resources [44]. In addition to having limited testing access, these groups are also often those at higher risk of poor outcomes from COVID-19 [45, 46]. Finally, individuals with asymptomatic SARS-CoV-2 infection often have no reason to seek testing, and hence their infections often go undiagnosed [47]. Together, limited test volumes, maldistribution of testing, and asymptomatic infections mean that many SARS-CoV-2 infections have gone undiagnosed.

1.3.5 Antibody testing for SARS-CoV-2

Adaptive immunity is pathogen-specific. As a result, testing for components of the adaptive immune response can indicate previous SARS-CoV-2 infection or vaccination [7]. Antigen-specific B and T cell responses can be evaluated using techniques like flow

cytometry, but high-throughput equivalents of these assays are still being developed [48]. In contrast, high-throughput antibody testing is economical and feasible, and serological assay platforms enabling it are broadly available [49].

In most serological assays, the antigen of interest – say, the SARS-CoV-2 spike protein – is affixed to a solid support [8, 50]. Then, a patient sample is added, and any anti-spike antibodies in this sample will bind to the fixed spike protein. A detectable secondary antibody (i.e., an antibody that recognises IgG antibodies, with an enzyme or fluorophore attached) is then added, and will bind to any anti-spike IgG antibodies that are present [50]. By measuring the amount of enzyme or fluorescence present, we can infer the concentration of anti-spike IgG antibodies (i.e., the antibody titer). With minor modifications to this procedure, we can also detect other antibody isotypes (e.g., IgM, IgA, total Ig) or anti-nucleocapsid antibodies [50].

For SARS-CoV-2 antibody assays, enzyme-linked immunosorbent assay refers to the laboratory-based version of this procedure where an enzyme is attached to the secondary antibody for detection, and chemiluminescent immunoassay to the laboratory-based procedure where a fluorophore or chromophore is conjugated to the secondary antibody [51]. Lateral flow serological assays package similar principles in a portable and rapid format, trading these advantages lower assay performance [7].

1.4 Use and evolution of serosurveys during COVID-19

Seroprevalence studies, or serosurveys, are population-based surveillance investigations that are crucial to understand the true extent of infection and immunity in a population [52–54]. These studies estimate the population prevalence of antibodies by administering serological assays to a representative sample from the target population.

To estimate seroprevalence in the general population, serosurveys either randomly sample households or community members, or take the seroprevalence in a specific group as a proxy for the general population [53, 55, 56]. Such proxy groups include blood donors, which enable the use of banked samples, and health service attendees, as a rapid option. These serosurveys are typically designed as one-time cross-sectional, repeated cross-sectional, or prospective cohort studies.

We describe the design and interpretation of SARS-CoV-2 serosurveys during two major eras of the COVID-19 pandemic: before vaccination and after broad vaccine availability.

1.4.1 Serosurvey interpretation early in the pandemic

Before broad vaccine availability, SARS-CoV-2 serosurveys are relatively straightforward to interpret because SARS-CoV-2 antibodies could be assumed to result from infection. In this scenario, serosurveys provide the best available information to estimate the true extent of past SARS-CoV-2 infection [57].

By estimating the prevalence of antibodies in the population, these studies capture infections in individuals who were asymptomatic or did not have access to diagnostic testing. This is in contrast to indicator-based surveillance – while cases and deaths are ubiquitous and important to monitor, these measures typically underestimate the true burden of infection due to under-reporting and underdiagnosing of asymptomatic and mild cases [58, 59].

Comparing seroprevalence between different sociodemographic groups can provide an important indication of which groups have been disproportionately affected [57, 60]. Seroprevalence data are also the best indicator of the proportion protected from reinfection – a key parameter for mathematical models of infectious disease [61].

Finally, serosurveys can also help inform vaccine trials and distribution before broad vaccine availability. Regions with low seroprevalence and a high incidence of new cases are promising sites for vaccine trials. Moreover, because serosurveys indicate which groups have been disproportionately infected by SARS-CoV-2, they can be used to inform vaccine distribution strategies [62].

1.4.2 Serosurvey design after broad vaccine availability

After vaccination becomes broadly available, it is crucial to differentiate seropositivity resulting from past infection and vaccination. This stratification enables public health officials to continue to monitor infection rates.

SARS-CoV-2 serological assays generally detect antibodies against either the spike or nucleocapsid protein. Infection elicits both anti-spike and anti-nucleocapsid antibodies. In contrast, some vaccines in use today use spike protein antigens (e.g., Moderna, Pfizer, AstraZeneca) and therefore elicit only anti-spike antibodies, while others use whole virus antigens (e.g., COVAXIN, Sinovac) and trigger both spike and nucleocapsid protein antibodies [63].

If only vaccines triggering anti-spike antibodies have been used, anti-spike seroprevalence results from infection and vaccination, whereas anti-nucleocapsid seroprevalence results from infection alone. Investigators in these countries can estimate the true extent of infection by measuring seroprevalence from infection alone. This is possible in serosurveys that measure both anti-spike and anti-nucleocapsid antibodies, using an algorithm outlined in Figure 1.1. In countries where inactivated virus vaccines have seen broad use, it is possible to isolate seroprevalence from infection by using vaccine coverage data (Section 3.4.4), but doing so is more analytically complex and relies on several assumptions.

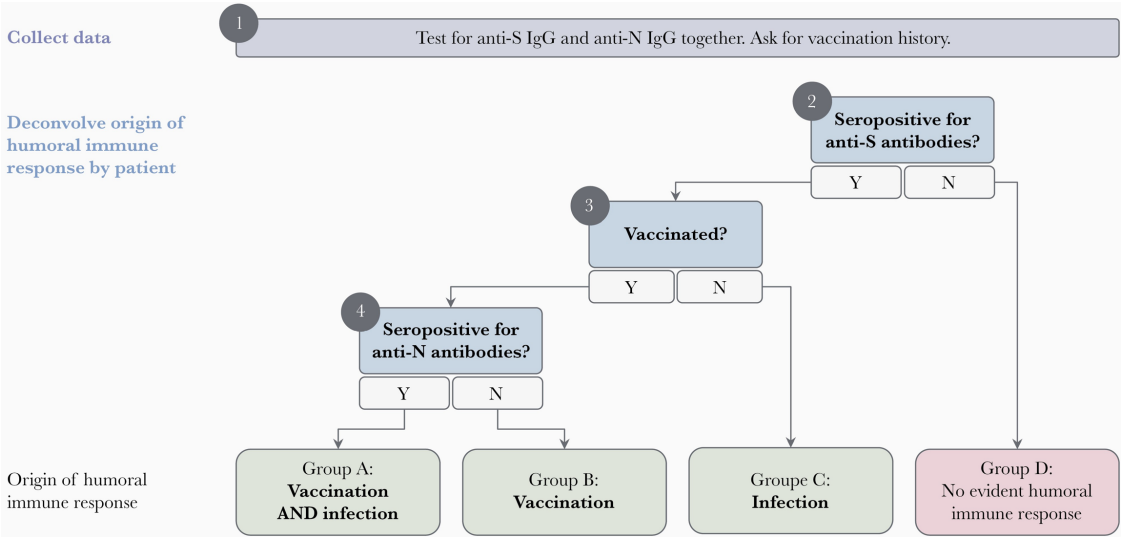


Figure 1.1: Conceptual decision tree for distinguishing between infection-induced and vaccine-induced antibodies. This decision tree is only applicable for populations where exclusively S protein antigen vaccines have been used. Individuals negative for anti-spike antibodies have likely been neither infected nor vaccinated (Group D). Among unvaccinated individuals who are anti-spike positive, seropositivity likely results from prior infection (Group C). In vaccinated individuals who are seropositive for anti-spike antibodies, investigators should determine whether individuals were previously infected based on anti-nucleocapsid seropositivity: anti-spike positive and anti-nucleocapsid negative individuals were likely vaccinated only (Group B), while anti-spike and anti-nucleocapsid positive individuals were likely both infected and vaccinated (Group A). Diagnostic history can also confirm prior infection if available; however, the absence of a past diagnosis does not preclude prior infection, considering imperfect recall of past infection and limited diagnostic test availability [64].

1.4.3 Serosurvey interpretation after broad vaccine availability

The applicability and public health value of seroprevalence studies also shifts after broad vaccine availability. In the vaccine era, serosurveys can be used for efficient vaccine allocation. Early in vaccine distribution, when vaccine supply is limited, it is crucial to distribute vaccine doses as efficiently as possible to protect the most vulnerable people. Vaccine prioritisation has largely been informed by mortality and exposure risk, which are both key considerations. To further improve distribution efficiency, public health officials can use serosurveys to identify regions with large uninfected and unvaccinated populations, and consider these for prioritisation to improve the marginal impact of each dose [62]. Additionally, previously infected individuals who receive n vaccine doses are likely to have similar immune responses to infection-naive individuals after $n + 1$ doses [65]. This observation suggests that reduced-dose regimens could be used in regions

where many individuals have been infected, enabling subsequent doses or the requisite healthcare capacity to be routed to the areas where they are most useful.

Serosurveys can also be used to estimate the true extent of protection in a population. Vaccine coverage data has often been used to estimate population SARS-CoV-2 protection. However, past SARS-CoV-2 infection is also associated with reduced risk of subsequent infection [66–69]. Seroprevalence data can therefore be used with vaccine coverage data to understand how many people in a population are protected from reinfection from all causes – both as a result of past infection and vaccination [14]. Incorporating seroprevalence data into estimates of protection is especially important in countries where there have been many past infections, or where vaccine coverage is low. This information can thus inform decisions to strengthen or relax infection and prevention control measures, and guides assessments of the impact of vaccination programs and public health and social measures on population immunity to SARS-CoV-2, as the disease progresses towards endemicity [70].

Finally, serosurveys can be used to inform the timing of subsequent vaccinations. SARS-CoV-2 is becoming an endemic pathogen [71]. Seroprevalence data can inform tailored vaccination strategies to manage COVID-19 as an endemic disease. SARS-CoV-2 vaccination is the only safe way to boost immunity. However, vaccination has its costs – both at the individual level, in terms of lost productivity, and at the population level, including the financial cost to procure vaccines and the health capacity cost to administer them. These costs are especially salient in limited-resource parts of the world, where COVID-19 competes with other public health priorities [72, 73]. Just as seroprevalence studies enable population protection to be estimated, so too can they be used to examine how immunity changes and wanes over time [74]. This is especially true for serosurveys that use scalable quantitative assays or surrogate neutralisation assays [75], which enable mapping between antibody levels and levels of protection. Understanding how population protection will change over time can help inform if or when additional doses of vaccination are necessary, enabling effective targeting of public health resources [76].

1.4.4 Limitations and caveats

When interpreting seroprevalence data, it is crucial to keep in mind our knowledge of SARS-CoV-2 antibody immunity, and the design and quality of the seroprevalence study providing this data [77].

Not all individuals who have been previously infected will be seropositive. This could result from an individual having a false negative antibody test result, that individual not developing a detectable antibody response after infection or being sampled before that response could develop [21], or that individual's antibodies already having waned below detectable levels [26]. These phenomena can cause seroprevalence studies to underestimate the true burden of past SARS-CoV-2 infection [78], especially when anti-nucleocapsid antibody tests are used, or when six or more months have passed since the early cases of COVID-19 [25].

False positive antibody test results can also occur, particularly with patients who have antibodies for related coronaviruses like the SARS (severe acute respiratory syndrome) virus or the MERS (middle eastern respiratory syndrome) virus [79]. High false positive rates can lead to overestimation of the true infection burden, especially where the true seroprevalence is low [80].

Finally, robust serosurvey design is also crucial. Seroprevalence should be considered in the context of the population sampled and how sampling was done. Some studies use non-probability sampling methods (e.g., convenience samples), meaning that seroprevalence estimates are not representative of the population at large [81]. Certain groups can sometimes be underrepresented in serosurveys, including individuals of low socioeconomic status, racial minorities, and those in rural areas [81]. Serosurvey authors often also fail to correct their seroprevalence estimates for the differences between their study population or population demographics, or for the characteristics of the tests they used [9]. We will review these factors in more detail in Chapter 6.

1.4.5 Reporting and synthesis

For the first 1.5 years of the pandemic, there was no standard reporting format for SARS-CoV-2 serosurveys. Many serosurveys early in the pandemic did not adequately report

the test they used, test sensitivity and specificity, sampling method, or even the sample size [81]. Some of these reporting challenges persist today, despite our publication of the ROSES-S reporting guidelines in mid-2021 [77].

Synthesising data across these studies was a major challenge early on. This heterogeneity in serosurvey quality and reporting highlights the importance of synthesising evidence from serosurveys.

1.5 Evidence synthesis during health emergencies

1.5.1 Systematic reviews and meta-analyses

The gold standard for evidence synthesis is the systematic review and meta-analysis [82]. Systematic literature reviews are considered primary medical research, and are characterised by clear inclusion criteria and a comprehensive search strategy defined *a priori*. Articles identified through the search strategy are screened, predefined data elements extracted, and included studies analysed [82]. Much of this work happens in duplicate to improve rigour. In meta-analysis, findings from studies in a systematic review are pooled, enabling the synthesis of evidence across many studies [82].

Systematic reviews prioritise rigour over speed in synthesising medical research findings. This key strength is also their key failing during health emergencies, as the median systematic review takes 61 weeks to be published after the final search is conducted, and 33 weeks from final search to publication is considered rapid [83]. These reviews are also often never updated, meaning that their findings can lose relevance quickly in the face of new evidence [84]. Moreover, delays in the process of summarising evidence can jeopardise public health responses during emergency health crises, where the body of primary literature may quickly accumulate.

Living systematic reviews, which describe a process to routinely update a SR after initial publication, have been proposed to accelerate the integration of time-sensitive evidence [85, 86]. However, updates to living systematic reviews are typically published only periodically (e.g., every 3-6 months) and as such, still represent a mostly static form of knowledge translation that may quickly lose relevance when more evidence is released [87, 88]. Moreover, findings from living systematic reviews are most often shared in research

papers or guidelines, which make it difficult for users to access the evidence directly, draw upon it for secondary analysis, and monitor key changes in real time [89, 90].

1.5.2 Dashboards

Dashboards and databases create a common source of knowledge and act as a foundational resource for other analyses, particularly in epidemiology [91, 92]. These dashboards give their target audience – be it public health experts, policymakers, or members of the lay public – the ability to visualise and manipulate public health data [89, 90]. Dashboards are often coupled with databases accessible via application programming interface (API), enabling people worldwide to build end-to-end pipelines from data to visualisation for responsive public health inference. Collectively, these public dashboards and databases reduce duplication of effort in gathering raw data and make information more accessible to the user [91, 92].

Some dashboards are continuously updated and intend to give their users a picture of the constantly evolving public health landscape. Dashboards of this sort have been crucial in the COVID-19 scientific and public health response – for example, *Our World in Data*'s vaccination dashboard, *Johns Hopkins University*'s indicator-based surveillance dashboard and its local analogues, and the *Oxford COVID-19 Government Response* tracker [3, 93–95]. Another kind of dashboard has new data uploaded infrequently, if ever, and aims to permit their users to easily explore data that would otherwise have been opaque [96].

Dashboard development relies on access to aggregated data in a standardised and cleaned format. Most dashboards do this by synthesising and joining already available datasets [96], or by sourcing their data through regular and up-to-date sources of clear indicators [3, 97]. On the other hand, seroprevalence data, mostly reported in the academic literature, requires dedicated resources for manual data abstraction, data homogenisation, and critical appraisal.

1.5.3 The future of pandemic monitoring

The COVID-19 pandemic has been accompanied by an explosion of publications, preprints, and reports [98]. Given the urgency of the pandemic, important results are increasingly

being posted in rapid and less conventional ways, with preprints in particular driving the global discourse [99].

This explosion in data has also held true for serosurvey reports. Complicating matters is the fact that these data are reported in disperse fashion, across published articles, preprints, government reports, presentations, news media, and even social media [81]. Synthesising key findings from serosurvey data requires synthesising evidence across heterogeneous data sources and making those findings easily accessible.

1.6 Aims of this thesis

At the start of the COVID-19 pandemic, we recognised the opportunity seroprevalence data presented, as well as the challenges that would hinder its impact. The group I lead, SeroTracker, is conducting a continuously updated systematic review and meta-analysis of all SARS-CoV-2 seroprevalence studies globally [9, 81]. We use the data gathered from this systematic review to build SeroTracker.com, a dashboard and data platform for SARS-CoV-2 serosurveys [100].

In this thesis, I engineer the SeroTracker SARS-CoV-2 data platform, develop novel statistical models and evidence synthesis tools for SARS-CoV-2 serosurveillance, and use these to provide a comprehensive analysis of serosurveillance, infection, and immunity during the COVID-19 pandemic.

Chapter 2 describes the design and engineering of the novel SeroTracker data platform, which combines a continuous systematic review with a seroprevalence-specific data architecture and an up-to-date dashboard to synthesise information from all SARS-CoV-2 seroprevalence studies worldwide.

In Chapter 3, we use SeroTracker's data to synthesise data from seroprevalence studies globally, conducting three increasingly sophisticated evidence syntheses: the first in early 2020, to identify lessons from early seroprevalence data; the second bringing together studies published in all of 2020; and the third providing a comprehensive picture of how seroprevalence has changed globally and over time.

This data is not without its limitations. Chapter 4 discusses the limitations of seroprevalence data as surveillance tools during COVID-19, with a particular emphasis on the

substantial delays before studies were published, on biases introduced by heterogeneous and poorly performing serological assays, and by sparse data availability.

Chapter 5 moves to address these limitations. We design a sophisticated hierarchical Bayesian modelling framework for up-to-date estimation of infection and immunity by integrating indicator-based and population-based surveillance data. We validate this modelling framework on simulated data and then apply it to Canadian data.

In Chapter 6, we reflect on the challenges we faced conducting evidence synthesis during the COVID-19 health emergency. We describe three strategies we developed to improve the speed and quality of evidence synthesis: an automated tool to assess risk of bias, evaluating natural language processing-based abstract screening in the real world, and thinking in a paradigm that emphasises research products over research papers.

Finally, Chapter 7 synthesises this body of work, situates it in broader context, and proposes next steps. The chapter also provides recommendations for how to better integrate public health with the engineering mindset in future public health crises. We close by proposing a centralised data submission and accession platform for population-based surveillance studies, to enable more effective responses to future infectious disease outbreaks.

2

Engineering the SeroTracker dataset and dashboard

2.1 Motivation

Early on in the pandemic, unified sources for COVID-19 cases and deaths rapidly emerged [3, 97]. However, there was no unified source for seroprevalence data, as complex reports of seroprevalence studies were dispersed across peer-reviewed papers, preprints, government reports, and media [81].

To address this need, we built SeroTracker as a unified dashboard and data platform for SARS-CoV-2 seroprevalence studies. SeroTracker combines the rigour and breadth of a living systematic review with the up-to-date and accessible nature of epidemiological dashboards. SeroTracker has catalogued 3,875 seroprevalence studies as of 1 November 2022, is updated each week, and has seen extensive use by governments, researchers, and journalists.

Chapter 2 describes SeroTracker and its development: the living systematic review we used to identify seroprevalence studies, our seroprevalence study database and core data elements, and our publicly accessible dashboard.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Arora RK, Rocco S, Van Wyk J, . . . , Buckeridge D. 2020. SeroTracker: an up-to-date

SARS-CoV-2 seroprevalence dashboard. *The Lancet Infectious Diseases. Rahul's contributions:* conceptualization of manuscript; data analysis and preparation of visualizations; writing original draft; revising draft.

Bobrovitz N*, **Arora RK*,** ..., Papenburg J**, Cheng MP***. 2021. Global seroprevalence of SARS-CoV-2 antibodies: a systematic review and meta-analysis. *PLOS ONE. Rahul's contributions:* sole responsibility for analytical methods development and analysis; contributions to search, screening, and data extraction methods. Additionally, conceptualization of manuscript; defining research question and objectives; writing original draft; revising draft, all in equal portion with Bobrovitz N.

Bergeri I*, Whelan M*, Ware H*, Subissi L*, ..., Bobrovitz N**, **Arora RK**,** Van Kerkhove MD**. 2022. Global epidemiology of SARS-CoV-2 infection: a systematic review and meta-analysis of standardized population-based seroprevalence studies, Jan 2020-Apr 2022. *PLOS Medicine.*

Yan T*, Guo F*, Rocco S, ..., **Arora RK.** 2022. SeroTracker: the development of a novel continuous systematic review dashboard and data platform for SARS-CoV-2 serosurveillance. In preparation. *Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

2.2 Systematic review of SARS-CoV-2 serosurveys

We focus first on SeroTracker's systematic review, which captures data about all SARS-CoV-2 seroprevalence studies published worldwide each week. The objective of this review was to capture a a complete sample and which was listed among the best COVID-19 evidence syntheses by the *COVID-END Network*.

Following best practices for evidence synthesis, we preregistered this systematic review in PROSPERO (the International Prospective Register of Systematic Reviews; ID CRD42020183634). In published papers [9, 81, 101, 102], we also reported this review per PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [82].

2.2.1 Data sources and search strategy

Many seroprevalence studies shared their results through the peer-reviewed literature and preprints. Our systematic review captured these seroprevalence studies through database searches, which we repeated each week.

- For published research articles, we searched MEDLINE, Embase, and Web of Science. On these platforms, we searched for COVID-19 related articles using platform-recommended strategies. We narrowed these searches using search terms related to seroepidemiology and antibody testing identified by infectious disease specialists.

- For preprints, we searched Europe PubMed Central, which indexes several preprint databases including medRxiv and bioRxiv. Search terms paralleled those used for the peer-reviewed literature.

Many serosurveys are not reported in these databases [81]. To capture the broadest possible dataset and limit publication bias, we also included serosurveys reported in the gray literature (e.g., serosurvey reports from national public health agencies or research institutes). We employed four additional search strategies to identify these reports:

- For international and national health agency reports, we searched the websites of several health policy organisations (e.g., World Health Organization, European Centres for Disease Control, National Institutes of Health) for recurring COVID-19 reports. When we identified regularly updated reports (e.g., Public Health England's weekly COVID-19 serosurveillance reports), we set up a recurring search to identify and include subsequent data.

- To capture reports by other institutions, we searched Google News to identify news reports alluding to serosurvey findings. When Google News revealed a potentially eligible report, we carried out targeted Google searches to locate and screen the article in question.

- We consulted with international experts via email to identify additional literature after all sources had been searched.

- We also invited users of the SeroTracker dashboard to submit links to public serosurvey reports that were not yet captured.

2.2.2 Study inclusion criteria

We included all SARS-CoV-2 serosurveys in humans. We defined a serosurvey as the serological testing of a defined population over a specified period to estimate the prevalence of SARS-CoV-2 antibodies [52, 53]. Included studies had to report a sample size, sampling date, geographic location of sampling, and an estimate of seroprevalence. Our full inclusion criteria are listed in Table 2.1.

We excluded studies conducted only in people who had been previously diagnosed with COVID-19 or vaccinated for COVID-19. Such studies generally aim to examine the kinetics of immunity or study protection from reinfection, and do not provide meaningful estimates of seroprevalence. We also excluded dashboards that were not associated with a defined serosurvey, as well as case reports, case-control studies, randomised controlled trials, and reviews.

Two independent reviewers independently screened identified abstracts for inclusion. For articles which passed this abstract screen, two independent reviewers evaluated the full-text article to make a final determination on inclusion. These reviewers were undergraduate students, graduate students, or research assistants at Canadian universities.

As of 1 November 2022, SeroTracker had included 3,875 eligible seroprevalence studies. These were conducted in 141 countries and territories, with over 33,000,000 participants in total. Up-to-date figures are available at SeroTracker.com/Explore [100].

Table 2.1: Inclusion criteria for our systematic review of SARS-CoV-2 serosurveys.

Characteristic	Inclusion Criterion
Population and Condition	
Population	Humans
Condition	SARS-CoV-2 antibodies from previous infection or vaccination
Design and Reporting	
Types of evidence	Serosurveys – studies that collect and test specimens from a sample of a defined population over a specified time period to estimate the prevalence of SARS-CoV-2 antibodies
Study designs	Cross-sectional, repeated cross-sectional, and cohort studies; measurements either at a single time point or at repeated time points
Sampling methods	Any
Outcome measures	An estimate of seroprevalence – the proportion of the sample with detectable SARS-CoV-2 antibodies
Reporting	Must report number of participants enrolled (denominator), study sampling end week, and the locations where the study took place
Provenance	
Sources	Published academic articles; preprints; government and non-governmental organisation reports
Languages	Any. SeroTracker reviewers were proficient in English, French, Portuguese, Spanish, and Cyrillic languages. Articles in all other languages were included if they could be translated in full using machine translation via Google Translate [103].

2.3 SeroTracker data extraction and database

With SARS-CoV-2 serosurveys described in research articles and reports, there has been substantial heterogeneity in the reporting of these studies. Key items are often not described, there is no standard format in which results are reported, and there is no standard taxonomy of variables [81].

To unify SARS-CoV-2 seroprevalence data, we built a standard database to extract these studies in the Airtable relational database tool. To do so, we reviewed early SARS-CoV-2 serosurveys to identify the data elements that they reported, examined seroepidemiologic studies for other diseases to identify the data elements which would be crucial to deliver valuable analysis, and developed a SARS-CoV-2 seroprevalence study

data schema. Data is extracted into our relational database by a first reviewer and verified by an independent second reviewer, with disagreements resolved by consensus. These reviewers were the same reviewers who conducted abstract and full text screening.

The schema we developed is summarised below and visualised in the Entity-Relationship diagram in Figure 2.1. A full schema for each data table is available in Appendix A. We later built on this schema to help develop the World Health Organization's (WHO) ROSES-S global standard reporting guidelines for SARS-CoV-2 serosurveys [77].

1. Sources (Table A.1). Includes sources of seroprevalence data that meet the SeroTracker inclusion criteria. Results from the same seroprevalence study were often published in multiple venues – for example, in a government report, preprint, and published article. In these cases, we used the most credible source as the version of record (full details: Section A.1).

   Characteristics of each source include author name, date, web address. Each source describes one or more SARS-CoV-2 seroprevalence studies.

2. Studies (Table A.3). Studies estimating SARS-CoV-2 seroprevalence in a defined population over a defined time period. By this definition, some sources described multiple methodologically distinct serosurveys, and therefore were linked to multiple studies (full details: Section A.2).

   Characteristics of each study include study design, sampling method, and inclusion criteria. Each study provides one or more seroprevalence estimates.

   (a) Countries (Table A.2). To contextualise seroprevalence data, we gathered information about the demographics and COVID-19 dynamics in the country in which it was conducted.

   (b) Risk of bias (Table 6.1). The quality of SARS-CoV-2 seroprevalence study reporting and design varies widely. We critically appraised each study using the SeroTracker-RoB tool (Section 6.2), which builds on the Joanna Briggs Institute Critical Appraisal Checklist for Prevalence Studies.

3. Seroprevalence estimates (Table A.4). Each study contains one or more seroprevalence estimates in a particular subgroup, at a particular time, and using a particular assay. We extracted the summary seroprevalence estimate from all seroprevalence studies. We additionally extracted sex, age, time, serological assay, and vaccination subgroups from studies of the general population (full details: Section A.3).

    (a) Serological assays (Table A.5). Each estimate was generated using one or more serological assays. Characteristics include manufacturer, assay technology, antibody isotype detected, and antibody target (i.e., whether the assay detects anti-spike or anti-nucleocapsid antibodies).

    (b) Serological assay evaluations (Table A.7). Specificity and sensitivity estimates for each serological assay, catalogued from manufacturer evaluations and standardized independent evaluations.

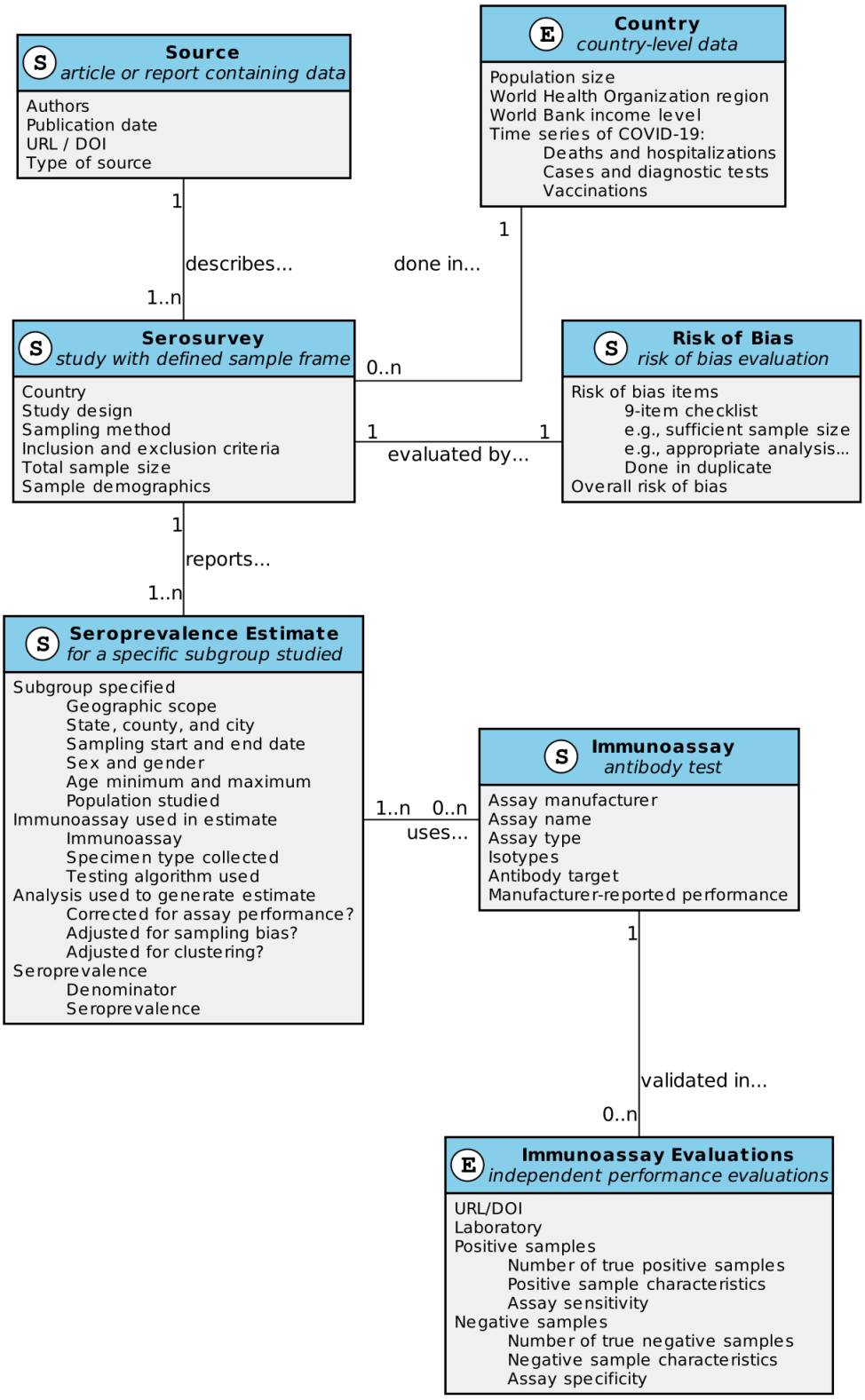


Figure 2.1: SeroTracker entity-relationship diagram. Each box represents a table in the database. The S or E at upper left represent whether the table is primarily sourced from SeroTracker or External data, respectively. Lines represent relationships between tables and their cardinality (e.g., lines that read $1-1..n$ represent one-to-many relationships).

2.4 The SeroTracker dashboard

We developed the SeroTracker.com dashboard with the objective of providing a unified and comprehensive source for SARS-CoV-2 seroprevalence data to public health, researchers, and the lay public. This dashboard draws its data from the systematic review described above. Here, we describe SeroTracker's development philosophy, its evolution, and its current state.

2.4.1 Development process and philosophy

We developed SeroTracker following an iterative and user-centered engineering design process [104]. We used several strategies to reach users, including asking users to self-identify in future engagement when downloading our data, soliciting feedback from attendees at presentations we delivered, direct outreach to individuals who publicly stated they were using our data, and working directly with partner and client organisations. We engaged these users through regular user interviews and user surveys, using these to identify usability challenges and new features of interest.

Our user-centered design approach ultimately led to numerous cycles of iteration which improved all aspects of the tool. Key improvements include updates to the map interface, refined ability to analyse and visualise seroprevalence data, and adding programmatic data accessibility to serve emerging user needs. We will discuss these evolutions and how they led to SeroTracker's current state.

2.4.2 Explore page and mapping

SeroTracker's Explore page is designed to help users explore findings from seroprevalence studies across the world. In the first version of SeroTracker's Explore page, each country was colored based on the seroprevalence of studies conducted in that country (i.e., based on the point estimate from meta-analysing all available seroprevalence studies from that country; Figure 2.2). However, this version of SeroTracker was potentially misleading. It gave equal weight to studies with different populations, designs, and time points. For example, if a country had a single health care worker study in March 2020 which reported seroprevalence 10%, that country would have the same color as another country,

also with a single study that reported seroprevalence 10%, but where that study was done in May 2020 in the general population. Moreover, shading the entire country provided the impression that the estimated seroprevalence was representative of the country as a whole, while some seroprevalence studies were done in single cities or in nonrepresentative populations.

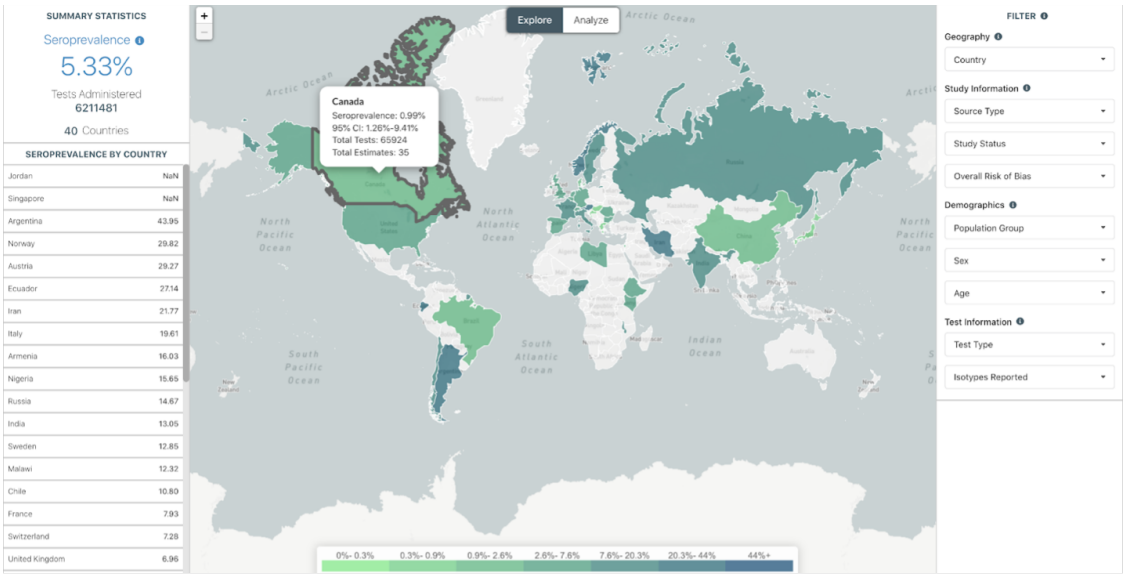


Figure 2.2: May 2020 Explore page heatmap design. Color indicates point estimate from meta-analysis of all available seropervalence studies from that country.

After receiving this user feedback, we rapidly developed a new design. Rather than summarising national seroprevalence across studies, we summarised national studies in the modal and incorporated a binary map to communicate the geographic availability of data (Figure 2.3). We also filtered the displayed studies by default, including only those studies that we considered sufficiently representative – national or regional-scale studies of the general population or blood donors, and excluding those studies reported in media articles. While this page addressed initial user criticism, further user interactions surfaced a need to more easily understand the provenance, location, and methodology of each individual study.

Today, SeroTracker's Explore page lets users rapidly identify and explore findings from seroprevalence studies around the world (Figure 2.4). The present Explore page represents each study as a pin on the map. The filters on this page enable users to identify studies from a particular time period, population, or age group. Hovering over

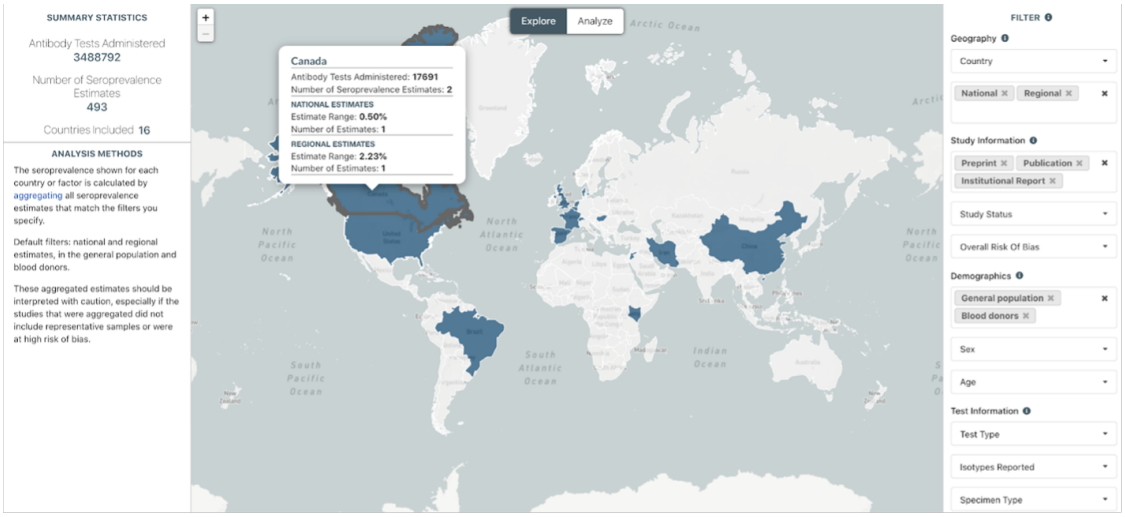


Figure 2.3: August 2020 Explore page binary design. Shaded countries have seroprevalence data available.

a country reveals details about the studies done in that country, just as clicking on a study reveals key findings from that study (Figure 2.5). This functionality enables public health officials to understand the seroprevalence studies done in their country and use these to inform public policy.

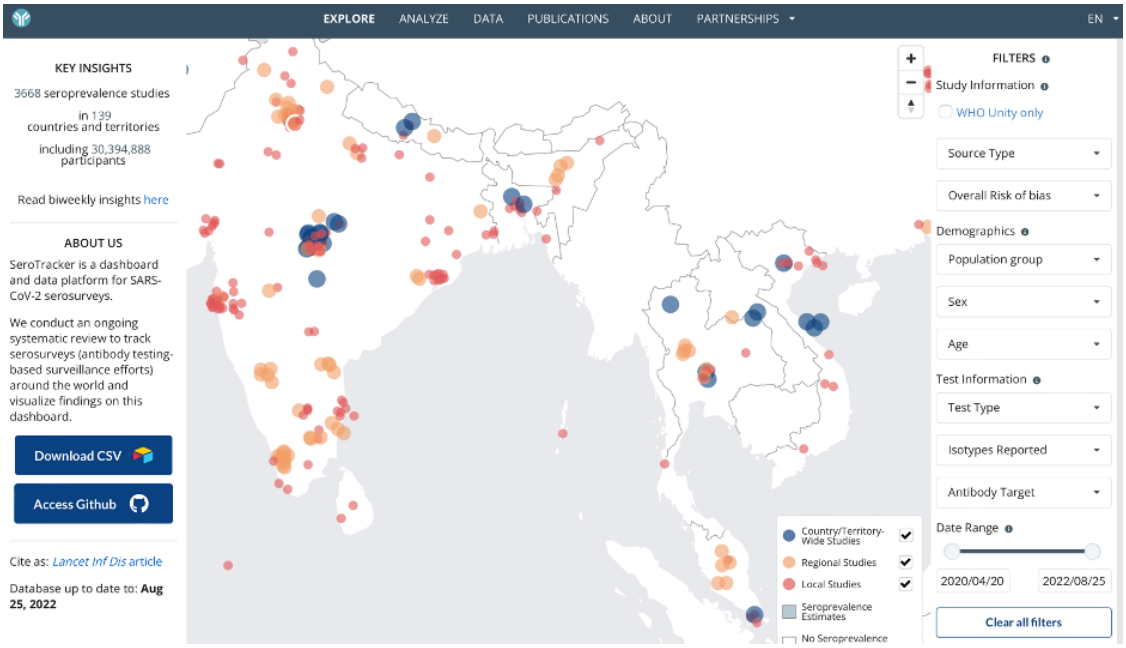
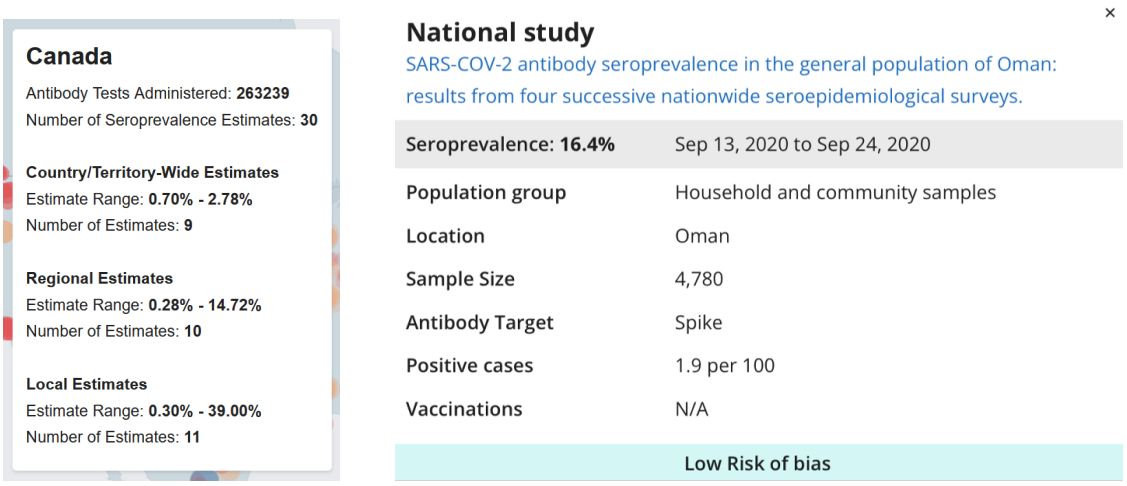


Figure 2.4: Current SeroTracker Explore page. Colored dots on the map indicate seroprevalence studies, with blue, orange, and red dots, indicating studies with national, regional, and local scope, respectively. Text at upper left indicates number of studies, countries, and participants included. Filters at right enable users to select only those seroprevalence studies which meet certain criteria. Site available at SeroTracker.com/explore.



(a) Country detail modal (b) Study detail modal

Figure 2.5: Current SeroTracker Explore page modals. Modal (a) appears when hovering over a country, and indicates the number of studies done in that country as well as the range in seroprevalence estimates from all national, regional, or local studies. Modal (b) appears when selecting a study, and highlights key information about that study including its web address, dates, sample size, summary seroprevalence estimate, and risk of bias. Site available at SeroTracker.com/explore.

The Explore page is custom-built. We mirror our Airtable database to a performant PostgreSQL database to serve our dashboard (Figure 2.6), doing so via a tailored Extract, Transform, Load pipeline. We built a Flask application to serve the data from the database to our dashboard. In turn, our React application hosts the SeroTracker website on the client and includes mapping functionality built with Mapbox. Figure 2.6 shows this data architecture.

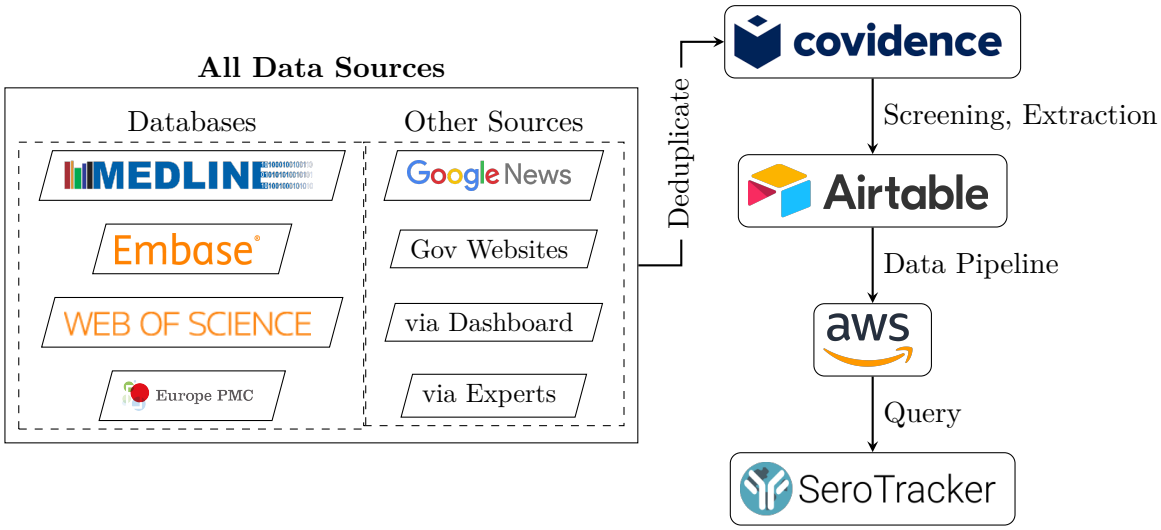


Figure 2.6: SeroTracker data architecture. We identify studies from data sources, screen them for inclusion, and extract them into a structured relational database, which then serves the SeroTracker dashboard.

2.4.3 Analyse page and charts

SeroTracker's Analyse page, which synthesises data from SARS-CoV-2 seroprevalence studies, underwent an analogous evolution. Initially, SeroTracker's Analyse page showed a single powerful graph, allowing users to filter studies and view the seroprevalence estimates from these studies aggregated by variable of interest (Figure 2.7).

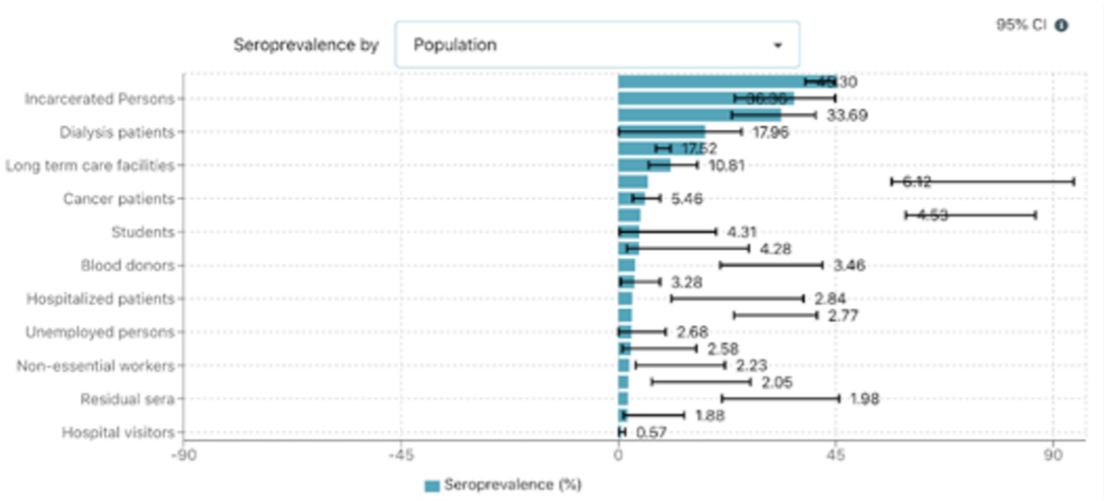


Figure 2.7: May 2022 Analyse page. Shows seroprevalence estimates aggregated by a variable of interest – in this case, the population in which studies were done.

However, in permitting users to filter and aggregate studies at their discretion, users

could also produce misleading visualisations. Moreover, user interviews revealed that the average website user was not interested in producing their own visualisations through this tool. Instead, they wanted a specific, common set of questions answered.

Today, the Analyse page provides up-to-date and interactive versions of analyses that answer key policy questions. Many of these visualisations were initially produced as one-off analyses for partners or for research papers, and then adapted into up-to-date visualisations to serve our users. For example, one such visualisation compares seroprevalence, confirmed cases, and vaccine coverage across geographies (Figure 2.8. This analysis demonstrates how seroprevalence has changed over time, the role of infection and vaccination in driving these factors, and the considerable heterogeneity in seroprevalence trajectories and drivers between countries. Appendix B describes the other visualisations on SeroTracker's Analyse page.

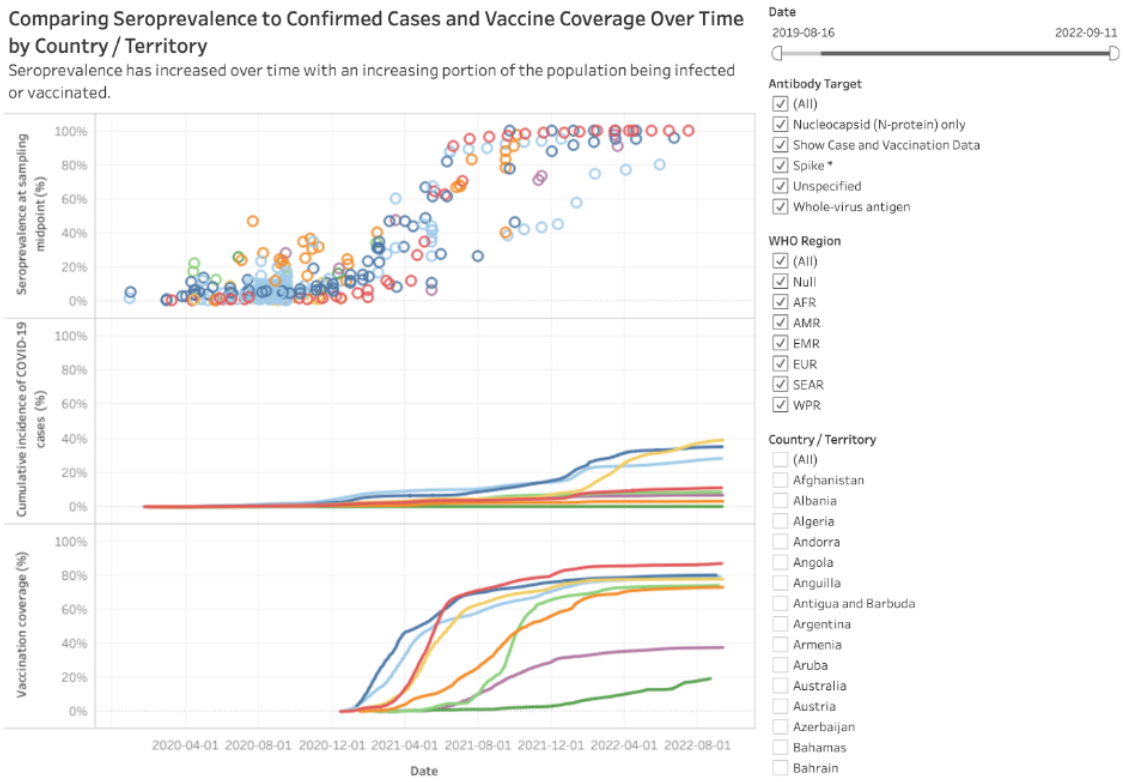


Figure 2.8: Current SeroTracker Analyse page. seroprevalence, confirmed cases, and vaccine coverage across geographies, demonstrating how increases in seroprevalence correlate with the extent of infection and vaccination, and how each of these factors differ dramatically between countries and territories. One of several visualisations on SeroTracker.com/analyze.

2.4.4 Data page, process, and availability

Our systematic review process and data page evolved in tandem to enable us to produce robust analyses and provide this data to our stakeholders. These evolutions were often driven by the pandemic itself. For example, in early 2021, COVID-19 vaccines became broadly available. As a result, we updated our data schema to include the proportion of the surveyed population vaccinated, whether the study assayed for antibodies against the SARS-CoV-2 nucleocapsid protein (anti-nucleocapsid) or spike protein (anti-spike).

Other evolutions were driven by practical considerations. When the number of studies was small, we collected highly granular data. As the rate of study publication increased, extracting data at this granularity began to impede timely data updates. We identified that up-to-date data and data on the general population was a priority to our partners, leading us to standardise and simplify the data we collected. This included only extracting data on the essential subgroups discussed in Section A.3 and creating standardised vocabularies for each field to ensure consistency.

SeroTracker's data page, shown in Figure 2.9, enables our users to download our data for their own use. Initially, we made a limited set of data fields accessible on our website. However, we rapidly realised that an important subset of our users was highly technical, and capable of exploiting our full data to conduct sophisticated analysis. This led us to share our full dataset, first allowing data downloads, and later allowing analysts and modellers to programmatically access data.

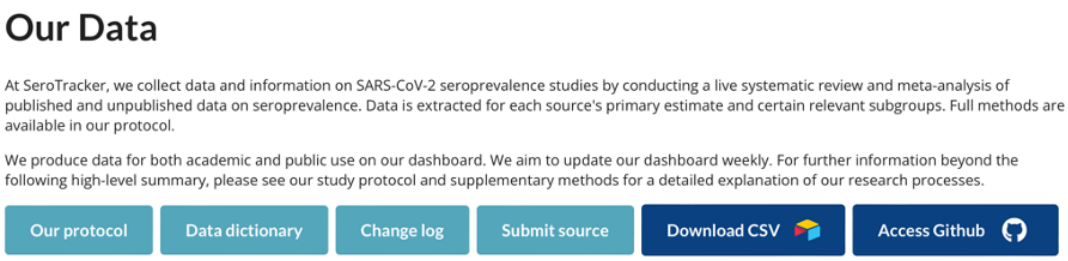


Figure 2.9: Current SeroTracker Data page. Details SeroTracker's methods and provides data download functionality, enabling users to employ SeroTracker data in their own work. Site available at SeroTracker.com/data.

2.4.5 SeroTracker's users

SeroTracker's users segment into three main categories: researchers, global policymakers, and journalists. Researchers in academia and industry use seroprevalence data as a window into the true incidence of COVID-19 infections and into humoral immunity against SARS-CoV-2. Key users here include Canada's COVID-19 Immunity Task Force, the WHO Modelling Group, and the Institute of Health Metrics and Evaluation (modelling true SARS-CoV-2 infections [105, 106]); Imperial College London (modelling infection fatality rate [107]); the Public Health Agency of Canada, Robert Koch Institute, and United States Centres for Disease Control (global and in-country seroprevalence data monitoring [108]); and the COVID-19 Vaccines Global Access (COVAX) initiative and Merck (vaccine trial site selection [109]).

Global consortia are particularly interested in seroprevalence data to understand the extent of infection and immunity worldwide, shaping their policy recommendations. Users in this category include the WHO's Unity and Solidarity II global serosurveillance initiatives (where SeroTracker forms the backbone of the data pipeline [110]), the Bill and Melinda Gates Foundation (informing investments in serosurveys in low- and middle-income countries), and the Independent Panel for Pandemic Preparedness and Response (understanding the early spread of COVID-19 through serological evidence from late 2019 and early 2020).

Journalists and consultants have used SeroTracker's seroprevalence data to help their readers and clients appreciate the underascertainment of SARS-CoV-2 infections and markers of immunity. Users in these industries include The Economist (excess mortality worldwide [111]), the New York Times (seroprevalence and underreporting in India [112]), as well as McKinsey & Company and the Boston Consulting Group (charting a course out of the COVID-19 pandemic [113]).

2.5 Discussion

We describe the development of SeroTracker, an integrated continuous systematic review and dashboard for SARS-CoV-2 seroprevalence studies. SeroTracker unified SARS-CoV-2 seroprevalence data during the COVID-19 pandemic, facilitating its use by key stake-

holders. SeroTracker demonstrates the power of integrating the rigour and breadth of living systematic reviews with the accessibility and secondary use potential of dashboards – a novel approach that can accelerate our scientific and policy responses to public health crises.

SeroTracker may be one of the first of many efforts to integrate effective knowledge translation with rigorous evidence synthesis. To our knowledge, the only other dashboard with a similar approach also emerged during the COVID-19 pandemic: the COVID-NMA, a collaboration led by the WHO and Cochrane review researchers that has created a living map and network meta-analysis of registered randomised trials for COVID-19 [114, 115].

In the process of creating and refining SeroTracker, three key lessons were crucial to our efforts, which we expect to be relevant to stakeholders seeking to create similar platforms: embracing user-centered design in our internal decisions, the importance of a focused scope, and developing the right tool for the right application.

2.5.1 Embracing user-centered design in internal decisions

The engineering literature extensively addresses the importance of the user perspective in product design. In our experience, user-centered design is essential beyond product design – it was a crucial element of how we designed our foundational systems. For example, our data schema was at the foundation of our website and of every analysis we could produce from our work. Adapting the design of this schema to include only elements required for user-defined analyses was crucial in enabling us to produce many of the analyses in Chapters 3, 4, 5, and 6. In turn, excising irrelevant elements from this schema was essential to making important analyses possible with limited resources. Making even these internal decisions with the user in mind was crucial in our ability to produce value for users.

2.5.2 Ensuring a focused scope

Platforms like SeroTracker take significant time and investment to create, particularly with the heterogeneity of data sources that must be integrated and the complexity of interpreting data. Considering this, it is crucial to clearly define such a platform's objective. Otherwise, a project's efforts can easily diverge from that key goal.

We experienced this divergence ourselves during our first year building SeroTracker. We were focused on publishing our own analyses and inundated with stakeholder invitations for presentations. Given limited resources, this led to us falling behind on updating our data – contrary to our mission to provide up-to-date SARS-CoV-2 seroprevalence data to inform public health decision-making. Upon realising this lag, we refocused on our mission, set weekly key performance indicators for extraction, and streamlined bottlenecks in data collection.

Prioritising user needs means that we have published less often than we otherwise might have. However, in our view, this trade-off has been worthwhile. For example, we learned from users that we should prioritise open and up-to-date data above all else. Providing this data has enabled many users to conduct their own novel analyses, magnifying the value of SeroTracker beyond what we could have achieved alone.

Much of the impact of platforms like SeroTracker and the COVID-NMA comes from their ability to act as an up-to-date foundational resource in a network of impactful research. We emphasise that groups conducting their own analyses should not come at the expense of keeping our core data up-to-date.

2.5.3 Choosing the right tool for the right application

Creating these platforms is resource intensive, and developing SeroTracker required considerable funding and a large group of reviewers. Keeping up with rapidly shifting methods and data interpretation in the course of a health emergency requires constant adaptation. Moreover, despite spending much time on SeroTracker's design, we have not yet been able to refine it to the point where nonexpert practitioners and laypeople can easily interpret its findings.

These limitations were worthwhile in the context of SeroTracker's research question: that is, what is the population prevalence of antibodies against SARS-CoV-2 in humans? This question is rapidly evolving, requiring timely intervention, and of substantial societal impact. Many evidence synthesis efforts do not fit into this category, and may be more efficiently served by other approaches. It is crucial to select the right tool for the right situation.

3

Meta-analysing SARS-CoV-2 seroprevalence data

3.1 Motivation

In Chapter 2, we developed a process to systematically identify all SARS-CoV-2 seroprevalence data globally. Using this data, we can now examine what these seroprevalence studies indicate about the global scale of SARS-CoV-2 infection and immunity.

During the COVID-19 pandemic, we conducted three global analyses of SARS-CoV-2 seroprevalence. This Chapter describes each of these analyses in turn. The first of these was a systematic review of seroprevalence studies including seroprevalence studies published through June 2020. The second was a meta-analysis including studies published through December 2020. These studies led to our definitive analysis of global SARS-CoV-2 seroprevalence studies through April 2022, which estimated the true extent of population infection and seropositivity during the first two years of the pandemic.

The citations for the meta-analyses described in this chapter follow. Co-first authors are denoted with * and co-senior authors with **.

Bobrovitz N*, **Arora RK*,**, Yan T, ..., Evans TG. 2020. Lessons from a rapid systematic review of early SARS-CoV-2 serosurveys. *medRxiv. Rahul's contributions:* sole responsibility for analytical methods development and analysis; contributions to search, screening, and data extraction methods. Additionally, conceptualization of manuscript; defining research question and objectives; writing original draft; revising draft, all in equal portion with Bobrovitz N.

Bobrovitz N*, **Arora RK*,** ..., Papenburg J**, Cheng MP***. 2021. Global seroprevalence of SARS-CoV-2 antibodies: a systematic review and meta-analysis. *PLOS ONE. Rahul's contributions:* sole responsibility for analytical methods development and analysis; contributions to search, screening, and data extraction methods. Additionally, conceptualization of manuscript; defining research question and objectives; writing original draft; revising draft, all in equal portion with Bobrovitz N.

Bergeri I*, Whelan M*, Ware H*, Subissi L*, ..., Bobrovitz N**, **Arora RK**,** Van Kerkhove MD**. 2022. Global epidemiology of SARS-CoV-2 infection: a systematic review and meta-analysis of standardized population-based seroprevalence studies, Jan 2020-Apr 2022. *PLOS Medicine.*

Lewis HC*, Ware H*, Whelan M*, Subissi L*, ..., Okeibunor J**, **Arora RK**,** Bergeri I**. 2022. SARS-CoV-2 infection in Africa: A systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021. *BMJ Global Health. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

3.2 Jan 2020 - June 2020 systematic review

In the first months of the pandemic, data from many seroprevalence studies was released. These could provide evidence to inform several questions that had emerged at that time. Was there substantial asymptomatic spread? How did the true extent of infection compare to what we had been able to diagnose? What did this imply for how close we were to herd immunity, which might stop the virus in its tracks? How deadly was COVID-19 – what was the infection fatality rate? How should the answers to these questions shape health policy?

We brought together the seroprevalence literature to help answer these questions. In early 2020, we were the first group to conduct a systematic review of SARS-CoV-2 seroprevalence studies [81].

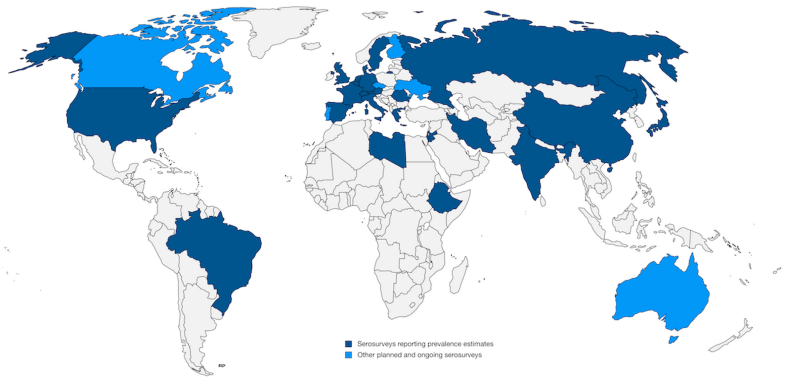


Figure 3.1: Map of planned, ongoing, and completed seroprevalence studies as of 26 June 2020.

As of 26 June 2020, 208 seroprevalence studies were being conducted in 40 countries (Figure 3.1). Serological testing results had been reported for 1,407,838 individuals, with an additional 3,968,706 to be enrolled in upcoming studies.

The best estimates indicated that the majority of the population has not been infected. Studies in seventeen countries had reported robust seroprevalence estimates that are at a low or moderate risk of bias RoB. In these studies, the median seroprevalence in the general population was 3.1%, with interquartile range (IQR) 1.6% to 6.0% [116]. This implied that the vast majority of the general population had not yet been infected, suggesting that the pandemic would likely be lengthy.

We also conducted our first comparison of confirmed case data and seroprevalence data, to examine the extent to which confirmed cases underestimated the true extent of infection (i.e., case ascertainment). On 11 May, Spain had confirmed 230,000 cases of COVID-19 using nucleic acid testing [3]. Meanwhile, a serosurvey of over 60,000 people in Spain ending the same day estimated 5% seroprevalence [117], which corresponds to an estimated 2.34 million SARS-CoV-2 infections nationally – ten times larger than the cumulative number of confirmed cases. This suggestion that many cases had been missed was a clear call to expand diagnostic testing for active disease, both for better surveillance and to prevent further chains of transmission.

This study was instrumental in shaping the future of seroepidemiological investigations for SARS-CoV-2 – for example, it directly influenced the development of the WHO UNITY standardised protocol [53]. Moreover, these analyses also underpinned

our later work, prompting further investigation of seroprevalence study results, their synthesis, and case ascertainment.

3.3 Jan 2020 - Dec 2020 meta-analysis

By the end of 2020, many more seroprevalence studies had been conducted. However, these studies continued to be conducted as research studies, with results reported across diverse sources [81]. At this point, synthesising serosurvey findings across these sources was crucial to measure the prevalence of SARS-CoV-2 antibodies worldwide, understand differences in past infection between groups, and inform the optimal distribution of COVID-19 vaccines [62]. We conducted a systematic review and meta-analysis of all SARS-CoV-2 seroprevalence studies published between 1 January 2020 and 31 December 2020 [9].

3.3.1 Characteristics of included studies

At the time, we identified 968 unique seroprevalence studies in 605 articles. Seventy-four countries across all *Global Burden of Disease* regions were represented among identified serosurveys 3.2). At this time, there was a relative paucity of seroprevalence studies (n = 221, 23%) conducted in low- and middle-income countries (Figure 3.2).

3.3.1.1 Seroprevalence from included studies

Among included studies, 590 (61%) estimated seroprevalence in the general population. The median seroprevalence in these studies was 4.6% (IQR 2.2 − 8.5%; Table 3.1). These studies included household and community samples ($n = 125$; median 6.0%, IQR 2.8 − 15.1%), residual sera ($n = 248$; median 4.0%, IQR 2.4 − 6.8%), and blood donors ($n = 54$; median 4.7%, IQR 1.4 − 6.8%). Across *Global Burden of Disease* regions, the median seroprevalence ranged from 1.0% (IQR 0.4 − 2.9%) in Southeast Asia, East Asia, and Oceania to 17.6% (IQR 8.8 − 26.8%) in South Asia (3.1). At the time, this low seroprevalence emphasised the importance of continued public health measures to reduce transmission until we achieved global vaccine uptake.

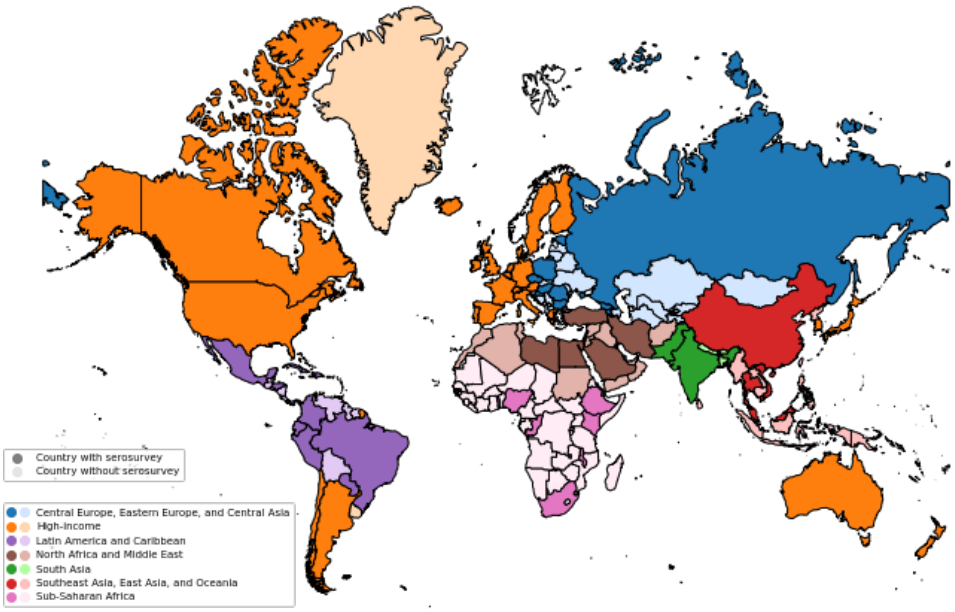


Figure 3.2: Map of serosurvey distribution by *Global Burden of Disease* region as of 31 December 2020. Each color represents a different region. Countries with dark shading had reported findings from at least one seroprevalence study at the time of our Jan 2020 - Dec 2020 meta-analysis. High-income countries are all included in the high-income region, identified in orange. Low- and middle-income countries are distributed across the remaining regions.

Table 3.1: Summary of general population seroprevalence studies as of 31 December 2020. Includes studies of the general population. Stratified by region, geographic scope, and RoB. Abbreviations: No.= number; L = low; M = moderate; H = high; U = unclear; N/A = not applicable

Characteristic	No. studies	No. countries	Median sample size (IQR)	Median seroprevalence (IQR)	Risk of bias
General population studies	590	57	987 (786-2,639)	4.6% (2.2-8.5%)	L: 4%, M: 62%, H: 27%, U: 6%
***Global Burden of Disease* region**					
Central & Eastern Europe, Central Asia	14	6	2,681 (992-3,037)	7.8% (2.3-20.5%)	L: 7%, M: 43%, H: 43%, U: 7%
High-income	453	28	985 (786-1,709)	4.4% (2.2-7.2%)	L: 3%, M: 65%, H: 27%, U: 5%
Latin America, Caribbean	57	10	900 (832-1,968)	6.8% (2.6-19.5%)	L: 5%, M: 70%, H: 18%, U: 7%
North Africa, Middle East	5	4	1212 (600-3,530)	12.9% (0.8-19.3%)	L: 20%, M: 40%, H: 20%, U: 20%
South Asia	35	2	3,000 (502-15,625)	17.6% (8.8-26.8%)	L: 17%, M: 43%, H: 20%, U: 20%
Southeast Asia, East Asia, Oceania	20	2	2,192 (434-18,024)	1.0% (0.4-2.9%)	L: 0%, M: 35%, H: 60%, U: 5%
Sub-Saharan Africa	6	5	528 (214-2,282)	14.6% (8.0-24.0%)	L: 17%, M: 33%, H: 50%, U: 0%
Scope					
National	83	32	4,297 (1,200-24,926)	4.5% (1.9-6.1%)	L: 10%, M: 64%, H: 19%, U: 7%
Regional	312	19	980 (802-1,106)	4.3% (2.3-7.6%)	L: 4%, M: 73%, H: 21%, U: 3%
Local	167	31	1,100 (752-2,547)	5.5% (2.1-14.8%)	L: 4%, M: 46%, H: 38%, U: 13%
Sub-local	28	14	500 (357-928)	7.2% (1.4-15.1%)	L: 0%, M: 36%, H: 61%, U: 4%
Risk of bias					
Low	25	13	4,151 (2,203-9,922)	8.2% (2.9-13.6%)	N/A
Moderate	367	42	985 (900-1,545)	4.7% (2.6-7.9%)	N/A
High	161	30	731 (313-2,415)	3.9% (1.2-9.4%)	N/A
Unclear	37	15	1,709 (774-8,006)	3.3% (1.5-11.0%)	N/A

3.3.2 Key limitations

Most studies in this analysis (46%) were at moderate RoB, with 52% at high or unclear RoB. Common reasons for higher RoB included missing test information, serological assays with low sensitivity and specificity, no correction for assay performance, lack of survey weighting, use of non-probability sampling, and use of nonrepresentative sample frames. Our results suggested that improved study quality (i.e., low RoB) may be associated with lower seroprevalence estimates (Table 3.1). Higher-risk studies may have been subject to selection bias that favoured groups at lower risk of infection.

Our summary estimates were also driven by high-income countries. The vast majority of data (77%) was from high-income countries, and studies from these countries tended to report lower seroprevalence estimates. The data we did have suggested importantly different dynamics in low- and middle-income countries. For example, we estimated the ratio between infections and confirmed cases in various *Global Burden of Disease* regions. However, while these estimates appeared much higher in low- and middle-income *Global Burden of Disease* regions (ranging from 24.1 : 1 in Central Europe, Eastern Europe, and Central Asia to 626.6 : 1 in Sub-Saharan Africa), they were based on very few data points (a single study apiece, for the regions highlighted here).

We had hoped to use this dataset to estimate seroprevalence globally and examine its changes in time. However, the poor quality of the data and its lack of representativeness made it inadvisable in this analysis. At the time, many seroprevalence studies were beginning to be conducted in low- and middle- income countries, and it seemed likely that these would soon fill the gap and enable further investigation.

3.4 Jan 2020 - Apr 2022 meta-analysis

During 2021, many regions experienced third and fourth waves of SARS-CoV-2 infection [110]. Concurrently, some countries vaccinated most residents, while others remain unable to achieve high vaccine coverage due to challenges with supply and uptake [93].

By this point, evidence was also emerging that anti-SARS-CoV-2 antibodies were highly predictive of immune protection against the original, Alpha, Beta, and Delta SARS-CoV-2 variants [31, 118]. As a result of this, seroprevalence studies could be

considered indicative of population levels of humoral immunity, and were therefore increasingly important to inform scenario modeling, public health planning, and national policies in response to the pandemic. Considering this epidemiological context, there was a clear need to synthesise seroprevalence studies to understand the shifting global dynamics and true extent of SARS-CoV-2 infection and humoral immunity.

At this time, several previous global systematic reviews of seroprevalence had been conducted, both by our group [9] and others [119–121]. However, the published literature at this time included only studies that sampled participants in 2020. These studies were also not able to examine changes in seroprevalence over time. Each of these previous systematic reviews emphasised that estimates of seroprevalence were difficult to compare systematically due to inter-study variation in sampling methods, testing methods, and reporting [9, 119, 120], underlining the importance of improved standardisation and study quality to enable more robust analysis.

The World Health Organization's UNITY Initiative aims to help produce harmonised and globally representative seroprevalence study results [122]. The WHO UNITY population-based, age-stratified seroepidemiological investigation protocol provides a standard study design and laboratory approach to general population seroprevalence studies. WHO UNITY and its partners have supported the implementation of UNITY by providing financial and technical resources, including a well-performing serological assay. As of September 2021, the UNITY protocol had been implemented in 74 countries globally, including 51 low- and middle-income countries [122].

In 2021, results from UNITY studies began to become broadly available, with many well-conducted seroprevalence studies providing robust estimates of seroprevalence [123–125]. Many of these studies were in low- and middle-income countries, where seroprevalence data is especially important due to limited contact tracing and diagnostic testing capacity [122]. The use of the UNITY protocol increased the pool of robust, comparable seroprevalence data, enabling a more comprehensive understanding of SARS-CoV-2 infection and humoral immunity globally [53].

SeroTracker partnered with WHO to conduct a systematic review and meta-analysis of all seroprevalence studies worldwide aligned with the UNITY protocol, including aligned studies whether or not they received direct support from WHO. We aimed to:

1. Estimate global and regional SARS-CoV-2 seroprevalence over time
2. Assess the extent of undetected infection, by estimating global and regional case ascertainment
3. Elucidate demographic differences in seroprevalence through meta-analysis
4. Identify study design and country-level factors associated with seroprevalence

3.4.1 Seroprevalence data

We identified data through the SeroTracker systematic review as described in Section 2.2, including all articles published between 1 January 2020 and 5 May 2022. For this analysis, we additionally captured early results from WHO UNITY-supported seroprevalence studies that were directly reported to WHO and SeroTracker. Study authors submitted their results into a standard Excel template designed to match the same data extracted during routine extraction of published studies. Templates were verified by two independent reviewers and we conducted follow-up to complete information with study investigators where needed.

We included studies aligned with the World Health Organization standardised UNITY protocol [53], aiming to minimise possible bias introduced by inter-study heterogeneity, poor assay performance, and nonrepresentative sampling methods. To be considered a UNITY-aligned seroprevalence study, a study must have estimated seroprevalence in the general population. Restricting inclusion to direct population samples such as household surveys, would have led to very little data in some regions and times, as these studies are expensive and difficult to conduct. Thus, household and community samples were included, as well as studies where a robust sampling frame was described that approximates to a wider population, such as individuals attending routine medical services (blood donors, pregnant mothers, primary care attendees, routine laboratory investigations) or people residing in slum dwellings. For studies conducted in the 63

vulnerable countries defined by the *Global Humanitarian Response Plan for COVID-19* [126], we included seroprevalence studies in specific patient populations, if the condition in question was widespread in the general population. We excluded studies sampling specific closed populations (e.g., prisons, care homes, or other single-institution populations).

Considering the robust sampling frame requirements, both random and non-random (i.e., convenience, sequential, quota) sampling methods were included. For convenience samples to be included, they were required to have a clear and defined sampling frame – i.e., studies recruiting volunteers were not included. As a result, the vast majority of included convenience samples were in blood donors or residual sera samples. We excluded studies that excluded people previously diagnosed with or vaccinated against COVID-19 after initial sampling.

To be included studies also had to use serological assays with at least 90% sensitivity and 97% specificity, as reported by the manufacturer or study authors through an independent evaluation of the test used. We made an exception to this criterion for studies done in *Global Humanitarian Response Plan* countries, where lower-performing assays were sometimes the only accessible option due to ongoing humanitarian emergencies.

3.4.2 Contextualising data

For each study, we gathered case, vaccination, and public health and social measure data to put seroprevalence into context. We drew confirmed case data from the WHO COVID-19 dashboard, matched to nine days before the midpoint of each study to account for time from infection to seroconversion [127, 128]. We defined the transmission phase by identifying periods of peak transmission in the country of each study – that is, local maxima in the smoothed seven-day rolling average of daily confirmed cases. We then classified the study sampling dates as before the peak of the first SARS-CoV-2 wave, after the peak of the first SARS-CoV-2 wave, or after the peak of the second SARS-CoV-2 wave in the country of the study.

We identified vaccination data from *Our World in Data*, timed fourteen days before the study's midpoint [129]. Variant of concern figures were based on weekly counts of SARS-CoV-2 sequences submitted to the *GISAID* repository [130].

Finally, we identified the public health and social measures (PHSM) stringency index value, sourced from a global London School of Hygiene and Tropical Medicine-WHO collaboration [131]. This index ranges from 1-10, and aggregates six indicators, including masks, international travel, gatherings, schools, workplaces and businesses, and movement. For each study, we calculated the average PHSM index in the corresponding country between the date of the 100th reported case in that country and 14 days before the study midpoint.

3.4.3 Included studies

We identified 173,430 titles and abstracts in our search spanning from 1 January 2020 to 2 May 2022 (Figure 3.3). Of these, 2,581 articles were included in full text screening. These included 513 seroprevalence data sources containing studies aligned with the UNITY protocol: 480 published (94%) and 33 aggregated results from collaborators (6%). The 513 sources contained a total of 965 unique seroprevalence studies.

Over 1,500 articles included in full text screening were subsequently excluded because they did not contain studies compatible with the UNITY protocol (Figure 3.3). The main reasons for articles' exclusion at this stage was having an incorrect sample frame for the scope of this analysis. Considering our focus on seroprevalence in the general population, we excluded 1,073 articles of exclusively health care workers, close contacts of confirmed cases, or other specific closed populations. We also excluded 374 articles not meeting our predefined assay quality performance threshold.

A total of 52% (100/194) of WHO Member States and four WHO areas and territories, across all 6 WHO regions, were represented among the seroprevalence studies included in the descriptive analysis (Figure 3.4). 23 of 47 Member States were represented in the WHO African Region; 11 of 21 Member States and one territory in the Eastern Mediterranean Region; 13 of 35 Member States and one territory in the Region of the Americas; 39 of 53 Member States and two territories in the European Region; 6 of 11 Member States in the South-East Asian Region; and 8 of 27 Member States in the Western Pacific Region.

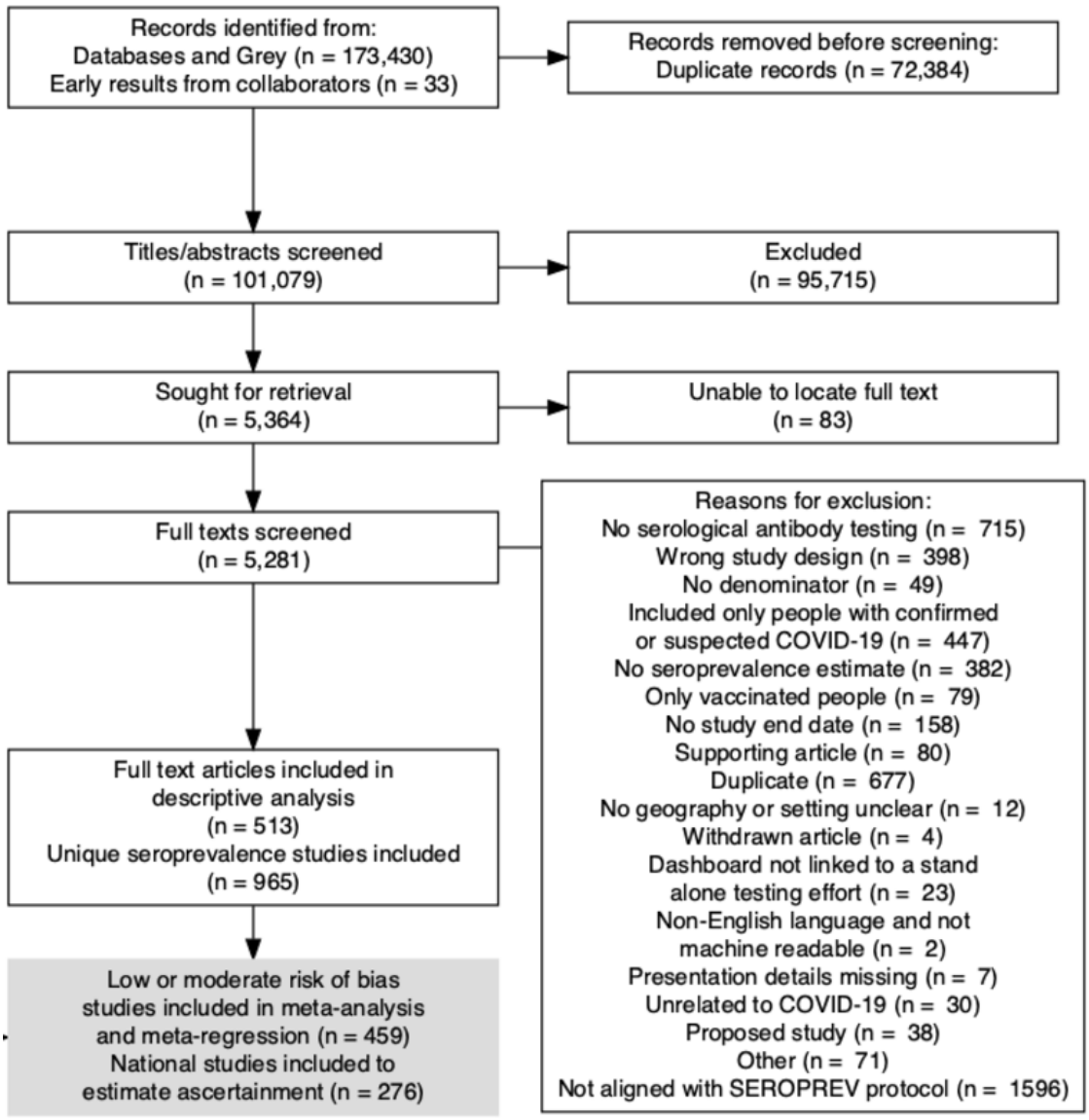


Figure 3.3: Flow diagram of included studies.

Table 3.2 describes the characteristics of included studies. Data from 61 of 134 low- and middle-income countries and from 36 of 63 vulnerable countries in the *Global Humanitarian Response Plan for COVID-19* were included. A large proportion of studies included in the descriptive analysis were conducted in low- and middle-income countries (397/965, 41%) and in *Global Humanitarian Response Plan countries* (206/965, 21%; Table 3.2). Of studies included in the meta-analysis and meta-regression, these proportions were 30% (137/459) and 14% (66/459), respectively. 20 of the 66 (30%) meta-analysed studies in *Global Humanitarian Response Plan* countries had test sensitivity below 90%

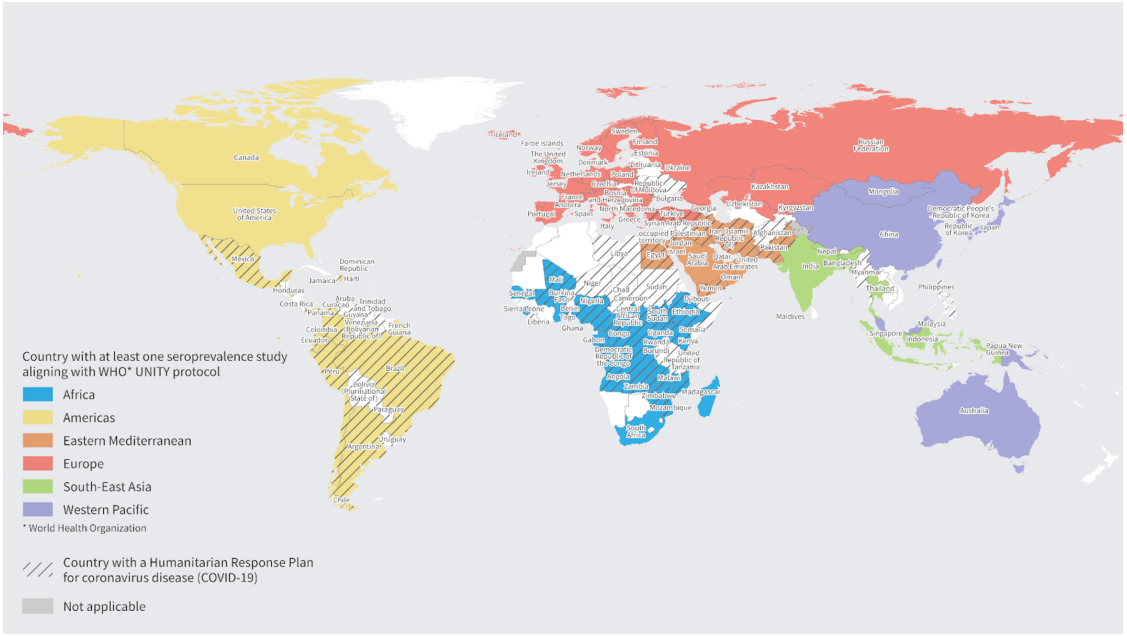


Figure 3.4: WHO member states with seroprevalence data identified. Colors represent WHO regions. Colored countries have at least one UNITY-aligned seroprevalence study included in the Jan 2020 - Apr 2022 analysis. Diagonal lines indicate *Global Humanitarian Response Plan* countries. The small text over each country states the name of that country.

or specificity below 97%, and were included due to the test performance exception.

Among the 965 studies included in the descriptive analysis, 42% (402/965) reported results at a local level, 36% (345/965) at a national level, and 23% (218/965) at a subnational level (Table 3.2). The most common sampling frame was household and community samples (52%, 500/965) and the most common sampling method was probability sampling (53%, 515/965). Within household-based samples, only 86/500 studies (17.2%) used convenience sampling (Table 3.2).

The majority of studies (734/965, 76%) had no vaccination at the sampling midpoint date in the country of the study (Table 3.2). Very few studies (14/965, 1.5%), conducted in only 4 countries (Canada, Japan, United Kingdom, and the United States of America), sampled participants during 2022.

Table 3.2: Characteristics of included studies, Jan 2020 - May 2022.

Dataset	**Dataset 0**	**Sub-dataset 1**	**Sub-dataset 2**

Contains	**All studies**	**Low or moderate RoB; national or subnational scope**	**Low or moderate RoB; national only scope**
Used for	**Descriptive analysis**	**Seroprevalence over time and associated factors**	**Case ascertainment**
Number of studies	**N = 9,651**	**N = 4,591**	**N = 2,761**
Income level			
Low-income country	101 (10%)	26 (5.7%)	24 (8.7%)
Lower-middle-income country	130 (13%)	52 (11%)	25 (9.1%)
Upper-middle-income country	166 (17%)	59 (13%)	23 (8.3%)
High-income country	568 (59%)	322 (70%)	204 (74%)
Vulnerable countries (*Global Humanitarian Response Plan*)			
Vulnerable country	206 (21%)	66 (14%)	29 (11%)
WHO Region			
African Region	171 (18%)	47 (10%)	38 (14%)
Region of the Americas	244 (25%)	117 (25%)	34 (12%)
Eastern Mediterranean Region	44 (4.6%)	21 (4.6%)	17 (6.2%)
European Region	402 (42%)	233 (51%)	172 (62%)
South-East Asian Region	65 (6.7%)	26 (5.7%)	5 (1.8%)
Western Pacific Region	39 (4.0%)	15 (3.3%)	10 (3.6%)
Geographic scope			
Local	402 (42%)	0 (0%)	0 (0%)
Subnational	218 (23%)	183 (40%)	0 (0%)
National	345 (36%)	276 (60%)	276 (100%)
Study population			
Blood donors	193 (20%)	86 (19%)	69 (25%)
Residual sera	197 (20%)	98 (21%)	53 (19%)
Household and community samples	500 (52%)	256 (56%)	142 (51%)
Pregnant or parturient women	60 (6.2%)	15 (3.3%)	11 (4.0%)
Persons living in slums	5 (0.5%)	1 (0.2%)	0 (0%)
Multiple general populations	8 (0.8%)	3 (0.7%)	1 (0.4%)
Representative patient population	2 (0.2%)	0 (0%)	0 (0%)
Sampling method			
Convenience sampling	254 (26%)	66 (14%)	29 (11%)

Probability sampling	515 (53%)	330 (72%)	198 (72%)
Sequential sampling	178 (18%)	56 (12%)	44 (16%)
Quota sampling	18 (1.9%)	7 (1.5%)	5 (1.8%)
Overall RoB			
Low	183 (19%)	124 (27%)	82 (30%)
Moderate	483 (50%)	335 (73%)	194 (70%)
High	299 (31%)	0 (0%)	0 (0%)
Percent vaccinated at sampling midpoint			
0%	734 (76%)	308 (67%)	158 (57%)
Above 0% up to 5%	70 (7.3%)	29 (6.3%)	22 (8.0%)
Above 5% up to 10%	15 (1.6%)	9 (2.0%)	7 (2.5%)
Above 10%	146 (15%)	113 (25%)	89 (32%)

Most (50%, 483/965) studies were rated moderate RoB (Figure 3.5). The most common reasons for elevated RoB was lack of correction for assay performance or sampling characteristics, use of a non-probability sampling method, or use of a sampling frame that did not exactly approximate the general population (Figure 3.5).

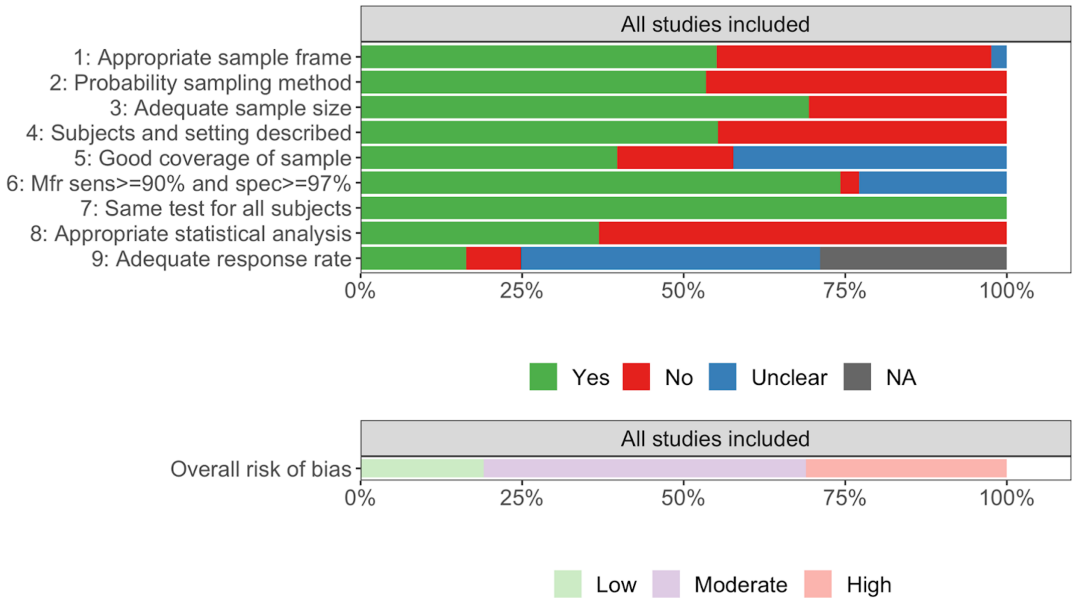


Figure 3.5: Risk of bias assessment for included studies

We selected the most representative and high-quality studies for analysis. As a result, we used only subnational or national studies rated low or moderate RoB to estimate seroprevalence in the general population over time, and identify factors associated with

seroprevalence (sub-dataset 1). We used only national studies rated low or moderate RoB to estimate case ascertainment (sub-dataset 2). These datasets are described in Table 3.2.

3.4.4 Global and regional estimates of seroprevalence over time

The drivers of seroprevalence vary considerably between regions and over time. SARS-CoV-2 antibodies result from infection or vaccination. The extent of infection and vaccination has increased considerably over time. Moreover, the dynamics of infection and vaccination are considerably different between different global regions, and within these regions between countries with different resources. We therefore aimed to produce estimates of global and regional seroprevalence over time.

We estimated seroprevalence over time for each WHO region by meta-analysis of seroprevalence studies in each country, with population weighting to generate regional and global estimates. The Region of the Americas and the European Region are highly heterogeneous; as a result, we divided each of these by World Bank income category into high-income country and low- and middle-income country subregions. We pooled high-income and low- and middle-income countries together in the Eastern Mediterranean and Western Pacific regions due to the lower number of studies, and in the Africa and Southeast Asia regions because the only two high-income countries in these regions had no studies.

3.4.4.1 Country-level estimates of seroprevalence over time

We first produced monthly estimates of seroprevalence in each country, doing so by meta-analysing studies by country in 12-week rolling windows. For each month starting 1 January 2020, we first identified the subnational and national studies in each region within a 12-week moving window centred on that week. If there were multiple studies in the window, we pooled the studies via meta-analysis. We expect seroprevalence to vary between different studies in the same rolling window, due to differences in study time point, target population, sampling, and antibody measurement between studies. Considering this heterogeneity, we pooled studies using random effects meta-analysis, which permits the true effect size to vary between studies.

For random-effects meta-analysis, we take the observed number of seropositive individuals in study i to be y_i among n_i participants (i.e., $\widehat{SP_i} = \frac{y_i}{n_i}$); the true effect for study i to be SP_i; and the grand mean seroprevalence across studies to be SP, with its logit as μ_{SP} and with corresponding between-study variance τ^2. Then, our meta-analysis model is a binomial-normal two-level hierarchical model with logit link, specified as follows:

$$\begin{aligned} y_i &\sim \text{Binomial}\left(n_i, SP_i\right) \\ \text{logit}\left(SP_i\right) &= \mu_{SP} + \alpha_i \\ \alpha_i &\sim \text{Normal}\left(0, \tau^2\right) \end{aligned}$$

We can recover SP from μ_{SP} by inverting the link function: $SP = \text{expit}\left(\mu_{SP}\right)$, where $\text{expit}(x) = \text{logit}^{-1}(x) = \frac{1}{1+e^{-x}}$.

We obtain the maximum likelihood fit for this model using the `metafor` package in R [132, 133]. With `metafor`, we also estimate standard errors for μ_{SP} that account for between-study variance τ^2, and standard errors of prediction for μ_{SP} that account for both between-study variance τ^2 and sampling variance σ^2. We quantify heterogeneity using the I^2 statistic, which captures the proportion of variance due to between-study variation τ^2 as opposed to sampling error σ^2 – that is, $I^2 = \frac{\tau^2}{\tau^2+\sigma_i^2}$ [134].

3.4.4.2 Aggregating country estimates into regional estimates

We produced monthly regional estimates by taking weighted averages of country estimates by population, ensuring that country contributions to these estimates are proportional to country population. Let seroprevalence be denoted SP, population size N, countries c, and regions r. We used weights $w_c = \frac{N_c}{\sum_c N_c}$. With these weights, we calculated regional estimates as:

$$SP_r = \sum_c SP_c \cdot w_c$$

We calculated the standard error SE_r of the regional estimates from the meta-analysis estimated standard errors of the country estimates, SE_c, using the laws of variance as follows:

$$SE_r = \left(\sum_c w_c^2 \cdot SE_c^2\right)^{\frac{1}{2}}$$

In turn, we produced 95% confidence intervals for the mean seroprevalence estimate, reflecting uncertainty in the summary effect size [135]. These were calculated using the normal distribution: $\mu_{SP} \pm 2 \cdot SE_r$. We also calculated 95% prediction intervals to give a range for the predicted parameter value in a new study, doing so using the corresponding standard errors for prediction intervals.

3.4.4.3 Aggregating regional estimates into global estimates

Finally, we aggregated regional estimates into global estimates. We produced monthly global estimates where estimates were available for a majority of regions. To maximise representation of all regions, we calculated monthly global estimates as a population-weighted average of regional estimates, and carried forward the last available regional estimates to September 2021 where necessary. The methods we used to aggregate regional estimates into global estimates directly parallel those we used to aggregate country estimates into regional estimates, as described above.

We then aimed to capture the trend in seroprevalence in each WHO region and globally. To do so, we took the weighted point estimates and fit smoothing splines. To anchor the trend line in the early pandemic period before the first serosurveys were conducted, we specified seroprevalence to be less than 0.1% before the $500th$ confirmed case of COVID-19 in each region. This smoothing is for data visualisations only; no numerical results were presented from this stage.

3.4.4.4 Estimating seroprevalence attributable to infection

Finally, we produced estimates of seroprevalence from infection, doing so by following the above procedure using exclusively estimates of seroprevalence from infection SP_I. We considered all seroprevalence estimates in 2020 to be estimates of SP_I.

For 2021 seroprevalence data, we accounted for seroprevalence study antibody target and the type of vaccines used to generate estimates of seroprevalence attributable to infection. In countries administering only vaccines using spike protein antigens, such as the Moderna, Pfizer, and AstraZeneca vaccines, only infection could trigger the production of anti-nucleocapsid antibodies. In these countries, we considered studies that detected anti-nucleocapsid seroprevalence to be seroprevalence from infection SP_I.

We disregarded studies detecting anti-spike seroprevalence, as anti-spike antibodies could have resulted from either infection or vaccination.

In countries administering inactivated vaccines that may generate both anti-spike and anti-nucleocapsid responses (e.g., Sinopharm, Bharat Biotech), we considered all seroprevalence studies to be seroprevalence from infection or vaccination SP_{IV}. We assumed that the probability that an individual had been infected and the probability that they had been vaccinated were independent of one another. This yields the following formula [136, 137], which we used to calculate SP_I from SP_{IV}:

$$
\begin{aligned}
SP_{IV} &= SP_I + SP_V - SP_I \cdot SP_V && \text{law of independent events} \\
SP_I &= \frac{SP_{IV} - SP_V}{1 - SP_V} && \text{solving for } SP_I \\
SP_I &= \frac{SP_{IV} - v}{1 - v} && \text{substituting vaccination rate } v \text{ for } SP_V
\end{aligned}
$$

In the above formula, v is the fraction of people in the country who were ever vaccinated for SARS-CoV-2 – that is, $v = \frac{N_{vax}}{N_{pop}}$.

3.4.4.5 Findings

Figure 3.6) shows the results from this analysis, including overall seroprevalence and seroprevalence from infection as estimated above, as well as case and variant of concern data to put these results into context. In September 2021, global seroprevalence from infection or vaccination (overall seroprevalence) was 59.2%, [95% CI $56.1-62.2\%$], having risen substantially from 7.7% [CI $5.7-10.3\%$] in June 2020. A substantial proportion of this was due to infection: in September 2021, global seroprevalence attributable to infection was 35.9% [CI $29.5-42.7\%$, prediction interval $22.8-51.4\%$] (Figure 3.6).

Increases in seroprevalence resulted from infection in some region and vaccination in others (Figure 3.6). Regional analyses began in January 2020 and ended in February 2021-March 2022 depending on when seroprevalence studies in each region sampled participants. Seroprevalence rose steeply in some regions with limited vaccination coverage (Figure 3.6). In September 2021, overall seroprevalence was 82.2% [$75.9-87.2\%$] in the South-East Asian Region (increased 8.9-fold since June 2020). Similarly, in December

2021, overall seroprevalence was 86.7% [84.6 − 88.5%] in the African region (increased 24.8-fold since June 2020).

In contrast, steep seroprevalence increases in other countries were in large part attributable to vaccination (Figure 3.6). In March 2022, overall seroprevalence was 95.9% [92.3−97.8%], but infection-induced seroprevalence was only 47.9% [41.0−54.9%] in high-income countries in the European Region. Similarly, March 2022 overall seroprevalence was 99.8% [99.7 − 99.9%] in high-income countries in the Region of the Americas, but infection-induced seroprevalence was only 33.7% [31.6 − 36.0%] (Figure 3.6).

In the meta-analyses by country with at least 2 studies, 75% (188/250) showed considerable heterogeneity, as defined by $I^2 \in [75\%, 100\%]$ [135].

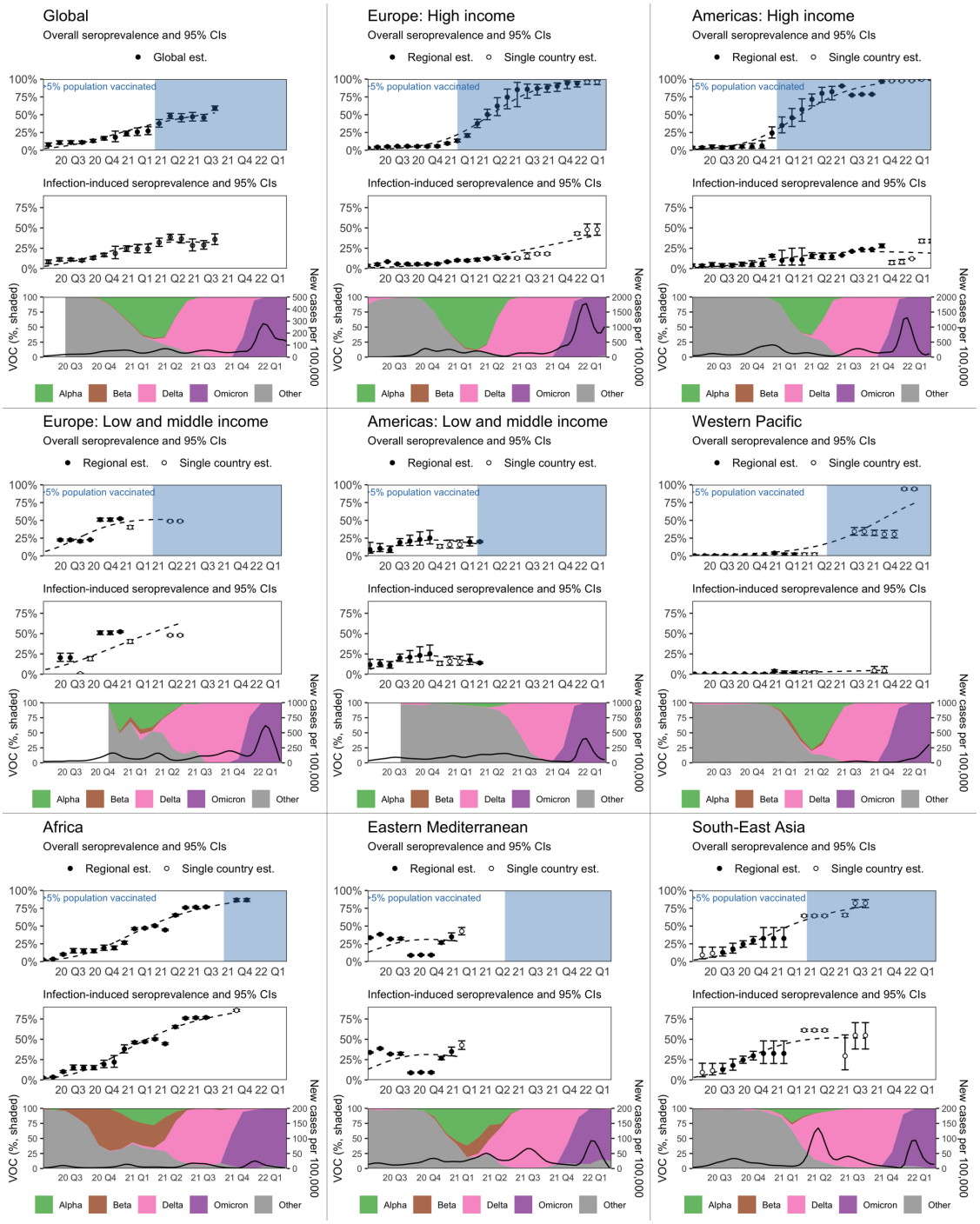


Figure 3.6: Estimated seroprevalence globally and by WHO region from January 2020 to March 2022. Est: estimate; CI: confidence interval. Boxes show global analysis (top left) and eight WHO regional analyses. Each box contains the following panels. *Top and middle panel:* Weighted point estimates and 95% CIs of overall (top) and infection-induced (middle) seroprevalence by meta-analysing studies in 12-week rolling windows. Dashed line visualises the trend over time. *Lower panel, left axis:* Shaded areas represent the relative frequency of major variants of concern, based on weekly counts of SARS-CoV-2 genomes submitted to *GISAID* [130]. *Lower panel, right axis* Black line: smoothed time series of new confirmed cases per 100,000.

3.4.5 Determining the extent of case ascertainment

Measuring case ascertainment is crucial for surveillance. Case data helps us understand the burden of COVID-19. However, case-based surveillance is subject to substantial ascertainment bias. Many infected individuals are never diagnosed with COVID-19 due to asymptomatic infection or limited diagnostic test availability or accessibility. Estimating the extent of underascertainment A is crucial to properly interpret case data by quantifying the number of infections I that go undetected for each confirmed case C.

3.4.5.1 Analytical methods

To estimate case ascertainment, we first repeated the methods described in Section 3.4.4 to generate estimates of infection-induced seroprevalence for each WHO region using only national studies. We used the results to estimate the ascertainment $A_t = \frac{I_t}{C_t}$. While we cannot directly measure I, we take SARS-CoV-2 seroprevalence to approximate the cumulative incidence of infection. We lag seroprevalence data for 9 days with respect to cases, accounting for the 14- day delay typical between infection and seroconversion [138], as well as the 5-day delay typical between infections and confirmation of a case [139, 140]. We then estimate the average ascertainment $\overline{A}$ up to time T as

$$\overline{A_T} = \frac{\sum_{t=1}^{T} I_t}{\sum_{t=1}^{T} C_t} \approx \frac{S_T \cdot N}{\sum_{t=1}^{T-9} C_t}$$

3.4.5.2 Findings

Table 3.3 shows the median seroprevalence to cumulative incidence ratio for each WHO region and income group during each quarter that sufficient data was available for. Our findings demonstrate that confirmed cases substantially underestimate infections. The median global ratio was 51.3 infections derived from seroprevalence to 1 reported case (51.3 : 1) in 2020 Quarter Three, suggesting that around 1.9% of infections were detected, and 10.5 : 1 in 2021 Quarter Three (9.5% of infections detected; Table 3.3). In 2020 Quarter Three, the median ratio ranged from 3.4 : 1 in high-income countries in the Region of the Americas (29.4% of infections detected) to 219.6 : 1 in the Eastern Mediterranean Region (0.5% of infections detected). In 2021 Quarter Three, this ranged

Table 3.3: Seroprevalence to cumulative incidence ratios by WHO region, income level, and quarter. Estimated using national studies. LMICs: low- and middle-income countries. HICs: high-income countries. NA indicates that national studies not available. Seroprevalence studies that sampled participants in 2021 were adjusted for antibody target and vaccination rate to calculate seroprevalence attributable to infection.

		Estimated seroprevalence to case ratios: median [range]									
WHO region	**Income**	**2020 Q3**		**2020 Q4**		**2021 Q1**		**2021 Q2**		**2021 Q3**	
Africa	LMICs	82.2	[43-104.9]	152.9	[115.5-156.9]	154.4	[145.9-211.2]	185.5	[131.2-226.1]	176.7	[164.1-189.2]
Americas	LMICs	22	[6.1-27.5]	18.3	[18.3-18.3]	NA		NA		NA	
	HICs	4.6	[2.5-6.5]	2.6	[2-2.9]	2.3	[2.2-2.4]	2	[1.9-2.2]	1.8	[1.8-1.8]
Eastern	LMICs	219.6	[36.9-425.2]	28.8	[20.8-56.7]	59.3	[56.8-61.8]	NA		NA	
Mediterranean	HICs	8.8	[8.5-27.2]	26.2	[25.3-27]	NA		NA		NA	
Europe	LMICs	84.5	[71.8-97.2]	44.4	[35.5-53.3]	25.8	[25.8-25.8]	11.5	[11.3-11.8]	NA	
	HICs	8.1	[5.5-15.4]	1.9	[1.5-4.3]	2.1	[1.8-2.1]	1.6	[1.6-1.8]	1.9	[1.7-2.1]
Southeast Asia	LMICs	30.1	[22.4-37.8]	40.3	[33.5-48.3]	37.5	[37.5-37.5]	NA		NA	
Western Pacific	LMICs	45.4	[29.4-338.9]	44.3	[38.7-51.6]	34.5	[31.2-37.9]	NA		NA	
	HICs	3.4	[2.9-4]	3.1	[3-3.2]	NA		NA		NA	
Global	**All**	**51.3**	**[39.6-62.9]**	**22.7**	**[20.9-24.5]**	**17.9**	**[15-20.7]**	**13.5**	**[13.5-13.5]**	**10.5**	**[9.3-11.7]**

from 1.8 : 1 in high-income countries in the Region of the Americas (55.6% of infections detected) to 176.7 : 1 in the African region (0.6% of infections detected; Table 3.3).

3.4.6 Meta-analysis of seroprevalence differences between groups

Examining differences in seroprevalence between demographic groups is crucial to understanding differences in the extent of past infection and humoral immunity. However, comparing seroprevalence estimates across serosurveys of different groups is confounded by the true extent of SARS-CoV-2 infection in a region, the assay used, and various study design factors. Instead, we calculate the relative difference in seroprevalence between demographic groups for each study, reducing the impact of confounding variables, and meta-analyse these effect sizes across studies.

3.4.6.1 Analytical methods

To quantify population differences in SARS-CoV-2 seroprevalence, we conducted random-effects meta-analysis of prevalence ratios. First, we identified studies with seroprevalence estimates for age and sex subgroups. We calculated the prevalence ratio PR between each age group and adults aged 20-29, and between males and females.

Let $\widehat{SP}_{G,i}$ be the estimate of true seroprevalence for group G in study i. Then, we can estimate the seroprevalence ratio between group B and group A, $PR_{B/A,i}$, as follows:

$$\widehat{PR}_{B/A,i} = \frac{\widehat{SP}_{B,i}}{\widehat{SP}_{A,i}}$$

As a measure of relative difference, prevalence ratios are more suitable to measure effect sizes across a range of prevalence values than measures of absolute difference like the prevalence difference $\widehat{PD}_{B/A,i} = \widehat{SP}_{B,i} - \widehat{SP}_{A,i}$ [141]. Moreover, prevalence ratios are risk ratios, and can be more directly interpreted than the seroprevalence odds ratio, $\widehat{POR}_{B/A,i} = \frac{\widehat{SP}_{B,i}}{1-\widehat{SP}_{B,i}} / \frac{\widehat{SP}_{A,i}}{1-\widehat{SP}_{A,i}}$ [142].

We expect the extent of differences in seroprevalence to vary across studies, countries, and contexts. Considering this heterogeneity, we pool prevalence ratios using random-effects meta-analysis, an evidence synthesis technique that permits the true effect size to vary between studies. Considering that $PR \in (0, \infty)$, we would like to use a model on the log scale.

A Poisson model with log link would be a conventional model for an outcome on this scale. However, our outcome variable PR is not a count. We employ a Poisson model with offset to address this. Poisson offset models are particularly well-suited for proportion, ratio, or rate outcome measures.

In a conventional Poisson generalised linear model for a count outcome y, we would place a linear model on $\log(\mathbb{E}\, y)$. Instead, in a Poisson offset model, we have a ratio outcome $\frac{\gamma}{\alpha}$ ($\gamma \in \mathbb{Z}_0^+$, $\alpha \in \mathbb{Z}^+$) and place a linear model on $\log\left(\frac{\mathbb{E}\,\gamma}{\alpha_i}\right)$, which resembles the original outcome. Rearranging gives the full model equation:

$$
\begin{aligned}
\gamma_i &\sim \text{Poisson}(\lambda_i) && \text{Poisson likelihood on count } \gamma \\
\log\left(\frac{\mathbb{E}\,\gamma}{\alpha_i}\right) &= \mu_i + \mathbf{x_i} \cdot \beta && \text{Model resembling original outcome } \frac{\gamma}{\alpha} \\
\log\left(\frac{\lambda_i}{\alpha_i}\right) &= \mu_i + \mathbf{x_i} \cdot \beta && \text{Substitute rate parameter for expectation} \\
\log \lambda_i &= \log \alpha_i + \mu_i + \mathbf{x_i} \cdot \beta && \text{Log-linear model with offset}
\end{aligned}
$$

Similarly, we can express prevalence ratios as a ratio of integers as follows:

$$
\begin{aligned}
\widehat{PR}_{B/A,i} &= \frac{\widehat{SP}_{B,i}}{\widehat{SP}_{A,i}} \\
&= \frac{y_{B,i}}{n_{B,i}} \div \frac{y_{A,i}}{n_{A,i}} \\
&= \frac{y_{B,i} \cdot n_{A,i}}{y_{A,i} \cdot n_{B,i}}
\end{aligned}
$$

We then construct a Poisson-normal hierarchical meta-analysis model, with outcome $y_{B,i} \cdot n_{A,i}$ and using $y_{A,i} \cdot n_{B,i}$ as our offset:

$$
\begin{aligned}
y_{A,i} \cdot n_{B,i} &\sim \text{Poisson}\,(\lambda_i) \\
\log(\lambda_i) &= \log\,(y_{A,i} \cdot n_{B,i}) + \mu + \alpha_i \\
\alpha_i &\sim \text{Normal}\left(0, \tau^2\right)
\end{aligned}
$$

We can recover PR by inverting the link function: $PR = e^{\mu}$. We fit this model using the `metafor` package in R [132].

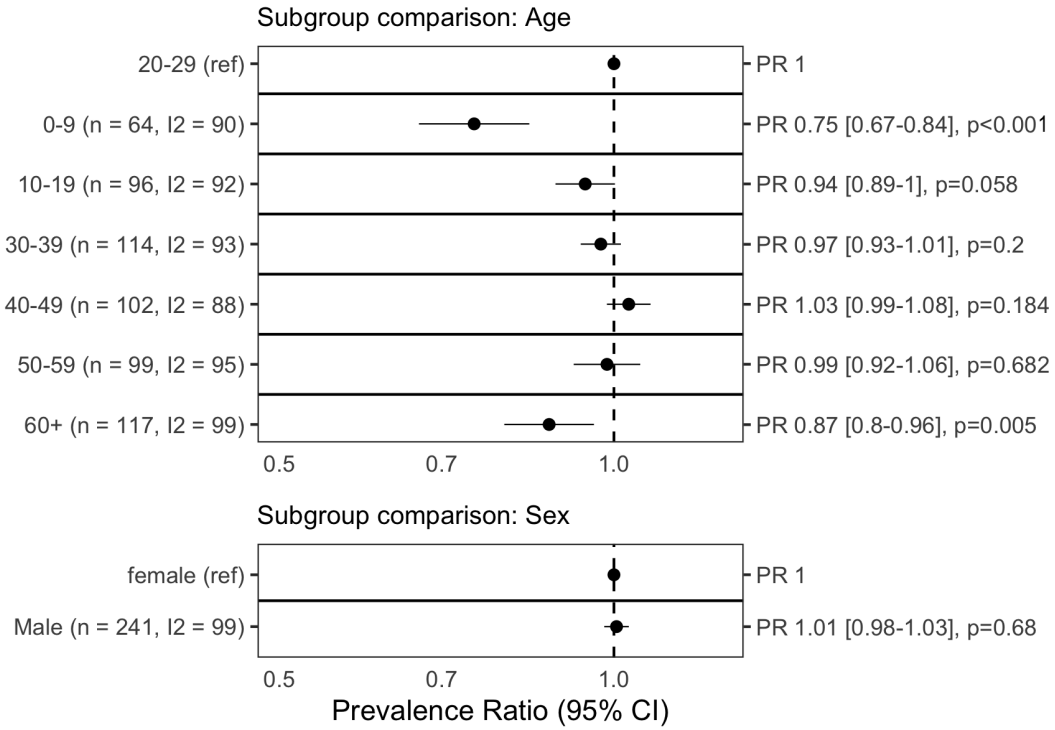


Figure 3.7: Meta-analysis of seroprevalence ratios between demographic groups within studies. Each row shows results from a different random-effects meta-analysis. In each analysis, the seroprevalence ratio between the two specified demographic groups was calculated for each study then meta-analysed across studies. n: number of studies; I^2: metric used to quantify heterogeneity; PR: prevalence ratio.

3.4.6.2 Findings

We observed differences in seroprevalence by age but not sex (Figure 3.7). Within studies, compared to the reference category of adults aged 20-29 years, seroprevalence was significantly lower for children aged 0-9 years (prevalence ratio 0.75, 95% CI $[0.67-0.84]$, $p < 0.001$) and adults aged 60+ years (0.87 $[0.80-0.96]$, $p = 0.005$). There were no significant differences between other age groups nor between males and females (Figure 3.7).

3.4.7 Meta-regression to examine study design and geographic factors

As Figure 3.5 shows, there is considerable variation in seroprevalence study design and quality. Moreover, studies of seroprevalence in the general population have been done using numerous study designs (different general population sample frames), in countries with different characteristics (WHO region, World Bank income level), and with different COVID-19 contexts (extent of spread, PHSM stringency). Understanding the role of

these study design and geographic factors in shaping the heterogeneity of seroprevalence is crucial to be able to interpret seroprevalence studies in context and evaluate the effect of PHSM.

3.4.7.1 Analytical methods

To examine these drivers, we conducted a random-effects multivariable meta-regression. To focus on factors associated with seroprevalence from infection, we included studies where less than 5% of the national population was vaccinated two weeks before the sampling midpoint date.

We defined independent predictors *a priori*. Continuous covariates included cumulative incidence of reported cases per hundred in the country of the study $C_{\text{per 100}}$. We obtained data on total confirmed SARS-CoV-2 infections, and population size that geographically matched the study populations nine days before the study end date, to reflect the time period between COVID-19 diagnosis and seroconversion [139, 140, 143]. We also included the average global PHSM index $PHSM$, ranging from 0 (least stringent) to 10 (most stringent), averaged between the day of the 100th COVID-19 case in the country and 14 days before the study start date [144].

Categorical predictors included WHO region WHO (reference group: Region of the Americas), country World Bank income class WBI (reference group: high-income), sampling frame SF (reference: household and community samples), and study scope SS (reference: national). These predictors were one-hot encoded, with the one-hot vector for predictor θ denoted $\mathbf{e}_\theta$ and the corresponding coefficient vector denoted $\boldsymbol{\beta}_\theta$.

Our ultimate outcome variable of interest is $\widehat{SP_i} = \frac{y_i}{n_i}$, which lends itself well to a Poisson offset model. We use a Poisson model here, rather than a Beta model, so that model coefficients can be directly interpreted as risk ratios. This yields a two-level Poisson-Normal hierarchical model, including Poission offset n_i, a study-level random-

effect, shared intercept β_0, as well as the predictors specified above:

$$
\begin{aligned}
y_i &\sim \text{Poisson}\left(\lambda_i\right) \\
\log(\lambda_i) &= \log\left(n_i\right) + \alpha_i + \beta_0 \\
&\quad + \beta_C \cdot C_{\text{ per 100},i} + \beta_{PHSM} \cdot PHSM_i \\
&\quad + \boldsymbol{\beta}_{WHO} \cdot \boldsymbol{e}_{WHO[i]} + \boldsymbol{\beta}_{WBI} \cdot \boldsymbol{e}_{WBI[i]} \\
&\quad + \boldsymbol{\beta}_{SS} \cdot \boldsymbol{e}_{SS[i]} + \boldsymbol{\beta}_{SF} \cdot \boldsymbol{e}_{SF[i]} \\
\alpha_i &\sim \text{Normal}\left(0, \tau^2\right)
\end{aligned}
$$

We fit this model using the `glmer` function from the `lme4` package in R [133, 145–147].

3.4.7.2 Findings

In the multivariable meta-regression, 329 studies remained after applying our inclusion criteria for pre-vaccination studies. The full model is reported in Figure 3.8.

We examined two study design factors: study scope and sample frame. Subnational studies reported higher seroprevalence estimates compared to national studies (PR 1.27 $[1.02 - 1.59]$, $p = 0.03$) (Figure 3.8). There were no significant differences between seroprevalence in studies that used blood donors (PR 1.04 $[0.77 - 1.40]$, $p = 0.79$) or residual sera (PR 1.08 $[0.83 - 1.41]$, $p = 0.55$) as proxies for the general population.

The two country characteristics we included were WHO region and World Bank income levels. Compared to the Region of the Americas, higher seroprevalence estimates were reported in the European Region, (PR 1.35 $[1.02 - 1.78]$, $p < 0.04$) with lower seroprevalence estimates reported in the Western Pacific Region (PR 0.17 $[0.10 - 0.29]$, $p < 0.001$; Figure 3.8). Compared to high-income countries, higher seroprevalence estimates were reported by low- income (PR 7.33 $[3.49 - 15.41]$, $p < 0.001$), lower-middle-income (PR 7.33 $[3.49 - 15.41]$, $p < 0.001$), and upper-middle-income countries (PR 3.97 $[2.88 - 5.49]$, $p < 0.001$).

Finally, we included two COVID-19 context factors: the cumulative incidence of confirmed cases and PHSM stringency. Higher cumulative incidence of reported cases was associated with higher seroprevalence (PR for one-unit increase 1.39 $[1.30 - 1.49]$, $p < 0.001$), while more stringent PHSM measures before the study were associated

with lower seroprevalence (PR for one-unit increase on 0-10 scale: 0.89 [0.81 − 0.98], $p = 0.02$; Figure 3.8).

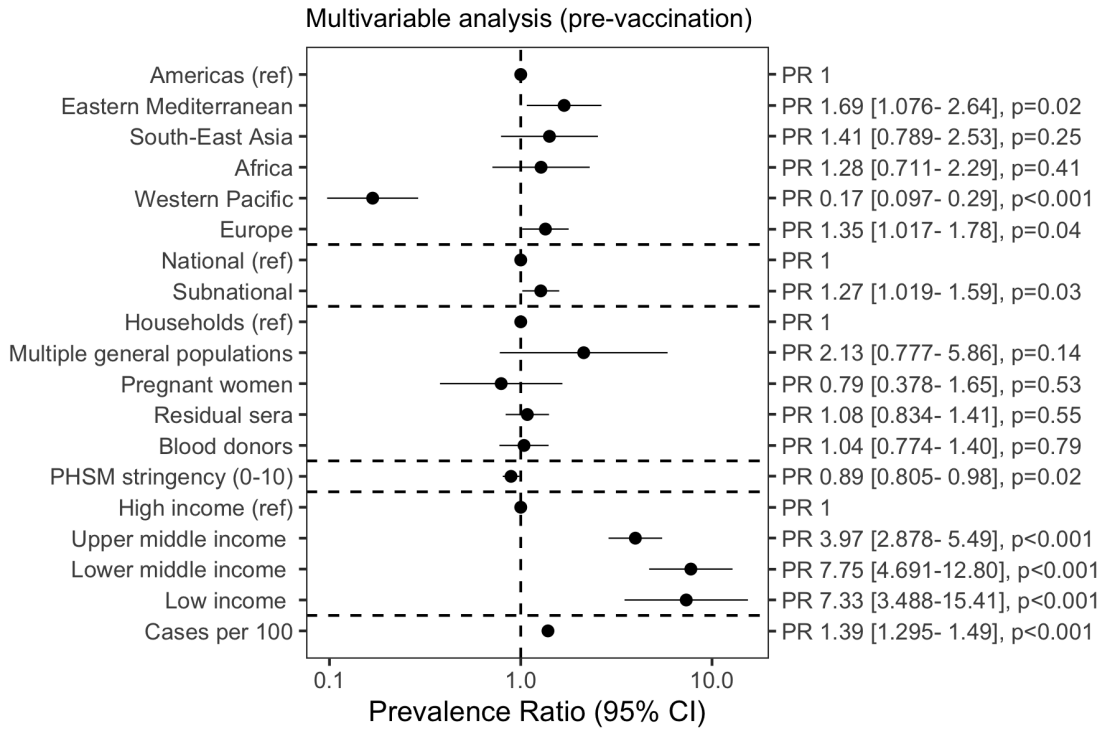


Figure 3.8: Meta-regression to identify study design, country, and COVID-19 context factors associated with seroprevalence pre-vaccination. PR = prevalence ratio. We fit a Poisson generalised linear mixed model with log link, including studies where less than 5% of the national population was vaccinated. Public health and social measure data was encoded as a continuous variable from 0 (least stringent) to 10 (most stringent). Confirmed cases per 100 was encoded as a continuous variable. All other variables were encoded as factors, with reference groups specified in the figure. The variation between studies explained only by fixed effects (i.e., marginal R^2) was 62.9%.

3.4.8 SARS-CoV-2 seroprevalence in Africa

Africa is experiencing unprecedented challenges due to the Coronavirus disease 2019 (COVID-19) pandemic, both directly and from the compounding effects of other health, economic, and social factors [148, 149]. Reported data indicates a less severe disease profile in Africa compared to other regions globally: fewer cases, proportionally fewer patients with severe outcomes and death, and proportionally more asymptomatic cases [110, 150, 151].

Precisely estimating COVID-19 cumulative incidence in Africa remains problematic due to challenges in contact tracing, routine surveillance systems and laboratory testing

capacities, and strategies in many countries. Furthermore, Africa is a large, complex and heterogeneous continent with a range of different economies, countries impacted by humanitarian crises, vulnerable population groups, and unique public health challenges. With low vaccine coverage in Africa (17% as of 9 February 2022 [93]), using seroprevalence data to understand the true dynamics of SARS-CoV-2 infection and case underascertainment is key to map the extent of protection in the general population and inform the public health response.

We generated seroprevalence-based epidemiological profiles for countries in Africa, paralleling the regional profiles shown above. We highlight four of these in Figure 3.9, from countries with notable longitudinal cohort or repeated cross-sectional seroprevalence studies: Malawi, Kenya, Ghana, and South Africa. Each of these indicate sharp increases in seroprevalence in the general population over time. SARS-CoV-2 seroprevalence is high and heterogenous in Africa, suggesting greater population exposure to SARS-CoV-2 and potential protection against COVID-19 severe disease than has been previously indicated by confirmed case data and vaccine coverage. As such, PHSM and vaccination strategies tailored to local settings and specific populations are warranted.

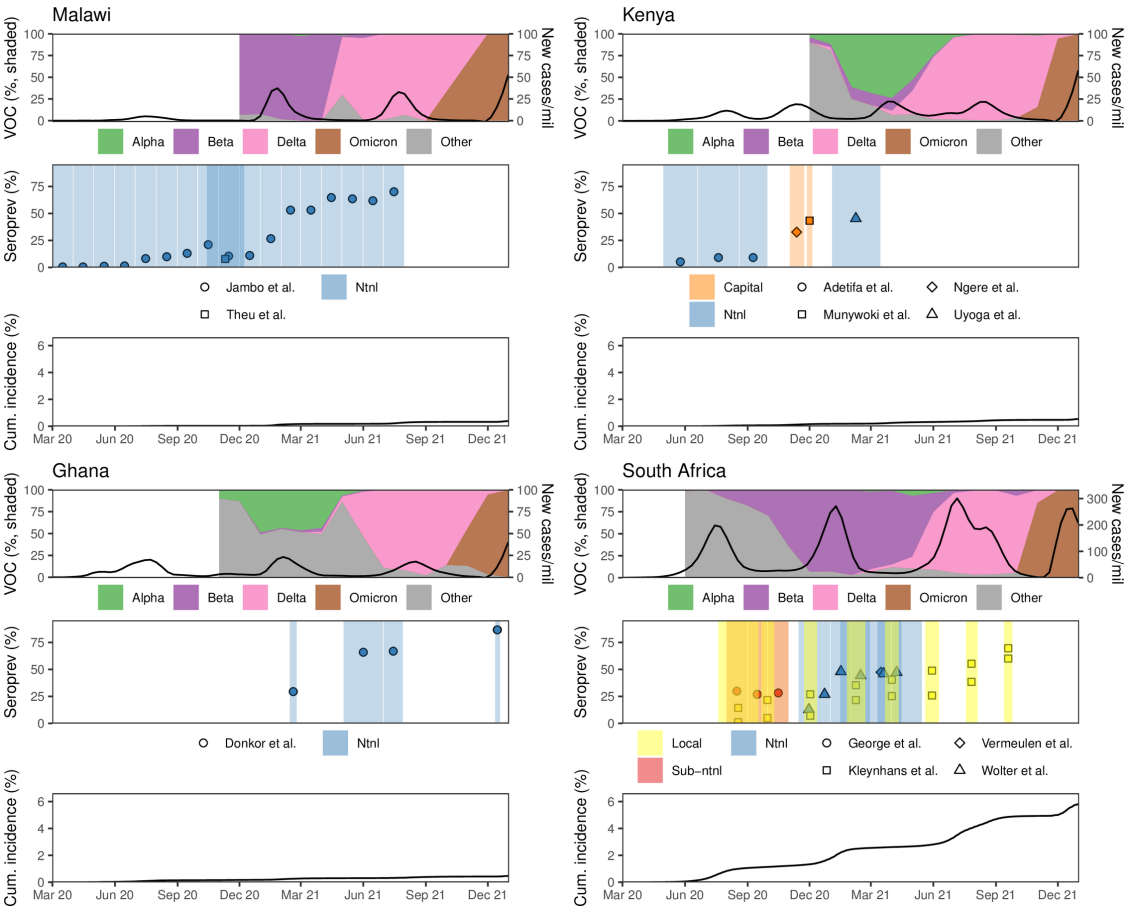


Figure 3.9: Reported seroprevalence, variants of concern, and cumulative incidence by selected countries over time, March 2020 - December 2021. Upper panel, left axis: Shaded areas represent the relative frequency of major variants of concern circulating, based on weekly counts of hCoV-19 genomes submitted to *GISAID* we have aggregated by month. Weeks with fewer than 10 total submissions in a given country were excluded from the analysis. Upper panel, right axis: Daily confirmed cases reported to WHO on a national level per million people, smoothed using local regression (LOESS). Middle panel: each point is an individual seroprevalence study, and identical shapes represent studies originating from the same cohort or repeated cross-sectional data source. Lower panel: cumulative incidence of confirmed cases per 100 people.

3.5 Discussion

We synthesised data from over 950 seroprevalence studies worldwide (43% from low- and middle-income countries) published up to May 2022, providing global and regional estimates of SARS-CoV-2 seroprevalence over time, with substantial representation of regions with limited available seroprevalence data. We estimate that approximately 59.2% of the global population had antibodies against SARS-CoV-2 in September 2021 (35.9% when excluding vaccination). Global seroprevalence has risen considerably over time, from 7.7% in June 2020.

Our monthly timeline of seroprevalence by region provides estimates of evolving immunity during the global pandemic. Regional variation is driven by differences in the extent of SARS-CoV-2 infection and vaccination. We observed increases in seroprevalence following the emergence of variants in regions with available data (e.g., 6% (July 2020) to 41% (April 2021) in the African region following the Beta variant and 12% (February 2021) to 75% (August 2021) in the South-East Asian Region following the Delta variant), demonstrating the substantial number of infections caused by more transmissible variants. In high-income countries, the increases in overall seroprevalence were driven by increased vaccine coverage in early 2021 (e.g., 6% (January 2021) to 95% (August 2021) in high-income countries in the Region of the Americas and 7% (January 2021) to 72% (August 2021) in high-income countries in the European Region), while we also observed increases in infection-induced seroprevalence following the Omicron variant (e.g., 7% (December 2021) to 34% (March 2022) in high-income countries in the Region of the Americas and 18% (October 2021) to 48% (March 2022) in high-income countries in the European Region). Another potential driver for regional variation is the potential detection of P. falciparum or other common cold coronaviruses which have high cross-reactivity to SARS-CoV-2. This has been remarked upon in the literature [152–154], and may impact seroprevalence estimates in areas of Africa or elsewhere with a high incidence of malaria.

Our results add global representation and principled estimation of changes in seroprevalence over time as compared to previous evidence syntheses [120, 121, 155]. These estimates are similar to estimates of true infections by global epidemiological models. For example, our global estimate of seroprevalence attributable to infection (35.9%) is similar to the Institute of Health Metrics and Evaluation cumulative infection incidence estimate of 42.8% on 15 September 2021 [156]. Our analysis provides an orthogonal estimate based solely on seroprevalence data, using a method that has the added value of being easily interpretable and with no assumed parameters.

We provide evidence of considerable case underascertainment, indicating that many cases of SARS-CoV-2, are not captured by case-based surveillance systems which may be based on testing for symptomatic infection or simply have limited access to testing [157]. There was wide variation in underascertainment in all regions, income groups, and

over time, with higher ratios consistently observed in low- and middle-income countries compared to high-income countries. Our ratios of seroprevalence to reported cases in late 2020 were comparable to other studies for the Region of the Americas, European Region, and South-East Asian Region [119–121, 158]. Our estimates of seroprevalence to cumulative incidence ratios for the African, Western Pactific, and Eastern Mediterranean Regions are entirely novel, with no other analyses we found having systematically estimated ascertainment through seroprevalence in these regions. Moreover, estimates of true infections from epidemiological models suggest the high levels of underascertainment suggested by this study are plausible [156].

We also provide more granular evidence of significant variation in infection by age by 10-year band. Children aged <10 years, but not children aged 10-19 years, were less likely to be seropositive compared to adults aged 20-29 years; similarly, adults aged >60 years, but not those aged 30-39 years, 40-49 years, or 50-59 years, were less likely to be seropositive than adults aged 20-29 years. These findings add nuance and granularity to differences in seroprevalence by age observed by other studies [119]. There are several possible explanations for lower seroprevalence in children: milder infections, which are generally associated with lower antibody titers [159]; school closures; and ineligibility for vaccination. Lower seroprevalence in adults aged 60+ years could be explained by immunosenescence that can lead to quicker seroreversion [160], higher mortality, and hence a lower proportion of individuals with evidence of past infection, gaps in vaccine access, or more cautious behaviour resulting in fewer infections in this age group.

Our multivariable model suggests higher seroprevalence estimates were reported by low- and lower-middle-income countries compared to high-income countries, with the highest seroprevalence in lower-middle-income countries (pre-vaccination). Potential explanations for this result are multifaceted and include weaker health system functionality and performance, lower capacity to isolate, and less stringent use of and ability to effectively implement PHSM. This is also consistent with others' findings [120]. This model also suggests that an increase in overall PHSM stringency was associated with lower seroprevalence. This and other work have shown that the use of PHSM was associated with reduced SARS-CoV-2 infections, especially when implemented early and limiting

population mobility [161–163]. Our model also found that subnational studies have higher estimates than national studies. One hypothesis for this is that subnational studies are often concentrated on cities or areas with denser populations, which may contribute to increased transmission of the virus. Further research is needed to validate this hypothesis.

About 40% of studies used blood from donors and residual sera as sampling frames to estimate seroprevalence in the general population. It has previously been shown that blood donors are disproportionately young, White, college graduates, employed, physically active, and never-smokers [164, 165]. In multivariable meta-regression, our results suggest that these studies report seroprevalence estimates similar to studies of household- and community-based samples, with other covariates held equal. This finding suggests that it may be reasonable to use these proxy sampling frames to obtain reasonable estimates of general population seroprevalence. To do so, investigators should use large sample sizes, ensure adequate coverage of important subgroups (e.g., age, sex, race/ethnicity, socioeconomic status), and use robust survey weighting methods.

3.5.1 Strengths and limitations

Our regional and global meta-analysis estimates are timely, robust, and geographically diverse, with estimates from all WHO regions. We present high-quality and comparable data, enabled by the laboratory and epidemiological standardisation enabled by the UNITY protocol, as well as the analysis of only studies assessed to have low or moderate RoB using a validated RoB tool [158]. Despite this effort, there are still methodological differences between the meta-analysed studies that may reduce their comparability. For example, 14% of studies in our analysis dataset (66/459, 18 of which were household-based) were convenience samples, which are less representative than population-based probability samples. To limit this bias, we required Unity-Aligned convenience samples to have a clearly defined sample frame (i.e., sampling of volunteers excluded). Our RoB evaluation also included subjective review of the demographic breakdown in the study, coverage of subgroup estimates, and author comments on representativeness of the sample, such that the most nonrepresentative studies were rated high RoB and excluded from analysis.

A few limitations should be described. First, although we conducted meta-regression to explore heterogeneity of the included studies, there remained some residual heterogeneity that could not be explained quantitatively — likely driven by differences in disease transmission in the different countries and time points that serosurveys were conducted. Second, we did not account for waning antibody levels or seroreversion, meaning that the present work likely underestimates the extent of past infection, and therefore underestimates case ascertainment. Third seroprevalence studies are cumulative, meaning that results reflect all COVID-19 countermeasures implemented up to the time of participant sampling, and thus, we cannot isolate the contributions of particular PHSM. Fourth, while we screened study eligibility based on high assay performance criteria, different serological assays may yield varying results, which should be taken into account when interpreting seroprevalence data. Fifth, at certain points in time, our meta-analysis estimates were driven by studies from specific countries — either very populous countries (i.e., South-East Asian Region: India, Region of the Americas high-income countries: USA, Region of the Americas low- and middle-income countries: Brazil, Western Pacific Region: China), or countries in regions with otherwise scarce data during the time in question (e.g., Eastern Mediterranean Region: 2 countries in early 2021). We also could not produce global estimates for late 2021/early 2022 due to the delays between when studies conducted their sampling and when these results were later published.

An additional category of limitations relates to our estimation of seroprevalence due to infection. When estimating seroprevalence from infection in countries which used vaccines triggering anti-N responses, we made the simplifying assumption that the probability an individual was infected is independent of the probability that they were vaccinated. However, this assumption may not be valid in reality, considering the protective effect of vaccination against infection and health-related behaviors that affect both the probability of vaccination and infection. If this assumption does not hold, this would create bias in our estimates of seroprevalence from infection. For example, if people who are infected are less likely to be vaccinated, we overestimate the seroprevalence due to both infection and vaccination, and thereby underestimate seroprevalence due to infection alone.

Deriving estimates of seroprevalence from infection required us to make an assumption about the joint distribution of vaccination and infection — a quantity for which data is not globally available. Of the assumptions we could have made — that infection and vaccination are mutually exclusive, that all infected people are also vaccinated, that all vaccinated people are also infected, or that vaccination and infection are independent — this is the choice that yielded the most reasonable estimates. These estimates could be further improved by using data about the joint distribution of infection and vaccination, in countries where it is available, or by using estimates of vaccine efficacy.

3.5.2 Implications and next steps

Population-based seroprevalence studies primarily give a reliable estimate of the exposure to SARS-CoV-2 infection in the general population. Where studies measure antibodies quantitatively, they can also estimate correlates for protection against infection [118]. The risk of reinfection with the Omicron variant is much higher than for previous variants, in both SARS-CoV-2 infected and vaccinated individuals [166]. This implies that the presence of SARS-CoV-2 antibodies may no longer be a correlate of protection against infection for Omicron. However, seroprevalence estimates remain indicative of some protection against severe disease, as cellular immunity is unlikely to be affected even in the case of highly mutated variants such as Omicron [167]. Research on the protective effectiveness of seropositivity against disease severity, and the degree and timing of protection against Omicron, is limited and warrants further investigation.

A key challenge in implementing serosurveillance has been timeliness of study implementation, data analysis, and reporting. As such, it will be important for public health decision-makers to prioritise investment and establish emergency procedures to facilitate timely study implementation early on in future outbreak or variant emergence responses as part of overall surveillance strategy. There is also a need to continue to build national capacity with WHO and other partners to rapidly enable high-quality study implementation and communication of findings in a format friendly to decision-makers. Globally standardised and quality seroprevalence data continue to be essential

to inform health policy decision-making around COVID-19 control measures, particularly in capacity-limited regions with low testing capacity and vaccination rates.

4

Delays and biases challenging SARS-CoV-2 seroprevalence estimation

4.1 Early methodological challenges and motivation

Fundamentally, seroprevalence studies are designed to contribute to effective population-based surveillance. To do so, it is crucial that they have the characteristics of effective surveillance systems, which have previously been discussed in Section 1.2.2. Meeting these characteristics requires conducting investigations that are timely, robust, and representative. Unfortunately, some SARS-CoV-2 seroprevalence studies have faced difficulty meeting each of these criteria.

Our early 2020 systematic review of seroprevalence studies, previously briefly discussed in 3.2, identified nearly 100 seroprevalence studies already done, and more than that amount ongoing or proposed (Figure 4.1). However, this systematic review also began to reveal some of the key methodological challenges in these studies [81].

Despite the substantial amount of data already available, there was substantial variation in study design and quality. Only one-quarter of initiatives reporting results used an appropriate sample frame and random sampling. Many of the studies used sampling designs that may oversample younger adults (e.g., recruiting through Facebook advertisements), healthy individuals (e.g., blood donors [168]), and individuals of higher socioeconomic status (e.g., daytime drive-through testing [169]). Few estimates were

available for low-income countries. As a result, no estimates had low risk of bias (RoB; 21% moderate RoB; 79% high or unclear RoB; Figure 4.1).

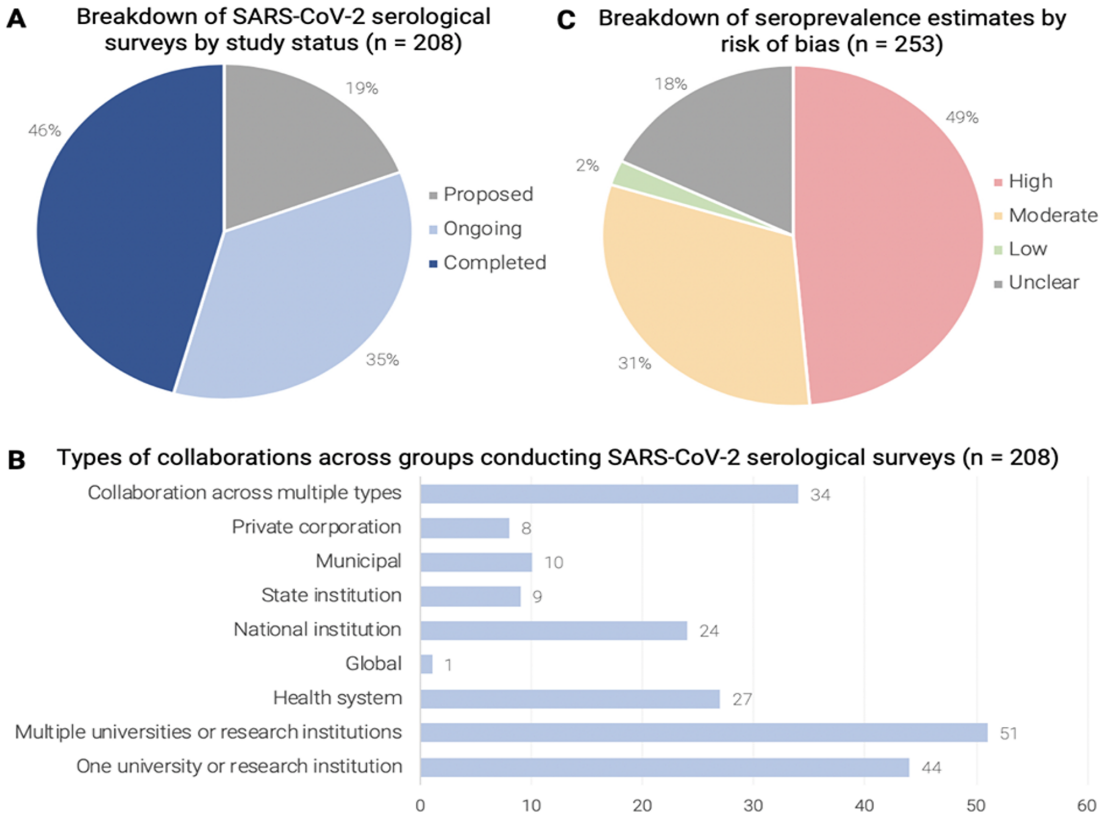


Figure 4.1: Characteristics of SARS-CoV-2 seroprevalence studies.

At the time, the reported seroprevalence across studies spanned 0% to 61.0% – a staggering range, considering that SARS-CoV-2 had been circulating for just months. While this wide variation could be attributed to true differences in viral spread, it could also be a consequence of the methodological flaws of these studies. Compare the results of studies at different RoB available at the time: of all studies available to 26 June 2020, the high and unclear RoB studies reported median seroprevalence 4.2% (IQR 1.5−8.8%, total range 0−61.0%), while the low and moderate RoB studies reported median seroprevalence 3.1% (IQR $1.6 - 6.0\%$, total range $0 - 22.6\%$). These results suggested that the urgency to examine seroprevalence may have compromised the methodological rigour and hence validity of some early seroprevalence estimates.

In parallel, the vast majority of seroprevalence studies were being done by academic institutions (Figure 4.1). Research sometimes prioritises rigour over speed, and

the research nature of many seroprevalence studies foreshadowed the challenges with their timeliness that we would later observe. Moreover, very few estimates were available for low-income countries, highlighting the coverage limitations that would continue through the pandemic.

These challenges sometimes persisted well beyond the early days of the pandemic and limited the ability of seroprevalence investigations to be useful surveillance tools. Here, we conduct two investigations of these challenges, focusing first on the poor timeliness of many seroprevalence studies, and second on the measurement bias introduced by poorly performing and heterogeneous serological assays.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Bobrovitz N*, **Arora RK*,**, Yan T, ..., Evans TG. 2020. Lessons from a rapid systematic review of early SARS-CoV-2 serosurveys. *medRxiv. Rahul's contributions:* sole responsibility for analytical methods development and analysis; contributions to search, screening, and data extraction methods. Additionally, conceptualization of manuscript; defining research question and objectives; writing original draft; revising draft, all in equal portion with Bobrovitz N.

Donnici C*, Ilincic N*, ..., **Arora RK.** 2022. Timeliness of reporting of SARS-CoV-2 seroprevalence results and their utility for infectious disease surveillance. *Epidemics. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

Ma X, Li Z, Whelan M, ..., **Arora RK.** 2022. Serological assays used in SARS-CoV-2 seroprevalence surveys worldwide: an analysis of assay features, testing algorithms, and performance. *Vaccines. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

4.2 Timeliness of seroprevalence study reporting

4.2.1 Motivation

During the COVID-19 pandemic, seroprevalence estimates have primarily been generated through numerous uncoordinated stand-alone research studies instead of systematic ongoing public health surveillance efforts, raising questions about their public health impact [9, 70, 100]. For these seroprevalence studies to be effective when used for public health surveillance, it is important that they have the attributes of effective surveillance systems (Table 4.1), which include timeliness, representativeness, and validity, among others [10, 18]. However, many of these attributes are challenging to realise through standalone one-time studies, particularly when findings are shared in research manuscripts published as peer-reviewed articles.

Peer review can delay the availability of data by months. The scientific publication process has been criticised for these delays [170–172], as these have an impact on public health responses, and can also hinder secondary analysis, modeling, and global comparisons. Journals are also not designed for the routine reporting of surveillance data, and may not even consider updated results from a previously reported seroprevalence study as sufficiently novel for publication.

To expedite dissemination of results, some researchers have turned to more rapid and accessible platforms, such as news and media [173], government reports [174], and preprints [175]. However, the generalisability and validity of such non-peer-reviewed evidence has been questioned [176, 177]. It remains unclear whether these alternative platforms do indeed lead to faster reporting compared to scientific journals, and whether the ability to bypass peer review has resulted in prolific publication of weaker evidence.

We aimed to determine the timeliness of SARS-CoV-2 seroprevalence studies in providing information useful for public health surveillance. First, we described the timeliness of SARS-CoV-2 seroprevalence reporting by publication venue. Second, we also investigated whether more timely reporting compromises other facets of effective surveillance, by examining relationships between timeliness, data quality, and representativeness.

4.2.2 Defining timeliness, representativeness, and data quality

We focused on timeliness as a key determinant of the effectiveness of seroprevalence data for public health surveillance, examining timeliness alone and in relation to representativeness and data quality. Other attributes of effective public health surveillance that could not be evaluated based on seroprevalence study reports, and not examined in detail here, include simplicity, flexibility, acceptability, stability, sensitivity, and positive predictive value. Table 4.1 provides definitions of each attribute and the measures used to operationalise it in the present study.

Table 4.1: Attributes and measures of effective public health surveillance used in this study. The measures for data quality and validity were derived from Serotracker's Risk of Bias tool (Section 6.2), which is a modified version of the Joanna Briggs Institute Critical Appraisal Checklist for Prevalence Studies [178].

Attribute	Definition	Measures used in this study
Timeliness	Refers to the time between any two steps in the surveillance system. For example, the time between the onset of a health event and reporting of the event to a public health agency [18]	• How rapidly were results published after participants were sampled? The time elapsed between the last date of participant sample collection (sampling end date) and the publication date • For studies released via multiple platforms in succession (e.g., first as a government report, second as a preprint, finally as a journal article), we used the first date the results were available
Data quality (validity)	Refers to the validity of data in the surveillance system, which is influenced by the performance of assays [10], statistical methods, and surveillance methods (i.e., study type, geographical scope, etc.)	• Were valid methods used to identify seropositivity? Did the serology test used meet the WHO standards for Emergency Use Authorisation for COVID-19 serology tests (sensitivity $\geq$ 90% and specificity $\geq$ 97%) • Was sampling error corrected for? If non-probability sampling was used, were statistical adjustments or weighting of sample demographics performed?
Representativeness	Refers to the ability of the surveillance system to accurately describe the health event. This is achieved by considering its distribution in the population [18].	• Was the sample frame representative of the target population? • Was data analysis conducted with sufficient coverage of the identified sample? To evaluate coverage, we examined if sample demographics aligned with target population demographics (age, sex, ethnicity)

4.2.3 Dataset and descriptive analysis

We examined all 1,844 seroprevalence studies in the SeroTracker database that were published between 1 January 2020 and 31 December 2021.

The majority (59%) of studies were first released as peer-reviewed journal articles, followed by preprints (24.2%), institutional reports (7.81%), news articles (6.24%), and presentations or conference abstracts (2.66%). Presentation or conference materials and media reports had higher RoB than other publication venues. Only 5.1% of presentation

or conference materials and 9.6% of media reports had a low or moderate RoB. Larger proportions of low or moderate RoB studies were reported in peer-reviewed journal articles (32%), preprints (42%), and institutional reports (51%).

Across all publication venues, median time to publication was 154 days (IQR 64−255). The shortest median time to publication was 0 days for a media report, while the longest was 556 days for a peer-reviewed article.

4.2.4 Timeliness by study publication venue

We examined the relationship between timeliness and publication venue. This was a time-to-event analysis, where the event in question was the publication of a study. For each study, we define t_0 as the time when the study finished sampling participants, as the theoretical first date a study could have been published. We define $S(t)$, the survival cumulative distribution function, as the probability that a given study remains unpublished in any venue t days after sampling participants (i.e., t days after t_0). Then, the publication cumulative distribution function is $F(t) = 1 - S(t)$, with its probability distribution function being $f(t) = \frac{\mathrm{d}F}{\mathrm{d}t}$.

We visualise these relationships with the Kaplan-Meier estimator, an unbiased non-parametric estimator of $S(t)$ [179]. Let p_i represent the number of studies published in any venue on day i, and n_i be the number at risk – the number of studies that still remained unpublished just before day i. The Kaplan-Meier estimator is then:

$$\begin{aligned}\hat{S}(t_i) &= \hat{S}(t_{i-1})\left(1 - \frac{p_i}{n_i}\right) \\ &= \prod_{t_i \leq t}\left(1 - \frac{p_i}{n_i}\right)\end{aligned}$$

Figure 4.2 shows Kaplan-Meier plots for timeliness stratified by publication venue.

Timeliness varied considerably by publication venue (Figure 4.2). Media reports (median 12 days; IQR $3-25$) were released significantly faster than institutional reports (median 18 days; IQR $2-45$, $p = 0.015$ by log-rank test). Both media and institutional reports were published significantly faster than studies released in all other publication venues ($p < 0.001$). Preprints (median 101 days; IQR $49-180$) were released faster than presentation or conference materials (median 187 days; IQR $43-295$; $p = 0.0033$),

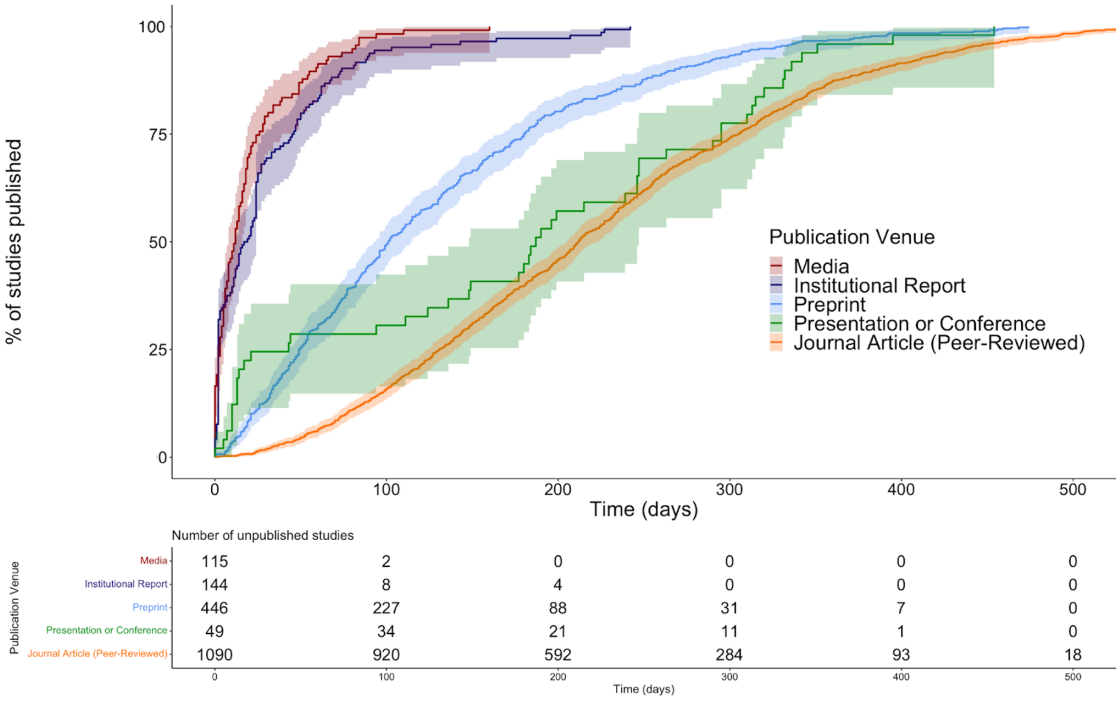


Figure 4.2: Kaplan-Meier curve and risk table for time to publication by publication venue. Timeliness curves are plotted with 95% confidence intervals (shaded area).

and both venues released study results significantly faster compared to peer-reviewed journal articles (median 212 days; IQR $131 - 305$; $p < 0.001$ for preprints and $p = 0.049$ for presentations).

To assess the role of preprints in expediting the release of data, we compared the median time to publication for peer-reviewed journal articles that first appeared as preprints and peer-reviewed journal articles that were not preprinted. There was no significant difference in time to publication between these groups ($n = 230$, $p = 0.3$).

We show that peer-reviewed manuscripts are released particularly slowly, with a median time to publication of about seven months. While many medical journals have expedited publication processes for COVID-19 research [176], our study demonstrates continued delay in the publication of seroprevalence findings. This raises the question of whether peer-reviewed journals are fit for purpose for reporting surveillance and seroprevalence findings. This is particularly true given that some journals may see routine updates on seroprevalence as not sufficiently novel to be published, potentially introducing publication bias.

The median preprint was reported close to four months faster than the median published article, but still nearly three months slower than the median government or institutional report. While preprints are enabling more rapid dissemination of information compared to publications, their median time to publication is still over three months, suggesting that the process of preparing an academic manuscript in the first instance introduces substantial delays. This suggests that preprint platforms may themselves not be suitable for routine surveillance reporting.

Interestingly, our analysis shows that it is possible for a seroprevalence study to be both timely and robust. Institutional reports, which were published rapidly, had a greater proportion of low RoB studies compared to preprints and published articles.

4.2.4.1 Relationship between timeliness and other study characteristics

We also examined the relationship between timeliness and each study characteristic (i.e., geographic scope, sample frame, and WHO region), each measure of representativeness, and each measure of validity.

To quantitatively assess these relationships, we conducted Cox regression. To do so, we first take the hazard probability distribution function $h(t)$, the probability of article publication at time t given that a study remained unpublished until time t - that is, $h(t) = \frac{f(t)}{S(t)}$. Then, Cox regression makes the assumption that the hazard probability distribution function for any two studies a and b is proportional to one another [180], implying that the hazard ratio $\frac{h_b(t)}{h_a(t)}$ is equal to a time-independent constant $\lambda > 0$. In Cox regression, we place a linear model on the log of the hazard ratio (HR). For a study a, a covariate vector $\mathbf{x}_a$ with coefficient vector $\boldsymbol{\beta}$, and a baseline hazard function across studies $h_0(t)$, this model is as follows [180]:

$$\log\left(\frac{h_a(t)}{h_0(t)}\right) = \mathbf{x}_a^\top \boldsymbol{\beta}$$

We examined the following categorical predictors: publication venue (reference: peer-reviewed journal article), study geographic scope (reference group: national), WHO region (reference group: Region of the Americas), sampling frame (reference: household and community samples), overall RoB (reference group: low), and four individual

items evaluating data representativeness and validity, as defined in Table 4.1 (reference: criterion met). We one-hot encoded these categorical predictors.

We included all significant predictors on univariate prediction in a multivariate Cox regression, excluding overall RoB due to its colinearity with individual RoB items assessing data quality and validity. Predictor coefficients were quantified using the corresponding hazard ratio (HR), with higher HR indicating faster time to publication.

Table 4.2 shows the full results from this analysis, including univariate Cox regressions conducted separately for each individual predictor, and an overall Cox regression conducted with all predictors significant on univariate regression. On univariate regression, we observed significant differences in timeliness by publication venue (overall model $p < 0.001$ per Wald test), sample frame ($p < 0.001$), WHO region ($p = 0.002$), sample frame representativeness ($p = 0.001$), sample coverage ($p < 0.001$), assay quality ($p < 0.001$), and appropriate sampling ($p < 0.001$), but not by geographic scope ($p = 0.3$; all Table 4.2).

In the multivariate regression, publication venue was a significant independent predictor of timeliness. Compared to peer-reviewed journal articles, preprints (HR 2.24 [2.0 − 2.52], $p < 0.001$), media (HR 17.0 [13.3 − 21.8], $p < 0.001$), and institutional reports (HR12.4 [10.1 − 15.1], $p < 0.001$) were associated with faster publication (Table 4.2). Presentation or conference materials were not associated with more timely dissemination in comparison to journal articles ($p = 0.1$).

Studies that sampled blood donors or residual sera did not differ in timeliness when compared to studies that investigated household or community samples ($p = 0.2$); however, studies sampling healthcare workers (HR 0.81 [0.70 − 0.94], $p = 0.004$) and other special populations (HR 0.74 [0.64 − 0.86], $p < 0.001$) took significantly longer to publish than studies that sampled the general population (Table 4.2). The use of non-probability sampling methods was associated with faster publication (HR 1.2 [1.03 − 1.41], $p = 0.02$).

There were no significant associations between timeliness and the WHO region a study was conducted in. Further, there were no significant associations between timeliness and study representativeness or data validity, with the exception of the item evaluating if there was appropriate sampling or population adjustment (Table 4.2).

Table 4.2: Cox regression table. Hazard ratios > 1 mean that the comparator group was published faster than the control group.

Reference group	Comparison group	Single variable Cox regression		Multivariable Cox regression	
		HR [95% CI]	*p* value	HR [95% CI]	*p* value
Publication Venue: Journal article (peer-reviewed)	Media	20.4 [16.4-25.3]	< 0.001***	17.0 [13.3-21.8]	< 0.001***
	Institutional Report	12.2 [10.1-14.8]	< 0.001***	12.4 [10.1-15.1]	< 0.001***
	Preprint	2.3 [2.0-2.5]	< 0.001***	2.2 [2.0-2.5]	< 0.001***
	Presentation or Conference	1.3 [1.0 - 1.8]	0.05	1.3 [0.96 - 1.7]	0.09
Sample Frame: Household and community samples	Blood donors or residual sera	0.76 [0.65 - 0.88]	< 0.001***	0.88 [0.74 - 1.05]	0.15
	Healthcare workers and caregivers	0.58 [0.51 - 0.66]	< 0.001***	0.81 [0.70-0.94]	0.004**
	Other special populations	0.58 [0.51 - 0.66]	< 0.001***	0.74 [0.64-0.86]	< 0.001***
	Multiple populations	0.69 [0.58 - 0.82]	1.7e-05***	0.90 [0.74 - 1.08]	0.26
WHO Region: Region of the Americas	African Region	1.13 [0.89 - 1.43]	0.30	1.17 [0.92-1.49]	0.19
	Eastern Mediterranean Region	0.77 [0.62 - 0.95]	0.02*	0.84 [0.68-1.05]	0.14
	European Region	0.85 [0.77 - 0.95]	0.004**	0.95 [0.85-1.06]	0.33
	South-East Asian Region	1.09 [0.91 - 1.31]	0.35	1.03 [0.86-1.25]	0.73
	Western Pacific Region	0.87 [0.72 - 1.06]	0.17	0.90 [0.74-1.10]	0.30
Was the sample representative of the target population: Yes	No	0.96 [0.87-1.06]	0.43	1.10 [0.99 -1.22]	0.09
	Unclear	1.30 [1.11-1.53]	0.001**	1.07 [0.90-1.27]	0.45
Was there appropriate sample coverage: Yes	No	0.83 [0.71-0.98]	0.03*	0.85 [0.72-1.00]	0.06
	Unclear	1.04 [0.93-1.16]	0.48	1.01 [0.90 -1.14]	0.82
Did the assay used meet WHO standards: Yes	No	1.09 [0.91-1.32]	0.35	1.01 [0.83-1.22]	0.92
	Unclear	1.24 [1.13-1.36]	< 0.001***	1.08 [0.98-1.19]	0.14
Was sampling appropriate: Yes	No	0.77 [0.67-0.89]	< 0.001***	1.20 [1.03-1.41]	0.02*
Geographical Scope: National	Regional	0.88 [0.75 - 1.03]	0.099	Excluded	Excluded
	Local	0.93 [0.82 - 1.05]	0.23	Excluded	Excluded
Overall RoB: Low	Moderate	1.13 [0.87 - 1.47]	0.34	Excluded	Excluded
	High	1.12 [0.87 - 1.44]	0.38	Excluded	Excluded
	Unclear	2.22 [1.65 - 2.99]	< 0.001***	Excluded	Excluded

In our analysis, studies of healthcare workers and other special populations took longer to publish compared to studies of the general population. This is in part because studies of the general population were typically regional or national studies, conducted by government affiliated groups, and disseminated via institutional or government reports, whereas studies of special populations were largely done by academic groups and released as research articles. However, these delays are problematic considering the importance of seroprevalence to inform best practices in high risk settings, such as hospitals. There is a clear need for rapid release of the findings from these studies to enable their use for public health.

There were no differences in the timeliness of low, high, and moderate RoB studies on univariate analysis, suggesting that faster publication of valid study results is possible. However, studies with limited information to evaluate bias (unclear RoB) were published significantly faster. Many of these studies were published via news and media reports, which may account for the lack of data needed to evaluate bias. News and media outlets should endeavour to link to extended reports provided by investigators, even if not peer-reviewed or preprinted on a formal platform, in order to maximise dissemination of crucial study context.

4.2.5 Discussion

Our analysis shows that many SARS-CoV-2 seroprevalence studies have not reported their findings in a timely fashion. With a median 154 days between sampling and reporting, there are considerable challenges in using these data for public health decision-making or impactful secondary analysis. Studies reported in preprints and peer-reviewed articles were much slower to be released compared to studies on other publication platforms, emphasising delays introduced by the academic writing and publishing process that make seroprevalence studies less useful for public health decision-making and secondary analysis. However, we also show that it is possible to quickly release robust seroprevalence results: government or institution reports were more timely and had better data validity and representativeness compared to academic manuscripts. This suggests that there

are opportunities to improve the timely reporting of strong seroprevalence studies, and thereby improve their value for public health surveillance.

The slow reporting of SARS-CoV-2 seroprevalence studies overall emphasises limitations in their relevance for public health action. The landscape of infection and immunity can change drastically in the median 154 days from the end of sampling to results release [181]. Notably, some of the results from these studies are made available to public health agencies directly before being released publicly — for example, many studies of blood donors and residual sera in Canada. While this improves the ability of the agency in question to act on the data, the closed sharing of results hinders interpretation and action by other stakeholders. Firstly, public health agencies who the data has not been shared with (e.g., federal authorities, for studies done at a state/province level), which limits the coordination between levels of government that is crucial in a pandemic setting [182]. Secondly, academic research groups, who have done secondary analysis and modeling that has generated key information during the pandemic [183]. Finally, global synthesis and comparison initiatives: where this has been carried out for seroprevalence, these delays have caused limitations in the synthesis that can be done [70].

Collectively, these findings point back to a fundamental divide between the way in which research studies are ordinarily done and performing effective public health surveillance. Surveillance systems generate information for action, which may not be novel; in contrast, research generates information for knowledge, and is conventionally communicated via the peer-reviewed literature. The academic literature – whether peer-reviewed or preprint – may not be fit for purpose as a way of communicating surveillance information. However, because many public health agencies have limited resources and serosurveillance expertise, they have often had to rely on delayed results from intermittent studies conducted by academics, which are often delayed and not systematic. This approach is problematic, and accelerating peer-review processes may be unlikely to solve issues of timeliness.

Continuous serosurveillance, where governments perform routine serological sampling with rapid reporting, would address many of the challenges that the present study identifies. As an example, the REACT 2 programme involves repeated serological testing with

an established analytical framework, thereby allowing for sampling, analysis, reporting, and data sharing [184]. Continuous serosurveillance provides standardised, up-to-date data, and avoids the need to extrapolate data over major gaps between sampling periods. To achieve this, public health agencies would need to be resourced to perform unique, ongoing, and systematic serosurveillance — either independently or in partnership with academic groups. However, these resources are often lacking, meaning that there are comparatively few examples of mature, well-funded, and ongoing serosurveillance systems run by public health agencies [185].

Our analysis had several strengths. We aimed to identify all seroprevalence studies publicly reported in 2020 and 2021, providing complete coverage of seroprevalence studies in the first two years of COVID-19 pandemic. Our inclusion of studies across all geographic regions, publication venues, populations studied, and study designs enables comprehensive analysis. Moreover, the multivariate Cox regression conducted here enables us to isolate the association between timeliness and the covariates of interest, including study characteristics, measures of data quality, and measures of representativeness.

Some limitations of our approach should also be kept in mind. First, we were not able to determine the duration of key steps in the reporting process for each article — for example, to analyse samples, analyse data, and prepare, peer review, copy-edit, and typeset the report. Greater granularity here would help inform tailored suggestions to expedite reporting. Second, some news and conference articles did not report an end date for sampling and had to be excluded, while we had to approximate sampling end date for studies reporting imprecise sampling dates (e.g., 'between March and April'). Third, studies that were reported directly to public health agencies and never published could not be identified or included in this analysis. Fourth, we did not assess for effect modification on the Cox models. If one of the included covariates was an effect modifier, we may have observed differential effects of other covariates depending on the effect modifier, and the parameters estimated in the Cox model may not be accurate.

4.3 Serological assays and measurement error

4.3.1 Motivation

Serological assays enable measurement of antibodies in seroprevalence studies. We quantify the performance of a serological assay in terms of its sensitivity γ, the proportion of true positive samples (i.e., containing SARS-CoV-2 antibodies) that are correctly classified seropositive, and its specificity δ, the proportion of true negative samples classified seronegative. In response to expanding demand for serosurveys, many SARS-CoV-2 serological assays were developed, mobilised, and adopted since the beginning of the pandemic. However, the performance of these assays has varied broadly, and many seroprevalence studies have used poorly performing assays.

Serological assays with imperfect performance can introduce measurement bias into seroprevalence studies. To illustrate this, in a seroprevalence survey with y positive samples of n total samples, the proportion of positive samples $SP = \frac{y}{n}$. However, for sensitivity $\gamma < 100\%$, y will erroneously exclude some true positives, and for specificity $\delta < 100\%$, y will erroneously include some true negatives, biasing the estimate of δ.

In our early 2020 systematic review, we noted that just 3% of studies used antibody tests that met the 90% sensitivity and 97% specificity threshold recommended by the World Health Organization. Serological test performance varied broadly, from $63.3\% - 100\%$ sensitivity and $89.2\% - 100\%$ specificity. This broadly varying assay performance raises the concern that SARS-CoV-2 seroprevalence estimates are potentially biased by imperfect or inconsistent assay performance, especially in cases where no statistical adjustments are made to account for test performance.

An important complicating factor is the fact that sensitivity γ and δ are not fixed properties of the assay. Rather, these properties depend on the population that an assay will be deployed in. Further complicating this, assay performance is most often not evaluated in the field – it is typically estimated in the laboratory, with a reference panel that may not represent the population where assays will be deployed. This imperfect correspondence introduces a bias termed spectrum bias [186].

Commercial assays constitute the vast majority of assays used in serosurveys, and manufacturers of these commercial assays self-certify their products by evaluating their

performance [187]. Such evaluations were usually done in the early pandemic, with reference panels using small sets of true positive samples drawn from patients with confirmed symptomatic COVID-19, and true negative samples excluding patients with potential cross-reactivity due to infection with other viruses [188]. The lack of endemic samples representing the demographics and endemic pathogens in a study area introduces spectrum bias [186]. The reference panels used in manufacturer estimates can be particularly prone to producing high estimates of sensitivity and specificity, potentially leading to biased estimates of γ and δ that propagate to biased estimates of seroprevalence.

Several other groups besides manufacturers conduct their own evaluations of serological assays, each using reference panels with varying compositions. As a result, performance evaluations from different sources can often disagree with one another, and with manufacturer evaluations, particularly as there is no gold standard for reference panel composition. This inconsistency raises the risk that seroprevalence study investigators will use misspecified seroprevalence estimates, which is of particular concern given the lack of independent evaluations for certain serological assays.

To interpret serosurvey data appropriately, it is important to understand how much estimates from different assay validations differ, to understand the extent to which assay performance may be misspecified by investigators, and in turn to understand the extent to which misspecified assay performance may bias seroprevalence estimates.

Considering this, we set out to:

1. Describe what serological assays have been used in SARS-CoV-2 serosurveys during the COVID-19 pandemic

2. Comprehensively examine assay sensitivity and specificity estimates and how they differ between evaluators

3. Assess the influence of performance misspecification on seroprevalence estimates

4.3.2 Serological assays used in seroprevalence studies

There is a broad and diverse array of SARS-CoV-2 serological assays, with hundreds of these assays commercially available. We document these assays in the SeroTracker

database, with a description of relevant data fields documented in detail in Table A.5. We also link each seroprevalence estimate to the assay(s) employed to generate that estimate.

A small fraction of these have seen disproportionate use. We set out to examine what serological assays have been used in population-based SARS-CoV-2 seroprevalence studies during the COVID-19 pandemic. We included all 1,807 seroprevalence studies in SeroTracker that were published between 1 January 2020 and 19 November 2021. This represents the time window during which most serological assays were introduced.

Table 4.3: Features of commercial and self-developed serology assays used by studies. Abbreviations as follows: LFIA: lateral flow immunoassay, IFA: immunofluorescence assays, ELISA: enzyme-linked immunosorbent assays; CLIAs: chemiluminescent immunoassays. NRL-WHO: Australian National Reference Laboratory for the World Health Organization, US FDA: United States Food and Drug Administration, FIND: Foundation for Innovative New Diagnostics, NL CIDC: Netherlands Center for Infectious Disease Control.

Assay characteristics	**Commercial assays (N = 192)**		**Self-developed assays (N = 380)**	
	n	**%**	**n**	**%**
Assay Type				
ELISA	60	31.3	261	68.7
LFIA	75	39.1	0	0.0
IFA	5	2.6	17	4.5
CLIA	30	15.6	3	0.8
Neutralisation	0	0.0	52	13.7
Other	22	11.5	47	12.4
Rapid Diagnostic Test?				
Rapid diagnostic test	103	53.6	17	4.5
Non-rapid diagnostic test	89	46.4	363	95.5
Antibody Target				
Spike	55	28.6	48	12.6
Nucleocapsid	37	19.3	37	9.7
Multiple targets	38	19.8	171	45.0
Unknown	62	32.3	124	32.6
Antibody Isotypes Detected				
IgG-only	52	27.1	149	39.2
IgG and IgM	103	53.6	31	8.2
Total antibody	22	11.5	38	10.0
Other or missing	15	7.8	162	42.6

Among these serosurveys, we identified 192 commercial assays and 380 self-developed assays (Table 4.3). Of the 192 identified commercial assays, 31.3% were enzyme-linked immunosorbent assays, 39.1% were lateral flow immunoassays, 15.6% were chemilumines-

cent immunoassays, 2.6% were immunofluorescence assays, and 15.6% were other types. Rapid diagnostic tests (typically including lateral flow immunoassays and immunofluorescence assays) accounted for 53.6% (103/192) of all commercial assays. Of the 380 self-developed assays, most were enzyme-linked immunosorbent assays (68.7%, Table 4.3). Only 4.5% (17/192) of self-developed assays were developed as rapid diagnostic tests.

80.7% of studies used a single serological assay, while 19.3% used a testing algorithm involving multiple assays (Table 4.3). 45.0% of studies using self-developed assays used multiplex detection to recognise multiple antibody targets.

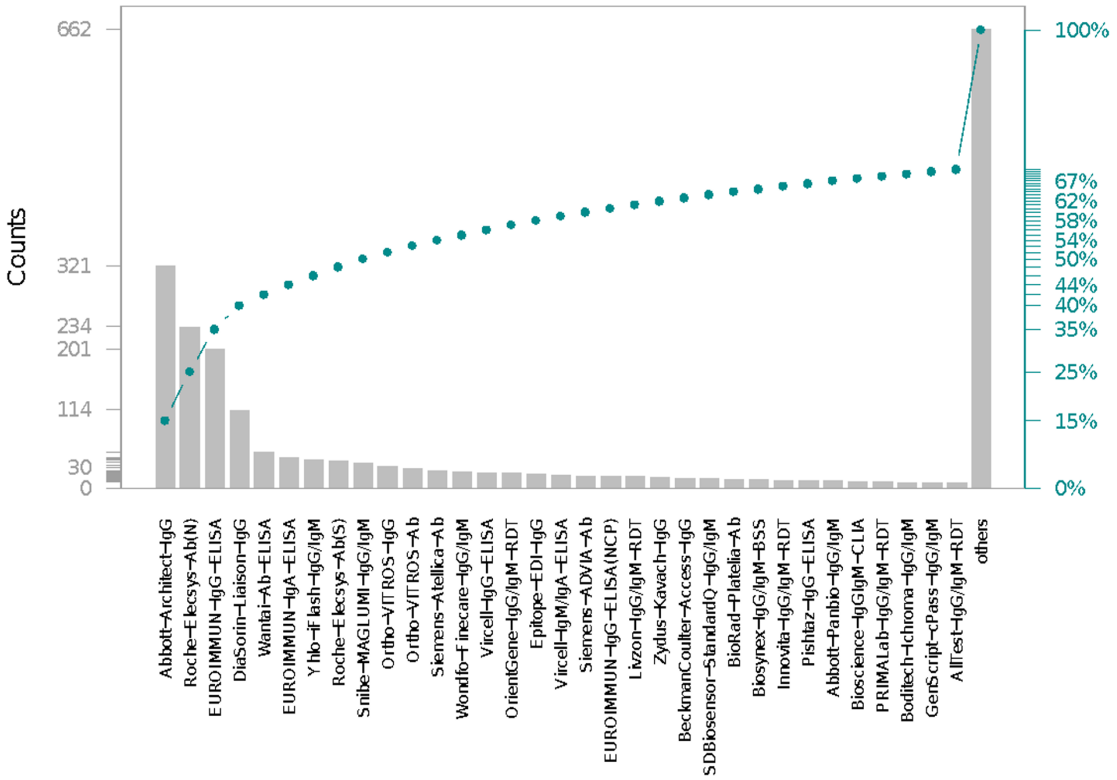


Figure 4.3: Pareto chart of serological assay use. Plot shows the frequency of serological assay use in 1,807 serosurveys. Assays are sorted by frequency of use. The most common 25 immunoassays are shown individually, with the remainder grouped together.

The deployment of commercial serological assays in seroprevalence studies followed a power law distribution, with the most common 25 assays accounting for 67.0% of total commercial assay use in seroprevalence studies (Figure 4.3), and the most common 50 assays accounting for 91.4% of use.

4.3.3 Comparison of assay performance evaluations

Many different groups conducted SARS-CoV-2 assay validations. The testing environment, validation procedure, and reference panel varied widely between groups conducting validation, in turn leading to considerable potential variation in sensitivity γ and specificity δ. We sought to systematically evaluate the agreement between these assay performance evaluations.

We included assay performance data from three sources: manufacturers, independent serosurvey investigators, and third-party laboratory validations. For each evaluation, we extracted information on the evaluator, the composition and size of the reference panel, and the estimated γ and δ (full schema in Table A.7). The sources we included were as follows:

- Manufacturer estimates of sensitivity and specificity, produced for assay product self-certification and available in product inserts.
- Independent serosurvey investigator validations reported alongside seroprevalence studies; these were typically either independent evaluations using readily available samples at serosurvey sites, or field evaluations designed to minimise spectrum bias by using a reference panel that resembled the study population.
- Third-party validations that conducted large-scale, head-to-head evaluations on a common reference panel. We identified five of these: from the Australian National Laboratory for the WHO ([189]), the US Food and Drug Administration [190], the Netherlands Center for Infectious Disease Control [191], the Doherty Institute [192], and the Foundation for Innovative New Diagnostics [193]. Details of each evaluation and its reference panel are in Table A.6.

4.3.3.1 Descriptive assessment

We systematically compared performance data for the 50 commercial assays that were most frequently used in seroprevalence studies, comparing these across manufacturer, independent, and third-party evaluations, and against the World Health Organization targets for emergency use (i.e., $\gamma \geq 90.0\%$, $\delta \geq 97.0\%$ [194]).

The mean manufacturer-evaluated γ was 97.8% and the mean δ was 99.7%; as a result, 78% of the most 50 frequently used serological assays met the WHO standard based on manufacturer evaluations (Figure 4.4). In contrast, only 46% and 54% of the same 50 assays met these criteria based on third-party and independent evaluations, respectively. Chemiluminescence immunoassays demonstrated higher and more reliable performance across all three evaluation sources than enzyme-linked immunosorbent assays, lateral flow immunoassays, and immunofluorescence assays (Figure 4.4).

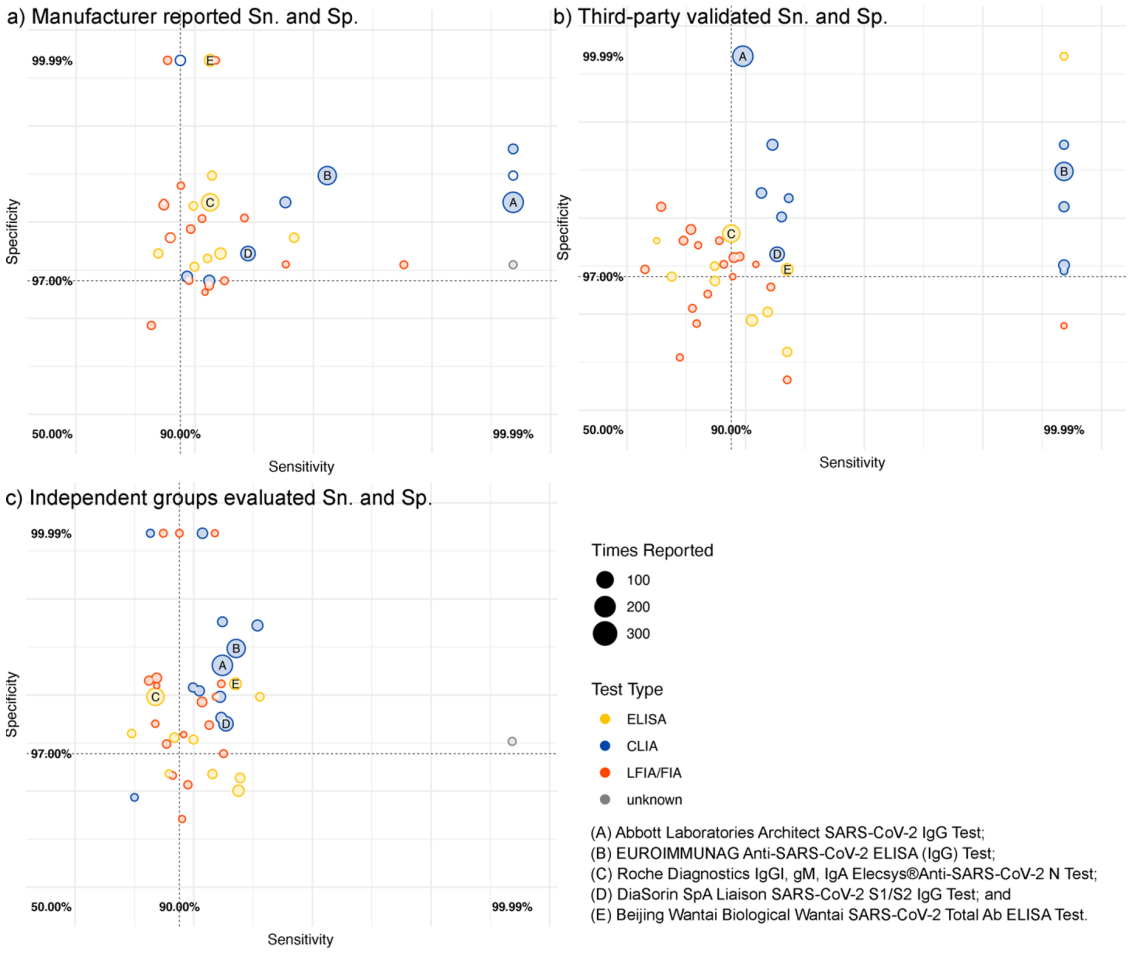


Figure 4.4: Comparison of sensitivity and specificity from manufacturer, third-party, and independent group evaluations. Includes the 50 most frequently used commercial serological assays. Sensitivity γ is plotted on a $\log(1-\gamma)$ scale and specificity δ is plotted on a $\log(1-\delta)$ scale, to make clear differences at high sensitivity and specificity values. The vertical and the horizontal lines indicate the WHO thresholds for Emergency Use Authorisation for COVID-19 serological tests: $\gamma > 90.0\%$ and $\delta 97.0\%$. Assays on the upper right area of each panel meet the WHO criteria for emergency use based on the dataset in question.

4.3.3.2 Statistical modelling of assay performance

To examine potential systematic differences in assay performance by evaluation source, we developed a mixed-effects Beta regression model. We denote an estimate of sensitivity for a given assay a with evaluation source ES as $\widehat{\gamma}_{a,ES}$. We encoded evaluation source ES as a categorical fixed effect, with groups including manufacturer (reference); independent field evaluation; and third-party laboratory validation, with each individual laboratory nested within that category. Other fixed effects included in the model were isotype detected ISO, assay type AT, and antibody target TGT. All categorical predictors were one-hot encoded, with the one-hot vector for predictor $PRED$ specified as $\boldsymbol{e}_{PRED}$ We included a random effect for each assay a, with between-assay variance τ^2.

Then, our model is a Beta-normal two-level hierarchical model with logit link, specified for sensitivity as follows:

$$\begin{aligned}
\widehat{\gamma}_{a,ES} &\sim \text{Beta}\left(\mu_{a,ES}, \kappa\right) \\
\text{logit}\left(\mu_{a,ES}\right) &= \alpha_a + \beta_0 + \boldsymbol{\beta}_{ES} \cdot \boldsymbol{e}_{ES} \\
&\quad + \boldsymbol{\beta}_{ISO} \cdot \boldsymbol{e}_{ISO[a]} + \boldsymbol{\beta}_{AT} \cdot \boldsymbol{e}_{AT[i]} + \boldsymbol{\beta}_{TGT} \cdot \boldsymbol{e}_{TGT[i]} \\
\alpha_a &\sim \text{Normal}\left(0, \tau^2\right)
\end{aligned}$$

For estimates of specificity $\widehat{\delta}_{a,ES}$, we use a parallel approach with the same predictors. We fit models for sensitivity and specificity independently.

In this model, manufacturers systematically overestimate sensitivity γ by 5.4% (95% CI: $2.6 - 8.2\%$, $p < 0.001$) compared to third parties and by 6.3% ($3.9 - 8.7\%$, $p < 0.001$) compared to independent evaluators (Table 4.4). Manufacturers similarly systematically overestimate specificity, but by a comparatively smaller amount: 2.8% ($1.8 - 3.9\%$, $p < 0.001$) compared to third parties and by 1.2% ($0.2 - 2.1\%$, $p = 0.016$) compared to independent evaluators (Table 4.4).

There is also considerable heterogeneity between specific third-party evaluations. For example, the Foundation for Innovative New Diagnostics evaluation systematically estimated lower assay sensitivity by 22.8% ($18.9 - 26.7\%$, $p < 0.001$; Table 4.4), potentially due to its positive reference panel composition which included many samples < 14 days

Table 4.4: Predictors of assay sensitivity and specificity, estimated with mixed-effects Beta regression. Sensitivity and specificity were modeled separately with random effects for individual serological assay. Ref indicates the levels of categorical predictors used as reference categories. Abbreviations as follows: LFIA: lateral flow immunoassay, IFA: immunofluorescence assays, ELISA: enzyme-linked immunosorbent assays; CLIAs: chemiluminescent immunoassays. NRL-WHO: Australian National Reference Laboratory for the World Health Organization, US FDA: United States Food and Drug Administration, FIND: Foundation for Innovative New Diagnostics, NL CIDC: Netherlands Center for Infectious Disease Control.

Variable	**Sensitivity**			**Specificity**		
	Est. (%)	95% CI (%)	*p*	Est. (%)	95% CI (%)	*p*
Source of Sensitivity or Specificity Estimate						
Manufacturer	ref.	-	-	ref.	-	-
Independent	-6.3	(-8.9, -3.7)	<0.001*	-1.2	(-2.1, -0.2)	0.016*
Third-Party Lab	-5.4	(-8.2, -2.6)	<0.001*	-2.8	(-3.9, -1.8)	<0.001*
NRL-WHO	2.2	(-2.8, 7.2)	0.39	-6.7	(-8.6, -4.7)	<0.001*
US FDA	0.5	(-2.7, 3.7)	0.739	-2.4	(-3.7, -1.2)	<0.001*
FIND	-22.8	(-26.7, -18.9)	<0.001*	-2.6	(-4.2, -1.0)	0.002*
NL CDC	-1.8	(-5.9, 2.4)	0.409	-2.1	(-3.6, -0.5)	0.011*
Doherty	-7.1	(-12, -2.3)	0.004*	-2.7	(-4.6, -0.8)	0.005*
Isotype						
IgG only	ref.	-	-	ref.	-	-
IgG and IgM	-1.5	(-6.1, 3.2)	0.538	-0.8	(-3.1, 1.6)	0.519
Total Antibody	2.8	(-1.6, 7.2)	0.217	0.8	(-1.6, 3.2)	0.516
Other	-3.8	(-9.9, 2.3)	0.225	-1.4	(-4.6, 1.7)	0.371
Assay Type						
ELISA	ref.	-	-	ref.	-	-
LFIA	-1.6	(-6.4, 3.1)	0.5	-0.8	(-3.2, 1.7)	0.533
IFA	2.7	(-6.4, 11.7)	0.565	-2.1	(-6.7, 2.5)	0.381
CLIA	0.0	(-4.0, 4.0)	0.995	0.4	(-1.7, 2.6)	0.699
Other	-1.4	(-10.9, 8.2)	0.778	2.4	(-1.6, 6.4)	0.246
Antibody Target						
Spike	ref.	-	-	ref.	-	-
Nucleocapsid	-4.3	(-8.1, -0.5)	0.026*	-1.0	(-2.8, 0.8)	0.294
Multiple	0.3	(-3.7, 4.2)	0.901	-2.4	(-4.5, -0.3)	0.022*
Unknown	-5.4	(-8.7, -2.1)	0.001*	-1.0	(-2.7, 0.6)	0.217

post-symptom onset that may not have seroconverted (Table A.6). Analogously, the Australian National Reference Laboratory evaluation systematically estimated lower assay specificity by 6.7% ($4.7 - 8.6\%$, $p < 0.001$; Table 4.4), potentially due to its inclusion of a

large number of cross-reactive specimens within its negative reference panel (Table A.6).

We observed significant differences in assay sensitivity by antibody target, with anti-nucleocapsid assays reporting lower sensitivity by 4.3% ($0.5 - 8.1\%$, $p = 0.026$; Table 4.4), which corroborates knowledge on the lower sensitivity of anti-nucleocapsid assays and anti-nucleocapsid seroreversion [26, 188]. Similarly, assays detecting multiple targets reported lower specificity by 2.5% ($0.3 - 4.5\%$, $p = 0.022$), an expected result of detecting antibodies against a less specific antigen. There were no significant differences in assay sensitivity and specificity between isotypes detected and assay detection technology.

4.3.3.3 Interpretation

The third-party head-to-head evaluations included here reported robust reference panels, including positive samples consisting of combinations of high-titer, mid-titer, and low-titer samples on N- and S- antibody targets that reflect more accurately the full-time course of infection [190, 195]. The design of these positive panels mirrors the complexity of antibody detection in real settings. Cross-reactivity to other viral infections was also consistently assessed in negative panels.

We show that manufacturer evaluations systematically overstate assay performance in comparison to independent and third-party evaluations. This corroborates results from previously published interassay comparisons [51, 196–199]. These results mean there may be more false positives and negatives than would be expected given manufacturer-verified test evaluations, which may impact seroprevalence results and interpretation.

The consistency of assay performance across different sources of validation has been difficult to identify in the literature. We show substantial disagreement between different third-party evaluations, particularly with regards to assay sensitivity. This emphasises the importance conducting assay evaluations using a standard international reference panel, and reporting comparable quantitative results from these evaluations in binding antibody units measured against the World Antibody Titre Standard [200].

4.3.4 Impact of misspecification on seroprevalence estimates

Manufacturer overestimation of sensitivity and specificity has the potential to bias seroprevalence estimates. We sought to assess the extent to which this misspessification could introduce such bias.

4.3.4.1 Simulation methods

We performed a simulation analysis. We used three simulated assays: the first with high performance ($\gamma = 95.0\%$, $\delta = 99.0\%$), the second with good performance ($\gamma = 90.0\%$, $\delta = 97.0\%$), and the third with moderate performance ($\gamma = 80.0\%$, $\delta = 87.0\%$).

We then simulated $1,000$ scenarios, where true prevalence SP varied in 0.1% increments from $0.0 - 99.9\%$. In each of these scenarios, we conducted a simulated seroprevalence survey, calculating the apparent seroprevalence $\widehat{SP}$ as follows: [201]:

$$\underbrace{\widehat{SP}}_{\text{apparent prevalence}} = \underbrace{\overbrace{SP}^{\text{true prevalence}} \cdot \overbrace{\gamma}^{\text{sensitivity}}}_{\text{true positives}} + \underbrace{\overbrace{(1-SP)}^{\text{proportion seronegative}} \cdot \overbrace{(1-\delta)}^{\text{false positive rate}}}_{\text{false positives}}$$

Then, we assumed that study investigators attempted to correct this apparent seroprevalence $\widehat{SP}$ for imperfect assay sensitivity and specificity using the Rogan-Gladen estimator [201]. However, we assumed that they attempted this correction based on misspecified estimates of sensitivity $\widehat{\gamma} = \gamma + \epsilon_\gamma$ and specificity $\widehat{\delta} = \delta + \epsilon_\delta$, where ϵ_γ and ϵ_δ measure the amount by which γ and δ are overestimated, respectively. Their correction would then have followed the following equation:

$$SP_{est} = \frac{\widehat{SP} + \widehat{\delta} - 1}{\widehat{\gamma} + \widehat{\delta} - 1}$$

The error in their estimated seroprevalence is then $\Delta_{SP} = SP_{est} - SP$. We used a threshold of $|\Delta_{SP}| > 5\%$ for a considerable bias introduced by this error in a population prevalence estimate [196].

4.3.4.2 Findings

The findings from this simulation are shown in Figure 4.5. In this figure, each facet shows the three different assays we simulated, with high, good, and moderate performance, respectively. The shaded areas show the extent of bias in seroprevalence estimation

Δ_{SP}, with area color indicating the extent of misspecification ϵ_γ and ϵ_δ and vertical lines showing the true seroprevalence values SP for which misspecified seroprevalence estimates lead to bias $|\Delta_{SP}| > 5\%$.

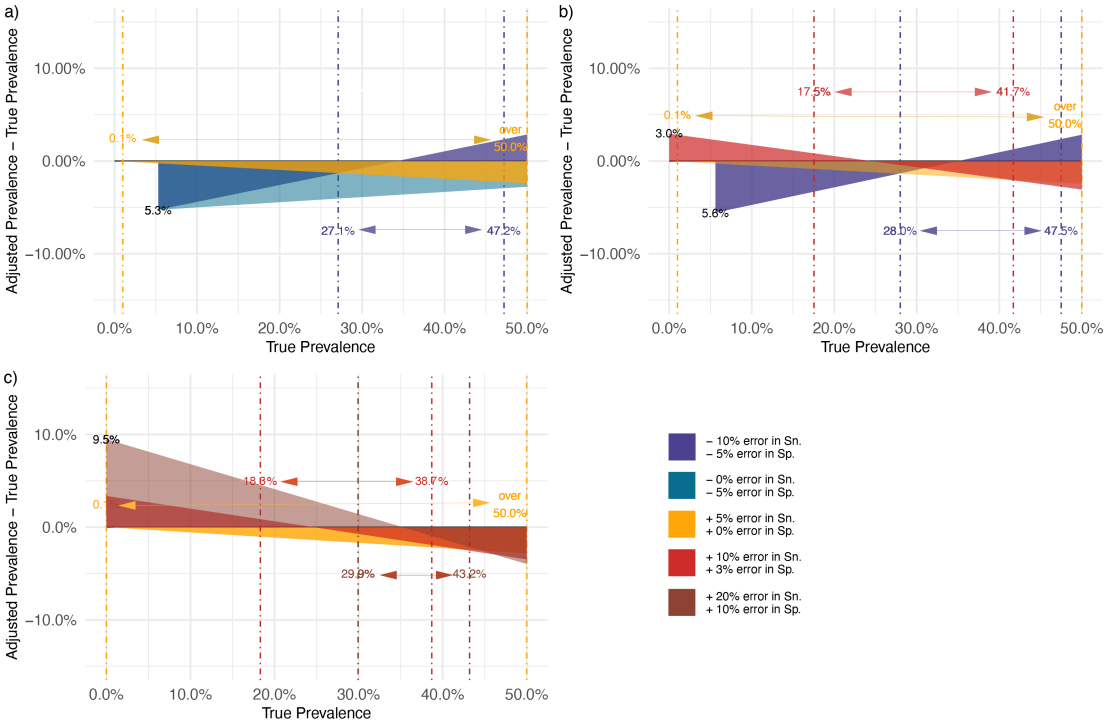


Figure 4.5: Consequences of correcting seroprevalence estimates using biased estimates of sensitivity and specificity. Simulation-based analysis. Serological assays with a) high performance ($\gamma = 95.0\%$, $\delta = 99.0\%$); b) good performance ($\gamma = 90.0\%$, $\delta = 97.0\%$); and c) moderate performance ($\gamma = 80.0\%$, $\delta = 87.0\%$). We plotted Δ_{SP} vs SP for different levels of misspecification ϵ_γ and ϵ_δ, as indicated in colours. The vertical lines indicate the interval within which $|\Delta_{SP}| > 5\%$.

This analysis reveals that misspecification of assay sensitivity and specificity can lead to considerable spectrum bias in seroprevalence estimates, particularly for low or high true SP values and large misspecification of γ and δ. Overestimating γ by 5% does not have a considerable effect on prevalence estimates for any of the simulated assays (Figure 4.5). However, overestimating γ by 10% and δ by 3% leads to considerable bias in SP_{est} for $SP < 18.3\%$ or $SP > 38.7\%$ (assays with moderate performance), or for $SP < 17.5\%$ or $SP > 41.5\%$ (assays with good performance). These findings caution against accepting aggregate sensitivities and specificities reported by assay manufacturers, favouring robust independent or third-party evaluations on representative populations.

4.3.5 Discussion

In examining global serosurveys published since the beginning of the pandemic, we found that 192 unique commercial and 380 unique self-developed serological assays were used. 50 commercial assays were used across 91% of SARS-CoV-2 seropervalence studies. Performance evaluations for the same assay varied widely between different sources. This variation in assay evaluations may have immpacted seroprevalence estimate validity and bias by underestimating or overestimating estimates by up to 9.5%.

Third-party evaluations are essential for more objective estimates of sensitivity and specificity, enabling retrospective adjustment of seroprevalence data and selecting candidate assays for new studies. However, there is substantial disagreement even within third-party evaluations. Moreover, these evaluations typically only target frequently used commercial assays, which were mostly developed and distributed in high-income countries.

This situation necessitates that study investigators validate assays not included in these evaluations. These independent evaluations better reflect the study geography, demographic context, epidemiological timecourse, and variant landscape, minimising spectrum bias. The value of such evaluations is corroborated by several independent findings about variation in serological assay performance: reduced sensitivity several months post-symptom onset as antibodies wane, [188, 202]; viral mutations potentially resulting in decreased assay performance [203]; and varying antibody dynamics that may impact test performance in specific populations including individuals from Sub-Saharan Africa, young adults, and pregnant women [152, 204–206]. However, such independent evaluations can be laborious, expensive, and difficult to conduct. Only a small proportion (6.9%) of author groups conducted their own assay pre-study validation before conducting their serosurvey. Fewer still described the evaluation panels and method they used.

We note some limitations in our analysis. We focus here on seroprevalence studies, but there are other important applications of serological assays. For example, we excluded longitudinal cohort studies of antibody kinetics, as these are done exclusively in confirmed COVID-19 cases and vaccinated individuals. Our findings may not apply to these areas of serological research. Second, our findings apply to population-based contexts and may not translate to the patient or clinical level, if serological assays are used to guide patient

care. Finally, while we made our best effort to identify assays used in seroprevalence studies and the performance of assays from different sources, some data was missing and excluded from analysis. For instance, not all studies cited the assay or gave a comprehensive description of the test used.

In conclusion, we found a large group of assays with highly varying performance was used across SARS-CoV-2 seroprevalence studies. This diversity in serological assays available may impact the interpretation and reliability of seroprevalence estimates, particularly given interassay heterogeneity but also variations in evaluations of the same assay. It is crucial to conduct third-party assay evaluations employing common reference panels to enable robust assay selection, and subsequently to carry out independent field evaluations with population-specific reference panels to minimise spectrum bias in seroprevalence studies.

Seroprevalence estimates available to us are often delayed, biased, and sparse: delayed, in the months it can take for results to be shared after participants are sampled; biased, as a result of measurement error and sampling error alike; and sparse, both in space and time. Correcting for these biases would enable more useful insights to be derived from this data.

5

Bayesian modelling to estimate the true extent of SARS-CoV-2 infection

5.1 Background

Two key sets of questions arise in monitoring the population dynamics of an infectious disease. First, what is the true extent of that disease's spread in the population: how many people are infected, in what areas, and at what times? Second, what is the true extent of immunity against that disease, how is it distributed, and how is it changing? Accurately answering these questions is crucial to inform public health decision-making.

Nowhere has this been more apparent than during the COVID-19 pandemic. Considerable data on confirmed COVID-19 cases, deaths, and SARS-CoV-2 seroprevalence has been generated, and yet the true dynamics of infection and immunity remain unclear. This state of affairs highlights the challenge of integrating heterogeneous surveillance data to understand infection and immunity.

There are two sorts of surveillance data that are particularly useful to inform our understanding of population infectious disease dynamics. The first sort is indicator-based surveillance data, where standardised measures of an infectious disease are monitored and centrally reported [17]. Classic examples include confirmed cases, hospitalisations, and deaths. More recently, wastewater surveillance has been used to measure pathogen

material shed by infected individuals in the water supply [207]. This data provides regular measurements of infectious disease spread.

However, indicator-based surveillance typically underestimates the true extent of disease. For an infected individual to become a confirmed case, they must have access to a diagnostic test and their positive test result must be reported [43]. This is often not the case where access to testing is limited, as with COVID-19 PCR testing in early 2020 [208], or with self-testing without a centralised reporting mechanism, as with COVID-19 lateral flow testing in many high-income countries today [209]. Deaths data is similarly subject to under-reporting [210], as only reported deaths with the infectious disease attributed as the cause of death are included in that indicator.

The second sort of surveillance data applicable here is population-based surveillance data. In contrast to indicator-based surveillance, population-based surveillance aims to directly measure infection and immunity at a population level. Some examples of population-based surveillance include seroprevalence studies, which estimate the proportion of the population with antibodies against an infectious disease, and infection prevalence studies, which estimate the proportion actively infected with a pathogen [16]. These surveillance data can provide more representative estimates of infection and immunity. However, they can be expensive to conduct, and data from these studies is often sparse, infrequent, irregular, and subject to reporting delays [9, 211].

Indicator- and population-based surveillance are complementary measures: the former is regular but biased, while the latter is irregular but representative. Integrating these two sources of data is therefore crucial to monitor the true extent of infection and immunity.

Several groups have used these data sources to develop models of true infections for COVID-19, with three notable such efforts outlined in Table 5.1. These models are important contributions, but each has considerable limitations. Some use reported death data, but do not pair this data with population-based surveillance data to account for undercounting the number of deaths [212]. Others were designed for the pre-2021 pandemic, and are not suited to the modern environment of vaccines and variants [213]. While the Institute of Health Metrics and Evaluation model comes closest to addressing these limitations, even it does not adequately treat measurement error or bias in the

SeroTracker-sourced serosurvey data it relies on, does not propagate uncertainty, or indeed, does not fully make clear its methods [214].

We can exploit the complementary advantages of indicator- and population-based surveillance data to model infection and immunity and to produce up-to-date estimates relevant for decision-making. We propose to develop a framework that synthesises indicator- and population-based surveillance data to understand infection and immunity, for COVID-19 and future pandemics alike. Learning from the strengths and limitations of others' efforts, our model:

1. Integrates both indicator- and population-based surveillance data, recognising their complementary nature, and is adaptable to different sorts of such data

2. Builds on SeroTracker's best-in-class seroprevalence data, as our most comprehensive window into true infections and immunity today, and properly reckons for its nuances, such as heterogeneous data availability, different antibodies measured, seroreversion, and sampling biases

3. Explicitly models the relationship between indicator-based surveillance and infections, and allows this relationship to vary based on its determinants (e.g., region, testing capacity, and variants)

4. Comprehensively accounts for sources of uncertainty in a jointly estimated unified and transparent multivariate parametric Bayesian model

In this chapter, we motivate and develop the model architecture, demonstrate its ability to fit simulated data, and apply it to reconstruct the true dynamics of COVID-19 infection and immunity using case and seroprevalence data in Canada.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Arora RK, Chen J, ..., Clifton DA, Buckeridge DL. 2022. Modelling SARS-CoV-2 infections using indicator- and population-based surveillance data. In preparation. *Rahul's contributions:* conceptualization of manuscript; design of modelling approach; conducting simulation and real-world data analysis; writing original draft; revising draft.

Table 5.1: Summary of three models of true SARS-CoV-2 infections.

Model	Structure	Data Used	Description	Limitations
O'Driscoll et al. [213]	Statistical	Deaths and seroprevalence	Uses seroprevalence to estimate cumulative infections and in turn the age-specific infection fatality ratio. Infers true infections from age-specific infection fatality ratio and age-stratified deaths	Ignores seroreversion, assuming that seroprevalence equals cumulative incidence of infection. Similarly, assumes a constant infection fatality ratio over time; however, bot variants and vaccines lead to time-varying changes in the infection fatality ratio today. Data from under 50 serosurveys
Imperial College London [212]	Stochastic SEIR	Deaths	Applies age-specific infection fatality ratios to reported death data to infer extent of infection	Reported death data are also subject to considerable undercounting, especially in low- and middle-income countries. No attempt to estimate undercounting by triangulating between different sources of data. Specifies, and does not learn, infection fatality ratio
Institute of Health Metrics and Evaluation [214]	Multi-step statistical	Cases, deaths, and hospitalisations. Seroprevalence from SeroTracker	Corrects seroprevalence data for vaccination and seroreversion to estimate cumulative incidence of infections. Uses cumulative infection data to estimate the infection fatality ratio, infection hospitalisation ratio, and case ascertainment ratio. Applies these ratios to smoothed case, death, and hospitalisation data time series to estimate infections	No reckoning for serosurvey measurement error or risk of bias. Does not use age-structured death data; assumes death age distribution from population age distribution and infection fatality ratio. Multistep approach does not propagate uncertainty. Black box model: extreme complexity, opaque communication, and lack of code availability make it challenging to understand what factors drive model estimates

5.2 Objective and workflow

The objective of this model is to infer the true, unobserved number of SARS-CoV-2 infections; directly interpret these estimates, both in terms of what they reveal about the burden of infection, the proportion of infections that were not detected, and how these quantities varied over time and between regions; and use these estimates to complete our understanding of SARS-CoV-2 seroprevalence, at times and in places where seroprevalence data is not available.

To establish this model, provide evidence of its validity, and interpret it, we follow a model development workflow inspired by the Bayesian workflow outlined in Gelman et al., Grinsztajn et al., and elsewhere [215–217]:

1. Model development. Establish a generative model with a structure and priors matching our domain knowledge.

2. Model simulation. Validate the model's realism and ability to fit simulated data from the same generative process.

    (a) Generate simulated data. Simulate latent infections from this model. Simulate the observed data that we would expect, given simulated infections and the specified generative model.

    (b) Prior predictive checks. Evaluate whether the simulations match our domain knowledge about this infectious disease.

    (c) Simulated data recovery. Given simulated observed data, evaluate whether the model can infer the true infections process and parameters.

3. Application to real-world data. Fit the model to real-world data from over two years of the pandemic.

    (a) Evaluate model diagnostics. Assess Hamiltonian Monte Carlo fit diagnostics to ensure that the model is accurately fitting the posterior.

    (b) Posterior predictive checks. To evaluate whether the model generative process captures the generative process underlying real-world data, compare the posterior predictive distribution from the model to real-world data.

(c) Apply the model and interpret its parameters. Use the model to infer the infections process underlying the real-world Canadian data and estimate seroprevalence over time. Interpret these estimates to understand the implied dynamics of SARS-CoV-2 infection and immunity

5.3 Bayesian inference and numerical Bayesian computation

Here, we provide an overview of Bayesian inference and numerical methods for Bayesian computation, building towards Hamiltonian Monte Carlo methods for sample-efficient high-dimensional Bayesian inference.

In a typical Bayesian inference problem, we have parameters θ, observed values x, a model for the likelihood of the data conditioned on parameter values $P(x \mid \theta)$, and the prior probability on model parameters $P(\theta)$. Given these values, we are interested in estimating the posterior probability of the parameters given the data $P(\theta \mid x)$ [218]. Bayes' theorem gives:

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)}$$

Bayesian inference generally aims to estimate the distribution of the posterior $P(\theta \mid x)$, enabling full analysis of the posterior distribution, a reckoning of the uncertainty in posterior estimates of parameter values, and the use of this distribution in subsequent estimation [218].

The Bayesian full probability modelling approach contrasts with others that aim to provide optimal point estimates $\hat{\theta}$ for the parameters θ – perhaps those point estimates that satisfy $\hat{\theta} = \text{argmax}_\theta P(x \mid \theta)$ (as in maximum likelihood estimation), that maximise the posterior $\hat{\theta} = \text{argmax}_\theta P(x \mid \theta)P(\theta)$ (as in maximum a posteriori estimation), or that minimise a specified cost function (as with most approaches commonly described as machine learning).

5.3.1 Numerical methods for Bayesian inference

In Bayes' theorem, the evidence $P(x)$ provides a normalising constant for the posterior, ensuring that $\int_\theta P(\theta \mid x)\mathrm{d}\theta = 1$ [218].

However, the evidence $P(x)$ is not known and is difficult to estimate. One way to estimate $P(x)$ is through marginalisation over θ: $P(x) = \int_\theta P(x \mid \theta)P(\theta)\mathrm{d}\theta$. However, the marginalisation integral can rarely be solved analytically and is too computationally expensive to solve numerically [218].

Instead, we often approximate the posterior by noting that $P(\theta \mid x) \propto P(x \mid \theta)P(\theta)$ [218]. Considering that we know $P(\theta \mid x)$ up to a proportionality factor, we can estimate it using:

- Monte Carlo sampling. Estimate $P(\theta \mid x)$ by repeatedly sampling from the likelihood times the prior $P(x \mid \theta)P(\theta)$ [219]. Discussed in detail in Section 5.3.2.

  Monte Carlo-based sampling strategies are guaranteed to converge to the target distribution with sufficient iterations, but are often computationally expensive [219].

- Variational inference. Estimate $P(\theta \mid x)$ by defining a variational family of parameterised distributions $\mathcal{F}_\Gamma = \{f_\gamma \mid \gamma \in \Gamma\}$, and finding the distributional parameters γ such that $f(x)$ best approximates $P(x \mid \theta)P(\theta)$ [220]. Variational inference often involves minimising the reverse Kullback-Liebler divergence: $\hat{\gamma} = \mathrm{argmin}_{\gamma\in\Gamma} KL(f_\gamma(x) \mid\mid P(x \mid \theta)P(\theta))$.

  Variational inference is well-suited for efficient inference, but is prone to bias – particularly if the variational family used does not resemble the posterior [220].

In choosing between these strategies for the work in this chapter, we note the complexity of the models that we will develop, and the importance of accurate estimation. In such a case, the accuracy of Monte Carlo-based approaches is crucial.

5.3.2 Monte Carlo Bayesian inference

5.3.2.1 Monte Carlo

Monte Carlo methods use repeated sampling to approximate integrals. Take the integral $\int f(x)P(x)\mathrm{d}x$, where $P(x)$ is the probability distribution function for x, and take $x_1...x_S$ to be samples drawn from $P(x)$ such that $x_s \sim P(x)$. Then, the Monte Carlo approximation of this integral is as follows [219, 221]:

$$\int f(x)P(x)\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

This approximation is unbiased, and has variance that shrinks with $\frac{1}{S}$ [219, 221].

To use this method, we need to generate S samples $x_1...x_S$ such that $x_s \sim P(x)$, but where we only have $\tilde{P}(x) \propto P(x)$. Some strategies to generate these samples yield identical and independently distributed samples, such as rejection and importance sampling, while others yield autocorrelated samples, such as Markov Chain Monte Carlo methods [219].

In rejection sampling, we sample from $P(x)$ by identifying a proposal distribution $Q(x)$ which has similar shape to $P(x)$, and scaling it by k such that $\forall X : kQ(X) \geq \tilde{P}(x)$ [218].

Then, we can take S samples from $P(x)$ as follows, for $i \in \{1...S\}$ [218]:

1. Generate a sample. Sample $x_i \sim Q(x)$
2. Decide whether to keep or reject the sample. Take $h \sim \text{Uniform}[0, kQ(x_i)]$.
   (a) Is $h \leq \tilde{P(x_i)}$? If so, keep x_i, which is from $P(x)$
   (b) If not, reject x_i as it is not from $P(x)$; resample x_i

   This rejection step ensures that we do not oversample areas where $kQ(x) > \tilde{P}(x)$.

Note that if $Q(x)$ is a poor approximation for $\tilde{P}(x)$, k will be large, and the majority of samples will be rejected. Identifying an appropriate $Q(x)$ is especially challenging for complex, high-dimensional distributions [222].

Rejection sampling frequently rejects proposed samples. Importance sampling can emerge as an attempt to address this inefficiency: instead of discarding samples from

$Q(x)$ that are not in $P(x)$, can we reweight all samples from $Q(x)$ by their relative importance in $P(x)$?

We can rewrite the target integrand, $\int f(x)P(x)\mathrm{d}x$, as $\int f(x)\frac{P(x)}{Q(x)}Q(x)\mathrm{d}x$. We approximate this integral using the Monte Carlo estimator, but this time over $Q(x)$ [218, 221]:

$$\begin{aligned}\int f(x)P(x)\mathrm{d}x &= \int f(x)\frac{P(x)}{Q(x)}Q(x)\mathrm{d}x && \text{multiplying and dividing by } Q(x)\\ &\approx \sum_{s=1}^{S} f(x_s)\frac{P(x_s)}{Q(x_s)},\ x_s \sim Q(x) && \text{Monte Carlo integration over } Q(x)\end{aligned}$$

Recall that we often only know $P(x)$ up to a constant (i.e., $\tilde{P}(x)$). We can modify our importance sampling algorithm to account for this. For each sample s, we can define a different weight for that sample based on $\tilde{P}(x)$: weight each sample by $W_s = \frac{\tilde{P}(x_s)}{Q(x_s)}$. Then, we divide by the sum of our weights to normalise. With this, our importance sampling equation becomes:

$$\int f(x)P(x)\mathrm{d}x \approx \frac{\sum_{s=1}^{S} W_s f(x_s)}{\sum_{s=1}^{S} W_s},\ x_s \sim Q(x) \qquad \text{weighting and normalisation}$$

For sample s, if $Q(x_s)$ is small for $P(\tilde{x}_s)$ large, then the denominator summand $W_s = \tilde{P}(x_s)Q(x_s)^{-1}$ will be large, resulting in high-variance estimates for $\int f(x)P(x)\mathrm{d}x$ [218, 221]. This is common for heavy-tailed or complex $P(x)$ that are poorly approximated by light-tailed or simple $Q(x)$.

5.3.3 Markov Chain Monte Carlo

The models we will consider in this dissertation have complex posteriors and often exist in high-dimensional parameter spaces, making rejection sampling inefficient and importance sampling high variance [222]. This is in part because rejection and importance sampling generate independent and identically distributed samples, meaning that subsequent samples ignore information embedded in previous samples about the posterior.

In contrast, Markov Chain Monte Carlo techniques generate chains of autocorrelated samples that efficiently find and explore the posterior [219]. These techniques Markov chain with the posterior as its stationary distribution, and yields autocorrelated samples $x_1 ... x_S$ from that Markov chain to be used in Monte Carlo estimation.

For a Markov chain to have the posterior as its stationary distribution, it must have a continuous state space. Such Markov chains have transition probability densities that define state transitions [219]. $T(x_{n-1}, x_n)$ gives a probability density for x_n given x_{n-1}, with $\int T(x_{n-1}, x_n)\,\mathrm{d}x_n = 1$. Moreover, if $x_{n-1} \sim D(x_{n-1})$, the state at x_n is $\int D(x_{n-1}) \cdot T(x_{n-1}, x_n)\,\mathrm{d}x_n - 1$. These Markov chains also satisfy the Markov property, where the probability of a state depends only on the previous state – i.e., $x_n \perp\!\!\!\perp x_{n-2}, ..., x_1 \mid x_{n-1}$ [219].

A Markov chain with transition density $T(x_{n-1}, x_n)$ may have a unique stationary distribution S, which remains the same when T is applied: $S(x_n) = \int S(x_{n-1})T(x_{n-1}, x_n)\mathrm{d}x_{n-1}$. Moreover, if the Markov chain is reversible with respect to a distribution D (that is, $D(x_{n-1}) \cdot T(x_{n-1}, x_n) = D(x_n) \cdot T(x_n, x_{n-1})$), then D is the stationary distribution for that Markov chain [219]. If we could construct a Markov chain with the posterior of interest as its stationary distribution, we could then efficiently generate autocorrelated samples from that posterior.

The Metropolis algorithm constructs a Markov Chain with stationary distribution equivalent to an arbitrary target distribution $P(x)$ from a proportional distribution $\tilde{P}(x)$ [219]. Starting with a sample x_n, we draw a proposed value for our next step χ_{n+1} from a proposal function $Q(x_n, \chi_{n+1})$ [219]. For now, we will set $Q(x_n, \chi) \sim \text{Normal}(x_n, \Sigma)$.

If this proposal is more probable than our current location ($\tilde{P}(\chi_{n+1}) \geq \tilde{P}(x_n)$), we move to it for our next step - that is, we set $x_{n+1} = \chi$. If the proposal is less probable ($\tilde{P}(\chi_{n+1}) < \tilde{P}(x_n)$), we move to the proposal with probability $\frac{\tilde{P}(\chi_{n+1})}{\tilde{P}(x_n)}$, and remain in the same location otherwise by setting $x_{n+1} = x_n$ [219]. Note that $\frac{\tilde{P}(\chi_{n+1})}{\tilde{P}(x)} = \frac{P(\chi_{n+1})}{P(x)}$ – the unknown constant of proportionality cancels.

The resulting transition probability is $T(x_{n+1}, x_n) = \text{Normal}(x_n, \Sigma)\min(1, \frac{P(\chi_{n+1})}{P(x)})$. This transition probability distribution corresponds to a Markov chain that is reversible with respect to $P(x)$ and hence has $P(x)$ as its stationary distribution. The full Metropolis-Hastings algorithm modifies the transition probability to remain reversible for asymmetric Q, where $Q(x_{n+1}, x) \neq Q(x, x_{n+1})$.

Classical Monte Carlo approaches draw S independent and identically distributed samples, giving an unbiased estimator with variance that shrinks with $\frac{1}{S}$ [219]. However,

Markov chain Monte Carlo yields chains of autocorrelated samples, and each autocorrelated sample may not contain as much information as an independent sample. For this reason, we define the effective sample size N_{eff} as the number of equivalent independent samples, and the variance of autocorrelated estimates shrinks with $\frac{1}{N_{eff}}$ [219]. A chain of samples with lower autocorrelation contains more information and hence has higher N_{eff}.

5.3.4 Hamiltonian Monte Carlo

The probability mass contained within a subset of a probability distribution depends not only on the probability density of that set but also on the volume of that set. In exploring the posteriors of probability distributions, we are interested not in the areas of highest probability density, but those of highest probability mass [222].

By definition, the mode of a probability distribution has the highest probability density. However, that mode has little volume – the ball surrounding it has a volume that decreases relative to the rest of the space as the dimensionality increases, meaning that in high dimensions very little probability mass is concentrated at the mode. Rather, in these spaces, the typical set of the distribution (that is, the subset with highest probability mass) is concentrated in regions with moderate probability density and moderate volume – often in a thin shell that surrounds, but is away from, the mode of the distribution [222].

In the Metropolis family of algorithms, we generate autocorrelated samples from the posterior. However, it is particularly challenging to construct Metropolis proposals that efficiently explore high-dimensional distributions. Let's imagine that we use a normally distributed Metropolis proposal. If the proposal distribution has high variance, the proposal is likely to move away from the thin shell of the typical set and will most often be rejected. Proposal distributions with low variance may stay on this typical set, but the samples from these chains will have high autocorrelation and low N_{eff} [222]. Efficiently sampling from high-dimensional distributions requires an algorithm that actively exploits the structure of their typical sets.

Hamiltonian Monte Carlo efficiently samples from high-dimensional distributions [222–224]. To motivate our choice of this algorithm, we will start by considering the surface given by $-\log \tilde{P}(\theta \mid x)$, and place a particle on this surface. We would like this particle

to move around the surface so that sampling from its trajectory yields samples from the target distribution.

If we have this particle follow a gravitational force and move against the gradient of this surface, we get gradient descent. In high dimensions, this will find the mode of the posterior, but rapidly move away from the typical set which is away from the mode [222].

Instead, we give this particle a metaphorical kick around the surface by giving it momentum. If this particle enters areas where the probability density is too low, it is moving uphill, and will ultimately have to turn around to return to areas of higher density. If this particle enters the mode of the distribution, it would have substantial kinetic energy, and would rapidly return to areas of higher probability mass. This particle's behavior actively exploits the structure of the posterior to explore the typical set.

Formalising this idea, we take the parameters θ to represent the position of our particle on this surface, and ϖ to represent its momentum. We write the joint density $P(\theta, \varpi \mid x)$ as $P(\varpi \mid \theta)\tilde{P}(\theta \mid x)$ [222, 223, 225]. The Hamiltonian for this system is then:

$$\begin{aligned} H(\theta, \varpi) &= -\log P(\theta, \varpi \mid x) \\ &= \underbrace{-\log P(\varpi \mid \theta)}_{\text{kinetic energy } T(\varpi,\theta)} + \underbrace{-\log \tilde{P}(\theta \mid x)}_{\text{potential energy } V(\theta)} \end{aligned}$$

This gives a Hamiltonian system with the dynamics described above: where the potential energy $V(\theta)$ varies as $-\log \tilde{P}(\theta \mid x))$, increasing as probability density decreases, and where the kinetic energy $T(\varpi, \theta)$ depends on the momentum ϖ assigned to the particle.

The equations of motion in this Hamiltonian system are:

$$\begin{aligned} \frac{\mathrm{d}\theta}{\mathrm{d}t} &= \frac{\partial H}{\partial \varpi} & \frac{\mathrm{d}\varpi}{\mathrm{d}t} &= -\frac{\partial H}{\partial \theta} \\ &= \frac{\partial}{\partial \varpi}(T(\varpi, \theta) + V(\theta)) & &= -\frac{\partial}{\partial \theta}(T(\varpi, \theta) + V(\theta)) \\ &= \frac{\partial T}{\partial \varpi} & &= -\frac{\partial T}{\partial \theta} - \frac{\partial V}{\partial \theta} \end{aligned}$$

These equations are reversible in time, giving a continuous stochastic process in θ, ϖ that is reversible with respect to $P(x)$ and thus has the desired stationary distribution [222, 223, 225].

In Hamiltonian Monte Carlo, we pick a distribution for the kinetic energy $P(\varpi \mid \theta)$. Then, for each sample, we take the starting position of parameters θ from the previous

sample, endow them with a momentum $\varpi \sim P(\varpi \mid \theta)$, and solve Hamilton's equations to yield a trajectory across θ, ϖ. Sampling θ from this trajectory yields samples from the posterior [222, 223, 225].

The probabilistic programming language `Stan` provides a syntax to specify a statistical model and to sample from it using an efficient implementation of Hamiltonian Monte Carlo, based on the No U-Turn Sampler [225, 226]. Stan's implementation of this algorithm solves Hamilton's equations using the leapfrog integrator, a numerical integrator which prevents numerical errors from accumulating over time. Stan's implementation proceeds in two phases: a warmup phase, in which the kinetic energy distribution $P(\varpi \mid \theta)$ is selected and the leapfrog integrator's parameters are tuned; and a sampling phase, which yields samples from the target [226].

Hamiltonian Monte Carlo is well-suited for Bayesian models with high-dimensional and complex posteriors. We use Hamiltonian Monte Carlo as implemented in `Stan` to model SARS-CoV-2 infection by integrating indicator- and population-based surveillance data.

5.4 Model architecture

The model we present here has three components. The first component is a latent process for infections with an autocorrelated structure, which encodes our knowledge about the growth, decay, and autocorrelation structure of infections. The ultimate goal of our model is to infer this infections process using two different sorts of imperfect measures. These measures comprise the second and third components: indicator-based surveillance measures (e.g., reported cases, hospitalisations, deaths, and wastewater), which are regular but delayed and biased measures of the infections process; and population-based surveillance measures (e.g., seroprevalence, point prevalence), which have complementary strengths in providing more representative but much less frequent measures of this process. The overall model architecture is shown in Figure 5.1.

Throughout this model, we use g to indicate geographical region (in the Canadian data, province or territory), N_{pop_g} to indicate the population of that region, t to indicate time for point measures, and T to indicate time for cumulative measures (Table 5.2).

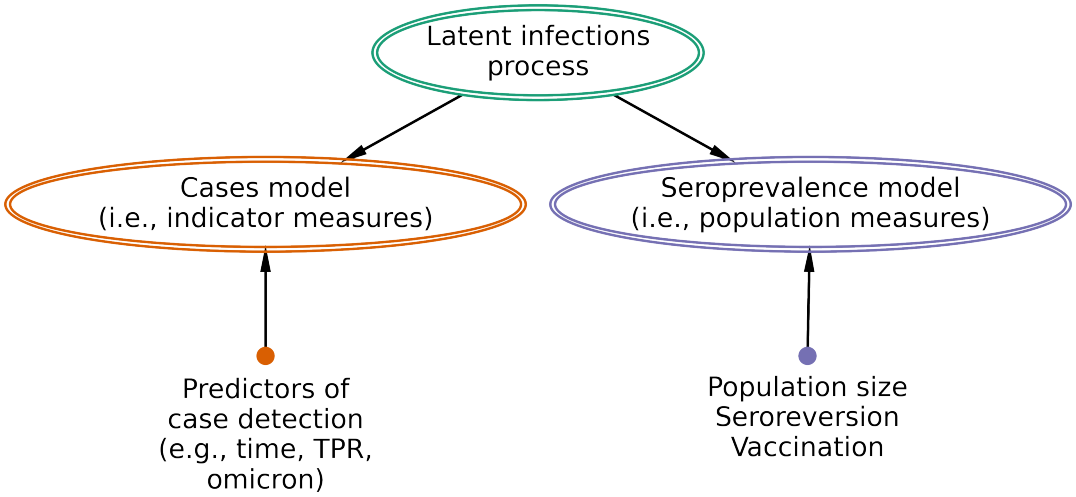


Figure 5.1: Overall model architecture. Double ellipses indicate sub-models, which may include parameters only (as in the infections process) or parameters as well as likelihoods (as in the case and seroprevalence sub-models). Dots indicate the data used to inform each of these sub-models. Edges indicate conditional dependencies between sub-model components.

5.4.1 Infections process

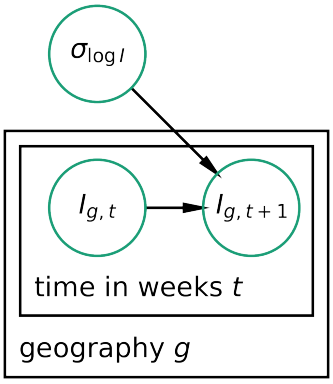
The true number of infections I is unobserved and key to this model. We know that infections in a given week result from infections in previous weeks, and that infectious diseases are characterised by exponential growth and decay. To reproduce these observations, we place a normal random walk prior on $\log I$, implying a multiplicative autocorrelation structure on infections I that allows for exponential growth. This approach shares similarities with other models for positive time series, such as log-linear Poisson autoregression [227].

The standard deviation of this random walk $\sigma_{\log I}$ reflects how much we expect I to change per unit time. For example, $\sigma_{\log I} = 0.5$ places 95% of our prior belief that $\log(I_{g,t+1})$ is within two standard deviations of $I_{g,t}$, for geographical region g and time t. In turn, this implies that I_{t+1} will change by no more than a factor of $e^{2 \cdot \sigma_{\log I}} \approx 2.71$, 95% of the time.

We can inform our prior for this parameter using other knowledge about the infectious disease being modelled. For SARS-CoV-2, the basic reproduction number R_0 (i.e., the number of transmissions that result from a single infection) ranges from 2.79 for the ancestral strain to 8.2 for the Omicron variant [33, 228]. Estimates of the generation

time T_g (i.e., the time between successive infections) vary from from 4.7 to 5.5 days [229]. This suggests a maximum growth rate by a factor of $\frac{R_0}{T_g} = \frac{2.79 \text{ to } 8.2}{\text{generation}} \cdot \frac{1 \text{ generation}}{5 \text{ days}} \cdot 7 \text{ days} =$ 3.9 to 11.48 each week. Considering that non-pharmaceutical interventions and immunity constrain this growth rate, our prior for $\sigma_{\log I}$ is Normal$(0.7, 0.2)$, which implies that the number of infections will change by a factor of at most $e^{2 \cdot 0.7} \approx 4$, 95% of the time.

A diagram of the infections model is given in Figure 5.2, with parameters as in Table 5.2. The diagrams of model architecture in this chapter follow standard notation for Bayesian graphical models, representing this model as a directed acyclic graph [230]. Nodes in this graph represent data with likelihoods (shaded circles), unobserved parameters (unshaded circles), and fixed parameters or covariates without likelihoods (dots). The edges between nodes represent conditional dependencies between parameters, where the node at the tip of the edge depends on the node at the tail. Finally, the boxes in the model diagram are termed plates [230]. Each plate represents a certain group, and each variable within the plate is repeated for each level of the group. For example, the variable $I_{g,t}$ in Figure 5.2 is repeated $N_t \cdot N_g$ times, for each combination of $g \in N_g$ and $t \in N_t$.



$$\log I_{g,t+1} \sim \text{Normal}(\log I_{g,t}, \sigma_{\log I}) \quad \text{random walk}$$
$$\sigma_{\log I} \sim \text{Normal}(0.7, 0.2) \quad \text{prior; from } R_0 \text{ and } T_g$$

Figure 5.2: Infections model diagram and specification. White circles indicate unobserved parameters. Edges indicate conditional dependencies between model components.

Table 5.2: Model indices and variables used in the infections model.

Variable	Values	Definition
INDICES		
t	$1, \ldots, N_t$	Time since the start of the period in question, in weeks; used for point measures
T	$1, \ldots, N_t$	Time since the start of the period in question, in weeks; used for cumulative measures
g	$1, \ldots, N_g$	Geographic region (in Canada, province or territory)
DATA		
N_{pop_g}	$\mathbb{Z}^+$	Population of region g
PARAMETERS		
$I_{g,t}$	$\mathbb{R}^+$	Number of new infections in region g and time point t
$\sigma_{\log I}$	$\mathbb{R}^+$	Standard deviation for random walk prior on the distribution of infections over time

5.4.2 Indicator-based surveillance measures: reported cases

The infections process is not directly observed. This makes indicator-based surveillance measures (e.g., reported cases, hospitalisations, and deaths) particularly important to understand the infections process. However, there are several considerations in relating these to the infections process. First, they are delayed realisations of that process. Second, they are related to infections by a scaling factor (e.g., the probability that an infection is detected, the infection fatality rate), and these scaling factors vary based on various determinants, including regions and time. Here, we present a specific model for the number of reported cases $C_{g,t}$, with ideas generalisable to other population-based surveillance measures. The data and parameters for this model can be found in Table 5.3.

In this model, cases C are always lower than infections because not every infected individual has access to diagnostic testing. To capture this, we take cases to be the product of infections I with a parameter $\rho \in (0, 1)$ that represents the probability that an infection will be detected as a case. We recognise that cases will lag infections, because it takes time for symptoms to manifest, for diagnostic testing, and for positive results to be reported. Finally, we use a negative binomial likelihood for cases. Our choice of a negative binomial likelihood reflects the fact that observed cases exhibit overdispersion relative to a Poisson likelihood – that is, their distribution reflects more

Table 5.3: Variables in model for cases and indicator-based surveillance.

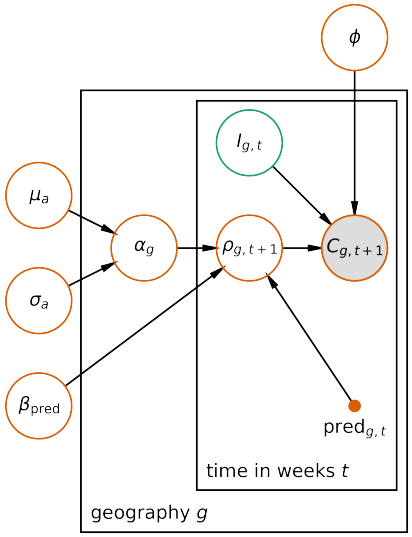
Variable	Values	Definition
DATA		
$C_{g,t}$	$\mathbb{Z}^+$	Observed cases in region g at time t
$T_{g,t}$	$\mathbb{Z}^+$	Diagnostic tests done in region g at time t
o_t	$\{0, 1\}$	Whether the Omicron variant was dominant at time t
Δ_{IC}	$\mathbb{R}^+$	Infection case delay
TRANSFORMED DATA		
$TPR_{g,t}$	$[0, 1]$	Test positivity rate: $C_{g,t}/T_{g,t}$
$\widetilde{TPR}_{g,t}$	$\mathbb{R}$	Rescaled and centered TPR: $\text{logit}(TPR_{g,t}) - \text{logit}(\text{mean}(TPR))$
PARAMETERS		
ϕ	$\mathbb{R}^+$	Overdispersion in observed cases
$\widetilde{\alpha_g}$	$\mathbb{R}$	Latent normal for $\alpha_{\rho,g}$ in non-centered parameterisation
μ_α	$\mathbb{R}$	Grand mean for region-specific intercept for logit ρ
σ_α	$\mathbb{R}^+$	Standard deviation for region-specific intercept for logit ρ
β_t	$\mathbb{R}$	Coefficient for time term in model for logit ρ
β_{TPR}	$\mathbb{R}^-$	Coefficient for $\widetilde{TPR}$ term in model for logit ρ
β_o	$\mathbb{R}$	Coefficient for Omicron variant term o in model for logit ρ
TRANSFORMED PARAMETERS		
α_g	$\mathbb{R}$	Region-specific intercept for logit ρ
$\rho_{g,t}$	$(0, 1)$	Proportion of infections detected as cases in region g at time t

variance than would be expected given a Poisson likelihood with a fixed rate parameter. This distribution is equivalent to a Poisson likelihood where the rate parameter is not fixed but rather follows a gamma distribution, therefore introducing additional variance and allowing for this overdispersion.

The probability of case detection ρ varies across time and regions. We introduce a logit-linear model for $\rho_{g,t}$ which allows both time-invarying and time-varying predictors of ρ. Recognising the differences in testing policy and public health response between jurisdictions, we include geography as a time-invarying predictor. We introduce a hierarchical geography-specific intercept α_g, allowing variation with partial pooling across geographies, and use a non-centered parameterisation for this hierarchical model to ease model fit.

We also include three time-varying predictors. First, variation by time, recognising

that diagnostic testing capacity was very low at the start of the pandemic and then rapidly increased. We introduce a time predictor as $\log(t+k)$, yielding a monotonically increasing relationship in t. The offset k accounts for time since the start of the pandemic, and is 10 weeks here considering that we include data starting from 17 February 2020 while the first SARS-CoV-2 infections were likely in early December 2019 [231]. Second, variation by the burden on the testing system. We use the test positivity rate $TPR_{g,t} = \frac{C_{g,t}}{T_{g,t}}$ as a predictor indicating greater burden, which implies reduced access to testing or likelihood of increasing testing [232]. Third, the dominance of the Omicron variant o_t, which in Canada caused a paradigm shift in testing policy towards at-home lateral flow tests, which are not centrally reported [209, 233]. The resulting model is shown in Figure 5.3, with parameters in Table 5.3.



$$
\begin{aligned}
C_{g,t+\Delta_{IC}} &\sim \text{NB}(I_{g,t} \cdot \rho_{g,t}, \phi) && \text{likelihood} \\
\phi &\sim \text{Gamma}\left(2, \frac{1}{4}\right) && \text{limited overdispersion} \\
\Delta_{IC} &= 1 \text{ week} && \text{infection-case delay} \\
\\
\text{logit}(\rho_{g,t}) &= \alpha_g && \text{logit-linear model} \\
&\quad + \beta_t \cdot \log(t+10) \\
&\quad + \beta_{TPR} \cdot \widetilde{TPR_{g,t}} \\
&\quad + \beta_o \cdot o_t \\
\\
\mu_a &\sim \text{Normal}(0, 3) && \text{hierarchical model for } \alpha_g \\
\sigma_a &\sim \text{Cauchy}^{+}(0, 3) && \text{weakly regularising prior} \\
\widetilde{\alpha_g} &\sim \text{Normal}(0, 1) && \text{latent normal for NCP} \\
\alpha_g &= \mu_\alpha + \widetilde{\alpha_g} \cdot \sigma_\alpha && \text{reconstruct } \alpha_g \text{ in NCP} \\
\\
\beta_t &\sim \text{Normal}(0, 0.2) && \text{constrained scale} \\
\beta_{TPR} &\sim \text{Normal}^{-}(0, 2) && \text{inverse relationship} \\
\beta_o &\sim \text{Normal}(0, 2) && \text{uninformative prior}
\end{aligned}
$$

Figure 5.3: Model diagram for cases and indicator-based surveillance. NCP: non-centered parameterisation. White circles indicate unobserved parameters. Gray circles indicate observed quantities. Dots indicate input data used to inform models and treated as fixed. Edges indicate conditional dependencies between model components.

5.4.3 Population-based surveillance measures: seroprevalence

Indicator-based surveillance is not sufficient alone to infer the infections process. Without additional information, it is very difficult to infer the scaling factors such as ρ that relate indicator-based surveillance and infections.

This highlights the importance of population-based surveillance – efforts which systematically sample the population as a whole, including population-based testing for antibodies (i.e., seroprevalence studies) and for active infection (i.e., infection prevalence studies). These studies are expensive to conduct and often infrequently available; nevertheless, the data that is available can be very helpful in pinning down the relationship between infections and indicator-based measures [9].

Here, we focus on seroprevalence data, considering their global and Canadian availability [234, 235], and the paucity of point prevalence survey data [236, 237]. Seroprevalence studies take a sample of the population and test them for antibodies, thereby estimating the fraction of individuals who presently have antibodies after previous infection or vaccination. These studies can measure antibodies that are either attributable to infection alone, or to both infection or vaccination, depending on the type of antibodies they measure and the vaccines in use. We outline our approach to each sort of seroprevalence measure below, with a full model diagram in Figure 5.4 and all variables in Table 5.4.

5.4.3.1 Mean seroprevalence attributable to infection

We first focus on modelling the true seroprevalence in a population attributable to infection, μ_{SPI}. To do so, we first note that seroprevalence is a cumulative measure – that is, a seroprevalence study will detect any individual with still-measurable antibodies. We index cumulative measures in time with T to distinguish them from the point measures indexed with t (Table 5.2). Then, we are interested in $N_{SPI_{g,T}}$ – the number of previous infections still seropositive at time T.

To model this quantity, we need to consider that antibody levels wane over time, and not all previously infected individuals will have antibodies [238–240]. We quantify this seroreversion with the parameter γ, where individuals infected on day 0 have a probability $\gamma_k \in [0, 1]$ of still being seropositive on day k. With this, we have $N_{SPI_{g,T}} =$

$\sum_{t=1}^{T-\Delta_{IS}} I_t \gamma_{T-t}$, where Δ_{IS} accounts for the delay between infection and detectable antibodies.

In the model here, we treat γ as fixed, using values drawn from an analysis of seroreversion in Canadian blood donors, and two studies on seroreversion in the assays used in a Canadian data [188, 241]. This analysis yielded a five-parameter logistic decay curve model, with parameters $\{a, b, c, d, g\}$: $\gamma_k = a + \frac{d-a}{\left[1+\left(\frac{k}{c}\right)^b\right]^g}$ [242]. We choose parameters to imply that about $d = 97\%$ develop antibodies postinfection, dropping to a nadir of $a = 25\%$ and with about half of seroreversion happening by $c = 48$; $b = 3.5$ and $g = 2$ are shape parameters.

While $N_{SPI_{g,T}}$ denotes the number of still-seropositive previous infections, we are interested in the proportion $\mu_{SPI_{g,T}}$ of individuals who are seropositive. In other words, if M infections occur in N individuals, what proportion p of unique individuals have been infected? For $M \ll N$, we expect most infections to occur in different individuals – i.e., $p \approx \frac{M}{N}$. However, if $M \gg N$, each individual can be infected no more than once, and so $p \approx 1$.

We can simplify this problem by assuming an individual is equally likely to be infected regardless of whether they have been previously infected. With this assumption, our situation reduces to resampling with replacement: for N objects (individuals), sampled (infected) M times, what proportion p of unique objects have been sampled? This is also known as the classical occupancy problem, and admits a closed-form solution: $p = 1 - \left[\frac{N-1}{N}\right]^m$ [243]. For N large, as with population sizes N_{pop_g}, we have $p \approx 1 - \exp\left(-\frac{M}{N}\right)$. Note that when $N = M$, this gives the classic $p = 1 - \frac{1}{e} \approx 0.632$ probability that an object will be sampled in bootstrapping. Returning to our original choice of, this gives $\mu_{SPI_{g,T}} = 1 - \exp\left(-\frac{N_{SPI_{g,T}}}{N_{pop_g}}\right)$.

One might assert that the probability of subsequent infection is lower if one has been previously infected. While there is evidence for this, we argue that in the current context, it is adequate to consider reinfections to be independent of past infection. Before the Omicron variant emerged, population seroprevalence in Canada was low, meaning that

$M \ll N$ and $p \approx \frac{m}{n}$ – regardless of the probability of reinfection. Second, once the Omicron variant became dominant, previous infection conferred much less protection against Omicron reinfection [244] – making the assumption of independence reasonable here too.

This gives the following model for $\mu_{SPI_{g,t}}$ (Figure 5.4):

$$
\begin{aligned}
N_{SPI_{g,T}} &= \sum_{t=1}^{T-\Delta_{IS}} I_t \gamma_{T-t} && \text{number of relevant infections still seropositive} \\
\mu_{SPI_{g,T}} &= 1 - \exp\left(-\frac{N_{SPI_{g,T}}}{N_{pop_g}}\right) && \text{assort seropositive infections to individuals} \\
\gamma_k &= 25\% + \frac{97\% - 25\%}{\left[1 + \left(\frac{k}{48}\right)^{3.5}\right]^2} && \text{five-parameter logistic model for seroreversion} \\
\Delta_{IS} &= 2 \text{ weeks} && \text{infection-seropositive delay}
\end{aligned}
$$

5.4.3.2 Mean seroprevalence attributable to vaccination

We now need to model true seroprevalence in a population attributable to vaccination, μ_{SP_V}. To do so, we consider the cumulative number of adults that have received at least one vaccine dose, $N_{vax_{g,T}}$. We use at least one vaccine dose as a single dose is sufficient to elicit an antibody response [245]. Recognising that the Canadian seroprevalence surveys are largely in blood donors, we consider only this adult population, because the minimum age to donate blood in Canada is 17, [246] and because vaccine eligibility was age stratified. Finally, we disregard seroreversion from vaccination in Canada, because most people who received one vaccine dose received subsequent doses that boosted antibody responses and prevented seroreversion.

To estimate μ_{SP_V}, we consider the proportion of individuals ever vaccinated $\frac{N_{vax_{g,T}}}{N_{pop_g}}$. Recognising that blood donors have different health behaviors than the general population [247], we adjust this with a parameter $OR_{vax_{BD}}$ that represents the odds ratio for vaccination in blood donors compared to all adults. The following model for $\mu_{SPV_{g,t}}$ results, where $\text{Odds}(p)$ indicates the odds function $\frac{p}{1-p}$:

$$
\begin{aligned}
\text{Odds}(\mu_{SPV_{g,T}}) &= \text{Odds}\left(\frac{N_{vax_{g,T-\Delta_{VS}}}}{N_{pop_g}}\right) \cdot OR_{vax_{BD}} && \text{sampling bias} \\
\Delta_{VS} &= 2 \text{ weeks} && \text{vaccination-seropositive delay} \\
\log OR_{vax_{BD}} &\sim \text{Normal}(0, 2) && \text{weakly regularising prior}
\end{aligned}
$$

5.4.3.3 Likelihoods for observed seroprevalence data

Finally, we establish likelihoods for observed seroprevalence from infection SP_I. These observations are subject to sampling and measurement biases, which lead to uncertainty in the true seroprevalence [158]. We use beta likelihoods to allow for this uncertainty: $\text{Beta}(\mu, \kappa)$, where $\mu_{SPI} \in (0, 1)$ indicates the distribution mean and κ_{SPI} indicates the precision. Higher κ can be interpreted as reduced seroprevalence study bias.

For seroprevalence due to infection or vaccination SP_{IV}, we make the simplifying assumption that the probability that an individual is infected μ_{SPI} is independent from the probability they are vaccinated μ_{SPI}, and combine the two using the law of independent events to get μ_{SPIV}. Finally, we also estimate a separate precision κ_{SPIV} for observed seroprevalence from infection and vaccination.

This yields the final component of the model, specified below:

$$SP_I \sim \text{Beta}(\mu_{SPI_{g,T}}, \kappa_{SP_I}) \qquad SP_I \text{ likelihood}$$

$$\kappa_{SP_I} \sim \text{Gamma}(4, 60) \qquad \text{broad prior on precision}$$

$$\mu_{SPIV_{g,T}} = \mu_{SPI_{g,T}} + \mu_{SPV_{g,T}} - \mu_{SPI_{g,T}} \cdot \mu_{SPV_{g,T}} \qquad \text{law of independent events}$$

$$SP_{IV} \sim \text{Beta}(\mu_{SPIV_{g,T}}, \kappa_{SP_{IV}}) \qquad SP_{IV} \text{ likelihood}$$

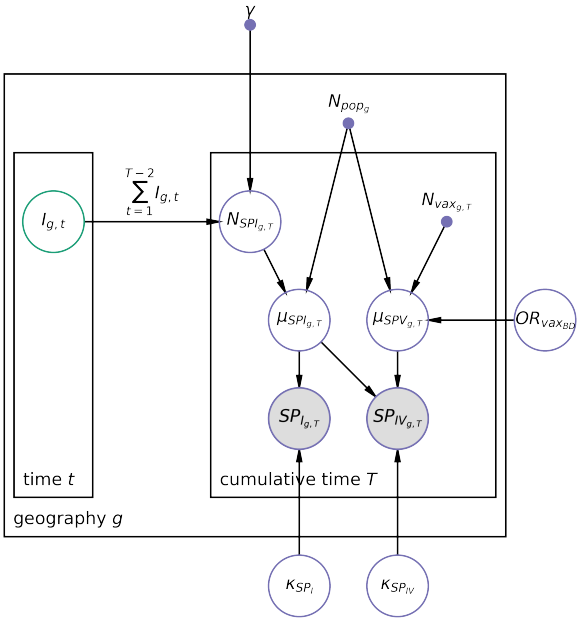
$$\kappa_{SP_{IV}} \sim \text{Gamma}(4, 60) \qquad \text{broad prior on precision}$$

5.5 Model simulation

Having defined the model and generative process of interest, we now aim to simulate a latent infections curve from this generative process, simulate the indicator- and population-based surveillance observations implied by this process, assess their verisimilitude, and evaluate whether the model can infer the true latent infections curve given these observations.

5.5.1 Generate simulated data

We simulate $n_g = 5$ geographies across $n_t = 100$ time points. To simulate data, we draw samples for all parameters from their prior distribution. We also specify the fixed values



$$N_{SPI_{g,T}} = \sum_{t=1}^{T-\Delta_{IS}} I_t \gamma_{T-t}$$

$$\mu_{SPI_{g,T}} = 1 - \exp\left\{-\frac{N_{SPI_{g,T}}}{N_{pop_g}}\right\}$$

$$\gamma_k = 25\% + \frac{97\% - 25\%}{\left[1 + \left(\frac{k}{48}\right)^{3.5}\right]^2}$$

$$\Delta_{IS} = 2 \text{ weeks}$$

$$\text{Odds}(\mu_{SPV_{g,T}}) = \text{Odds}\left(\frac{N_{vax_{g,T-\Delta_{VS}}}}{N_{pop_g}}\right) \cdot OR_{vax_{BD}}$$

$$\Delta_{VS} = 2 \text{ weeks}$$

$$\log OR_{vax_{BD}} \sim \text{Normal}(0, 2)$$

$$SP_I \sim \text{Beta}(\mu_{SPI_{g,T}}, \kappa_{SP_I})$$

$$\kappa_{SP_I} \sim \text{Gamma}(4, 60)$$

$$\mu_{SPIV_{g,T}} = \mu_{SPI_{g,T}} + \mu_{SPV_{g,T}} - \mu_{SPI_{g,T}} \cdot \mu_{SPV_{g,T}}$$

$$SP_{IV} \sim \text{Beta}(\mu_{SPIV_{g,T}}, \kappa_{SP_{IV}})$$

$$\kappa_{SP_{IV}} \sim \text{Gamma}(4, 60)$$

Figure 5.4: Model diagram for seroprevalence and population-based surveillance. White circles indicate unobserved parameters. Gray circles indicate observed quantities. Dots indicate input data used to inform models and treated as fixed. Edges indicate conditional dependencies between model components.

which the model does not contain a likelihood for. Some of these fixed values are constant across all runs of this model: γ_k, as defined above, and the delays $\Delta_{IC} = 1$, $\Delta_{IS} = 2$, $\Delta_{VS} = 2$. Others are data inputs and covariates that vary between model runs:

- N_{pop_g}: Independently drawn for all g from Normal$(1 \cdot 10^7, 3 \cdot 10^6)$
- $TPR_{g,t}$: We allow this to vary periodically over time, and set different patterns of variation by group. We specify a formula for the simulated TPR with these properties: $TPR_{g,t} = \frac{3}{g} \sin\left(\frac{t}{\pi} + g\right)$
- o_t: We define that the Omicron variant is present for all $t > 20$
- $V_{g,t}$: We define the proportion vaccinated to be 0% up to week 40, rising by 1% of the population each week to reach 60% of the population by week 100

Table 5.4: Variables in model for seroprevalence and population-based surveillance..

Variable	Values	Definition
DATA		
$SP_{I_{g,t}}$	$\mathbb{Z}^+$	Observed seroprevalence from infection in g at T; sparse
$SP_{IV_{g,t}}$	$\mathbb{Z}^+$	Observed seroprevalence from infection and vaccination in g at T; sparse
γ_k	$(0, 1)$	Probability of seropositivity k days after infection
$N_{vax_{g,T}}$	$\mathbb{Z}^+$	Number of adults ever-vaccinated against COVID-19 in g by T
Δ_{IS}	$\mathbb{R}^+$	Delay between infection and seropositivity
Δ_{VS}	$\mathbb{R}^+$	Delay between vaccination and seropositivity
PARAMETERS		
κ_{SP_I}	$\mathbb{R}^+$	Precision of observed $SP_{I_{g,t}}$
$\kappa_{SP_{IV}}$	$\mathbb{R}^+$	Precision of observed $SP_{IV_{g,t}}$
$OR_{vax_{BD}}$	$\mathbb{R}^+$	Odds ratio for vaccination in blood donors vs. all adults
TRANSFORMED PARAMETERS		
$N_{SPI_{g,T}}$	$(0, 1)$	Number of infections still seropositive in g up to T
$\mu_{SPI_{g,T}}$	$(0, 1)$	Seroprevalence from infection in g at T
$\mu_{SPV_{g,T}}$	$(0, 1)$	Seroprevalence from vaccination in g and at T
$\mu_{SPIV_{g,T}}$	$(0, 1)$	Seroprevalence from infection and vaccination in g at T

5.5.1.1 Simulated infections process

To simulate the underlying infections process, we modify the baseline infections process in Section 5.4.1 to ensure that the real-world infection rate is constrained by the size of the population. For example, in the UK Office of National Statistics infection survey [248], the baseline infection rate was never lower than 0.03% of the population and never higher than 7.60% of the population.

First, we expect a low infection rate at the start of the pandemic, so we centre the prior of $I_{t=1}$ around the minimum UK infections rate. Second, we place soft constraints on the log infections rate, encouraging it to increase if it is low relative to the minimum of the UK rate, and decrease if it is high relative to the UK rate. This yields the following

infections process, with Odds again denoting the odds function $\frac{p}{1-p}$:

$$\log(I_{g,t+1}) \sim N(\log(0.03\% \cdot N_{pop_g}), \sigma_{\log I}) \quad \text{minimum observed infection rate}$$

$$\mu_{\log I_{g,t+1}} = I_{g,t} \quad \text{multiplicative random walk}$$

$$+ \frac{\text{Odds}(0.03\%)}{\text{Odds}(I_{g,t}/N_{pop_g})} \quad \text{soft floor if infection rate low vs. observed min}$$

$$- \frac{\text{Odds}(I_{g,t}/N_{pop_g})}{\text{Odds}(7.60\%)} \quad \text{soft cap if infection rate high vs. observed max}$$

$$\log(I_{g,t+1}) \sim \text{Normal}(\mu_{\log I_{g,t+1}}, \sigma_{\log I}) \quad \text{normal model on } \log I$$

$$\sigma_{\log I} = 0.4 \quad \text{specified value for forward simulation}$$

Figure 5.5 shows the simulated infections realisation used for parameter recovery.

5.5.2 Simulated observations and prior predictive checks

We then simulate the observed cases and seroprevalence, drawing these from the model likelihood given the sampled parameters and simulated infections process. For seroprevalence studies, we specify that there are 15 seroprevalence studies done per region. Of these studies, 10 capture seroprevalence from infection, every 9 days from day 10 to day 91; 5 capture seroprevalence from vaccination, every 9 days from day 55 to day 91.

We conduct prior predictive checks on these simulations, to evaluate whether our prior choice and model architecture is realistic and reflects our domain knowledge of this infectious disease.

Figure 5.5 shows simulated cases alongside the infections realisation that they were generated from. As expected, we see low infections initially, and an oscillating pattern of infections over time that varies between geographies. There is a noisy correlation between the infections and case curves, reflecting that cases are a biased measure of infections and reflecting the time-varying factors that modulate the case-infection relationship. The magnitude of cases relative to infections (i.e., the probability of case detection) shows the intended region to region variation.

Figure 5.6 shows simulated seroprevalence alongside the corresponding cumulative infections realisation and proportion vaccinated. The seroprevalence from infection diverges from the cumulative infections curve, exhibiting the expected seroreversion. In this simulation, seroprevalence from infection and vaccination lags the proportion ever

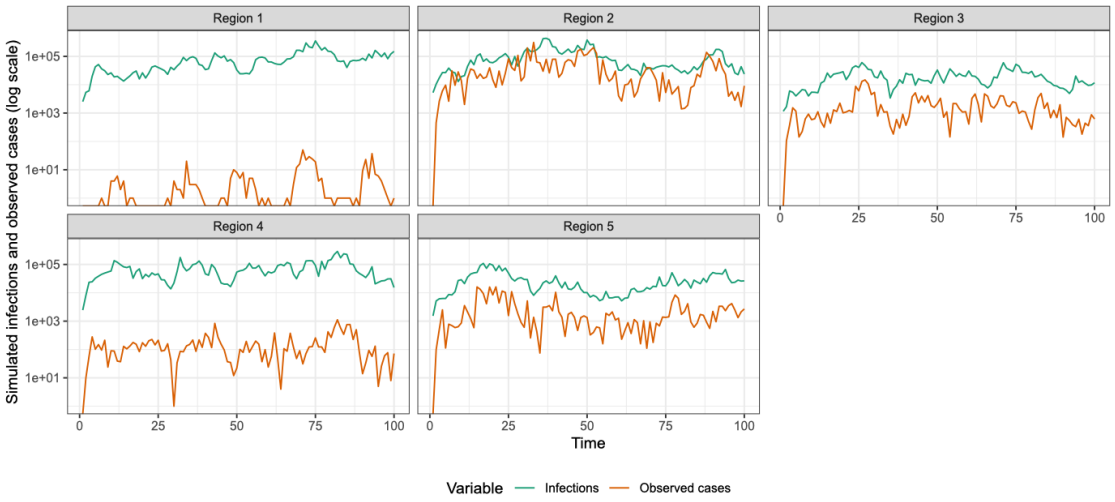


Figure 5.5: Simulated infections and observed cases. Faceted by geography, with each subplot showing a realisation of the infections process over time and the observed cases implied by these infections, specified parameters, and the model likelihood.

vaccinated, reflecting that the simulated $OR_{vax_{BD}} \approx 0.45$ implies less vaccination in serosurvey participants than the general population.

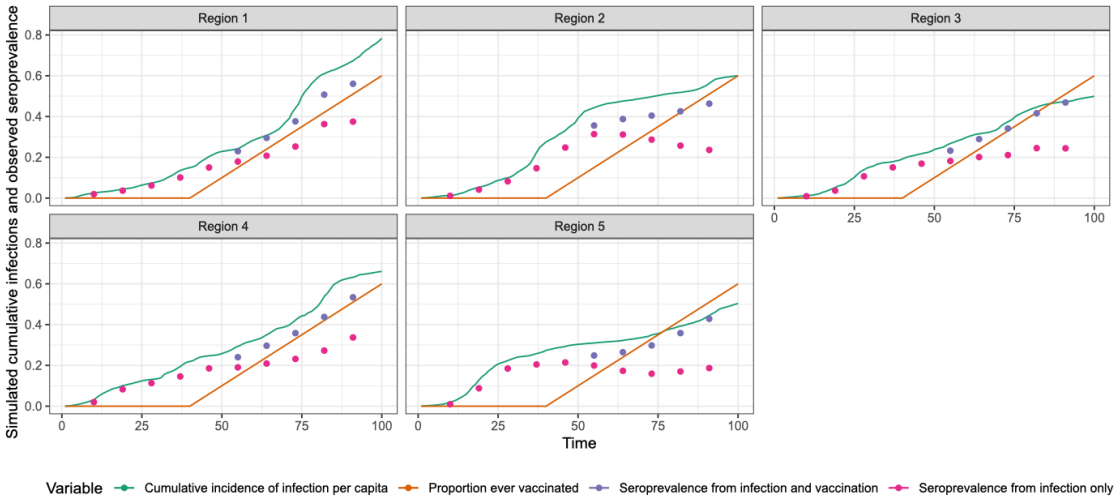


Figure 5.6: Simulated infections, proportion vaccinated, and observed seroprevalence. Faceted by geography, with each subplot showing the cumulative infections process over time and proportion ever vaccinated. Subplots also show the observed seroprevalence implied by these infections, specified parameters, and the model likelihood.

Overall, the infections process and the observed cases and seroprevalence are realistic and reflect our knowledge of this infectious disease, providing some confidence in our mode and the underlying generative process.

5.5.3 Simulated data recovery

We then aim to evaluate whether the model can infer the true infections process and parameters given observed cases and seroprevalence from the same generative process. To do so, we fit the model to the simulated cases and seroprevalence generated in Section 5.5.2. We specify this model in the probabilistic programming language `Stan` [226]. All components specified above (i.e., the infections, cases, and seroprevalence models) are included in a single model estimated jointly. Sampling is done using the No-U-Turn Sampler [249], tuning algorithm parameters to set the target acceptance probability `adapt_delta` to 0.99 from its 0.8 default and the maximum treedepth `max_treedepth` to 12 from its 10 default. We sample four Markov chains with random initial values, making 1,000 warmup and 1,000 sampling draws per chain for a total number of sampling draws N_{draw} of 4,000.

To ensure robust model fit, we examine four diagnostics (Figure 5.7). First, trace plots suggest efficient exploration of the posterior. These trace plots show rapid convergence of each chain to the typical set. The chains mix well (i.e., explore the same regions of the posterior parameter state). Moreover, they do not get stuck in particular regions of the posterior (Figure 5.7).

Efficient mixing is supported by the second diagnostic: the convergence statistic $\hat{R}$, which is the ratio of the variance of samples within the same chain to the variance of all samples across chains [250]. If all chains have converged to the same equilibrium distribution, then all samples will be drawn from that equilibrium distribution, hence $R = 1$. It is ideal for $\hat{R} < 1.01$, and, if not, desirable for $\hat{R} < 1.05$ [222]. The ideal $\hat{R}$ value was achieved for all but one parameter 5.7).

Third, we examine the effective sample size N_{eff} for each parameter, with reference to an ideal of 400 (100 per chain) and acceptable of 40 [251]. Here, the ideal value was achieved for almost all parameters (Figure 5.7).

Finally, we examine the number of divergent transitions. Divergent transitions occur when the numerically calculated solutions to Hamilton's equations diverge from their true solutions. This often occurs when Hamiltonian Monte Carlo is unable to explore certain

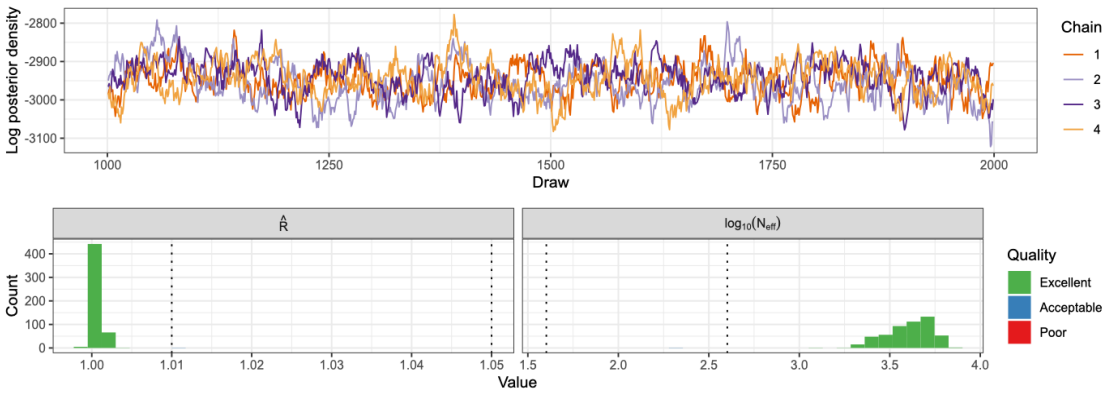


Figure 5.7: Trace plot and model diagnostics $\hat{R}$ and $\log_{10}(N_{eff})$ for parameter recovery. For trace plot, each colour is a different chain. For diagnostics, colours indicate diagnostic value relative to a quality threshold, indicated by the dotted lines. $\hat{R} < 1.01$ was considered excellent and ≤ 1.05 acceptable. $N_{eff} \geq 400$ was considered excellent and ≥ 40 acceptable.

regions of the posterior space, such as regions with high curvature. There was only one divergent transition among $4,000$ draws, suggesting a tractable posterior geometry.

Given satisfactory diagnostics, we now move to evaluate parameter recovery by comparing known values to posterior predictions for the latent time series of infections and for all relevant model parameters.

Figure 5.8 shows that the model recovers the true infections process well across regions. Most infections values fall within the model 95% credible interval, and the model captures the general shape of these curves well. There is little apparent overfit, considering that the model does not attempt to capture every fluctuation or extremum in the infections process. Moreover, the model confidence decreases towards the edges of the infections window, appropriately reflecting the lack of seroprevalence data and lagged case data available to the model at those times.

The model is also able to recover the true parameters from these data. Figure 5.9 shows recovery of the nine parameters in this model beyond the latent infections curve. For most of these parameters, the true parameter value is contained within the model-estimated 95% credible interval. The exceptions are β_t, α_4, and α_5, for which the true parameter value lies slightly outside that interval. Each of these deviations are in the model for $\rho_{g,t}$, and can be attributed to colinearity between the time, test positivity rate, and Omicron predictors creating slight challenges for identifiability.

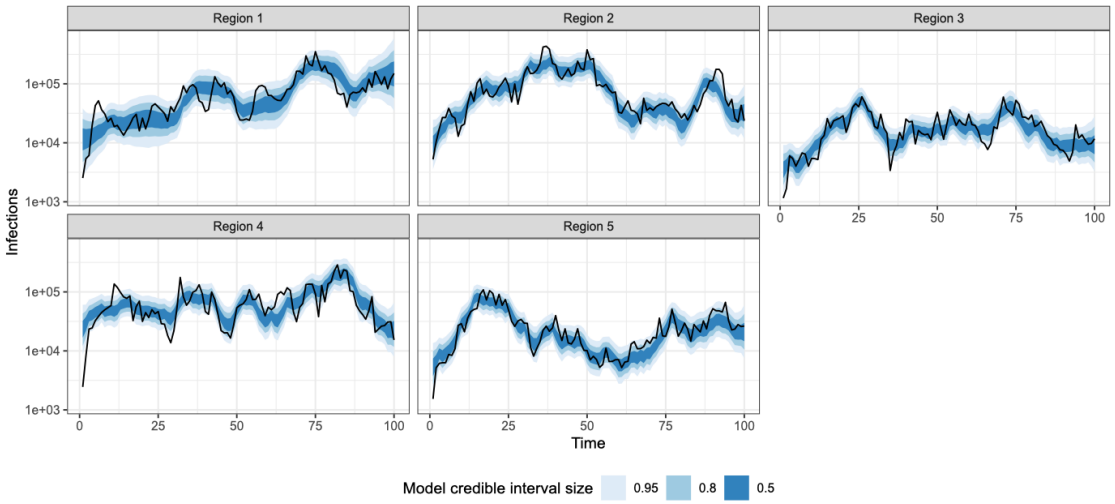


Figure 5.8: True infections curve and model estimated infections for parameter recovery. Estimates were generated by fitting the model to case and seroprevalence data implied by the shown true infections curve, other model parameters, and the specified generative process.

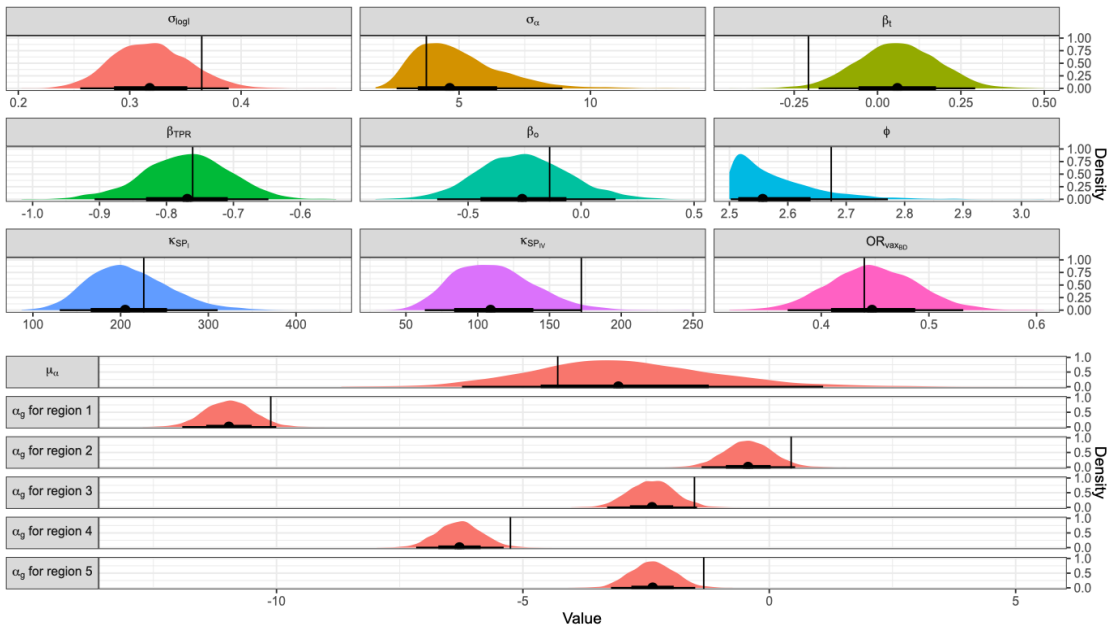


Figure 5.9: All parameters and region-specific intercepts for parameter recovery. Vertical lines indicate true parameter values. Densities are from model fit to simulated data, with points indicating the median estimate, thick lines indicating the 66% HDCrI, and thin lines indicating the 95% HDCrI.

Overall, the precise recovery of the infections process, and strong recovery across parameters, provide confidence that this model is capable of recovering its parameters when applied to data that reflect the same underlying generative process. With this in hand, we now proceed to apply this model on Canadian data.

5.6 Application to real-world Canadian data

To apply this model to real-world Canadian data, we first identify the relevant datasets, evaluate the sensibility of this model fit to real-world data, and then interpret model parameters.

5.6.1 Data

We included data from 17 February 2020 – roughly two weeks before community transmission of SARS-CoV-2 was confirmed in Canada – until 24 April 2022, after Canada's first Omicron wave had passed. For population sizes, we used estimates from Statistics Canada halfway through this time period – that is, from the second quarter of 2021 [252]. We used weeks as our unit of time, defining Monday as the first day of the week.

We obtained a time series of daily confirmed COVID-19 cases and diagnostic tests completed for each Canadian province and territory from the Canada COVID-19 Open Data Working Group database [253], which collects these data from government and public health authority reports and dashboards. We aggregated this dataset to the weekly level by summing cases C and diagnostic tests T in each group. We then calculated the test positivity rate for each province g and week t as $TPR_{g,t} = \frac{C_{g,t}}{T_{g,t}}$. To rescale TPR using the logit transformation required $TPR \in (0, 1)$. To achieve this, if more cases were discovered than tests done, we set the TPR to be the maximum TPR value below 1. Similarly, if no tests were done, we set the TPR to be the minimum non-zero TPR value. We also obtained the week during which the Omicron variant became dominant in Canada from CoVariants. This was the week of 13 December 2021 [254].

To obtain seroprevalence estimates for Canadian provinces and territories, we used SeroTracker [9, 100]. We included all seroprevalence studies estimating seroprevalence in the general population: for Canada, this included household surveys like Statistics Canada's Canadian COVID-19 Antibody and Health Survey and the AB-C study [255, 256], studies of blood donors including those from Canadian Blood Services and Hema-Quebec [257, 258], and studies done by each Canadian province if available [26, 259]. Studies done exclusively in children were excluded. We used the most granular time points available.

We assigned each seroprevalence study to the time point halfway between the dates it started and finished sampling participants, which we term its sampling midpoint date. The beta likelihood we used admits $SP \in (0, 1)$; for this reason, if seroprevalence was 0, we set seroprevalence to half of the minimum non-zero value. No COVID-19 vaccines were administered in Canada before 14 December 2020 [260], meaning all seroprevalence studies before this date measured seroprevalence attributable to infection SP_I. After 14 December 2020, COVID-19 vaccines were in use in Canada, all of which used spike protein-derived antigens and triggered only anti-spike antibodies [261]. For this reason, we considered anti-spike seroprevalence after this date to be from infection or vaccination SP_{IV}, while anti-nucleocapsid seroprevalence studies after this date continued to measure seroprevalence against infection SP_I.

For vaccination data, considering our interest in vaccination data in adults, we used the Public Health Agency of Canada's age-structured vaccine administration dataset [262]. From this dataset, we extracted the proportion of the adult population who had received at least one dose of any vaccine. Some cells were reported as $\geq 99\%$; for these, we imputed vaccine coverage of 99.5%. We then multiplied this by the adult population size in each province and territory (again, from Statistics Canada [252]) to obtain the number of adults vaccinated $N_{vax_{g,T}}$.

5.6.2 Model fit and diagnostics

We fit this model and assessed diagnostics described in Section 5.5.3. There were no divergent transitions among $4,000$ draws. Trace plots indicated good mixing (Figure 5.10). $\hat{R}$ was excellent for most parameters, acceptable for most others, and poor for very few. N_{eff} was excellent for most parameters and acceptable for all others (Figure 5.10).

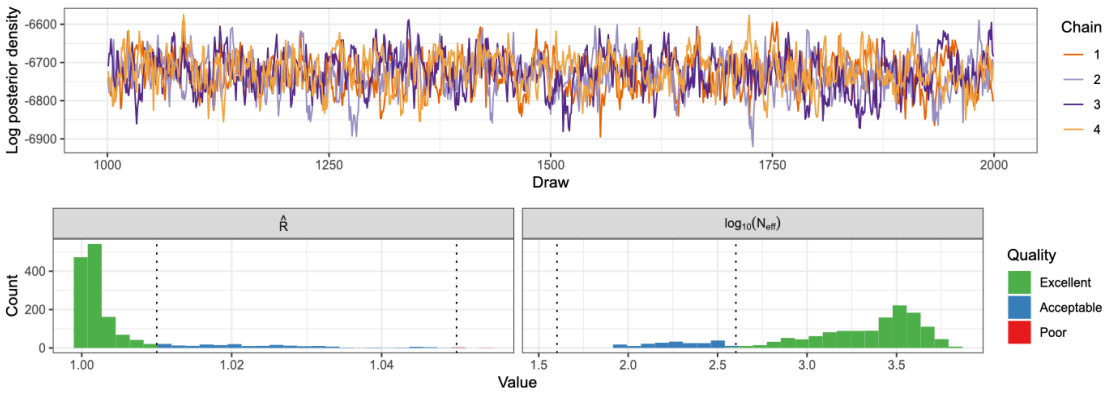


Figure 5.10: Trace plot and model diagnostics $\hat{R}$ and $\log_{10}(N_{eff})$ for fit to Canadian data. For trace plot, each colour is a different chain. For diagnostics, colours indicate diagnostic value relative to a quality threshold, indicated by the dotted lines. $\hat{R} < 1.01$ was considered excellent and ≤ 1.05 acceptable. $N_{eff} \geq 400$ was considered excellent and ≥ 40 acceptable.

5.6.3 Posterior predictive checks

With convergence established, we proceed to conduct posterior predictive checks. These checks evaluate the quality of model fit to data, reflecting the extent to which the model generative process can capture the generative process underlying real-world data.

The model posterior predictive distribution for cases tracks very well with observed cases C across Canadian provinces and territories (Figure 5.11), validating this aspect of model fit. This fit somewhat smooths local variation in observed C, and disregards atypical such variation (e.g., in Nunavut from December 2020 to April 2021), suggesting robust fit without overfit.

Similarly, for seroprevalence data from infection, we also see excellent correspondence between model predictions of seroprevalence and observed seroprevalence, for both SP_I (Figure 5.12) and SP_{IV}(Figure 5.13). Notably, for seroprevalence, the model is not overfitting by attempting to fit every seroprevalence observation. The fit disregards obvious outliers in seroprevalence, like the atypically high observed SP_I in British Columbia around March 2020 and in Manitoba in November 2020 (Figure 5.12), as well as the atypically low observed SP_{IV} in Ontario in June 2021. The absence of overfit seen here suggests robust model inference.

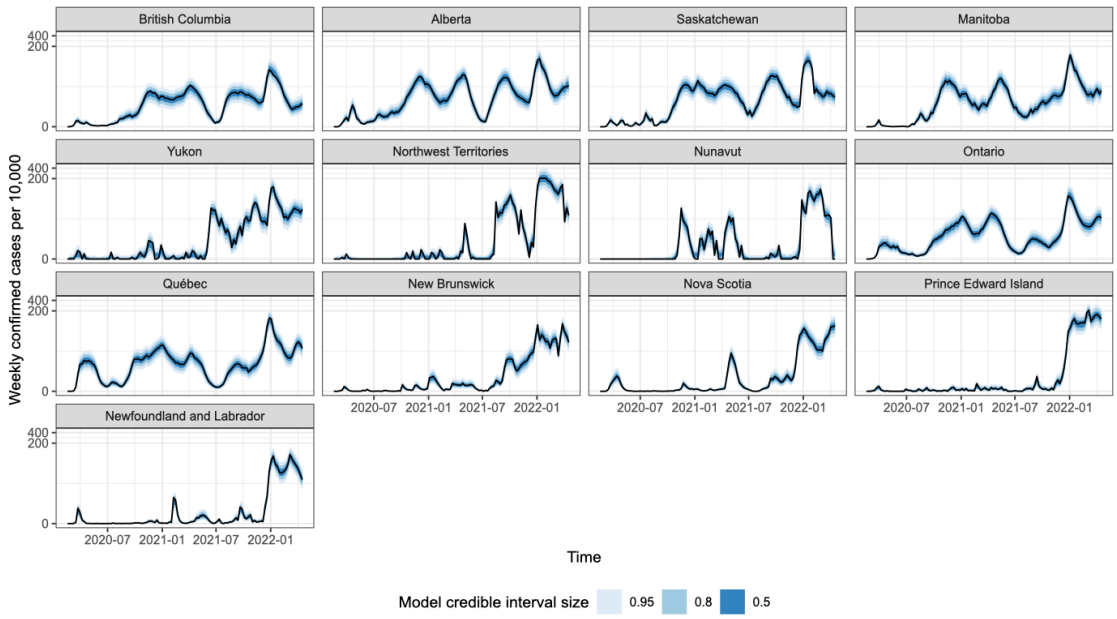


Figure 5.11: Posterior predictive distributions and observed values for cases C in Canadian data fit. Posterior predictive distributions in blue and observed values in black. Cases shown per $10,000$. log axis.

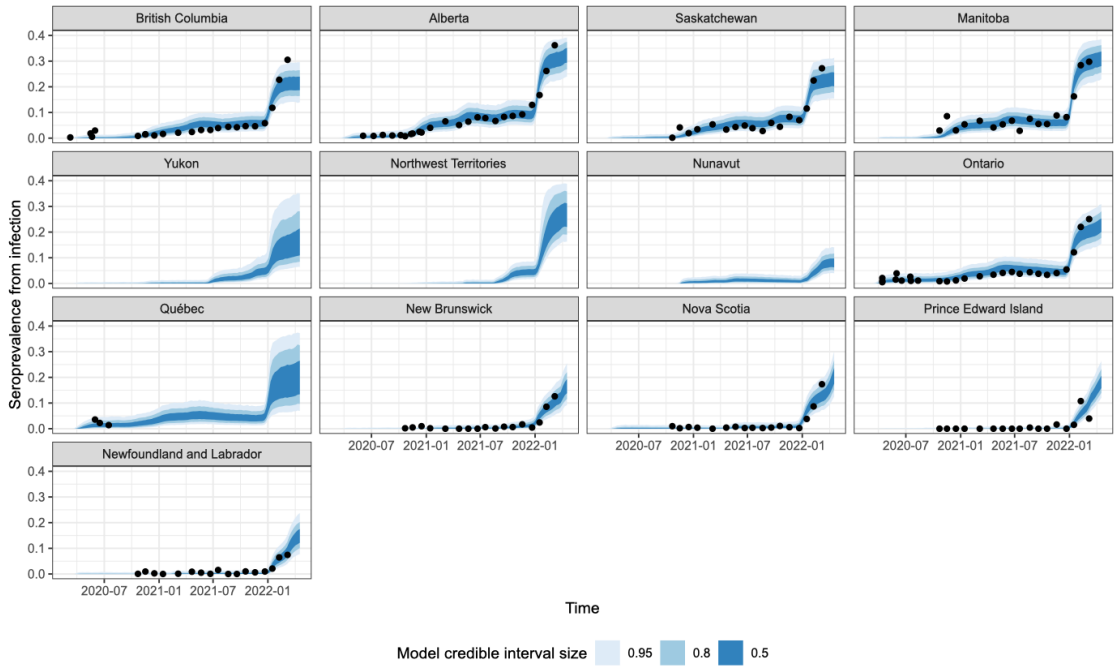


Figure 5.12: Posterior predictive distributions and observed values for seroprevalence from infection SP_I in Canadian data fit. Posterior predictive distributions in blue and observed values in black.

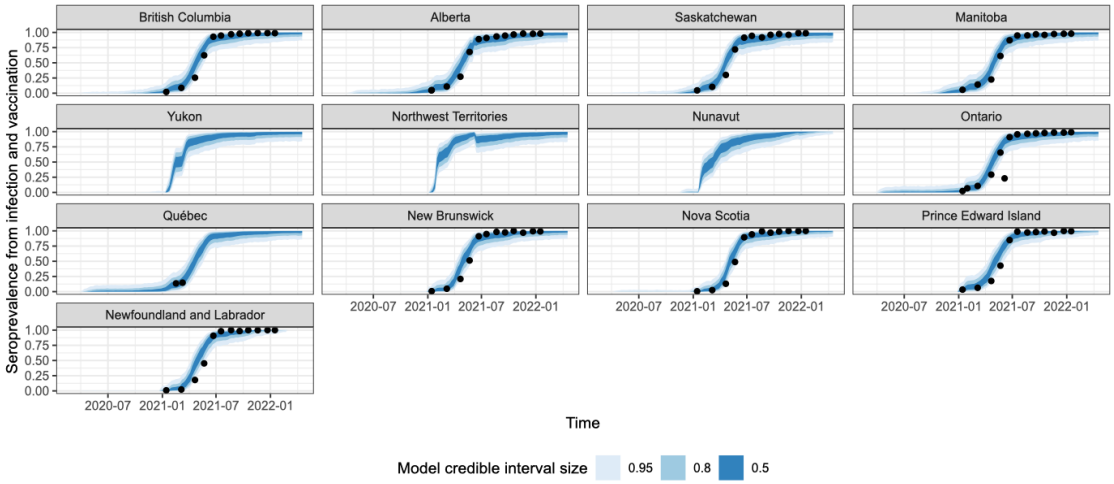


Figure 5.13: Posterior predictive distributions and observed values for seroprevalence from infection or vaccination SP_{IV} in Canadian data fit. Posterior predictive distributions in blue and observed values in black.

5.6.4 Interpreting model fit

We have shown that our model encodes a realistic generative process (per prior predictive checks), that it can recover the true parameters from simulated data drawn from that generative process (per simulated data recovery), that it has accurately fit the Bayesian posterior when fit to real-world data (per model diagnostics), and that its generative process captures the generative process underlying real-world data (per posterior predictive checks).

These steps validate this model and its fit to Canadian data, enabling us to deploy the model and interpret its parameters with confidence. We will now use this model to examine the dynamics of SARS-CoV-2 infection and immunity. Our analysis will focus on infections I and the probability of case detection ρ through three eras: the first COVID-19 wave in Canada; subsequent waves before Omicron; and the emergence of Omicron. We will also discuss seroprevalence predictions from the model. We refer to and interpret single parameters (e.g., α_g, κ_{SP_I}) throughout.

5.6.4.1 Estimated infections

This model reveals the dynamics of SARS-CoV-2 infection over time in Canada – a key, typically unobserved, parameter. Figure 5.14 shows model estimates of $I_{g,t}$.

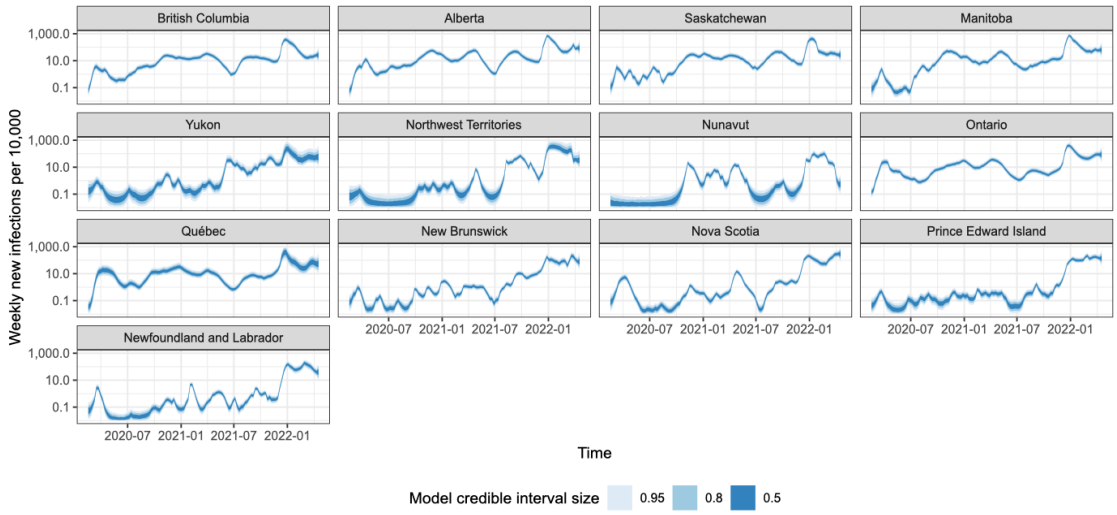


Figure 5.14: Modelled infections I for Canadian data fit. Infections per $10,000$; log axis.

The posterior 95% credible interval for $\sigma_{\log I}$ is $(0.610, 0.675)$ (Figure 5.16). This is close to, but notably different from and more precise than, our prior of Normal$(0.7, 0.2)$. In general, all parameters have posterior distributions quite different from their prior distributions, suggesting that the model posteriors are well-informed by data (Figure 5.16).

5.6.4.2 Infections and case detection in the first wave

One of the key questions that this model lets us answer is how big the first wave of COVID-19 (i.e., March 2020 to May 2020) was in Canada. The burden of infection during this wave was particularly unclear given the limited availability of diagnostic testing [208]. We see that in all Canadian provinces, this wave was sizable (up to about 30 infections per 10,000; Figure 5.14). In some provinces it was much lower than subsequent pandemic waves (e.g., British Columbia) whereas in Canada's largest provinces, Ontario and Quebec, it neared the size of subsequent pandemic waves despite case counts being much lower at that time. Notably, Canada's first confirmed COVID-19 case was in Ontario [263].

These estimates of infections I are informed by data on cases C and the proportion of infections detected as cases. During the COVID-19 pandemic, cases were extensively used in Canada to monitor the current burden of disease [264]. Examining the proportion of infections detected helps us investigate the quality of our case monitoring systems, biases inherent in them, and the factors that drive them.

Across all provinces and territories, the proportion of infections detected was very low in February 2020, ranging from 10% to 25% in most large provinces (Figure 5.15). These values mean that, for each case reported at this time, 4 to 10 infections occurred – reflecting the limited early availability of diagnostic tests. Diagnostic testing availability rapidly increased during the first wave [253], concomitant with the increase in the proportion of infections detected from March 2020 to April 2020. Reflecting this increase in time, the model estimates a positive time coefficient β_t in the logit-linear model for the proportion of infections detected ρ (Figure 5.16).

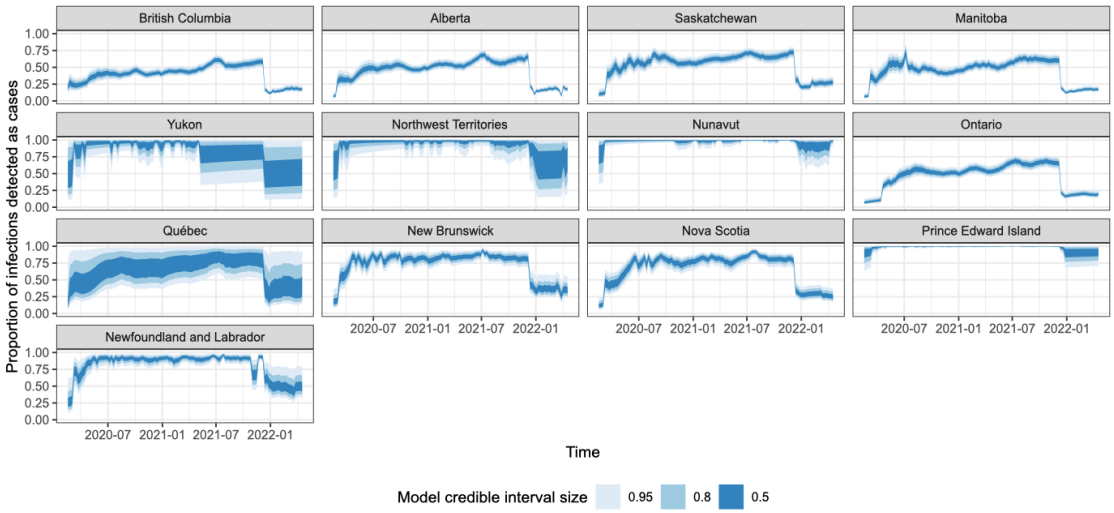


Figure 5.15: Modelled proportion of infections detected ρ for Canadian model.

5.6.4.3 Infections and case detection in the pre-Omicron era

We now turn our attention to COVID-19 after the first wave and before the Omicron era. These estimates reveal considerable geographic heterogeneity that reflect differences in public health policy. Three clusters emerge from model estimates of SARS-CoV-2 infection. The first cluster includes provinces in Western and Central Canada (i.e., British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, and Quebec), with populations from one to fifteen million. These populations pursued the mitigation strategy common to many high-income countries in the Anglosphere [265], and reveal the classic several waves of infection that characterised that strategy (Figure 5.14). Notably, different provinces had varying levels of success in mitigating infections. For example, British Columbia and Quebec both took earlier and more aggressive action to head off pre-Omicron pandemic waves [265]. The provinces also had lower peak infection rates – approximately 2.5-fold lower than the peaks in provinces where less aggressive action was taken (i.e., Alberta, Manitoba, and Ontario; Figure 5.14). The proportion of infections detected was generally moderate in these provinces during this era, ranging from 0.33 to 0.75, with some province-to-province variation 5.15.

The second cluster of interest is the Atlantic provinces – New Brunswick, Nova Scotia, Newfoundland and Labrador, and Prince Edward Island. These provinces have populations ranging from one hundred thousand to one million, are on the east coast

of Canada, and share fewer land borders with other Canadian provinces. Early in the pandemic (i.e., July 2020), these provinces recognised their geographical proximity to one another and their similarly aggressive public health responses, forming the so-called Atlantic Bubble to allow free movement without quarantine within the bubble [266]. Indeed, we see success in each of these territories in mitigating infection spread until November 2020 (Figure 5.14), when free travel was again limited in response to rising case counts in New Brunswick and Nova Scotia [266]. Between that point and the Omicron era, each of the Atlantic provinces was able to considerably limit SARS-CoV-2 infections, although to different extents. The lowest mean infection rates and highest mean proportion of infections detected were in Prince Edward Island, where the most stringent restrictions were imposed [267]; the intermediate infection rates and proportion of infections detected in Newfoundland and Labrador; and the highest mean infection rates and lowest average proportion of infections detected in New Brunswick and Nova Scotia, although the infection rates even in these Atlantic provinces remained much lower than those in the Western and Central provinces (Figures 5.14 and 5.15).

The final cluster here is the territories – Yukon, Northwest Territories, and Nunavut. These territories are all in Northern Canada, and have populations between thirty-five and fifty thousand. Considering the low populations in these regions, the infections curves here are characterised by particularly steep spikes and drop-offs which indicate specific outbreaks. The infection burden in the territories was generally low before the omicron era, punctuated by said outbreaks. The proportion of infections detected in the territories was also high during this time period (Figure 5.15).

5.6.4.4 Parameters for proportion of infections detected

The parameters in the logit-linear model for the proportion of infections detected ρ drive its variation. First, the province- and territory-specific intercepts α_g show considerable variation (Figure 5.16), paralleling the regional differences in the proportion of infections detected. The median estimate for $\sigma_\alpha = 2$, a moderate value suggesting notable differences in case detection between geographies but with some commonalities within Canada. Short-term variation in the proportion of infections detected is driven

by variation in the test positivity rate TPR, where a higher TPR indicates a greater burden on diagnostic testing capacity. The model-estimated coefficient for β_{TPR} was less than zero (Figure 5.16), indicating that a reduced proportion of cases is detected when burden on testing systems is high.

The final parameter in the model for cases is ϕ, which indicates the extent of overdispersion in the negative binomial likelihood. The median estimate for ϕ is 2.56 (Figure 5.16). This implies a variance a factor $1 + \frac{1}{\phi} = 1.39$-fold larger than the Poisson distribution with the same mean and indicating mild overdispersion.

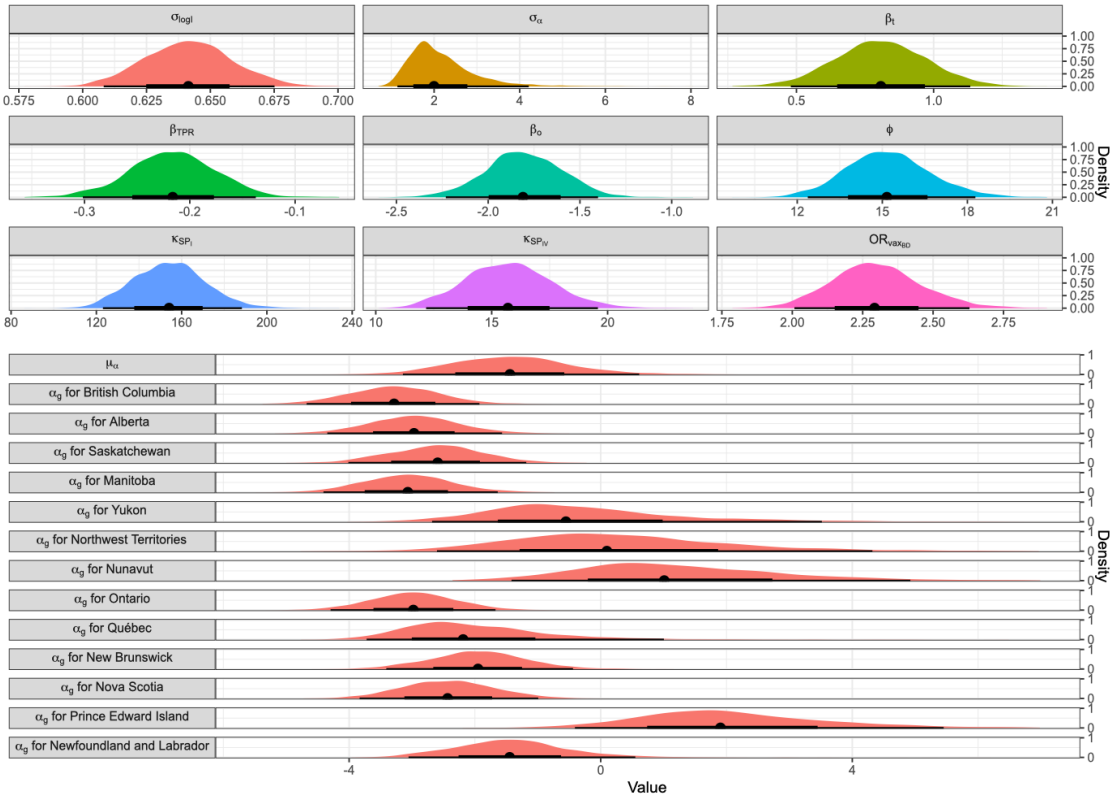


Figure 5.16: All parameters and region-specific intercepts for Canadian model. Densities are from model fit to simulated data, with points indicating the median estimate, thick lines indicating the 66% HDCrI, and thin lines indicating the 95% HDCrI.

5.6.4.5 Infections and case detection in the Omicron era

Omicron led to a regime change across all regions. The first case of the Omicron variant in Canada was reported on 28 November 2021 [268], with this variant rapidly becoming dominant by the week of 13 December 2021 [254]. The Omicron variant was much more transmissible and much more adept at evading immunity than previous variants [228,

244]. These characteristics were reflected in the considerable Omicron wave, which was much higher than previous COVID-19 waves. For example, in Alberta, this wave peaked at just below 10% of the population infected – 30-fold higher than Alberta's previous largest peak. These spikes led to a concomitant spike in seroprevalence from infection, from 8% in Nunavut up to 40% in Alberta.

Omicron itself was associated with a regime change in Canadian testing policy. At this time, many jurisdictions restricted PCR testing access to at-risk individuals [265]. This also coincided with the broad availability of home lateral flow testing [233], where infections detected from this testing were not centrally reported as cases [209]. Our model demonstrates the implication of this for the reliability of case counts: the proportion of infections detected dropped sharply across all provinces at the start of the Omicron era. Higher test positivity rates at this time are insufficient to explain the drop in proportion of infections detected during the Omicron era: β_o, the coefficient for the omicron era in the model for the proportion of infections detected, was estimated to be -1.80 (95% HDCrI -2.23 to -1.42), indicating a strong negative association between Omicron and the proportion of case detection.

5.6.4.6 Inferred seroprevalence

Finally, we can also use this model to gain insight into the true dynamics of SARS-CoV-2 immunity, as measured through the prevalence of SARS-CoV-2 antibodies. In addition to showing model fit to real-world seroprevalence data, Figures 5.12 and 5.13 also show model out-of sample predictions for seroprevalence. This includes seroprevalence in provinces where data was available but where multiple estimates conflicted with one another. Take Ontario and British Columbia before July 2020. Here, the model estimate of seroprevalence agrees with the lowest seroprevalence observations at each time where there are multiple observations.

Second, this includes seroprevalence prediction beyond the times where seroprevalence studies were available, considering the sizable lag between when seroprevalence studies sample participants and seroprevalence data are made available. We see these extrapolations in all Canadian provinces after March 2022, and these provide important

updates on seroprevalence: the stagnation in seroprevalence in Alberta, Saskatchewan, and Manitoba; and continued and sharpening increases in New Brunswick and Nova Scotia. Understanding these changing immune dynamics is crucial to both map the true extent of infection, in an era when testing systems are capturing a lower fraction of infection, and also to understand immunity itself and the expected future spread of the disease in this population.

Finally, and perhaps most obviously, this model also estimates seroprevalence for regions where there are no estimates (the territories) or very few (Quebec). This data has clear value in helping inform public health decision-making in these regions. Notably, the uncertainty bands for seroprevalence in geographies without this data are wider. This greater uncertainty is also reflected in estimates of the infections process I, the proportion of infections detected ρ, and the province-level coefficient for the proportion of infections detected α_g for these geographies (Figures 5.14 and 5.15). This reflects the limited population-based surveillance data needed to pin down the case-infections relationship in these regions; with this said, estimates for these regions remain informed by different region-specific predictors and by partial pooling with other regions.

The median precision of seroprevalence studies from infection κ_{SP_I} was 154. This parameter can be interpreted as the sample size of an unbiased study that would yield the same precision. In contrast, the median actual sample size of seroprevalence studies included was 848 (IQR $445, 2644$). The roughly 6-fold difference between these values reflects the sampling and measurement biases inherent in all seroprevalence studies. With this said, $\kappa_{SP_I} = 154$ still represents a reasonably precise and informative seroprevalence study.

Considering seroprevalence studies from vaccination, we first consider the relative odds of vaccination in blood donors $OR_{vax_{BD}}$. The median estimate for this quantity was 2.28, indicating that blood donors are much more likely to be vaccinated than adults at large – an observation in line with our initial hypothesis about the health-seeking behaviors in these groups. Even with this term considered, $\kappa_{SP_{IV}}$ was median 16, a roughly 60-fold difference from the median actual sample size, that suggests that seroprevalence studies from vaccination are much less informative in this model. This

can be explained by the regular time series of SP_I that remain available in Canada into the vaccine era. However, being able to integrate such seroprevalence data from vaccination would be particularly important if applying this model to the global context, where seroprevalence from infection cannot be easily separated from vaccination, and where there is a paucity of seroprevalence data altogether in some regions.

5.7 Discussion

We present here a flexible and interpretable framework for modelling true infections and immunity for an infectious disease, which integrates data from indicator-based and population-based surveillance alike. We validate our model by demonstrating that it can recover the parameters used to generate simulated data. We then apply it to Canadian data, first showing that the model fits well and its posterior captures Canadian data well without overfit, and then applying it to reconstruct the dynamics of true COVID-19 infections, the probability of case detection, and up-to-date country-wide seroprevalence.

A primary strength of this model is its flexibility on several dimensions. First, its flexibility in the scale of the data we apply it to. Here, we apply it to provincial and territorial data within one country, but its structure admits easy application to sub-provincial data or to global data. Second, its easy extensions to other sorts of data, including other indicators (deaths, hospitalisations) and other population-based measures (point prevalence). Third, the adaptability of the model for infections. This model can be kept relatively simple, as in the multiplicative random walk presented here. This could also be substituted for more complex models: for example, using a bounded random walk on count data [269], or a Gaussian process model for infections over time. This flexibility enables our model's adaptability across infectious diseases and also through the rapid developments of any given infectious disease outbreak.

The interpretability of this model is also notable. Each of the parameters and intermediate parameters have clear interpretations, and can be interpreted to provide insight about the underlying data generation processes – for example, the effect of the Omicron variant β_o on the proportion of infections detected ρ, or the fundamental rate of variation in infections $\sigma_{\log I}$.

There are also some notable limitations here. First, we used a fixed seroreversion curve rather than estimating that curve, due to nonidentifiability challenges in estimating such decay curves [270]. While our fixed parameters were based on the same population and same assays, it would be ideal to refine these estimates, or estimate this as part of a unified model. Second, we did not include seroreversion for vaccinated individuals. We know that this seroreversion occurs, and that subsequent vaccinations delay it. While our assumption of no seroreversion is reasonable soon after the primary and booster vaccination campaigns, accounting for said seroreversion will be crucial to apply this model over much longer time horizons. Third, the approach here requires some hand selection of predictors of the relationship between infections and indicator-based surveillance models – in this case, time, the test positivity rate, and the emergence of the Omicron variant. While these are reasonably simple to select for domain experts, they do limit the ability to apply this model directly without fine tuning. Finally, we applied this model to Canada, where case reporting can be thought somewhat reliable, and seroprevalence data is reasonably dense. Moreover, we applied this model in the pandemic context. Questions remain about this model's performance in more data-sparse applications and for endemic infectious disease.

6

Expediting rigorous evidence synthesis during health emergencies

6.1 Motivation

In Section 1.5, we described the role of the systematic review as the gold standard for robust evidence synthesis. Systematic reviews combine a comprehensive search with rigorous quality appraisal [271, 272], and have particular importance in health emergencies in light of a changing evidence base. We also discussed the delays that they introduce: the median systematic review takes 61 weeks to be published after the final search is conducted, and 33 weeks from final search to publication is considered rapid [83]. Living systematic reviews, intended to routinely update systematic reviews to address these limitations, have also fallen short: updates to a typical living systematic review are still only published once every 3-6 months [273]. The delay in the availability of these crucial syntheses substantially hinders their actionability and use for secondary analysis, especially during public health emergencies. It is therefore crucial to expedite evidence synthesis methods without sacrificing quality.

In developing SeroTracker, we sought to update our data every week to maximise its public health impact while maintaining the rigour of a systematic review. This was a substantial challenge, considering that our search yielded 800 to 1,000 new abstracts

each week to be screened in duplicate, yielding 25 to 40 full text articles to be extracted and critically appraised.

Here, we describe two tools that we developed to expedite our rigorous evidence synthesis effort in the context of the COVID-19 health emergency. First, the strategy we developed to automate risk of bias assessment as the most subjective and complex component of conducting data extraction, and the approach we used to develop this strategy that can be broadly generalised to other evidence synthesis efforts. Second, the in-house tool we built to apply natural language processing to accelerate abstract screening, and the real-world evaluation of it that we conducted to show that this tool could replace one of two human abstract screeners.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Bobrovitz N*, Noël K*, Li Z, . . . , **Arora RK.** 2022. SeroTracker-RoB: an approach to automating reproducible risk of bias assessment of seroprevalence studies. Resubmitted major revisions to *Research Synthesis Methods. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

Perlman-Arrow S*, Loo N*, Bobrovitz N, Yan T**, **Arora RK****. 2022. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Resubmitted minor revisions to *Research Synthesis Methods. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

6.2 Automating reproducible risk of bias assessment for seroprevalence studies

Seroprevalence studies have made important contributions to the pandemic response, but their methods and quality have varied widely [9, 70]. Accordingly, robust risk of bias (RoB) assessments have been crucial to identify trustworthy seroprevalence data for public health decision-making.

While there are several tools for RoB assessment of prevalence studies, there is no consensus on a definitive tool for use in evidence synthesis [274]. Even the most commonly used tools have heterogeneous evaluation criteria, are time-consuming to use, and yield RoB assessments with inherent variation given a reliance on evaluators' subjective judgments [178, 275, 276]. These limitations present a major challenge for rapid and reproducible synthesis of seroprevalence studies. Meta-epidemiological reviews of prevalence study evidence syntheses demonstrate that a unified, rapid, and reproducible approach to RoB assessment is needed [274].

The Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Prevalence Studies is a validated tool that is the most commonly used for prevalence studies and therefore could form the basis for such a unified approach [274]. However, the JBI tool provides a nine-item checklist with no provision for overall RoB assessment. As a result, this checklist has been operationalised differently by groups conducting evidence synthesis, with each effort defining their own approach to derive overall RoB assessments [277–279]. In some instances, this has resulted in markedly different RoB assessments for the same body of underlying literature [102, 277].

We aimed to develop an objective, rapid, and reproducible approach to RoB assessment for seroprevalence studies. In this Chapter, we describe the SeroTracker-RoB automated approach, which operationalises the JBI checklist for seroprevalence studies, automates much of the JBI checklist by applying objective criteria to routinely generated data, and automates overall RoB assessments using decision rules based on the checklist item responses. We evaluated the efficiency, coverage, and reliability of this approach, and developed a tool that implements this automated approach for use by other researchers.

6.2.1 Operationalising and automating the JBI checklist

SeroTracker has used the JBI checklist since its inception to critically appraise seroprevalence studies. We also selected the JBI checklist as the basis for this approach as a validated and commonly used critical appraisal tool for prevalence studies. The original JBI checklist involves reviewers judging nine categorical items based on data reported in a prevalence study, described in full in the first column of Table 6.1.

We operationalised the JBI checklist for seroprevalence studies, developing reviewer guidance for each criterion to make them more specific to the seroprevalence context, improve reproducibility, and ease the burden of decision-making (6.1). Specifically, four of the nine JBI checklist items (items 4, 6, 8, 9) were modified to add conditions that enabled them to be judged in a binary fashion. For example, Item 4 (study subjects and setting described in detail) was marked as 'Yes' if the average age and distribution of gender/sex was provided, and was marked 'No' otherwise. Item 8 (appropriate statistical analysis) was split in two parts ($8A$, $8B$) to differentiate studies that adjusted prevalence estimates for test characteristics (sensitivity/specificity) but not population demographics and vice versa.

We categorised the resulting items as relating to sampling biases, relating to measurement biases, or reporting items (Table 6.1). Sampling bias items captured the extent to which the seroprevalence estimated in the study sample risked being biased away from the seroprevalence in the target population (items 1, 2, 3, 5, 8B, and 9). Measurement bias items captured the extent to which the seroprevalence estimate reported in the study risked being biased away from the true seroprevalence (items 6, 7, and 8A). Lastly, reporting items were those related to reporting alone, which therefore had no bearing on study risk of bias (item 4).

Table 6.1: Risk of bias table fields. The first column describes the Joanna Briggs Institute Critical Appraisal Checklist for seroprevalence studies [178]. The second includes the SeroTracker-RoB operationalisation of this checklist. Checklist items are reorganised into three categories: those relating to sampling biases, those relating to measurement biases, and those related to reporting only. Item 8 was split into two sub-items: 8A, relating to measurement biases, and 8B, relating to sampling biases.

Checklist item	**SeroTracker-RoB operationalisation**
Checklist items relating to sampling biases	

Item 1: Was the sample frame appropriate to address the target population?	A serosurvey's sample frame is the set of individuals that it could have sampled from – that is, the full set of participants or specimens that could have been sampled. The target population is the population that the study aims to make inference about. • Yes: Sample frame described and approximated the target population (e.g., sampling households to estimate seroprevalence in the general population) • No: Sample frame did not describe or did not approximate the target population (e.g., sampling blood donors to estimate seroprevalence in the general population)
Item 2: Were study participants recruited in an appropriate way?	• Yes: Probability sampling • No: Non-probability sampling • Unclear: Sampling method unclear
Item 3: Was the sample size adequate?	Adequately powered serosurveys have the sample size required to detect low seroprevalence with reasonable precision – particularly important given that immunoassays have imperfect specificity and that the extent of SARS-CoV-2 infection has been very heterogeneous globally. We use a threshold of $n = 599$, which is the sample size required to detect a seroprevalence of 2.5% to a precision of 1.25% with 80% power. • Yes: $n \geq 599$ or sample size calculation provided and showed precision $\leq 1.25\%$ • No: $n \leq 599$ • Unclear: Sample size calculation not reported
Item 5: Was data analysis conducted with sufficient coverage of the sample?	To draw inference about the population, it is crucial that statistical analysis includes a sample that represents the population • Yes: In both main and subgroup analysis, the demographics of the sample analysed are at least somewhat representative of the population • No: Does not meet the standard for Yes Requires subjective review of demographic breakdown in the study, coverage within subgroup estimates, and review of author comments on representativeness of the sample
Item 8: Was there appropriate statistical analysis?	Sampling bias-specific sub-item 8B: Was there appropriate statistical analysis for target population and sample differences? • Yes: Conducted statistical adjustment for differences between population and sample or Items 1, 2, and 3 were all Yes • No: Did not conduct statistical adjustment for differences between population and sample and at least one of Items 1, 2, and 3 was No • Unclear: Analysis method unclear

Item 9: Was the response rate adequate, or was a low response rate managed properly?	• Yes: Response rate $\geq$ 60%, or the demographics of the sample reasonably approximated the demographics of the target population • No: Response rate $\leq$ 60% and the demographics of the sample did not reasonably approximate the target population • Unclear: Response rate not reported
Checklist items relating to measurement biases	
Item 6: Were valid methods used for the identification of the condition?	Seroprevalence studies use immunoassays to identify the presence or absence of antibodies. • Yes: The serological test used met the WHO UNITY criteria for serological assays (sensitivity $\geq$ 90%, specificity $\geq$ 97%)[53] • No: Sensitivity and specificity reported, but do not meet the criteria for Yes • Unclear: Sensitivity or specificity not reported Where multiple assays were used, their combined sensitivity and specificity must meet this threshold. Virus neutralisation assays were excepted from this requirement, as they constitute the gold standard in serological evaluation [280]
Item 7: Was the condition measured in a standard, reliable way for all participants?	• Yes: Same assay used for all participants • No: Assay varied between different participants • Unclear: No details about which participants were tested using which assay
Item 8: Was there appropriate statistical analysis?	Measurement bias-specific sub-item 8A: Was there appropriate statistical analysis for test performance? • Yes: Conducted statistical adjustment for differences between population and sample or Items 1, 2, and 3 were all Yes • No: Did not conduct statistical adjustment for differences between population and sample and at least one of Items 1, 2, and 3 was No • Unclear: Analysis method unclear
Checklist items relating to reporting	
Item 4: Were the study subjects and setting described in detail?	• Yes: Average age and gender breakdown provided • No: Does not meet the standard for Yes

As a first step towards reproducible RoB evaluations, we then sought to automate completion of each checklist item. We identified items that could be automatically completed based on key information that should be extracted as part of any critical appraisal or systematic review of seroprevalence studies [77], and included this in our

operationalisation (6.1).

Seven of the nine JBI checklist items were fully automated in this way. However, item 1 (whether the sample frame was representative of the target population) and item 5 (whether the characteristics of the sample were representative of the target population, in both the main and subgroup analyses) are judgements that require study, sample, and target population specific context. We thus did not automate these items, instead choosing to provide detailed guidance and illustrative examples to improve the reliability of manual assessment (Table 6.1). Additionally, a small proportion of studies used complex testing algorithms involving multiple assays with no clear overall sensitivity or specificity; for these studies the validity of methods to identify the condition (Item 6) was determined by manual review with clear guidance.

This yields a seroprevalence-operationalised version of the JBI Critical Appraisal Checklist for Prevalence Studies, which can almost entirely be automatically completed based on extracted data from a given seroprevalence study.

6.2.2 Developing decision rules for risk of bias assessment

There is no consensus on the most valid approach to automate RoB assessments [274]. Some studies have used summary scores, adding the number of criteria met for a given critical appraisal checklist and bucketing that score to assess RoB [277–279]. However, this approach weights each checklist item equally, which does not reflect the different implications of critical appraisal concepts for RoB. For example, in the context of the JBI checklist, marking six out of nine JBI items Yes could be achieved by 84 different combinations of item responses, each having different implications for RoB [280–282].

Weighted averages better reflect the relevance of each item, overcoming some limitations of simple summary scores. However, checklist items cannot always be considered independently in assessing RoB, thereby introducing complexity in the derivation of a weighted score. For example, it is more important to correct for antibody test performance (item $8B$ = Yes) when assay sensitivity and specificity are low (item 6 = No).

Considering combinations of items together may enable better automated RoB assessment. Some studies have trained deep learning algorithms to identify relevant text in

publications of randomised controlled trials and predict RoB assessments, with reasonable accuracy compared to human reviewers [280, 281]. However, these algorithms have largely been trained on small datasets and have limited interpretability and transparency.

Decision rules allow features to have different importance, permit interactions between checklist items, and are interpretable. For this reason, we moved to develop a decision rule that could be automatically applied to the ratings from the JBI checklist to generate an overall risk of bias assessment for each study. This assessment captures the risk that the seroprevalence estimates from this study are biased away from the true seroprevalence.

The decision rules were developed based on published guidance on estimating disease prevalence, reports on the evaluation of prevalence studies [274, 283], opinions of experts in evidence synthesis and infectious disease epidemiology, and the consensus of researchers at SeroTracker after evaluation of over 2,000 seroprevalence studies [9, 100]. The decision tree to classify overall study RoB is shown in Figure 6.1. The tree included 61 decision rules, which could lead to low ($n = 6$ rules), moderate ($n = 29$ rules), or high RoB ($n = 26$ rules).

The derived decision tree for overall RoB assessment was based on the extent of sampling and measurement biases present in a study, and was not affected by any reporting biases. Studies were considered to be:

- Low RoB if sampling biases were limited (i.e., appropriate sample frame, probability sampling, adequate sample size, and statistical adjustment for population characteristics), and measurement biases were also limited (i.e., adequate sensitivity and specificity and/or adjustment for test performance)

- Moderate RoB if they had a few sampling biases (e.g., inappropriate sample frame but probability sampling, adequate sample size, and statistical adjustment for population characteristics), or limited measurement biases

- High RoB if they had sampling biases (e.g., non-probability sampling and inadequate sample size), and measurement bias was considerable (i.e., poor sensitivity or specificity and lack of adjustment for test performance)

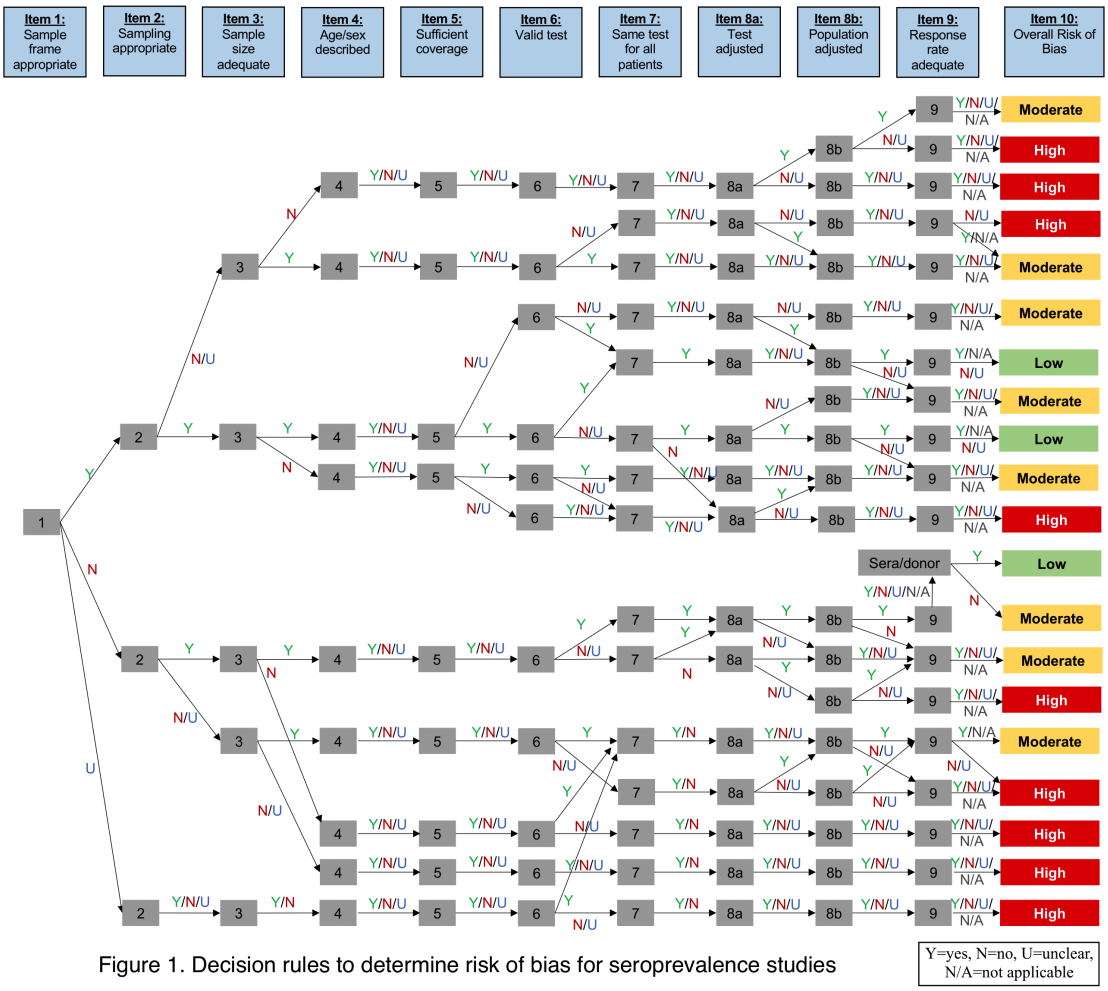


Figure 6.1: Decision rules to determine risk of bias for seroprevalence studies.

This decision rule approach provides a transparent and interpretable model to automate RoB assessment [282]. The tree structure captures interactions between checklist features, is easy to implement, and enables clear visualisation of which combinations of features are most important for RoB. The decision tree in the SeroTracker-RoB automated approach reveals sampling and measurement biases as the two key axes of bias in seroprevalence studies. Clusters of rules in this decision tree also make clear how features in each axis interact to determine study RoB.

6.2.3 Evaluating coverage, reliability, and efficiency

We evaluated three aspects of the SeroTracker-RoB approach: its coverage and ability to evaluate risk of bias for studies in the SeroTracker database, its reliability with respect to manual human review, and the efficiency improvements it offered.

6.2.3.1 Coverage evaluation

The dataset used for validation included all 2,070 studies in SeroTracker published between 1 January 2020 and 17 November 2021. To determine coverage, we attempted to apply the SeroTracker-RoB decision rules to each of these studies to assess the proportion for which the automated process yielded a RoB assessment. This approach yielded a RoB assessment of low, moderate, or high for 100% of studies in SeroTracker.

6.2.3.2 Reliability evaluation

We sought to examine the agreement between manual reviewer RoB assessments and the SeroTracker-RoB approach. Manual RoB assessments had previously been completed for each included study by two independent reviewers. We computed automated SeroTracker-RoB assessments for each study per the process above.

To evaluate reliability between the ordinal manual RoB assessment and the SeroTracker-RoB assessment, we calculated the intraclass correlation [284–286]. The underlying model for this intraclass correlation calculation was an ordered logistic regression model with random effects for the measurement tool and for the study. We also evaluated reliability between the two independent reviewers conducting manual assessment.

The intraclass correlation between the SeroTracker-RoB automated approach and consensus manual review was 0.77 (95% confidence interval [CI] $0.74 - 0.80$), indicating excellent reliability. For comparison, the intraclass correlation between the two manual independent reviewers was 0.74 (95% CI $0.71 - 0.76$). This intraclass correlation of 0.77 indicates excellent reliability [287] for the SeroTracker (ST)-RoB approach compared to manual review, emphasising the validity of this approach.

In traditional manual RoB assessment, there is imperfect agreement between even trained human reviewers [275, 288–290], as is reflected in the moderate intraclass correlation of 0.74 that we observed between two manual reviewers in the SeroTracker review. As such, manual review cannot be considered a perfect standard, and the 0.77 intraclass correlation between the SeroTracker-RoB automated approach and manual review in part reflects the heterogeneity and inconsistency of manual assessment. This highlights the benefit of an automated approach that yields reproducible assessments.

6.2.3.3 Efficiency evaluation

RoB assessment is a time-consuming component of conducting systematic reviews [289, 290]. As such, improving efficiency of the assessment process may help to reduce the burden of evidence synthesis.

We assessed efficiency by comparing time to complete manual RoB and automated SeroTracker-RoB assessments for 25 seroprevalence articles that were randomly selected from the SeroTracker dataset. Five SeroTracker systematic reviewers were timed as they each completed assessments for five articles. We compared the speed of manual versus the automated SeroTracker-RoB assessment for each article using a paired t-test.

Use of the SeroTracker-RoB automated approach resulted in significantly faster RoB assessments compared to the manual approach with a mean and standard deviation time of 0.80 (standard deviation 0.53) versus 2.93 (standard deviation 1.08) minutes per article, respectively ($p < 0.001$). The automated approach took less time than the manual approach for all 25 articles evaluated.

On average, 35 new studies are added each week to SeroTracker. The SeroTracker-RoB automated approach is over five times faster than traditional manual review for assessing seroprevalence study RoB, and would save nearly three hours of reviewer time each week compared to the manual approach. This is particularly important for living reviews and reviews during health emergencies, which need to process a high volume of information rapidly to enable action.

6.2.4 Discussion

We developed the SeroTracker-RoB automated approach for RoB assessments of seroprevalence studies. When validating this approach against the SeroTracker database of over $2,000$ seroprevalence studies, it was over five times faster than manual review, had perfect coverage, and had good agreement with manual review from two independent reviewers.

There have been several systematic reviews of SARS-CoV-2 seroprevalence studies to date [9, 70, 102, 119, 277–279]. Several of these reviews use the JBI checklist for critical appraisal, but they implement the checklist and judge overall RoB in different

ways [9, 70, 102, 277–279]. This approach may be valuable to other investigators seeking to conduct rapid and reliable RoB assessments for seroprevalence studies, and improve consistency between evidence synthesis experts — an important endeavor, given the ongoing value of SARS-CoV-2 seroprevalence studies for surveillance [291], and given the many (sero)prevalence studies conducted for other diseases and conditions. This tool has clear potential to enable rapid, robust, and reproducible RoB assessment in systematic reviews of prevalence studies. Developing decision rules may enable rapid and reproducible risk of bias assessment for evidence syntheses of other study designs.

Our study had several strengths. First, our validation dataset included thousands of heterogeneous SARS-CoV-2 seroprevalence studies across study designs, regions, and target populations. Second, the use of a decision rule approach enables transparent RoB assessment while also ensuring reproducibility and speed. Finally, the decision rules have a robust foundation in the validated and widely used JBI checklist, published guidance on estimating disease prevalence, opinions of experts, and the experiences of researchers that have conducted thousands of RoB assessments for seroprevalence studies.

This approach retains some limitations. Our approach still requires reviewers to make expert judgments on two JBI items. Moreover, the decision rule was derived in part using expert judgment. The transparency of this algorithm allows for scientific debate and further refinement of the decision rules, if needed. Finally, the generalisability of the SeroTracker-RoB decision rules beyond SARS-CoV-2 seroprevalence data is unclear. Applying these specific rules to seroprevalence studies for other pathogens or conditions may require additional validation or adaptation.

6.3 Real-world assessment of NLP-based screening

In systematic reviews, under 3% of abstracts screened are typically eligible for inclusion [292], making it time-consuming to parse through search results, and particularly useful to expedite this stage of the process. Software tools using natural language processing (NLP) models have been developed to accelerate screening, classifying abstracts for inclusion based on a sample trained by human reviewers [293–299].

The classification performance of such abstract screening models has been extensively studied and optimised [300, 301]. Some studies have also assessed efficiency, concluding that use of NLP models could improve screening efficiency – both in retrospective evaluations of completed reviews [295, 296, 299, 302], and in ongoing reviews [294, 303, 304].

Despite this, there has been little study of the process of implementing these tools in the real world. Assessments of implementation have focused almost exclusively on screening time, with only one study considering user satisfaction [305]. Furthermore, only one study has examined NLP-based screening in the context of a living systematic review [297]. These living reviews can regularly benefit from improved screening efficiency, and lend themselves well to the implementation of NLP tools: an initial search and manual screen can yield a large number of screened texts, which could serve as training data for a model that in turn expedites ongoing updates.

SeroTracker screens 800 to 1,000 abstracts each week, which is a key contributor to reviewer workload. Reviwers on our team identified the time-consuming nature of abstract screening in general. They also identified challenges with the Covidence platform SeroTracker uses for abstract screening [306], noting that it is difficult to track the number of abstracts screened, reverse misclicked votes, and determine inclusion.

To optimise the efficiency of our screening process, we developed an NLP-assisted software tool, and assessed the efficiency changes and usability of integrating this tool into our usual methods. We evaluated changes in the time taken to conduct screening, the accuracy of the screening process, the characteristics of included texts in our overall review, user interactions with the tool, and user satisfaction with the process. We also assessed different algorithms to combine user and tool screening results to determine how to best improve the screening process. By implementing this NLP-based tool in an ongoing living systematic review, we provide novel and comprehensive evidence regarding the feasibility of NLP for screening and its real-world performance benefits.

This effort was conducted as a quality improvement project [307, 308]. We received a research ethics board exemption for this project from the Conjoint Health Research Ethics Board at the University of Calgary, as quality assurance is not considered human subjects research under the TriCouncil Policy Statement 2018 (Chapter 2, Article 2.5 [309]).

6.3.1 Developing a NLP-enabled tool to expedite abstract screening

To improve SeroTracker's screening process based on the reviewer needs, we developed an NLP-enabled tool that adds additional features to Covidence. The tool included two NLP-based features to allow more efficient identification of text eligibility: (1) an inclusion recommendation indicator, which displays a confidence rating ranging from 'not recommended' to 'strongly recommended' in the form of a coloured circle beside the abstract title, and (2) a feature that annotates the key elements of observational study design (population, intervention, and outcome) in different colors in the abstract (Figure 6.2). The tool also incorporated features to streamline screening and improve user experience: (3) a screening progress tracker, (4) a button to undo a user's most recent votes on a text, (5) a feature that displays abstracts in a way that visually separates them by section headings (e.g., 'Background', 'Methods', etc.), and (6) a feature highlighting reviewer-specified keywords (Figure 6.2).

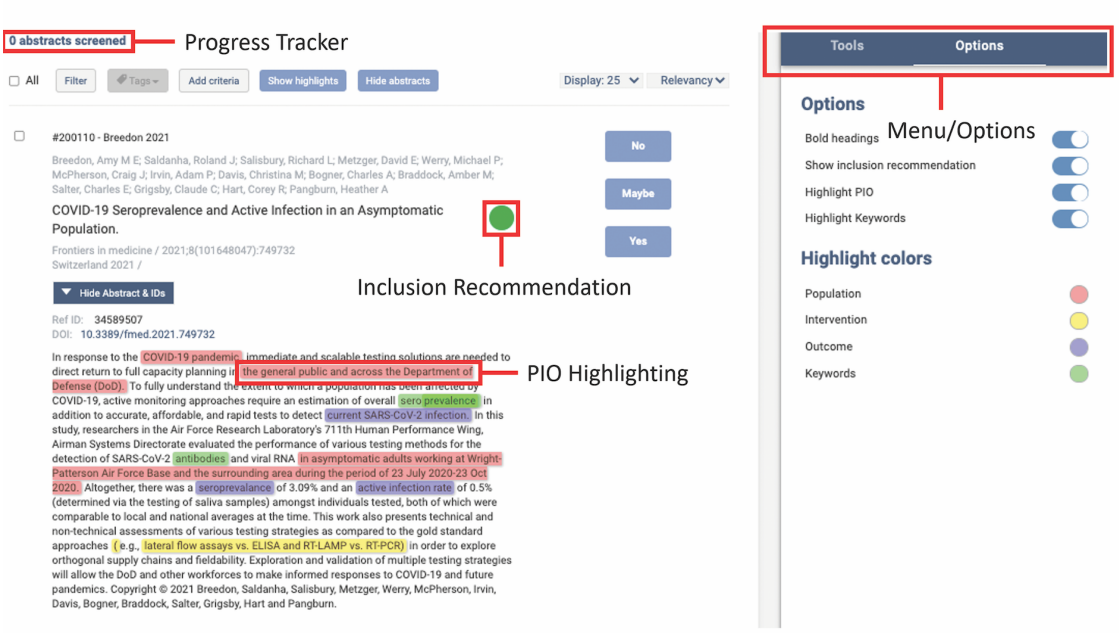


Figure 6.2: The NLP-enabled tool augmenting Covidence abstract screening. PIO: population, intervention, and outcome. Key features shown include progress tracking, inclusion recommendations, and population, intervention, and outcome annotation. This tool was implemented as a Google Chrome extension for Covidence. To reduce latency, we generated and cached inclusion recommendations and key element annotations on a remote Amazon Web Services server, and queried these as needed by the locally installed extension.

When selecting what NLP model architecture to use to develop both NLP-based

features, we reviewed previous models for systematic review inclusion prediction. Previously published models include Rayyan [298], DistillerSR [296], and ResearchScreener [294], which use naive Bayes or n-gram support vector machine approaches. Recently, transformer-based NLP models have shown particular promise for text screening [310]. Transformer models are typically pretrained on large corpuses of text then fine-tuned on a specific task. This results in broadly improved performance [311, 312]. Considering this observation, we used PubMedBERT – a domain-specific pretrained transformer-based NLP model extensively pretrained on the biomedical literature – as a baseline model for both NLP-based tool features [312].

We next identified the training, validation, and real-world evaluation datasets for this model. As of early July 2021, when the model was trained, SeroTracker reviewers had manually classified 38,674 abstracts. Of these, 95.2% were deemed ineligible in abstract screening (i.e., negative), 1.4% were rejected in full text review (also negative), and 3.4% were moved to the extraction stage (i.e., positive). Of these, 400 were set aside for real-world evaluation, 26% were set aside as a validation set, and 74% (i.e., 25,000 abstracts) were used in the training set.

To preprocess abstracts for the model, we concatenated the article title and abstract, and removed any newlines, tab characters, or other delimiters found in the data. If the resulting string exceeded the 512-token maximum for the model, we removed any excess. Considering our highly imbalanced dataset, we found that upsampling the positive samples in the training set by a factor of 25 significantly improved recall at the cost of some precision – a worthwhile trade-off, considering the nature of this screening task and the cost of omitting an abstract. We fine-tuned an ensemble of 5 pretrained models on independent random samples of 60% of the training set. We trained for 4 epochs, using a batch size of 32 and a learning rate of 10^{-5}. The predictions of each model were averaged to yield the final inclusion likelihood.

Table 6.2 summarises the performance of the tool evaluated on the validation set at various classification thresholds. We first report the 50% threshold, which is the default for most classification tasks. At this threshold, the F1-score was 0.818, with a better recall of 0.871.

Table 6.2: Operating characteristics of the tool at different confidence thresholds on the test set of abstracts. Abbreviations as follows: TN: true negatives; FP: false positives; FN: false negatives; TP: true positives; Prec.: precision; Rec.: Recall; F1: F1 score.

Threshold	TN	FP	FP	FP	Prec.	Rec.	F1
Standard classification (50%)	9,756	76	38	257	0.772	0.871	0.818
Weakly recommended (15%)	9,416	416	11	284	0.406	0.963	0.571
Somewhat recommended (35%)	9,718	114	29	266	0.700	0.902	0.788
Recommended (60%)	9,775	57	41	254	0.817	0.861	0.838
Strongly recommended (75%)	9,801	31	61	234	0.883	0.793	0.836

The calibration curve reveals that model inclusion likelihoods are not well calibrated (Figure 6.3). As a result, we move to use four recommendation thresholds, corresponding to 15%, 35%, 60%, and 75% model predictions, as opposed to directly reporting the inclusion recommendation score to the user. These thresholds are plotted in Figure 6.3, with performance characteristics enumerated in Table 6.2.

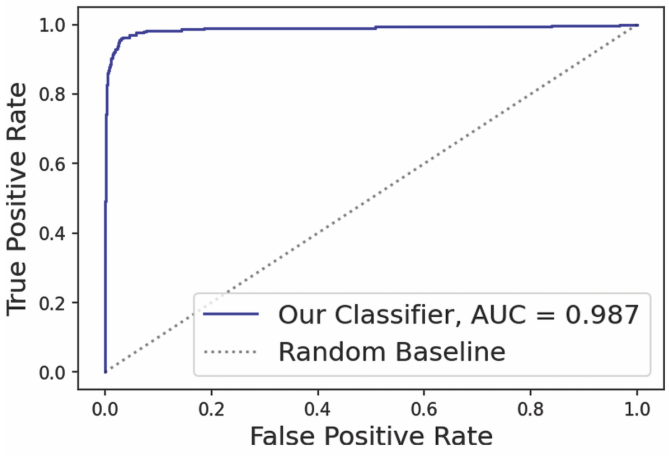


Figure 6.3: Calibration curve of the inclusion recommendation system. As evaluated on the set of validation abstracts.

At the lowest recommendation threshold, we achieved a recall of 0.963 at a precision of 0.406 (Table 6.2), implying that nearly all true positive are captured by the model. We chose this precision-recall trade-off to prioritise model sensitivity in this early-stage screening task, as abstracts that are not relevant can be excluded later. This threshold achieves an area under the receiver operating characteristic curve of 0.987 (Figure 6.4).

To highlight the population, intervention, and outcome of each article, we fine-tuned

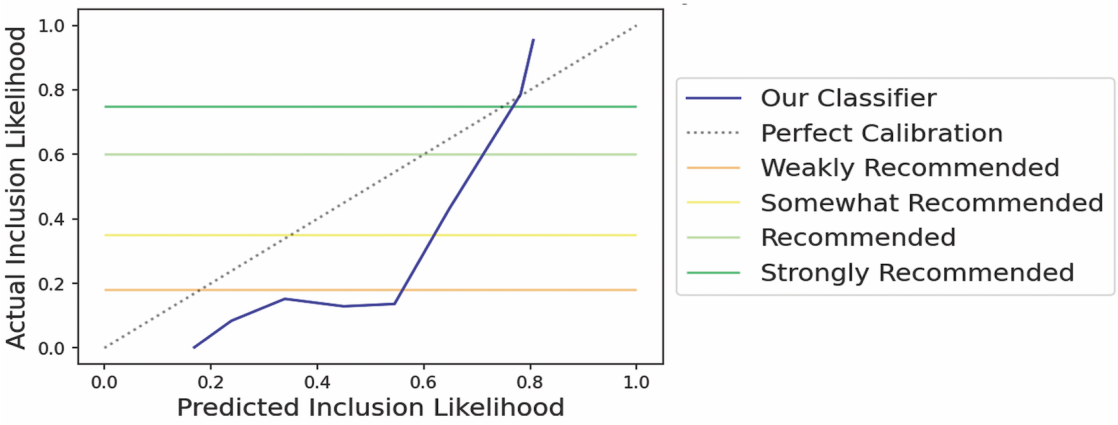


Figure 6.4: Receiver operating characteristic curve of the inclusion recommendation model. Shows inclusion threshold 15%. Evaluated on the set of validation abstracts.

the same PubMedBERT model on the EBM-NLP dataset, which contains 5,000 abstracts annotated with their population, intervention, comparison, and outcome components [313]. To highlight long abstracts, we split the abstracts into sets of sentences at most 512 tokens long, fed these sets into the model separately, and combined the results of the two sets. We highlighted the three contiguous blocks of text that had the highest average probability for each of population, intervention, and outcome.

6.3.2 Real-world evaluation of the NLP-enabled tool

We conducted a quality improvement project to assess the feasibility and impact of tool implementation on abstract screening. This project was conducted over a five-week period in three stages. In the first two weeks, reviewers conducted screening without the tool ('without-tool stage'). 200 pilot abstracts were randomly intermixed with SeroTracker's regularly weekly abstract screening load. We subsequently implemented a weeklong washout period, where no pilot abstracts were added, and reviewers installed and familiarised themselves with the tool. In the final two weeks, reviewers used the tool for screening ('with-tool stage') using the features they felt were most helpful. In this stage, 200 pilot abstracts were again randomly intermixed into the regular weekly search results.

We selected 400 abstracts from April 2021 (the 'pilot abstracts') to validate the performance of the tool in the real world. We chose this April 2021 date to minimise the chance that reviewers would recall these abstracts, while still keeping the abstracts contemporary to what would have been seen in a typical week when real-world evaluation

was done. Of these abstracts, we selected 91 that were extracted and 309 that were excluded at screening, full-text review, or extraction. We deliberately enriched for positive abstracts to improve class balance during real-world evaluation.

Three sets of reviewer votes on the pilot abstracts were collected: votes from the initial screen in April ('pre-project votes'), votes from the without-tool stage, and votes from the with-tool stage. We used pre-project votes as the reference standard for comparison.

We evaluated key process, outcome, and structure measures. Process measures, which relate to the day-to-day review processes, included efficiency and accuracy on abstract screening. Outcome measures included the overall results of the systematic review and user satisfaction with the tool. Finally, we evaluated one structure measure compared to how model inclusion recommendations and key votes were combined. A summary of the process, outcome, and structure measures studied, the approach to study each measure, and the results of each evaluation is available in Table C.2.

6.3.3 Process measures: screening efficiency and accuracy

Process measures included efficiency and accuracy metrics. We focus first on efficiency metrics: the screening time and the conflict rate with and without the tool.

The tool was associated with a 33.7% decrease in mean screening time per abstract from mean 21.45s (standard deviation 22.30s n = 2,852) to 14.22s (standard deviation 17.52s; n = 2,961, $p < 0.001$ by two-sided unequal-variances t-test). This reduction was similar when considering the pilot abstracts alone (mean 22.78s, standard deviation 22.59s, $n = 746$ without tool to mean 14.00s, standard deviation 18.04s, $n = 736$ with tool; $p < 0.001$). Included abstracts took longer to screen regardless of tool use (mean 45.67s, standard deviation 32.37s, $n = 119$ without tool vs. mean 41.25s, standard deviation 32.43s, $n = 159$ with tool; $p = 0.264$) indicating that the improved screening time was driven by faster abstract exclusion.

An increased number of conflicting votes decreases the efficiency of abstract screening, as a third reviewer must resolve these. We also assessed the tool's impact on conflict rates. The conflict rate decreased from $^{163}/_{1,960} = 8.32\%$ without the tool to $^{87}/_{2,388} = 3.64\%$ with the tool ($p < 0.001$, Fisher's exact test). A similar decrease was observed among

the pilot abstracts: $^{45}/_{400} = 11.25\%$ to $^{23}/_{400} = 5.75\%$ ($p = 0.007$). This conflict rate reduction reduces the mean number of votes required per abstract from $2 + {}^{8.32}/_{100}$ to $2 + {}^{3.64}/_{100}$, a 2.2% time savings. Including both the mean screening time and conflict rate improvements, we expect a 35.2% reduction in screening time per abstract, leading to 3.93 hours of reviewer time saved each week.

Process measures also included accuracy, measured using precision and recall of screening, as well as the performance of the tool's inclusion recommendations. We first note that there is inherent human error in the systematic review process. [314]. To evaluate precision and recall, we hence first calculated the baseline expected variability due to human error in screening, by comparing the included and excluded texts between the pre-project and without-tool stages. We then assessed whether the outcomes of the with-tool stage were within expected levels of variability.

When considering pilot abstracts that were included past full-text review, 74 of the 91 pre-project included abstracts were included without the tool and 82 with the tool (Table 6.3). This change was not significant ($p = 0.137$, Fisher's exact test), meaning accuracy was conserved.

Table 6.3: Precision, recall and F1-scores in all user-tool pairings for texts that were included in full-text screening. Reference: ground truth established in the pre-project stage. Abbreviations: OPOT: one person, one tool; OPOT-W/O: one vote from a human reviewer without tool access, one vote from the tool; OPOT-W: one vote from a human reviewer with tool access, one vote from the tool

Scenario	TN	FP	FP	FP	Prec.	Rec.	F1
Without	299	10	17	74	0.881	0.813	0.846
With	302	7	9	82	0.921	0.901	0.911
OPOT-W/O	299	10	12	79	0.888	0.868	0.878
OPOT-W	301	8	7	84	0.913	0.923	0.918
Tool-Only	301	8	6	85	0.914	0.934	0.924

6.3.4 Outcome measures: review results and tool usage

The first outcome measure evaluated was the tool's impact on review results, assessed by comparing summary descriptive statistics for included seroprevalence estimates in the pre-project, without-tool, and with-tool stages (Table 6.4). These remained fairly consistent across stages.

Table 6.4: Summary statistics of the seroprevalence estimates from the included texts in the pre-project, without-tool and with-tool stages.

Characteristic	Pre-Project (n=308)	Without Tool (n=287)	With Tool (n=297)
Geographic Scope			
Local	69 (22%)	50 (17%)	59 (20%)
Regional	220 (71%)	218 (76%)	219 (74%)
National	19 (6.2%)	19 (6.6%)	19 (6.4%)
World Bank Income Level			
High income	286 (93%)	269 (94%)	275 (93%)
Low income	2 (0.6%)	2 (0.7%)	2 (0.7%)
Lower-middle income	7 (2.3%)	7 (2.4%)	7 (2.4%)
Upper-middle income	13 (4.2%)	9 (3.1%)	13 (4.4%)
Risk of Bias			
High	89 (29%)	72 (25%)	78 (26%)
Low	10 (3.2%)	9 (3.1%)	10 (3.4%)
Moderate	204 (66%)	201 (70%)	204 (69%)
Unclear	5 (1.6%)	5 (1.7%)	5 (1.7%)
Sampling Method			
Non-probability	288 (94%)	268 (93%)	278 (94%)
Probability	20 (6.5%)	19 (6.6%)	19 (6.4%)

Our second outcome measure was user interaction with the tool. We assessed the frequency of voter agreement with the inclusion recommendations. Users typically agreed with the recommendations, with an area under the receiver operating characteristic curve of 0.956 (Figure 6.5a) and the highest F1-score for agreement (0.821) being at the 'somewhat recommended' threshold (Figure 6.5b). This suggests that users very strongly voted in alignment with the tools recommendations.

We also examined reviewer use of tool features.The inclusion recommendation feature was used by $^{7}/_{7}$ users, population, intervention, and outcome annotation by $^{5}/_{7}$, keyword highlighting by $^{3}/_{7}$, and the undo feature by $^{3}/_{7}$.

Finally, we surveyed users to understand satisfaction with the tool. After the project's conclusion, reviewers were sent a survey to rate the usefulness of features on a five-point scale. 9 of 13 members provided feedback (Table 6.5). Reviewers who used features rated

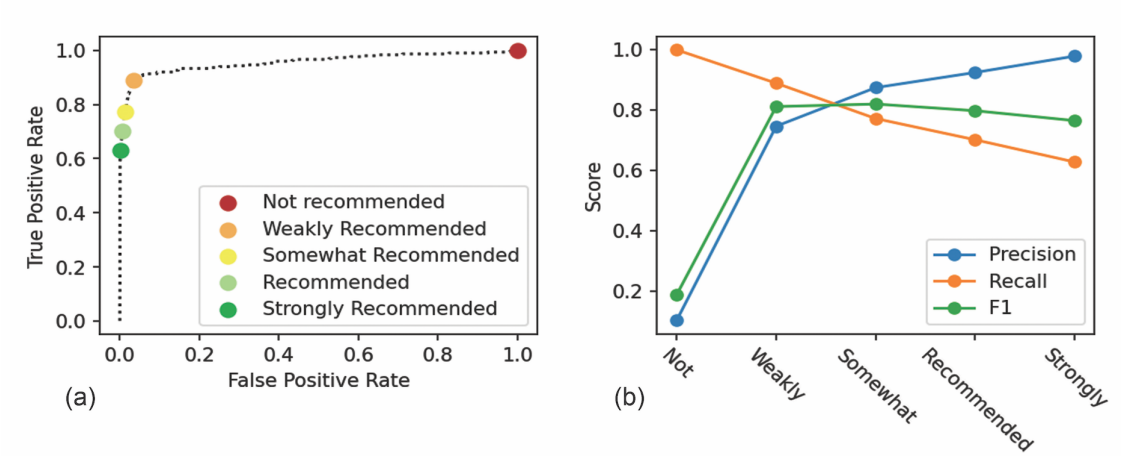


Figure 6.5: Operating characteristics of the tool compared to user votes. Comparison: true labels from user votes. (a) shows the operating curve and five confidence levels. (b) shows how the precision, recall, and F1-score vary with confidence threshold.

their usefulness on a five-point scale. The inclusion recommendation feature was voted most useful (mean score of 4.70/5), and keyword (3.88/5) and population, intervention, and outcome annotation (4.00/5) were least useful (Table 6.5).

Table 6.5: Tool feature usefulness as rated by nine users.

Feature	Mean usefulness (individual scores)
Undo	4.40 (3,4,5,5,5)
Inclusion recommendations	4.70 (4,4,5,5,5,5,5)
Bold headings	4.13 (3,3,3,4,5,5,5,5)
Keyword highlighting	3.88 (3,3,3,4,4,4,2,2)
population, intervention, and outcome annotation	4.00 (3,3,4,5,5)

Reviewers reported that the tool allowed them to rapidly identify key information bearing on abstract inclusion through the bold headings, population, intervention, and outcome annotation, and keyword highlighting. While many users found the inclusion recommendations useful, they noted that it could provide a sense of security that may lead users to blindly trust the tool. Users also noted that the population, intervention, and outcome annotation feature often highlighted irrelevant information, leading to its low adoption.

6.3.5 Structure measures: one-person one-tool voting model

Given the reliability of the tool's inclusion recommendations, we evaluated outcomes of an abstract screening process where one human reviewer was replaced with the tool's vote (OPOT: one-person and one-tool), in a scenario in which the remaining human reviewer did not have access to the tool (OPOT-W/O) and in one where they did (OPOT-W). We compared this to a tool-only system for abstract screening.

Results are reported in Table 6.3 for texts that were included in the abstract screen and for texts that were included in the full-text screen. Recall for included texts was improved from 0.813 without-tool to 0.868 in the OPOT-W/O model, and from 0.901 with-tool to 0.923 in a OPOT-W setting. OPOT would result in a further 48.6% reduction in screening time, for a potential total time saving of 72.2%. The tool-only screening scenario performed comparably to both OPOT and to two humans (Table 6.3).

6.3.6 Discussion

We improved our screening processes by implementing an NLP-based tool. Tool use significantly reduced both the mean time taken per abstract and the screening conflict rate, leading to an overall decrease in 3.93 hours of weekly screening time. The improved efficiency did not come at the cost of accuracy: precision and recall of the screening process, as well as summary statistics of the included estimates, were similar with and without the tool. Users provided positive feedback overall to using the tool.

One-person-one-tool and tool-only screening systems, which have the potential for increased time reductions, performed as well as two human reviewers. The results of the OPOT analyses are particularly promising, as this system can reduce screening time by over 70%. Though the tool-only system performed comparably well to the OPOT system in our analyses, further study may be required before the research community is prepared to adopt a screening practice without a human in the loop. We have since integrated OPOT as regular practice in our work.

The living systematic review in particular provides an ideal context for the implementation of such tools, as results of an initial manual review can be used as a training set to develop a tool that would expedite all future iterations. Other groups working on

living systematic reviews would benefit from integrating such tools. Our QI project had a robust design, evaluated a comprehensive set of metrics, and can be used as a model.

It is important to mention a few limitations of this work. First, the AB project design could have induced order effects. While there were very few instances where the same reviewer saw the same abstract twice, we could not definitively demonstrate that order effects did not influence votes, particularly for texts that were accepted into full-text screening and subsequently interacted with multiple times. Second, there is no definitive 'gold standard' for inclusion or exclusion of abstracts. We assumed screening labels from before the project to be accurate. Finally, we examined summary statistics of missed texts in Section 6.3.4, a meta-analysis was not performed to determine the impact on downstream analysis.

This study provides the first comprehensive analysis of the implementation of NLP technology in the screening stage of a living systematic review. Incorporating an internally developed tool at SeroTracker was feasible and significantly improved our processes by increasing screening efficiency while maintaining accuracy. User feedback was positive, leading to continued use of the tool in regular practice. The success of this tool at SeroTracker demonstrates that automation has a key role to play in allowing more up-to-date evidence synthesis, especially during health emergencies.

The ideas presented here accelerated SeroTracker's evidence synthesis effort, enabling us to update our systematic review of seroprevalence data weekly and make this data available directly to users to meet these needs. Collectively, this Chapter provides a collection of tools and ideas that can be used to accelerate knowledge synthesis across disciplines.

7 Discussion and future work

7.1 Motivation

Today's globalised world is highly susceptible to infectious threats. These are becoming even more likely to arise, due to climate change facilitating spillover events [315], and with the acceleration and democratisation of biotechnology [316, 317]. At the same time, our technological toolkit to detect and respond to these threats is constantly growing. Some illustrative examples of this include the advent of multiplexed antibody testing [318], metagenomic sequencing [319], CRISPR-based diagnostics [320], and wearable sensors [321].

This represents a delicate and constantly evolving balance between escalating threats and more effective responses. It is crucial that we tip this balance in favour of public health, which we can do by effectively using these new technologies. However, bringing these technologies into practice is a major bottleneck: the COVID-19 pandemic has demonstrated that our ability to effectively engineer the implementation of these technologies is just as important as the technology itself.

Today's infectious disease and public health landscape demands that we develop solutions to engineering problems. SeroTracker is one realisation of this engineering approach to public health. In this chapter, we summarise the aims and contributions of this thesis, place SeroTracker's work in the context of the broader literature, and

highlight promising directions for future inquiry that bridge engineering and public health to improve our ability to respond to future biological threats.

Work from this chapter appears in the following publications. Co-first authors are denoted with * and co-senior authors with **.

Yan T, **Arora RK.** 2022. Research products beyond the research paper: reflections on user-centered evidence synthesis from SeroTracker. *Frontiers in Research Metrics and Analytics. Rahul's contributions:* conceptualization of manuscript and its central points; first draft; revision of the drafted version, in equal proportion to Yan T.

Donnici C*, Ilincic N*, Cao C, ..., **Arora RK.** 2022. Timeliness of reporting of SARS-CoV-2 seroprevalence results and their utility for infectious disease surveillance. Resubmitted minor revisions to *Epidemics. Rahul's contributions:* conceptualization of manuscript and its central points; defining research question and objectives; methodology development; substantial rewriting and revision of the drafted version.

7.2 SeroTracker in context

This thesis draws together elements from three broader fields: public health surveillance data systems, systematic seroprevalence analysis, and computable evidence synthesis. Here, I will briefly summarise the SeroTracker work and its impact while putting it in the context of the broader epidemiology and health security literature. This Section focuses on the broader SeroTracker effort, recognising that each of the individual investigations described in this thesis has been discussed extensively with broader literature contextualisation in the Chapter where it was described.

7.2.1 Public health surveillance data sharing

SeroTracker represents one effort to improve public health surveillance data sharing. However, this data sharing remains an ongoing challenge. One of the early efforts to address this was in the development of the Morbidity and Mortality Weekly Reports journal, which was initially intended to help facilitate the dissemination of key public health and surveillance data [322]. Nearly 100 years have elapsed since that time, and yet a plurality of seroprevalence studies included in SeroTracker were extracted from

journal articles. Journals still remain one of the primary modes by which surveillance data are communicated in the modern world.

Nevertheless, there have been substantial recent evolutions in the sharing of certain types of surveillance data. Genomic surveillance data are a prime example of this. Take, for example, the advent of *GISAID* – the *Global Initiative on Sharing Avian Influenza Data*. This genomic data submission and accession repository made it routine for researchers and public health officials to deposit influenza genome sequences for global analysis [323]. *GISAID* has facilitated data sharing for seasonal influenza, several influenza pandemics, and now COVID-19 pandemic, during which GISAID data has been crucial to accelerate medical countermeasure development [324]. *GISAID* has also enabled the development of platforms like *NextStrain* that use this data to provide valuable analytics [325]

Population-based infectious disease surveillance has not seen the same data revolution. In part, this is driven by the different cultures underpinning genomic surveillance and clinical population-based surveillance. The genomic surveillance community has roots in bioinformatics and the computational sciences, fields in which open data sharing is highly encouraged and normalised and with the repositories to match [326–328]. In contrast, population-based surveillance has roots in clinical medical research, a field where patient privacy is paramount, and where researchers can be highly protective of their research datasets to establish primacy in publishing [329, 330] and where data sharing platforms are more nascent.

In this thesis, we aimed to modernise our approach to seroprevalence as a key facet of population-based infectious disease surveillance. To surmount the limited sharing of granular seroprevalence data during the COVID-19 pandemic, we extracted seroprevalence study results from public reports of research studies. Our secondary data extraction approach meant that we were able to bring together data from all publicly reported seroprevalence studies to develop a global and comprehensive dataset and develop the SeroTracker dashboard. This approach also limited us to information in public reports, meaning that our work was subject to the systematic delays in releasing study reports, could not fully account for the sampling and measurement biases inherent in these

seroprevalence studies, and lacked the access to participant-level results that would have enabled detailed further analysis. This represents a meaningful step towards improved data infrastructure, sharing, and access for population-based surveillance. There remains considerable additional work to be done before population-based surveillance data sharing can match the speed and granularity of genomic data sharing – before this data can be used to its full potential.

7.2.2 SARS-CoV-2 seroprevalence evidence syntheses

We were far from the only group to attempt syntheses of seroprevalence data during COVID-19. Several other groups conducted systematic reviews of SARS-CoV-2 seroprevalence globally [119–121, 279, 331] and in specific regions [277, 332]. Each of these efforts represented important contributions to the literature, with similar and sometimes complementary findings, and with varying degrees of timeliness and robustness. In many ways, these studies highlight the redundancy of efforts that can result from evidence synthesis, and the challenges of making these syntheses available as peer-reviewed journal articles where they cannot easily be used for secondary analysis.

Peer-reviewed scientific papers have become near-synonymous with credible scientific communication, both for primary research and for evidence syntheses [333]. These papers are a key form of scientific discourse, and peer review has an important role in helping validate research claims. However, they are also largely static outputs, and are not necessarily the best format for continuous updates, dynamic data exploration, or secondary data use. Journal publication can also be a slow process [211, 334], hindering responses to a rapidly changing scientific landscape.

Our key contributions relative to these other efforts are threefold. First, we conducted the first systematic review of these studies [81]. Second, our most recent analyses of seroprevalence map how SARS-CoV-2 immunity has changed over time in each global region due to infection and vaccination, unique among published studies [102, 234]. These findings influenced global health policy, being shared with the WHO's highest bodies including the World Health Assembly and the International Health Regulations Emergency Committee for COVID-19. Finally, our open data platform is built on a

standard data architecture, up-to-date and globally comprehensive, and easily accessible, which enables user interaction on the SeroTracker dashboard as well as others' use for their own analyses.

7.2.3 Seroprevalence dashboards and computable evidence synthesis

There were other notable efforts early in the pandemic to synthesise global seroprevalence data in the same up-to-date, interactive way. In April 2020, there was a public spreadsheet of seroprevalence studies being shared on Twitter [335], as a nascent version of this type of effort. Researchers at Stanford University attempted to develop `seroprevalence.com`, a now-defunct website that had the same aims as SeroTracker but which ceased operations due to the laboriousness of this effort [336]. To our knowledge, SeroTracker remains the only such data platform for seroprevalence studies worldwide, driving its use by global health, disease modelling, journalistic, and commercial stakeholders alike (Section 2.4.4.

Several countries embarked on their own efforts. The US National Cancer Institute set out to develop their own such platform, even referring to their effort as a SeroTracker in a generic sense after our own platform had already been made publicly available [337]. Their effort ultimately evolved into the SeroHub, which carries out much the same function as SeroTracker but with a focus on the United States, and which draws on SeroTracker data [338, 339]. SeroTracker itself developed this country-specific functionality for Canada, in partnership with the Public Health Agency of Canada, and for Germany [340, 341], in partnership with the Robert Koch Institute. Other countries developed their own such platforms based on the national seroprevalence studies they conducted [342].

Collectively, these efforts are early realisations of the emerging movement towards computable evidence synthesis, which extends far beyond the present pandemic. Computable evidence synthesis focuses on the use of structured study design and results data to accelerate evidence synthesis, enabling rapid identification of new evidence and potentially even continuous evidence surveillance [343]. There have been some efforts towards establishing computable evidence synthesis knowledge networks in this vein in

the past year [344]. Considering the nuance and detail of seroprevalence data, strategies from computable evidence synthesis hold promise in population-based surveillance data sharing strategies.

7.2.4 Enabling user-centered evidence synthesis

Computable evidence synthesis is a crucial step for evidence synthesis to keep pace with an ever-growing body of scientific literature [345]. Automation tools, like those we have developed here to automate literature identification and risk of bias assessment, are also key to enable this. However, effective evidence syntheses are not just rapid – they are also responsive to user needs.

This emphasises the importance of developing research products tailored towards particular user needs. Research papers are just one type of research product. Conceiving of SeroTracker as such a product, and the user-centered approach it led us to adopt, was crucial to our impact. The user-centered approach is also reflected in other widely-used health data projects, such as *OpenPrescribing* and *Our World in Data* [93, 96]. Ultimately, adopting this approach will make evidence synthesis more effective in its end goal — to help stakeholders act quickly on new information. Here, we discuss lessons learned in developing SeroTracker as a user-centered evidence synthesis product, each of which points at a key area of future work required for user-centric evidence synthesis.

First, it is important to document and learn from tacit knowledge. Each group building a user-centered evidence synthesis shares certain goals: regularly updated data, ongoing synthesis of that data, and usefulness to decision-makers. However, the process of building these products is not well documented. When we we started SeroTracker it was difficult to find practical knowledge on how other data platforms were built. Challenges we faced included managing large numbers of data extraction staff, creating a database that is human-friendly but that can also feed a dashboard, and actively engaging users in product development.

Dashboard development is sometimes described in the literature, but these articles tend to focus on the technical architecture of the system being presented, rather than key decision points in developing the tool [346]. We ultimately relied more on the management

literature, which carefully documents implicit knowledge [347]. Researchers should aim to document how they integrate data systems, visualisation, and decision-makers. To this end, we are now working to document our own learnings, which we hope will help future developers of user-centered evidence syntheses.

Second, evidence synthesis researchers should collaborate on common and computable data platforms. During the COVID-19 pandemic, thousands of systematic reviews were registered – for some research questions, more reviews than primary studies [348]. Of these, a small fraction had to be curated as the highest-quality efforts [349]. One driver of these efforts is the perceived role of publication metrics in hiring and promotion decisions [350]. These incentives can encourage researchers to prioritise producing publishable papers [351, 352], which indirectly conflicts with collaborative and user-focused strategies.

Through SeroTracker and other initiatives, we have seen the benefits of groups collaborating on widely accepted, open-access core datasets. Many different users can apply these datasets to their own research questions, reducing duplication of effort and enabling scientists to focus on value-added analyses. Data platforms of this sort have the potential to centralise the efforts of multiple groups, minimising redundancy and enabling each group to carry out their work more efficiently. Creating platforms that facilitate this collaboration and incentives is an important step for computable and impactful evidence synthesis [353].

Finally, it will be important to consider how to ensure rigour, efficiency, and effective impact evaluation for new evidence synthesis products [345]. It is essential to consider how to preserve the strengths of the existing scientific communication paradigm: the rigour of preregistration and reporting guidelines, the validation and legitimacy provided by peer review, and clear dissemination mechanisms by which knowledge users can learn of new relevant work [354, 355].

These strengths raise several questions for future work. What does it mean to peer-review a dashboard? What is the best way to measure the impact of these digital tools? Counting the number of pageviews or citations is simple, but it is much more challenging to measure the extent of impact realised through better policy decisions. How can we appropriately fund this sort of research infrastructure, which often does not directly

fit the mandate of traditional academic funding sources [356]? How do we make new products easily discoverable, to help researchers publicise their work and knowledge users find tools that are useful to them?

We expect user-centered evidence synthesis to become even more relevant as the volume of data grows, researchers become more technologically adept, and health emergencies demand us to close the time gap between data and decisions [345]. In the wake of COVID-19, we have the opportunity to move towards user-centered evidence syntheses purpose-built for maximum impact.

7.3 Future work

SeroTracker's successes were enabled by our interdisciplinary and pragmatic approach, bringing together best practices in engineering design and epidemiology to address pressing public health problems. This approach has considerable potential to address today's complex and multifaceted health challenges, and is crucial to enable emerging technology to be applied to those problems.

Here, we highlight three natural directions for future work emerging from this dissertation: extending on the infectious disease modelling framework developed in Chapter 5; developing a data submission and accession platform for population-based surveillance across infectious diseases; and evaluating emerging infection detection technologies for surveillance and early detection.

7.3.1 Extended Bayesian modelling framework for surveillance data

In Chapter 5, we present an interpretable framework to synthesise indicator- and population-based surveillance data to estimate infections and immunity. While we apply this model to COVID-19 data to understand the true extent of SARS-CoV-2 infection, this flexible model shows the potential to be expanded to include other surveillance measures (e.g., deaths, wastewater) and to be applied for other infectious diseases.

Some of these extensions are data-driven. A first extension would be to add age structure, other indicators like hospitalisations, or deaths to the model. While the model presented here is for reported cases, there are direct analogies to other indicators such as

hospitalisations and deaths. For example, the probability of detection parameter ρ for cases is directly analogous to the infection fatality ratio for deaths. Incorporating age structure would be particularly important in extending this model to hospitalisations and deaths, considering that the infection fatality ratio for COVID-19 varies across several orders of magnitude for different age groups [265].

Second, this model could be used to establish the relationships between other regular indicators and infections. These indicators could include wastewater data globally [17], or the prevalence of disease-specific symptoms as in the ZOE COVID-19 study [357]. Wastewater studies in particular have become very common over the past two years as leading indicators of infection [17], but these studies cannot yet be compared across different locations due to differences in wastewater systems, sample collection, and sample processing [358]. This model could be applied to normalise across wastewater datasets with reference to seroprevalence data as a population-based measure.

Other extensions focus on the model architecture. Third, a Gaussian process infections model could be used for the latent infections model. Using an established Gaussian process model model could permit the recovery of key infectious disease parameters like the time-varying force of infection [359], which are difficult to estimate but crucial for infectious disease.

A fourth direction involves evaluating the importance of each model component. We present here a reasonably complete model informed by domain knowledge. This model fits the data well. This said, the relative importance of each predictor or each component of the model is not clear. One could imagine running versions of this model with different components absent or present, and quantifying the importance of each component using a metric such as the absolute log likelihood.

Finally, this model could be invaluable in informing the design of future infectious disease surveillance systems. The model presented here is designed to address the limitations of our current surveillance data: lags and biases in indicator-based surveillance, and irregular and sparse population-based surveillance. Of course, it would be ideal to have the right data in place in the first instance. Many countries are beginning to consider how to revamp their surveillance systems in response to the COVID-19 pandemic. Knowing

what data is important to build these models, and the geographic coverage, regularity, precision, and timeliness needed in this data, is crucial to design surveillance systems sufficient to inform public health decision-making.

7.3.2 PathoTracker: unified data platform for population-based surveillance

In the wake of COVID-19, stakeholders are redoubling their investments in fit-for-purpose surveillance — moving away from selectively reported indicators like cases and deaths, to population-based surveillance which provides a representative picture of infectious disease burden. These fit-for-purpose surveillance strategies include representative antibody testing (serosurveillance), active diagnostic testing (point prevalence), and environmental monitoring (wastewater).

However, these efforts are fragmented — dozens of countries worldwide have announced such efforts with limited global coordination, preventing unified synthesis and analysis. Take the UK'S REACT studies [360, 361], active testing efforts by Canada's COVID-19 Immunity Task Force [362], the Bill and Melinda Gates Foundation's support of continuous serosurveillance in Sub-Saharan Africa, and the European Center for Disease Control's new surveillance plan [363]. Moreover, our work on SARS-CoV-2 evidence synthesis has emphasised the importance of open data practises, sharing early results, collation and analysis of timely data, and clear reporting guidelines [77, 102]. Data standards and global repositories for this data have the potential to enable powerful insights.

We will develop PathoTracker — a global open data repository for population-based surveillance of respiratory viruses. PathoTracker will enable submission of data from seroprevalence, point prevalence, and wastewater investigations. This computable evidence synthesis platform will enable global aggregation, comparison, and synthesis of infectious disease surveillance studies, providing the foundation for a new era of surveillance.

PathoTracker will directly build on SeroTracker, especially its data schema, analysis pipeline, and comprehensive database for serosurveillance. Learning from past proposals for centralised serological observatories, like the suggested World Serum Bank [364] and Global Serological Observatory [365], will also help inform the design and utility of the decentralised platform that we propose.

We will begin by developing pan-pathogen core data elements underpinning PathoTracker for seroprevalence, point prevalence, and wastewater studies. These will include the study metadata and aggregated results that are commonly reported, adding more granular data like antibody titres to enable sophisticated analysis. Once prototyped, we will conduct a Delphi process with global stakeholders and experts to iterate on these data elements and build consensus [366]. We will first develop these for serosurveillance, building on SeroTracker's data elements (Chapter 2) and the ROSES-S reporting guidelines. We will then rapidly expand our aperture to point prevalence studies, which have very similar data to seroprevalence studies, and then to wastewater.

We will then build the PathoTracker data submission, accession, and synthesis platform. We will include features for data submission following the core data elements outlined above, webpages for easy overviews of submitted data, online dashboards synthesising our data, and a data download form to enable analysis. Rapid and flexible approaches to peer review, such as crowdsourcing, could be built into such a repository to validate submitted data [367].

To bootstrap PathoTracker, we will include SeroTracker's database of seroprevalence studies within it. In parallel, we will engage groups supporting global seroprevalence standardisation and implementation (i.e., our partners at WHO UNITY), and groups beginning to implement these surveillance programs (as above), to understand their plans, align our efforts, and incentivise them to submit data.

Finally, we will analyse global pathogen transmission and immunity with PathoTracker data. PathoTracker's granular data will enable high-impact insights, including:

- Using granular antibody data to develop global models of infection, protection and immunity, building on the work in Chapter 5 to link seroprevalence data with the distribution of antibody titres at the individual level, the change in these over time, and what this implies for population-level protection

- Relating wastewater data to point prevalence. Wastewater data is an excellent indicator of disease burden. We will conduct analyses to clearly establish the correspondence between wastewater viral loads and point prevalence

- Linking PathoTracker and viral genomic data to enable analysis of linked genomic and non-genomic infectious disease surveillance data

Early development of PathoTracker is already underway.

7.3.3 Emerging detection technology assessment and prioritisation

Early detection of new outbreaks is a crucial surveillance function, helping expedite containment and countermeasure development. Here, we have focused extensively on antibody detection. Antibodies are crucial for immune surveillance, but are not well-suited to early detection of new outbreaks because of the inherent delays between infection and humoral immunity, and the challenges in detecting antibodies against uncharacterised pathogens [368].

Fortunately, the advances in biotechnology that create new threats may also prove to be our saving grace. The Apollo report sets out a list of emerging technologies that show considerable promise in pandemic prevention and response [369], from wearable devices to point-of-care diagnostics to metagenomic sequencing [370].

Taking these emerging technologies forward requires answering several key questions. First, what is the state of this technology today? Second, what do we know about the signals that this technology detects? For example, for wearable physiological sensors, what physiological changes are triggered by infection, and how do these changes differ between pathogens and across individuals? Third, what technology advancements are required for these tools to deliver their promised value? Using the same example, our recent evaluation of wearable sensors for COVID-19 mitigation identified improved detection specificity as a key requirement for their broad implementation [371]. Finally, epidemiological considerations in implementing these technologies; for example, which populations should be sampled (building on new work in genomic surveillance [372]), and at what coverage to detect outbreaks with what sensitivity (as for syndromic surveillance of other infectious diseases [373]).

We will systematically answer these open questions, using a combination of evidence synthesis and infectious disease modelling techniques. We will work with the research groups leading the development of the most promising technologies to further their

development, pairing their technical expertise with our epidemiological and systems thinking approach. By doing so, we will unlock the potential of emerging technologies for early detection.

7.4 Conclusion

With advancing technology, the world teeters on the edge between health security and health crisis. We now aim to continue SeroTracker's momentum to engineer pandemic modelling frameworks that synthesise surveillance data, build future surveillance data platforms, validate promising pandemic response technologies, and develop more effective evidence synthesis strategies. To address these complex and multifactorial problems, we will bring together the cutting edge of several disciplines to generate novel ideas, take a user- and impact-centered approach to turn these ideas into tangible products; and work closely with global partners across sectors, from academia to multilateral agencies to private industry. These projects will help tip the global balance towards health security.

Appendices

A

SeroTracker full data schema and tables

A.1 Sources of seroprevalence data

Sources that meet the SeroTracker inclusion criteria are extracted into the Source Table (Table A.1).

Table A.1: Source table fields.

Variable	Description	Data Type
Name of source	Title of article, report, or presentation	Free text
Type of source	Type of source: • Peer-reviewed article • Preprint • Report • Presentation • News article	Vocabulary
First author	First author of the source	Free text
Publication date	Date source was published	Date
URL	Web address where the source was found	Free text
DOI	Digital object identifier that uniquely identifies the source	Free text
Citation	Information needed to reference article	BibTEX File
Studies	Studies described in this source; one or more	Study table

To rapidly release findings from seroprevalence studies, results from the same seroprevalence study were often published in increasingly credible but increasingly delayed venues: for example, government report, then peer-reviewed article. In these cases, we

identified the most credible version of the source as the version of record, and archived less credible versions in our database. We used the following order of credibility, from most to least credible: peer-reviewed articles, preprints, reports, presentations, news articles.

A.2 Seroprevalence studies

We defined a seroprevalence study as an effort to test a defined population over a specified time to estimate the prevalence of SARS-CoV-2 antibodies. These studies were described in Table A.3.

Where a source contained multiple methodologically distinct serosurveys, we split it into multiple studies. This was most often because they reported nonoverlapping results for distinct time points (e.g., monthly studies of blood donors, where participants and methodology shifted between months) or for distinct populations (e.g., all students at a school, in addition to a convenience sample of the local community around the school).

A.2.1 Country data for each study

To put seroprevalence data in context, we gathered information about the demographics and COVID-19 dynamics in each country in which a serosurvey in our database was done. The fields gathered for each country are described in Table A.2.

Table A.2: Country table fields.

Variable	Description	Data Source
COUNTRY CHARACTERISTICS		
Population	Size of each country's population	2019 UN World Population Prospects [374]
WHO Region	Which of the six World Health Organization regions the country is in: • African Region • Region of the Americas • Southeast Asian Region • European Region • Eastern Mediterranean Region • Western Pacific Region	World Health Organization
World Bank Income Level	Country Income Level, as defined by the World Bank based on Gross National Income per capita (GNI) in $USD: • Low: GNI $\in [\$0, \$1{,}036)$ • Lower-middle: GNI $\in [\$1{,}036, \$4{,}046)$ • Upper-middle: GNI $\in [\$4{,}046, \$12{,}536)$ • High: GNI $\in [\$12{,}536, \infty)$	2020 World Bank country classifications by income level
HRP status	Whether the country is an emergency humanitarian setting	Global COVID-19 Humanitarian Response Plan [375]
COUNTRY-LEVEL TIME SERIES OF COVID-19 DATA		
Confirmed cases	Number of confirmed COVID-19 cases, typically based on diagnostic testing	*Our World in Data* [376]
Deaths	Number of deaths attributed to 2019	*Our World in Data* [376]
Diagnostic tests	Number of diagnostic tests carried out; typically PCR	*Our World in Data* [376]
Vaccinations	Number of individuals who have received at least one COVID-19 vaccine dose	*Our World in Data*

Table A.3: Study table fields.

Variable	Description	Data Type
Study Characteristics		
Study name	Name of study	Free text
Country	Country that the study was conducted in	Country table
Study design	Study design: • Single cross-sectional • Repeated cross-sectional • Prospective cohort • Retrospective cohort	Vocabulary
Geographic coverage	Study coverage: • National (country or territory) • Subnational/regional (state or province) • Local (county or municipality)	Vocabulary
Sampling method	Probability sampling: • Simple probability • Stratified probability • Entire sample Non-probability sampling: • Convenience • Self-referral • Sequential	Ontology
Inclusion criteria	Criteria required for participant inclusion	Free text
Exclusion criteria	Criteria sufficient for participant exclusion	Free text
Participant Demographics		
Total sample size	Number of individuals or specimens included	$\mathbb{Z}_0^+$
Observational unit	Most granular observational unit: • People (as in a household survey) • Samples (as in a study of blood donors)	Vocabulary
Number of females	Number of females included	$\mathbb{Z}^+$
Number of males	Number of males included	$\mathbb{Z}^+$
Median age	Median participant age	$\mathbb{R}^+$
Q1 participant age	25th percentile of participant age	$\mathbb{R}^+$
Q3 participant age	75th percentile of participant age	$\mathbb{R}^+$
Number vaccinated	Number vaccinated for COVID-19	$\mathbb{Z}_0^+$
Findings		
Estimates	Estimates of seroprevalence generated by study	Estimates table
Risk of bias	SeroTracker evaluation of risk of bias (low, moderate, high, or unclear)	Risk of bias table

A.2.2 Risk of bias table

The quality of SARS-CoV-2 seroprevalence study reporting and design varies widely [9, 81]. We critically appraised each study using the nine-item Joanna Briggs Institute Critical Appraisal Checklist for Prevalence Studies [178]. To operationalise this checklist, we added guidance on its application to SARS-CoV-2 seroprevalence studies. We also added an overall risk of bias item, capturing the risk that the seroprevalence estimates from this study are biased away from the true seroprevalence. The overall risk of bias was graded based on the individual risk of bias items, and graded low, moderate, high, or unclear.

The specific risk of bias items, our operationalisation of them, and the decision rule we used to assess overall risk of bias based on these items is described in Section 6.2. The full risk of bias table is included in the main text in Section 6.1.

A.3 Seroprevalence estimates

Each study in our database includes one or more seroprevalence estimates – estimates of seroprevalence in a particular subgroup, at a particular time, using a particular assay, and generated using a particular method of analysis. The schema for seroprevalence estimates is in Table A.4.

Table A.4: Estimates table fields.

Variable	Description	Data Type
Subgroup: Which group was studied?		
Geographic scope	Breadth of geography included in this estimate. Vocabulary as in Table A.3	Vocabulary
State	What state(s) does this estimate describe?	List
County	What count(ies) does this estimate describe?	List
City	What cit(ies) does this estimate describe	List
Sampling start	First date of specimen collection for samples included in this estimate	Date
Sampling end	Final date of specimen collection for samples included in this estimate	Date
Sex or gender	Sex or gender of included participants: • Male • Female • Other • All	Vocabulary

Age minimum	Participants were at least this age	$\mathbb{Z}_0^+$
Age maximum	Participants were at most this age	$\mathbb{Z}_0^+$
Population studied	Specific population studied: • General population – Household survey participants – Blood donors – Patients seeking routine medical care – ... • Other (specific) populations – Health care workers – Incarcerated individuals – ...	Ontology
Assay: How was seropositivity determined?		
Assay(s) used	Which serological assays were used to test for antibodies	Assays table link
Specimen type collected	What type of specimens were collected from participants: • Whole blood • Dried blood • Serum • Plasma	Vocabulary
Testing algorithm	If multiple assays were used, how specimens were deemed to be seropositive or seronegative • Same assay used on all specimens • One assay used on each specimen; different assays used on different specimens • Same two assays used on all specimens; specimens deemed seropositive if at least one assay was positive • Same two assays used on all specimens; specimens deemed seropositive if both assays were positive • Same assay used on all specimens, with positive samples evaluated on a confirmatory assay; specimens deemed seropositive if both assays were positive • Three or more assays	Vocabulary
Analysis: Which biases were statistically corrected for?		
Assay performance correction	Whether this seroprevalence estimate was corrected for imperfect test performance. Particularly important in settings with very low prevalence	Boolean

Sampling correction	Was this estimate corrected for differences between the sample and target population? Particularly important with low response rates or where sample and population demographics substantially differ	Boolean
Clustering correction	Was this estimate corrected for clustered samples (e.g., if all members of the same household were sampled)?	Boolean
SEROPREVALENCE: FOR THAT SUBGROUP, ASSAY, AND ANALYSIS		
Denominator	Sample size for subgroup	$\mathbb{Z}^+$
Seroprevalence point estimate	Point seroprevalence estimate	$[0, 1]$
95% interval lower bound	Author-reported 95% interval lower bound for the estimate	$[0, 1]$
95% interval upper bound	Author-reported 95% interval upper bound for the estimate	$[0, 1]$

We defined a standard set of subgroups for which we would extract seroprevalence estimates, if available. From all included studies, we extracted the summary seroprevalence estimate, as reported by the study authors. This estimate typically includes the entire sample and is generated using the full suite of serological assays used by the study. This estimate is also usually corrected for both assay performance and for differences between the sample and population, if the study authors conducted such corrections.

Studies of the general population are of particular interest. These are studies of populations that can reasonably be generalised to the population at large (e.g., households, blood donors, patients attending routine outpatient medical care), and provide highly relevant information to estimate infections and immunity in the population at large. For these studies, we extracted key additional subgroups that are relevant for decision-making:

- Sex subgroups (male, female, other)
- Age subgroups (10-year age bands, if possible; if not, available subgroups)
- Serological assay subgroups (anti-nucleocapsid, anti-spike assays)
- Vaccination subgroups (vaccinated, unvaccinated)

- Timing of specimen collection (for repeated cross-sectional and cohort studies)

A.3.1 Serological assays table

While our data includes several thousand seroprevalence studies, there are a few hundred SARS-CoV-2 serological assays in use, and a small fraction of these have seen disproportionate use. We documented which assays have been used to generate each seroprevalence estimate (schema in Table A.5).

These tables are further described in Section 4.3, which also examines assays used in SARS-CoV-2 seroprevalence studies, their performance, the level of agreement between different evaluations of their performance, and the impact of misclassification error on seroprevalence estimates.

Table A.5: Serological assay table fields.

Variable	Description	Data Type
PRODUCT		
Manufacturer	Manufacturer. Typically either a commercial manufacturer or public health laboratory	Free text
Assay name	Product name for the serological assay	Free text
Country	Country of assay development	Country table link
DETECTION CHARACTERISTICS		
Technology	• Enzyme-linked immunosorbent assay • Lateral flow immunoassay • Immunofluorescence assay • Chemiluminescence immunoassay • Neutralisation assay • Other	Vocabulary
Time to result	• Rapid diagnostic test • Laboratory diagnostic test	Vocabulary
Isotype	Antibody isotype(s) detected: • IgG • IgA • IgM • Total antibody	Vocabulary
Antibody target	Antigen that antibodies are detected against: • Spike protein (anti-spike) – S1 domain – S2 domain – Receptor-binding domain • Nucleocapsid protein (anti-nucleocapsid) • Other	Ontology
EVALUATION		
Evaluations	Evaluations of performance by the manufacturer or independent evaluators	Evaluations table link

A.3.2 Serological assay evaluations table

We also tabulate extensive data on assay performance. We included assay performance data from three sources: manufacturers, independent serosurvey investigators, and third-party laboratory validations. The sources we included were as follows:

- Manufacturer estimates of sensitivity and specificity, produced for assay product self-certification and available in product inserts.

- Independent serosurvey investigator validations reported alongside seroprevalence studies; these were typically either independent evaluations using readily available samples at serosurvey sites, or field evaluations designed to minimise spectrum bias by using a reference panel that resembled the study population.
- Third-party validations that conducted large-scale, head-to-head evaluations on a common reference panel. We identified five of these: from the Australian National Laboratory for the WHO ([189]), the US Food and Drug Administration [190], the Netherlands Center for Infectious Disease Control [191], the Doherty Institute [192], and the Foundation for Innovative New Diagnostics [193]. Details of each evaluation and its reference panel are in Table A.6.

For each evaluation, we extracted information on the evaluator, the composition and size of the reference panel, and the estimated γ and δ (Table A.7).

Table A.6: Reference panels of five third-party lab evaluations. The full cross-reactivity panel for the Doherty Institute and Australian National Reference Laboratory evaluations included specimens positve for malaria, Influenza A, Influenza B, cytomegalovirus, acute parvovirus B19, Epstein-Barr, mycoplasma, parainfluenza, C. psittaci, toxoplasma, rubella, Hepatitis A and B. Abbreviations as follows: LFIA: lateral flow immunoassay, IFA: immunofluorescence assays, ELISA: enzyme-linked immunosorbent assays; CLIAs: chemiluminescent immunoassays. NRL-WHO: Australian National Reference Laboratory for the World Health Organization, US FDA: United States Food and Drug Administration, FIND: Foundation for Innovative New Diagnostics, NL CIDC: Netherlands Center for Infectious Disease Control.

Evaluator	**Assay Types**	**Positive Sample**	**Negative Sample**
US FDA [190]	LFIAs, IFAs, ELISAs, CLIAs, and others others	Confirmed with PCR; disease severity unknown; time post-symptom onset unknown ($n = 30$)	Pre-pandemic negative serum and plasma (prior to 2020) without regard for clinical status ($n = 70$); cross-reactivity human immunodeficiency virus positive ($n = 10$)
FIND [193]	LFIAs ($n = 35$) and ELISAs ($n = 16$)	Confirmed with PCR; disease severity varies; n varying by time post-symptom onset (10 for day 0-3, 20 for day 4-7, 30 for day 8-14, 20 for day 15-28, 20 for day 29+)	Pre-pandemic samples (prior to 2018; target $n > 300$)
NL CIDC [191]	LFIAs, IFASs, ELISAs, and CLIAs ($n = 47$)	Confirmed with PCR; disease severity varies; time post-symptom onset stratified by before or after 14 days (n varies between test and site)	Pre-pandemic samples (prior to Dec 2019); syndromic patients with potentially cross-reactive microorganisms during pandemic (n varies)
Doherty [192]	LFIAs ($n = 27$)	Confirmed with PCR; disease severity unknown; time post-symptom onset varies ($n = 50 - 100$, varying by phase)	Pre-pandemic samples (2018 and 2019) and non-COVID-19 infections ($n = 92 - 100$); cross-reactivity panel ($n = 25$)
NRL-WHO [189]	LFIAs, ELISAs, and CLIAs (n = 40)	Confirmed with PCR or clinical diagnosis; disease severity unknown; time post-symptom onset $>$ 14 days post-PCR (n unknown)	Pre-pandemic samples (prior to Nov 2019) (n unknown); cross-reactivity panel ($n = 55$)

Table A.7: Serological assay evaluations table fields.

Variable	Description	Data Type
Context		
Laboratory	Name of the evaluating group	Free text
Category	• Manufacturer • Independent serosurvey investigator • Third-party laboratory	Vocabulary
True Positive Rate Evaluation		
True positive panel	Specimens used as true positives. Often from an individual with previous PCR-confirmed SARS-CoV-2 infection or specimens that test positive on neutralisation assay	Free text
Number of true positives	Used to estimate sensitivity	$\mathbb{Z}^+$
Sensitivity	Fraction of true positives classified correctly	$[0, 1]$
True Negative Rate Evaluation		
True negative panel	Specimens used as true negatives. Often stored blood samples from before 2020. Also often includes samples positive for antibodies against other, sometimes-related viruses (e.g., other pandemic coronaviruses, common cold coronaviruses, human immunodeficiency virus, influenza)	Free text
Number of true negatives	Used to estimate specificity	$\mathbb{Z}^+$
Specificity	Fraction of true negatives classified correctly	$[0, 1]$

B

SeroTracker Analyse page visualisations

The visualisations currently on SeroTracker's Analyse page are outlined below.

1. Comparing seroprevalence, confirmed cases, and vaccine coverage across geographies, demonstrating how increases in seroprevalence correlate with the extent of infection and vaccination, and how each of these factors differ dramatically between countries and territories; shown in Figure 2.8.

2. Principled estimates of the seroprevalence in each WHO region and how it has changed over time, following the methodology described in Chapter 3.

3. Case underascertainment: box plots showing the difference between confirmed COVID-19 cases and seroprevalence estimates.

4. Seroprevalence across populations: box plots showing how seroprevalence varies between population groups with different behaviors.

5. Seroprevalence literature by *Global Burden of Disease* region and over time: bar and line plots showing the increase in seroprevalence data over time.

6. Canadian visualisations: map showing these visualisations for Canadian studies; available at SeroTracker.com/Partnerships/Canada.

7. WHO visualisations: custom map for the World Health Organization at SeroTracker.com/Partnerships/Unity, showing all seroprevalence studies globally aligned with the UNITY population-based, age-stratified seroepidemiological investigations protocol [122]. This page is featured on the WHO COVID-19 dashboard [110].

C

Natural language processing for abstract screening: additional results

C.1 Summary of project measures, methods, and results

C.2 Additional analyses

C.2.1 Process measures: Inclusion recommendation performance on the pilot abstracts

We evaluated the inclusion recommendation feature's precision and recall on the pilot abstracts (Table C.1), using pre-project votes as ground truth. All the true positive pilot abstracts were at least weakly recommended, with $^{73}//_{91}$ being strongly recommended. The weakly recommended threshold gave the highest F1-score (0.905), with corresponding precision 0.827 and recall of 1.0 at this level.

Table C.1: Operating characteristics of the tool evaluated on the pilot abstracts at different inclusion thresholds. Reference: ground truth established in the pre-project stage.

Threshold	**TN**	**FP**	**FP**	**FP**	**Prec.**	**Rec.**	**F1**
Weakly Recommended	290	19	0	91	0.827	1.000	0.905
Somewhat Recommended	299	10	9	82	0.891	0.901	0.896
Recommended	301	8	14	77	0.906	0.846	0.875
Strongly Recommended	302	7	18	73	0.912	0.802	0.854

Table C.2: Key methods and results of process, outcome and structure measures evaluated.

Measure	Description	Rationale	Analysis	Key Result
Process Measures				
Efficiency	a) Mean time taken to screen an abstract with-tool compared to without-tool	Direct indicator of screening time reduction	Two-sided unequal variances t-test to compare screening time	Tool use was associated with a significant 33.7% decrease in mean time taken per abstract (21.45s to 14.22s).
	b) Conflict rate: proportion of abstracts requiring conflict resolving votes	Conflicting votes must be resolved by a third reviewer, increasing screening time	Fisher's exact test to compare conflict rates	Tool use was associated with a significant 2.3-fold reduction in conflict rates (8.32% to 3.64% of abstracts)
Accuracy	a) Assessment of screening recall with vs. without the tool	Ensuring that tool implementation does not detract from review results	Fisher's exact test to compare whether the proportion of included abstract is significantly different	There was no significant loss of sensitivity with the tool ($^{74}/_{91}$ without tool, $^{82}/_{91}$ with the tool)
	b) Evaluation of the tool's inclusion recommendations	Accuracy of inclusion recommendations on pilot abstracts	Operating characteristics	All included pilot abstracts met the 'weakly recommended' threshold, with $^{73}/_{91}$ meeting the 'strongly recommended' threshold. The highest F1-score was at the 'weakly recommended' threshold (0.905), with corresponding precision 0.827 and recall 1.0
Outcome Measures				

Results of overall review	Comparison of the characteristics of included seroprevalence estimates	To identify any differences in the results of a systematic review due to the tool	Compared summary statistics of the estimates included pre-project, without-tool and with-tool	There were consistent characteristics in estimates that were ultimately included at all three stages
User interaction with the tool	a) Evaluation of agreement between user votes and the inclusion recommendations	To assess whether users voted in alignment with the tool and whether they found it useful	Measured operating characteristics comparing tool predictions to user's votes	Users voted strongly in agreement with the tool (area under the receiver operating characteristic curve of 0.956 in Figure A1a
	b) Use of tool features	To assess which features were likely to contribute most to any changes observed	Descriptive summary of tool use	All 7 users used the inclusion recommendation feature, population, intervention, and outcome annotation was used by 5, and the keyword and undo features by 3
User satisfaction	Reviewer satisfaction with tool and its features	To determine whether SeroTracker reviewers were satisfied with the tool and drive further improvements	Circulated a user satisfaction survey and summarised its results	The inclusion recommendation feature was assessed as most useful (mean rating $^{4.7}/_{5}$) and keyword highlights the least (mean rating $^{3.88}/_{5}$)
Structure Measures				
'One-person-one-tool' (OPOT) model	Evaluation of operating characteristics if one human vote is replaced by the tool's vote	This application of the toolcould significantly reduce screening time and costs	Simulated one human vote, with access to the tool (W) or without (WO), alongside one model vote and assessed operating characteristics.	F1-score for included texts was improved from 0.846 for two humans without the tool to 0.878 in OPOT-W/O, and from 0.911 for two humans with the tool to OPOT-W.

'Tool-only' model	Evaluation of operating characteristics if both human votes are replaced by the tool's vote	To serve as a comparator group to the project and OPOT models	Assessed operating characteristics in this setting	The tool-only setting performed as well both OPOT and as two humans.

References

1. Cheng, M. P., Papenburg, J., Desjardins, M., *et al.* Diagnostic Testing for Severe Acute Respiratory Syndrome–Related Coronavirus 2: A Narrative Review. *Annals of Internal Medicine* **172,** 726–734. (2020) (June 2020).
2. World Health Organization. *COVID-19 Weekly Epidemiological Update* (World Health Organization, Nov. 24, 2020).
3. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20,** 533–534. (2020) (May 1, 2020).
4. Buitrago-Garcia, D., Egli-Gany, D., Counotte, M. J., *et al.* Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Medicine* **17,** e1003346 (Sept. 2020).
5. American Society for Microbiology. *Supply shortages impacting COVID-19 and non-COVID testing* (2021).
6. Lieberman-Cribbin, W., Tuminello, S., Flores, R. M. & Taioli, E. Disparities in COVID-19 Testing and Positivity in New York City. *American Journal of Preventitive Medicine* **59,** 326–332 (Sept. 2020).
7. Cheng, M. P., Yansouni, C. P., Basta, N. E., *et al.* Serodiagnostics for Severe Acute Respiratory Syndrome–Related Coronavirus 2. *Annals of Internal Medicine* **173,** 450–460. (2020) (June 4, 2020).
8. Meyer, B., Drosten, C. & Müller, M. A. Serological assays for emerging coronaviruses: Challenges and pitfalls. *Virus Research. Nidoviruses I* **194,** 175–183. (2022) (Dec. 19, 2014).
9. Bobrovitz, N., Arora, R. K., Cao, C., *et al.* Global seroprevalence of SARS-CoV-2 antibodies: A systematic review and meta-analysis. *PloS One* **16,** e0252617 (2021).
10. German, R. R., Lee, L. M., Horan, J. M., *et al.* Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports* **50,** 1–35, quiz CE1–7 (RR-13 July 27, 2001).
11. Division of Global Health Protection, Global Health & CDC. *Event-based Surveillance* (2022).
12. Lawrence, S. V. COVID-19 and China: A Chronology of Events (December 2019-January 2020), 54 (2020).
13. Adams, D. A., Thomas, K. R., Jajosky, R. A., *et al.* Summary of Notifiable Infectious Diseases and Conditions - United States, 2015. *MMWR. Morbidity and mortality weekly report* **64,** 1–143 (Aug. 11, 2017).
14. Gibbons, C. L., Mangen, M.-J. J., Plass, D., *et al.* Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health* **14,** 147. (2022) (Feb. 11, 2014).
15. Van Beneden, C. A., Arvay, M., Thamthitiwat, S. & Lynfield, R. in *Infectious Disease Surveillance* 93–108 (John Wiley & Sons, Ltd, 2013). (2022).

16. Menachemi, N., Yiannoutsos, C. T., Dixon, B. E., *et al.* Population Point Prevalence of SARS-CoV-2 Infection Based on a Statewide Random Sample - Indiana, April 25-29, 2020. *MMWR. Morbidity and mortality weekly report* **69,** 960–964 (July 2020).

17. Shah, S., Gwee, S. X. W., Ng, J. Q. X., *et al.* Wastewater surveillance to infer COVID-19 transmission: A systematic review. *The Science of the Total Environment* **804,** 150060 (Jan. 2022).

18. Groseclose, S. L. & Buckeridge, D. L. Public Health Surveillance Systems: Recent Advances in Their Use and Evaluation. *Annual Review of Public Health* **38,** 57–79 (Mar. 20, 2017).

19. Marshall, J. S., Warrington, R., Watson, W. & Kim, H. L. An introduction to immunology and immunopathology. *Allergy, Asthma, and Clinical Immunology : Official Journal of the Canadian Society of Allergy and Clinical Immunology* **14.** (2020) (Suppl 2 Sept. 12, 2018).

20. Alfego, D., Sullivan, A., Poirier, B., *et al.* A population-based analysis of the longevity of SARS-CoV-2 antibody seropositivity in the United States. *eClinicalMedicine* **36.** (2022) (June 1, 2021).

21. Lou, B., Li, T.-D., Zheng, S.-F., *et al.* Serology characteristics of SARS-CoV-2 infection since exposure and post symptom onset. *European Respiratory Journal.* (2020) (Jan. 1, 2020).

22. Iyer, A. S., Jones, F. K., Nodoushani, A., *et al.* Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. *Science Immunology* **5.** (2020) (Oct. 8, 2020).

23. Isho, B., Abe, K. T., Zuo, M., *et al.* Persistence of serum and saliva antibody responses to SARS-CoV-2 spike antigens in COVID-19 patients. *Science Immunology* **5.** (2020) (Oct. 8, 2020).

24. Wajnberg, A., Amanat, F., Firpo, A., *et al.* Robust neutralizing antibodies to SARS-CoV-2 infection persist for months. *Science.* (2020) (Oct. 28, 2020).

25. Dan, J. M., Mateus, J., Kato, Y., *et al.* Immunological memory to SARS-CoV-2 assessed for greater than six months after infection. *bioRxiv,* 2020.11.15.383323. (2020) (Nov. 16, 2020).

26. Bolotin, S., Tran, V., Osman, S., *et al.* SARS-CoV-2 seroprevalence survey estimates are affected by anti-nucleocapsid antibody decline. *medRxiv : the preprint server for health sciences,* 2020.09.28.20200915 (Sept. 2020).

27. Guo, L., Wang, G., Wang, Y., *et al.* SARS-CoV-2-specific antibody and T-cell responses 1 year after infection in people recovered from COVID-19: a longitudinal cohort study. *The Lancet Microbe* **3,** e348–e356. (2022) (May 1, 2022).

28. Gaebler, C., Wang, Z., Lorenzi, J. C. C., *et al.* Evolution of Antibody Immunity to SARS-CoV-2. *bioRxiv,* 2020.11.03.367391. (2020) (Nov. 5, 2020).

29. Turner, J. S., Kim, W., Kalaidina, E., *et al.* SARS-CoV-2 infection induces long-lived bone marrow plasma cells in humans. *Nature* **595,** 421–425. (2022) (July 2021).

30. Krammer, F. A correlate of protection for SARS-CoV-2 vaccines is urgently needed. *Nature Medicine* **27,** 1147–1148 (July 2021).

31. Khoury, D. S., Cromer, D., Reynaldi, A., *et al.* Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. *Nature Medicine* **27,** 1205–1211 (July 2021).

32. Gilbert, P. B., Montefiori, D. C., McDermott, A. B., *et al.* Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science (New York, N.Y.)* **375,** 43–50 (Jan. 7, 2022).

33. Liu, Y. & Rocklöv, J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *Journal of Travel Medicine* **28,** taab124 (Oct. 2021).
34. Mykytyn, A. Z., Rissmann, M., Kok, A., *et al.* Antigenic cartography of SARS-CoV-2 reveals that Omicron BA.1 and BA.2 are antigenically distinct. *Science Immunology* **7,** eabq4450. (2022) (June 23, 2022).
35. Fox, J. D. Nucleic acid amplification tests for detection of respiratory viruses. *Journal of Clinical Virology* **40,** S15–S23. (2020) (Oct. 2007).
36. Mackay, I. M., Arden, K. E. & Nitsche, A. Real-time PCR in virology. *Nucleic Acids Research* **30,** 1292–1305. (2020) (Mar. 15, 2002).
37. Watson, J., Whiting, P. F. & Brush, J. E. Interpreting a covid-19 test result. *BMJ (Clinical research ed.)* **369,** m1808 (May 12, 2020).
38. Lo, Y.-M. D., Mehal, W. Z. & Fleming, K. A. False-positive results and the polymerase chain reaction. *The Lancet* **332,** 679. (2020) (Sept. 17, 1988).
39. Jarrom, D., Elston, L., Washington, J., *et al.* Effectiveness of tests to detect the presence of SARS-CoV-2 virus, and antibodies to SARS-CoV-2, to inform COVID-19 diagnosis: a rapid systematic review. *BMJ Evidence-Based Medicine.* (2020) (Oct. 1, 2020).
40. Dinnes, J., Deeks, J. J., Adriano, A., *et al.* Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *The Cochrane Database of Systematic Reviews* **8,** CD013705 (2020).
41. World Health Organization. *Global partnership to make available 120 million affordable, quality COVID-19 rapid tests for low- and middle-income countries* (2020).
42. Patel, R., Babady, E., Theel, E. S., *et al.* Report from the American Society for Microbiology COVID-19 International Summit, 23 March 2020: Value of diagnostic testing for SARS–CoV-2/COVID-19. *mBio* **11.** (2020) (Apr. 28, 2020).
43. Marsh, T. *Over 67 million americans lack access to COVID-19 testing.* GoodRx Research. (2020).
44. Kim, E. J., Marrast, L. & Conigliaro, J. COVID-19: Magnifying the effect of health disparities. *Journal of General Internal Medicine* **35,** 2441–2442. (2020) (May 11, 2020).
45. Sze, S., Pan, D., Nevill, C. R., *et al.* Ethnicity and clinical outcomes in COVID-19: A systematic review and meta-analysis. *EClinicalMedicine* **0.** (2020) (Nov. 12, 2020).
46. Wiemers, E. E., Abrahams, S., AlFakhri, M., *et al.* Disparities in Vulnerability to Severe Complications from COVID-19 in the United States. *medRxiv.* (2020) (June 22, 2020).
47. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine.* (2020) (June 3, 2020).
48. Pira, G. L., Ivaldi, F., Starc, N., *et al.* Miniaturized and High-Throughput Assays for Analysis of T-Cell Immunity Specific for Opportunistic Pathogens and HIV. *Clinical and Vaccine Immunology* **21,** 488–495. (2020) (Apr. 1, 2014).
49. Loeffelholz, M. J. Evaluation of High-Throughput Serological Tests for SARS-CoV-2. *Journal of clinical microbiology* **58,** e02179–20 (Oct. 21, 2020).
50. Voller, A., Bartlett, A. & Bidwell, D. E. Enzyme immunoassays with special reference to ELISA techniques. *Journal of Clinical Pathology* **31,** 507–520. (2020) (June 1978).
51. Bastos, M. L., Tavaziva, G., Abidi, S. K., *et al.* Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *British Medical Journal* **370.** (2020) (July 1, 2020).
52. Clapham, H., Hay, J., Routledge, I., *et al.* Seroepidemiologic study designs for determining SARS-CoV-2 transmission and immunity. *Emerging Infectious Diseases* **26,** 1978–1986 (Sept. 2020).

53. World Health Organization. *Population-based age-stratified seroepidemiological investigation protocol for COVID-19 virus infection, 17 March 2020* WHO/2019-nCoV/Seroepidemiology/2020.1 (World Health Organization, 2020). (2022).

54. Gronvall, G., Connell, N., Farley, J. E., *et al. Developing a National Strategy for SARS-CoV-2 Serosurveys in the US* (2020). (2020).

55. Ward, H., Atchison, C., Whitaker, M., *et al.* SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. *Nature Communications* **12,** 905. (2022) (Feb. 10, 2021).

56. Uyoga, S., Adetifa, I. M. O., Karanja, H. K., *et al.* Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. *Science (New York, N.Y.)* **371,** 79–82 (Jan. 1, 2021).

57. Peeling, R. W. & Olliaro, P. L. The time to do serosurveys for COVID-19 is now. *The Lancet. Respiratory Medicine* **8,** 836–838. (2020) (Sept. 2020).

58. Lipsitch, M., Donnelly, C. A., Fraser, C., *et al.* Potential Biases in Estimating Absolute and Relative Case-Fatality Risks during Outbreaks. *PLoS Neglected Tropical Diseases* **9,** e0003846. (2022) (July 16, 2015).

59. Wu, S. L., Mertens, A. N., Crider, Y. S., *et al.* Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications* **11,** 4507. (2022) (Sept. 9, 2020).

60. Lai, C.-C., Wang, J.-H. & Hsueh, P.-R. Population-based seroprevalence surveys of anti-SARS-CoV-2 antibody: An up-to-date review. *International Journal of Infectious Diseases* **101,** 314–322. (2022) (Dec. 1, 2020).

61. Winter, A. K., Martinez, M. E., Cutts, F. T., *et al.* Benefits and challenges in using seroprevalence data to inform models for measles and rubella elimination. *The Journal of Infectious Diseases* **218,** 355–364 (July 2, 2018).

62. Bubar, K. M., Reinholt, K., Kissler, S. M., *et al.* Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science* **371,** 916–921. (2021) (Feb. 26, 2021).

63. Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nature Reviews Immunology* **21,** 73–82. (2021) (Feb. 2021).

64. Ripperger, T. J., Uhrlaub, J. L., Watanabe, M., *et al.* Orthogonal SARS-CoV-2 Serological Assays Enable Surveillance of Low-Prevalence Communities and Reveal Durable Humoral Immunity. *Immunity* **53,** 925–933.e4. (2021) (Nov. 17, 2020).

65. Planas, D., Veyer, D., Baidaliuk, A., *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596,** 276–280 (Aug. 2021).

66. Lumley, S. F., O'Donnell, D., Stoesser, N. E., *et al.* Antibody Status and Incidence of SARS-CoV-2 Infection in Health Care Workers. *New England Journal of Medicine* **384,** 533–540. (2022) (Feb. 11, 2021).

67. Abu-Raddad, L. J., Chemaitelly, H., Coyle, P., *et al.* SARS-CoV-2 antibody-positivity protects against reinfection for at least seven months with 95% efficacy. *eClinicalMedicine* **35.** (2022) (May 1, 2021).

68. Hall, V. J., Foulkes, S., Charlett, A., *et al.* SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). *The Lancet* **397,** 1459–1469. (2022) (Apr. 17, 2021).

69. Mulenga, L. B., Hines, J. Z., Fwoloshi, S., *et al.* Prevalence of SARS-CoV-2 in six districts in Zambia in July, 2020: a cross-sectional cluster sample survey. *The Lancet Global Health* **9,** e773–e781. (2022) (June 1, 2021).

70. Bergeri, I., Whelan, M., Ware, H., *et al. Global epidemiology of SARS-CoV-2 infection: a systematic review and meta-analysis of standardized population-based seroprevalence studies, Jan 2020-Oct 2021* (medRxiv, Dec. 15, 2021), 2021.12.14.21267791. (2022).

71. Telenti, A., Arvin, A., Corey, L., *et al.* After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596,** 495–504 (Aug. 2021).

72. Gebremeskel, A. T., Otu, A., Abimbola, S. & Yaya, S. Building resilient health systems in Africa beyond the COVID-19 pandemic response. *BMJ Global Health* **6,** e006108. (2022) (June 1, 2021).

73. Adamu, A. A., Jalo, R. I., Habonimana, D. & Wiysonge, C. S. COVID-19 and routine childhood immunization in Africa: Leveraging systems thinking and implementation science to improve immunization system performance. *International Journal of Infectious Diseases* **98,** 161–165. (2022) (Sept. 2020).

74. Wilson, S. E., Deeks, S. L., Hatchette, T. F. & Crowcroft, N. S. The role of seroepidemiology in the comprehensive surveillance of vaccine-preventable diseases. *Canadian Medical Association Journal* **184,** E70–E76. (2020) (Jan. 10, 2012).

75. Abe, K. T., Li, Z., Samson, R., *et al.* A simple protein-based surrogate neutralization assay for SARS-CoV-2. *JCI Insight* **5.** (2021) (Oct. 2, 2020).

76. Feikin, D. R., Higdon, M. M., Abu-Raddad, L. J., *et al.* Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *The Lancet* **399,** 924–944. (2022) (Mar. 5, 2022).

77. Van Kerkhove, M., Grant, R., Subissi, L., *et al.* ROSES-S: Statement from the World Health Organization on the reporting of seroepidemiologic studies for SARS-CoV-2. *Influenza and Other Respiratory Viruses* **15,** 561–568 (June 26, 2021).

78. Reese, H., Iuliano, A. D., Patel, N. N., *et al.* Estimated incidence of COVID-19 illness and hospitalization — United States, February–September, 2020. *Clinical Infectious Diseases.* (2020).

79. Ng, K. W., Faulkner, N., Cornish, G. H., *et al.* Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science,* eabe1107. (2020) (Nov. 6, 2020).

80. Gelman, A. & Carpenter, B. Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69,** 1269–1283. (2020) (Nov. 2020).

81. Bobrovitz, N., Arora, R. K., Yan, T., *et al.* Lessons from a rapid systematic review of early SARS-CoV-2 serosurveys. *medRxiv : the preprint server for health sciences,* 2020.05.10.20097451 (May 2020).

82. Liberati, A., Altman, D. G., Tetzlaff, J., *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *British Medical Journal* **339.** (2020) (July 21, 2009).

83. Sampson, M., Shojania, K. G., Garritty, C., *et al.* Systematic reviews can be produced and published faster. *Journal of Clinical Epidemiology* **61,** 531–536 (June 2008).

84. Shojania, K. G., Sampson, M., Ansari, M. T., *et al.* How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. *Annals of Internal Medicine* **147,** 224–233. (2022) (Aug. 21, 2007).

85. Macdonald, H., Loder, E. & Abbasi, K. Living systematic reviews at The BMJ. *BMJ* **370,** m2925. (2022) (July 30, 2020).

86. Elliott, J. H., Turner, T., Clavisi, O., *et al.* Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLOS Medicine* **11,** e1001603. (2021) (Feb. 18, 2014).

87. Siemieniuk, R. A., Bartoszko, J. J., Ge, L., *et al.* Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ* **370,** m2980. (2022) (July 30, 2020).

88. Allotey, J., Fernandez, S., Bonet, M., *et al.* Clinical manifestations, risk factors, and maternal and perinatal outcomes of coronavirus disease 2019 in pregnancy: living systematic review and meta-analysis. *BMJ* **370,** m3320. (2022) (Sept. 1, 2020).

89. Bir, A., Freeman, N., Chew, R., *et al.* Making Evidence Actionable: Interactive Dashboards, Bayes, and Health Care Innovation. *eGEMs* **7,** 40. (2020) (Aug. 5, 2019).

90. Concannon, D., Herbst, K. & Manley, E. Developing a Data Dashboard Framework for Population Health Surveillance: Widening Access to Clinical Trial Findings. *JMIR Formative Research* **3.** (2020) (Apr. 4, 2019).

91. Dixit, R. A., Hurst, S., Adams, K. T., *et al.* Rapid development of visualization dashboards to enhance situation awareness of COVID-19 telehealth initiatives at a multihospital healthcare system. *Journal of the American Medical Informatics Association* **27,** 1456–1461. (2022) (Sept. 1, 2020).

92. Berry, I., Soucy, J.-P. R., Tuite, A. & Fisman, D. Open access epidemiologic data and an interactive dashboard to monitor the COVID-19 outbreak in Canada. *CMAJ : Canadian Medical Association Journal* **192,** E420. (2022) (Apr. 14, 2020).

93. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., *et al.* A global database of COVID-19 vaccinations. *Nature Human Behaviour* **5,** 947–953. (2022) (July 2021).

94. Hale, T., Angrist, N., Goldszmidt, R., *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour* **5,** 529–538. (2022) (Apr. 2021).

95. *Ontario COVID-19 Data Tool* Public Health Ontario. (2022).

96. Curtis, H. J. & Goldacre, B. OpenPrescribing: normalised data and software tool to research trends in English NHS primary care prescribing 1998–2016. *British Medical Journal Open* **8.** (2020) (Feb. 23, 2018).

97. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. *Our World in Data Coronavirus Pandemic (COVID-19) Dataset* Our World in Data. (2020).

98. Gianola, S., Jesus, T. S., Bargeri, S. & Castellini, G. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. *PLOS ONE* **15,** e0240123. (2020) (Oct. 6, 2020).

99. Majumder, M. S. & Mandl, K. D. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *The Lancet Global Health* **8,** e627–e630. (2020) (May 1, 2020).

100. Arora, R. K., Joseph, A., Wyk, J. V., *et al.* SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. *The Lancet Infectious Diseases* **21,** E75–E76. (2020) (Aug. 4, 2020).

101. Bergeri, I., Whelan, M., Ware, H., *et al. Global SARS-CoV-2 seroprevalence: a systematic review and meta-analysis of standardized population-based studies from Jan 2020-May 2022* July 26, 2022. (2022).

102. Lewis, H. C., Ware, H., Whelan, M., *et al.* SARS-CoV-2 infection in Africa: a systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021. *BMJ Global Health* **7,** e008793. (2022) (Aug. 1, 2022).

103. Jackson, J. L., Kuriyama, A., Anton, A., *et al.* The Accuracy of Google Translate for Abstracting Data From Non-English-Language Trials for Systematic Reviews. *Annals of internal medicine* **171,** 677–679 (Nov. 2019).

104. Razzouk, R. & Shute, V. What Is Design Thinking and Why Is It Important? *Review of Educational Research* **82,** 330–348. (2022) (Sept. 1, 2012).

105. Barber, R. M., Sorensen, R. J. D., Pigott, D. M., *et al.* Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. *The Lancet* **399,** 2351–2380. (2022) (June 25, 2022).

106. Wang, H., Paulson, K. R., Pease, S. A., *et al.* Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet* **399,** 1513–1536. (2022) (Apr. 16, 2022).

107. Brazeau, N., Verity, R., Jenks, S., *et al. Report 34: COVID-19 infection fatality ratio: estimates from seroprevalence* (Imperial College London, Oct. 29, 2020). (2022).

108. *SeroTracker - COVID-19 Immunity Task Force* (2022).

109. Dull, P. & Cramer, J. *COVAX Clinical Development and Operations* The Global Health Network. (2020).

110. Organization, W. H. *WHO Coronavirus Disease (COVID-19) Dashboard* (2021).

111. How we estimated the true death toll of the pandemic. *The Economist.* (2022).

112. Gamio, L. & Glanz, J. Just How Big Could India's True Covid Toll Be? *The New York Times.* (2022) (May 25, 2021).

113. Charumilind, S., Matt Craven, Jessica Lamb, *et al. When will the COVID-19 pandemic end? | McKinsey* www.mckinsey.com. (2022).

114. Boutron, I., Chaimani, A., Devane, D., *et al.* Interventions for the prevention and treatment of COVID-19: a living mapping of research and living network meta-analysis. *Cochrane Database of Systematic Reviews.* (2022) (2020).

115. Nguyen, V. T., Rivière, P., Ripoll, P., *et al.* Research response to coronavirus disease 2019 needed better coordination and collaboration: a living mapping of registered trials. *Journal of Clinical Epidemiology* **130,** 107–116. (2022) (Feb. 2021).

116. SeroTracker. *SeroTracker Newsletter* (2022).

117. *ESTUDIO ENE-COVID19: PRIMERA RONDA ESTUDIO NACIONAL DE SERO-EPIDEMIOLOGÍA DE LA INFECCIÓN POR SARS-COV-2 EN ESPAÑA INFORME PRELIMINAR 13 DE MAYO DE 2020* (Instituto de Salud Carlos III).

118. Earle, K. A., Ambrosino, D. M., Fiore-Gartland, A., *et al.* Evidence for antibody as a protective correlate for COVID-19 vaccines. *Vaccine* **39,** 4423–4428. (2021) (July 2021).

119. Chen, X., Chen, Z., Azman, A. S., *et al.* Serological evidence of human infection with SARS-CoV-2: a systematic review and meta-analysis. *The Lancet Global Health* **9,** e598–e609. (2021) (May 1, 2021).

120. Rostami, A., Sepidarkish, M., Fazlzadeh, A., *et al.* Update on SARS-CoV-2 seroprevalence – Regional and worldwide. *Clinical Microbiology and Infection* **0.** (2021) (Sept. 25, 2021).

121. Byambasuren, O., Dobler, C. C., Bell, K., *et al.* Comparison of seroprevalence of SARS-CoV-2 infections with cumulative and imputed COVID-19 cases: Systematic review. *PLOS ONE* **16,** e0248946. (2021) (Apr. 2, 2021).

122. Bergeri, I., Lewis, H. C., Subissi, L., *et al.* Early epidemiological investigations: World Health Organization UNITY protocols provide a standardized and timely international investigation framework during the COVID-19 pandemic. *Influenza and Other Respiratory Viruses* **n/a.** (2021) (n/a Oct. 5, 2021).

123. Barrie, M. B., Lakoh, S., Kelly, J. D., *et al.* SARS-CoV-2 antibody prevalence in Sierra Leone, March 2021: a cross-sectional, nationally representative, age-stratified serosurvey. *medRxiv,* 2021.06.27.21259271 (Jan. 1, 2021).

124. Murhekar, M. V., Bhatnagar, T., Thangaraj, J. W. V., *et al.* SARS-CoV-2 seroprevalence among the general population and healthcare workers in India, December 2020-January 2021. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* **108,** 145–155 (July 2021).

125. Al-Abri, S. S., Al-Wahaibi, A., Al-Kindi, H., *et al.* SARS-COV-2 antibody seroprevalence in the general population of Oman: results from four successive nationwide seroepidemiological surveys. *International Journal of Infectious Diseases* **112,** 269–277. (2021) (Nov. 1, 2021).

126. *COVID-19 Report: Publications, Clinical Trials, Funding* `https://reports.dimensions.ai`. Dimensions.

127. Basto-Abreu, A., Carnalla, M., Torres-Ibarra, L., *et al.* Nationally representative SARS-CoV-2 antibody prevalence estimates after the first epidemic wave in Mexico. *Nature Communications* **13,** 589. (2022) (Dec. 2022).

128. Canadian Blood Services. *COVID-19 Seroprevalence Brief Report Report #21A: April 1-15, 2022, Survey* (May 17, 2022).

129. Mercado-Reyes, M., Malagón-Rojas, J., Rodríguez-Barraquer, I., *et al.* Seroprevalence of anti-SARS-CoV-2 antibodies in Colombia, 2020: A population-based study. *The Lancet Regional Health - Americas* **9,** 100195. (2022) (May 1, 2022).

130. GISAID. *GISAID EpiCov™ Database* Dec. 2021. (2022).

131. Paulino-Ramirez, R., Báez, A. A., Degaudenzi, A. V. & Tapia, L. Seroprevalence of Specific Antibodies against SARS-CoV-2 from Hotspot Communities in the Dominican Republic. *The American Journal of Tropical Medicine and Hygiene* **103,** 2343–2346. (2021) (Dec. 2, 2020).

132. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36,** 1–48 (2010).

133. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health* **22,** 153–160. (2021) (Nov. 1, 2019).

134. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21,** 1539–1558. (2021) (2002).

135. Higgins, J. & Green, S. *Cochrane Handbook for Systematic Reviews of Interventions* 5.1.0 (The Cochrane Collaboration, Mar. 2011).

136. Fernández-Rojas, M. A., Luna-Ruiz Esparza, M. A., Campos-Romero, A., *et al.* Seroconversion dynamic and SARS-CoV-2 seropositivity in unvaccinated population during the first and second outbreaks in Mexico. *Scientific Reports* **12,** 5241. (2022) (Dec. 2022).

137. For Health metrics {and} Evaluation (IHME), I. *COVID-19: Estimating the historical time series of infections* (IHME, University of Washington, Seattle, Washington, USA, 2021).

138. Guo, L., Ren, L., Yang, S., *et al.* Profiling early humoral response to diagnose novel coronavirus disease (COVID-19). *Clinical Infectious Diseases* **71,** 778–785. (2020) (July 28, 2020).

139. Wilson, N., Kvalsvig, A., Barnard, L. T. & Baker, M. Case-Fatality Risk Estimates for COVID-19 Calculated by Using a Lag Time for Fatality. *Emerging Infectious Disease journal* **26,** 1339 (2020).

140. Vanella, P., Wiessner, C., Holz, A., *et al.* The role of age distribution, time lag between reporting and death and healthcare system capacity on case fatality estimates of COVID-19. *medRxiv : the preprint server for health sciences,* 2020.05.16.20104117 (Jan. 2020).

141. Deeks, J. J. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **21,** 1575–1600. (2021) (2002).

142. Cummings, P. The Relative merits of risk ratios and odds ratios. *Archives of Pediatrics & Adolescent Medicine* **163,** 438–445. (2021) (May 4, 2009).

143. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA : the journal of the American Medical Association* **324,** 782–793. (2020) (Aug. 2020).

144. Organization, W. H. & of Hygiene {and} Tropical Medicine, L. S. *Public Health and Social Measures* (2021).

145. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* **67.** (2021) (2015).

146. Team, R. C. *R: A language and environment for statistical computing* 2021.

147. Canto-Osorio, F., Stern, D., Pérez-Ferrer, C., *et al.* Seroprevalencia de SARS-CoV-2 en niños y adolescentes mexicanos en edad escolar. *Salud Pública de México* **63,** 803–806. (2021) (Nov. 5, 2021).

148. Dhaliwal, M., Small, R., Webb, D., *et al.* Covid-19 as a long multiwave event: implications for responses to safeguard younger generations. *BMJ,* e068123. (2022) (Jan. 27, 2022).

149. Verhagen, W., Bohl, D. K., Cilliers, J., *et al. Long-Term Socio-Economic Impacts of COVID-19 in African Contexts | UNDP in Africa* (UNDP, Mar. 11, 2021). (2022).

150. Boakye-Agyemang, C. *Six in seven COVID-19 infections go undetected in Africa* WHO | Regional Office for Africa. (2022).

151. Cohen, C., Kleynhans, J., Gottberg, A. v., *et al.* SARS-CoV-2 incidence, transmission and reinfection in a rural and an urban setting: results of the PHIRST-C cohort study, South Africa, 2020-2021. *medRxiv,* 2021.07.20.21260855. (2021) (July 22, 2021).

152. Tso, F. Y., Lidenge, S. J., Peña, P. B., *et al.* High prevalence of pre-existing serological cross-reactivity against severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) in sub-Saharan Africa. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* **102,** 577–583 (Jan. 2021).

153. Steinhardt, L. C., Ige, F., Iriemenam, N. C., *et al.* Cross-Reactivity of Two SARS-CoV-2 Serological Assays in a Setting Where Malaria Is Endemic. *Journal of Clinical Microbiology.* (2022) (Apr. 14, 2021).

154. Iesa, M. A. M., Osman, M. E. M., Hassan, M. A., *et al.* SARS-CoV-2 and Plasmodium falciparum common immunodominant regions may explain low COVID-19 incidence in the malaria-endemic belt. *New Microbes and New Infections* **38,** 100817. (2022) (Nov. 1, 2020).

155. Bobrovitz, N., Arora, R. K., Boucher, E., *et al. A systematic review and meta-analysis of SARS-CoV-2 seroprevalence studies aligned with the WHO population-based sero-epidemiological 'Unity' protocol CRD42020183634* Oct. 5, 2021.

156. For Health metrics {and} Evaluation (IHME), I. *COVID-19 Estimates* (IHME, University of Washington, Seattle, Washington, USA, 2021).

157. Suthar, A. B., Schubert, S., Garon, J., *et al.* Coronavirus Disease Case Definitions, Diagnostic Testing Criteria, and Surveillance in 25 Countries with Highest Reported Case Counts. *Emerging Infectious Diseases* **28,** 148–156. (2022) (Jan. 2022).

158. Bobrovitz, N., Noël, K., Li, Z., *et al. SeroTracker-RoB: an approach to automating reproducible risk of bias assessment of seroprevalence studies* (medRxiv, Mar. 11, 2022), 2021.11.17.21266471. (2022).

159. Seow, J., Graham, C., Merrick, B., *et al.* Longitudinal observation and decline of neutralizing antibody responses in the three months following SARS-CoV-2 infection in humans. *Nature Microbiology* **5,** 1598–1607. (2021) (Dec. 2020).

160. Hansen, C. H., Michlmayr, D., Gubbels, S. M., Mølbak, K. & Ethelberg, S. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. *The Lancet* **397,** 1204–1212. (2021) (Mar. 2021).

161. Brauner, J. M., Mindermann, S., Sharma, M., *et al.* Inferring the effectiveness of government interventions against COVID-19. *Science* **371,** eabd9338 (2021).

162. Deb, P., Furceri, D., Ostry, J. & Tawk, N. The Effect of Containment Measures on the COVID-19 Pandemic. *IMF Working Papers* **20.** (2021) (Aug. 7, 2020).

163. Talic, S., Shah, S., Wild, H., *et al.* Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *BMJ* **375,** e068302. (2021) (Nov. 18, 2021).

164. Goldman, M., Steele, W. R., Di Angelantonio, E., *et al.* Comparison of donor and general population demographics over time: a BEST Collaborative group study. *Transfusion* **57,** 2469–2476 (Oct. 2017).

165. Patel, E. U., Bloch, E. M., Grabowski, M. K., *et al.* Sociodemographic and behavioral characteristics associated with blood donation in the United States: a population-based study. *Transfusion* **59,** 2899–2907 (Sept. 2019).

166. (HQ), W. H. & Programme, W. H. E. *Enhancing response to Omicron SARS-CoV-2 variant* Jan. 21, 2022. (2022).

167. Gao, Y., Cai, C., Grifoni, A., *et al.* Ancestral SARS-CoV-2-specific T cells cross-recognize the Omicron variant. *Nature Medicine* **28,** 472–476 (Mar. 1, 2022).

168. Erikstrup, C., Hother, C. E., Pedersen, O. B. V., *et al.* Estimation of SARS-CoV-2 infection fatality rate by real-time antibody screening of blood donors. *medRxiv.* (2020) (Apr. 28, 2020).

169. Bendavid, E., Mulaney, B., Sood, N., *et al.* COVID-19 antibody seroprevalence in Santa Clara County, California. *International Journal of Epidemiology,* dyab010. (2021) (Feb. 22, 2021).

170. Cornelius, J. L. Reviewing the review process: Identifying sources of delay. *The Australasian Medical Journal* **5,** 26–29 (2012).

171. Graf, C. *Timeliness: An Essential Area for Better Peer Review* (2022).

172. Huisman, J. & Smits, J. Duration and quality of the peer review process: the author's perspective. *Scientometrics* **113,** 633–650. (2022) (Oct. 1, 2017).

173. Rodriguez, A. MGH Coronavirus Case Study In Chelsea Will Help Researchers Understand Immunity. *CBS Boston.* (2021) (Apr. 23, 2020).

174. Region, T. S. *27 april: Lägesrapport om covid-19* (The Stockholm Region, Apr. 27, 2021). (2021).

175. Ward, H., Whitaker, M., Tang, S. N., *et al. Vaccine uptake and SARS-CoV-2 antibody prevalence among 207,337 adults during May 2021 in England: REACT-2 study* (July 18, 2021), 2021.07.14.21260497. (2022).

176. Horbach, S. P. J. M. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies* **1,** 1056–1067. (2022) (Aug. 1, 2020).

177. Ravinetto, R., Caillet, C., Zaman, M. H., *et al.* Preprints in times of COVID19: the time is ripe for agreeing on terminology and good practices. *BMC Medical Ethics* **22,** 106. (2022) (July 28, 2021).

178. Munn, Z., Moola, S., Riitano, D. & Lisy, K. The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. *International Journal of Health Policy and Management* **3,** 123–128. (2020) (Aug. 13, 2014).

179. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer* **89,** 232–238. (2022) (July 2003).

180. Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer* **89,** 431–436. (2022) (Aug. 2003).

181. Grabowski, F., Kochańczyk, M. & Lipniacki, T. *The spread of SARS-CoV-2 variant Omicron with the doubling time of 2.0–3.3 days can be explained by immune evasion* (medRxiv, Jan. 19, 2022), 2021.12.08.21267494. (2022).

182. Fos, P. J., Honoré, P. A. & Honoré, R. L. *Coordination of Public Health Response: The Role of Leadership in Responding to Public Health Emergencies* (2022) (IntechOpen, Apr. 21, 2021).

183. *COVID-19 Primer* `https://covid19primer.com/`. COVID-19 Primer.

184. *The REACT 2 programme* Imperial College London. (2022).

185. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22.** (2022) (Mar. 30, 2017).

186. Einhauser, S., Peterhoff, D., Niller, H. H., *et al.* Spectrum Bias and Individual Strengths of SARS-CoV-2 Serological Tests—A Population-Based Evaluation. *Diagnostics* **11,** 1843. (2022) (Oct. 6, 2021).

187. Department of Health and Social Care (DHSC). *Validating COVID-19 tests in the private market* (2020).

188. Peluso, M. J., Takahashi, S., Hakim, J., *et al.* SARS-CoV-2 antibody magnitude and detectability are driven by disease severity, timing, and assay. *Science Advances* **7,** eabh3409 (July 2021).

189. NRL Science of Quality. *WHO SARS-CoV-2 Test Kit Comparative Study* (2021). (2021).

190. U.S. Food and Drug Administration. *Independent Evaluations of COVID-19 Serological Tests* (2020). (2021).

191. Van den Beld, M. J. C., Murk, J.-L., Kluytmans, J., *et al.* Increasing the Efficiency of a National Laboratory Response to COVID-19: a Nationwide Multicenter Evaluation of 47 Commercial SARS-CoV-2 Immunoassays by 41 Laboratories. *Journal of Clinical Microbiology* **59,** e0076721 (Aug. 18, 2021).

192. Australian Government Department of Health Therapeutic Goods Administration. *Post-market evaluation of serology-based point of care tests* (2021). (2021).

193. FIND Diagnostics for All. *SARS-CoV-2 Test Performance* (). (2021).

194. World Health Organization. *Target product profiles for priority diagnostics to support response to the COVID-19 pandemic v.1.0* (2020). (2021).

195. Stein, D. R., Osiowy, C., Gretchen, A., *et al.* Evaluation of commercial SARS-CoV-2 serological assays in Canadian public health laboratories. *Diagnostic Microbiology and Infectious Disease* **101,** 115412 (Nov. 2021).

196. Pourhoseingholi, M. A., Vahedi, M. & Rahimzadeh, M. Sample size calculation in medical studies. *Gastroenterology and Hepatology from Bed to Bench* **6,** 14–17 (2013).

197. Whitman, J. D., Hiatt, J., Mowery, C. T., *et al.* Test performance evaluation of SARS-CoV-2 serological assays. *medRxiv : the preprint server for health sciences* (2020).

198. Vengesai, A., Midzi, H., Kasambala, M., *et al.* A systematic and meta-analysis review on the diagnostic accuracy of antibodies in the serological diagnosis of COVID-19. *Systematic Reviews* **10,** 155. (2021) (May 26, 2021).

199. Theel, E. S. Performance Characteristics of High-Throughput Serologic Assays for Severe Acute Respiratory Syndrome Coronavirus 2 with Food and Drug Administration Emergency Use Authorization: A Review. *Clinics in Laboratory Medicine* **42,** 15–29 (Mar. 2022).

200. World Health Organization. *WHO/BS.2020.2403 Establishment of the WHO International Standard and Reference Panel for anti-SARS-CoV-2 antibody* (Nov. 18, 2020). (2021).

201. Speybroeck, N., Devleesschauwer, B., Joseph, L. & Berkvens, D. Misclassification errors in prevalence estimation: Bayesian handling with care. *International Journal of Public Health* **58,** 791–795. (2020) (Oct. 1, 2013).

202. Takahashi, S., Greenhouse, B. & Rodríguez-Barraquer, I. Are Seroprevalence Estimates for Severe Acute Respiratory Syndrome Coronavirus 2 Biased? *The Journal of Infectious Diseases* **222,** 1772–1775 (Nov. 9, 2020).

203. Lippi, G., Adeli, K. & Plebani, M. Commercial immunoassays for detection of anti-SARS-CoV-2 spike and RBD antibodies: urgent call for validation against new and highly mutated variants. *Clinical Chemistry and Laboratory Medicine* (Dec. 16, 2021).

204. Emmerich, P., Murawski, C., Ehmen, C., *et al.* Limited specificity of commercially available SARS-CoV-2 IgG ELISAs in serum samples of African origin. *Tropical Medicine & International Health* **26,** 621–631 (June 2021).

205. Bottomley, C., Otiende, M., Uyoga, S., *et al.* Quantifying previous SARS-CoV-2 infection through mixture modelling of antibody levels. *Nature Communications* **12,** 6196 (Oct. 26, 2021).

206. Irwin, N., Murray, L., Ozynski, B., *et al.* Age significantly influences the sensitivity of SARS-CoV-2 rapid antibody assays. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases* **109,** 304–309 (Aug. 2021).

207. Hippich, M., Holthaus, L., Assfalg, R., *et al.* A Public Health Antibody Screening Indicates a 6-Fold Higher SARS-CoV-2 Exposure Rate than Reported Cases in Children. *Med* **2,** 149–163.e4 (Feb. 2021).

208. Cohen, J. *The United States badly bungled coronavirus testing—but things may soon improve* Feb. 2020.

209. Rader, B., Gertz, A., Iuliano, A. D., *et al.* Use of At-Home COVID-19 Tests - United States, August 23, 2021-March 12, 2022. *MMWR. Morbidity and mortality weekly report* **71,** 489–494 (Apr. 2022).

210. Karlinsky, A. & Kobak, D. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife* **10** (eds Davenport, M. P., Lipsitch, M., Lipsitch, M., Simonsen, L. & Mahmud, A.) e69336 (June 2021).

211. Donnici, C., Ilincic, N., Cao, C., *et al. Timeliness of reporting of SARS-CoV-2 seroprevalence results and their utility for infectious disease surveillance* Feb. 2022.

212. MRC Centre for Global Infectious Disease Analysis. *Imperial College COVID-19 LMIC reports.* Future scenarios of the healthcare burden of COVID-19 in low- or middle-income countries. (2021).

213. O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., *et al.* Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590,** 140–145. (2021) (Feb. 2021).

214. Institute for Health Metrics and Evaluation. *COVID-19: Estimating the historical time series of infections.* HealthData.org. (2021).

215. Gelman, A., Vehtari, A., Simpson, D., *et al.* Bayesian workflow. *arXiv:2011.01808.* (2020) (Nov. 3, 2020).

216. Javage. *Modern Statistical Workflow: What is Modern Statistical Workflow?* Aug. 2016.

217. Grinsztajn, L., Semenova, E., Margossian, C. C. & Riou, J. Bayesian workflow for disease transmission modeling in Stan. *Statistics in Medicine* **40,** 6209–6234 (Nov. 2021).

218. Gelman, A., Carlin, J. B., Stern, H. S., *et al. Bayesian data analysis* (CRC Press, 2013).

219. Robert, C. & Casella, G. *Monte Carlo statistical methods* (Springer New York, 2013).

220. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112,** 859–877 (Apr. 3, 2017).

221. Rasmussen, C. E. & Ghahramani, Z. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems,* 505–512 (2003).

222. Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat].* (2020) (July 2018).

223. Neal, R. M. *et al.* MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo* **2,** 2 (2011).

224. Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid Monte Carlo. *Physics Letters B* **195,** 216–222 (1987).

225. Hoffman, M. D., Gelman, A., *et al.* The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15,** 1593–1623 (2014).

226. Carpenter, B., Gelman, A., Hoffman, M. D., *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* **76,** 1–32 (2017).

227. Fokianos, K. & Tjøstheim, D. Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102,** 563–578 (Mar. 2011).

228. Liu, Y. & Rocklöv, J. The effective reproductive number of the Omicron variant of SARS-CoV-2 is several times relative to Delta. *Journal of Travel Medicine* **29,** taac037 (May 2022).

229. Hart, W. S., Miller, E., Andrews, N. J., *et al.* Generation time of the alpha and delta SARS-CoV-2 variants: An epidemiological analysis. *The Lancet Infectious Diseases* **22,** 603–610 (May 2022).

230. Ghahramani, Z. *Graphical models* (2022).

231. Huang, C., Wang, Y., Li, X., *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395,** 497–506 (Feb. 2020).

232. Chiu, W. A. & Ndeffo-Mbah, M. L. Using test positivity and reported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLOS Computational Biology* **17,** e1009374 (Sept. 2021).

233. Lindeman, T. Canada plays catchup on rapid testing amid Omicron surge. *The Guardian* (Jan. 2022).

234. Bergeri, I., Whelan, M., Ware, H., *et al.* Global epidemiology of SARS-CoV-2 infection: A systematic review and meta-analysis of standardized population-based seroprevalence studies, Jan 2020-Dec 2021 (Feb. 2022).

235. Tuite, A. R., Fisman, D., Abe, K. T., *et al.* Estimating SARS-CoV-2 Seroprevalence in Canadian Blood Donors, April 2020 to March 2021: Improving Accuracy with Multiple Assays. *Microbiology Spectrum* **10,** e02563–21 (Feb. 2022).

236. Franceschi, V. B., Santos, A. S., Glaeser, A. B., *et al.* Population-based prevalence surveys during the Covid-19 pandemic: A systematic review. *Reviews in Medical Virology,* e2200 (Dec. 2020).

237. Kantor, R. S., Greenwald, H. D., Kennedy, L. C., *et al.* Operationalizing a routine wastewater monitoring laboratory for SARS-CoV-2. *PLOS Water* **1,** e0000007 (Feb. 2022).

238. Stone, M., Grebe, E., Sulaeman, H., *et al.* Evaluation of commercially available high-throughput SARS-CoV-2 serological assays for serosurveillance and related applications. *medRxiv,* 2021.09.04.21262414. (2021) (Sept. 15, 2021).

239. Favresse, J., Eucher, C., Elsen, M., *et al.* Persistence of anti-SARS-CoV-2 antibodies depends on the analytical kit: A report for up to 10 months after infection. *Microorganisms* **9,** 556 (Mar. 8, 2021).

240. Schallier, A., De Baets, S., De Bruyne, D., *et al.* Assay dependence of long-term kinetics of SARS-CoV-2 antibodies. *Diagnostic Microbiology and Infectious Disease* **100,** 115403. (2021) (Apr. 22, 2021).

241. Bond, K., Williams, E., Nicholson, S., *et al.* Longitudinal evaluation of laboratory-based serological assays for SARS-CoV-2 antibody detection. *Pathology* **53,** 773–779 (Oct. 2021).

242. Gottschalk, P. G. & Dunn, J. R. The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry* **343,** 54–65 (Aug. 2005).

243. O'Neill, B. The Classical Occupancy Distribution: Computation and Approximation. *The American Statistician* **75,** 364–375 (Oct. 2021).

244. Altarawneh, H. N., Chemaitelly, H., Hasan, M. R., *et al.* Protection against the Omicron Variant from Previous SARS-CoV-2 Infection. *The New England Journal of Medicine,* NEJMc2200133 (Feb. 2022).

245. Bradley, T., Grundberg, E., Selvarangan, R., *et al.* Antibody Responses after a Single Dose of SARS-CoV-2 mRNA Vaccine. *New England Journal of Medicine* **384,** 1959–1961 (May 2021).

246. Canadian Blood Services. *ABCs of eligibility to donating blood* Canadian Blood Services.

247. Van den Hurk, K., Zalpuri, S., Prinsze, F. J., Merz, E.-M. & de Kort, W. L. A. M. Associations of health status with subsequent blood donor behavior—An alternative perspective on the Healthy Donor Effect from Donor InSight. *PLOS ONE* **12,** e0186662 (Oct. 2017).

248. *Coronavirus (COVID-19) Infection Survey headline results, UK - Office for National Statistics*

249. Hoffman, M. D. & Gelman, A. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo* Nov. 2011.

250. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472. (2021) (Nov. 1992).

251. Gelman, A., Carlin, J. B., Stern, H. S., *et al.* in *Bayesian Data Analysis* 3rd ed., 284 (Chapman and Hall/CRC, New York, July 2015).

252. Government of Canada, S. C. *Population estimates, quarterly* Mar. 2022.

253. Berry, I., O'Neill, M., Sturrock, S. L., *et al.* A sub-national real-time epidemiological and vaccination database for the COVID-19 pandemic in Canada. *Scientific Data* **8,** 173. (2021) (July 15, 2021).

254. Consortium, C. *Overview of SARS-CoV-2 variants by country* 2022.

255. CCAHS Investigators. *Canadian COVID-19 Antibody and Health Survey (CCAHS)* Statistics Canada. (2021).

256. Tang, X., Sharma, A., Pasic, M., *et al.* SARS-CoV-2 seroprevalence during the first and second pandemic waves in Canada. *SSRN.* (2021) (Aug. 12, 2021).

257. Saeed, S., Drews, S. J., Pambrun, C., *et al.* SARS-CoV-2 seroprevalence among blood donors after the first COVID-19 wave in Canada. *Transfusion* **61,** 862–872 (Feb. 23, 2021).

258. Lewin, A., Therrien, R., De Serres, G., *et al.* SARS-CoV-2 seroprevalence among blood donors in Québec, and analysis of symptoms associated with seropositivity: a nested case-control study. *Canadian Journal of Public Health* **112,** 576–586 (May 17, 2021).

259. Charlton, C. L., Nguyen, L. T., Bailey, A., *et al.* Pre-vaccine positivity of SARS-CoV-2 antibodies in Alberta, Canada during the first two waves of the COVID-19 pandemic. *Microbiology Spectrum* **9,** e0029121 (Sept. 3, 2021).

260. Jacobs, E. Canada Administers Its 1st COVID-19 Vaccine Shots. *NPR* (Dec. 2020).

261. Brown, P. E., Fu, S. H., Bansal, A., *et al.* Omicron BA.1/1.1 SARS-CoV-2 Infection among Vaccinated Canadian Adults. *New England Journal of Medicine* **0,** null (May 2022).

262. Of Canada, P. H. A. *Demographics: COVID-19 vaccination coverage in Canada - Canada.Ca* Jan. 2021.

263. *Coronavirus: Here's a timeline of COVID-19 cases in Canada - National | Globalnews.Ca* Mar. 2020.

264. Of Canada, G. *COVID-19 daily epidemiology update*

265. Wood, A., Smith, J., Sawatsky, J., *et al.* Variation in the Canadian provincial and territorial responses to COVID-19 (Mar. 2021).

266. News, A. C. C. *Atlantic bubble 2.0 expected to begin by April 19, premiers say | CBC News* Mar. 2021.

267. Polisena, J., Ospina, M., Sanni, O., *et al.* Public health measures to reduce the risk of SARS-CoV-2 transmission in Canada during the early days of the COVID-19 pandemic: A scoping review. *BMJ Open* **11,** e046177 (Mar. 2021).

268. Lao, D. *Canada finds first cases of Omicron COVID-19 variant in Ontario. Here's what we know | Globalnews.Ca* Nov. 2021.

269. Kim, H.-Y., Weiß, C. H. & Möller, T. A. Models for autoregressive processes of bounded counts: How different are they? *Computational Statistics* **35,** 1715–1736 (Dec. 2020).

270. Modrak, M. *Identifying non-identifiability* May 2018.

271. Garritty, C., Stevens, A., Hamel, C., *et al.* Knowledge Synthesis in Evidence-Based Medicine. *Seminars in Nuclear Medicine. Evidence Based Nuclear Medicine* **49,** 136–144. (2021) (Mar. 1, 2019).

272. Guyatt, G. H., Sackett, D. L., Sinclair, J. C., *et al.* Users' Guides to the Medical Literature: IX. A Method for Grading Health Care Recommendations. *JAMA* **274,** 1800–1804. (2021) (Dec. 13, 1995).

273. Millard, T., Synnot, A., Elliott, J., *et al.* Feasibility and Acceptability of Living Systematic Reviews: Results from a Mixed-Methods Evaluation. *Systematic Reviews* **8,** 325. (2021) (Dec. 14, 2019).

274. Borges Migliavaca, C., Stein, C., Colpani, V., *et al.* How are systematic reviews of prevalence conducted? A methodological study. *BMC Medical Research Methodology* **20,** 96. (2022) (Apr. 26, 2020).

275. Hartling, L., Hamm, M. P., Milne, A., *et al.* Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology* **66,** 973–981 (Sept. 2013).

276. Hoy, D., Brooks, P., Woolf, A., *et al.* Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *Journal of Clinical Epidemiology* **65,** 934–939 (Sept. 2012).

277. Chisale, M. R. O., Ramazanu, S., Mwale, S. E., *et al.* Seroprevalence of anti-SARS-CoV-2 antibodies in Africa: A systematic review and meta-analysis. *Reviews in Medical Virology* **n/a,** e2271. (2021) (n/a July 6, 2021).

278. Galanis, P., Vraka, I., Fragkou, D., Bilali, A. & Kaitelidou, D. Seroprevalence of SARS-CoV-2 antibodies and associated factors in healthcare workers: A systematic review and meta-analysis. *Journal of Hospital Infection* **108,** 120–134 (Feb. 2021).

279. Rostami, A., Sepidarkish, M., Leeflang, M. M. G., *et al.* SARS-CoV-2 seroprevalence worldwide: A systematic review and meta-analysis. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **27,** 331–340 (Mar. 2021).

280. Jüni, P., Witschi, A., Bloch, R. & Egger, M. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. *JAMA : the journal of the American Medical Association* **282,** 1054–1060 (Sept. 1999).

281. Sanderson, S., Tatt, I. D. & Higgins, J. P. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology* **36,** 666–676 (June 2007).

282. Greenland, S. & O'rourke, K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics (Oxford, England)* **2,** 463–471 (Dec. 2001).

283. Migliavaca, C. B., Stein, C., Colpani, V., *et al.* Quality assessment of prevalence studies: a systematic review. *Journal of Clinical Epidemiology* **127,** 59–68 (Nov. 2020).

284. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* **15,** 155–163 (June 2016).

285. McGraw, K. & Wong, S. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1,** 30–46 (1996).

286. Shrout, P. & Fleiss, J. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86,** 420–428.

287. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* **6,** 284–290 (1994).

288. Hartling, L., Milne, A., Hamm, M. P., *et al.* Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *Journal of Clinical Epidemiology* **66,** 982–993 (Sept. 2013).

289. Hartling, L., Ospina, M., Liang, Y., *et al.* Risk of bias versus quality assessment of randomised controlled trials: Cross sectional study. *BMJ (Clinical research ed.)* **339,** b4012 (Oct. 2009).

290. Hartling, L., Bond, K., Vandermeer, B., *et al.* Applying the Risk of Bias Tool in a Systematic Review of Combination Long-Acting Beta-Agonists and Inhaled Corticosteroids for Persistent Asthma. *PLOS ONE* **6,** e17242 (Feb. 2011).

291. European Centre for Disease Prevention and Control. *COVID-19 surveillance guidance* (Stockholm, 2021).

292. Borah, R., Brown, A. W., Capers, P. L. & Kaiser, K. A. Analysis of the Time and Workers Needed to Conduct Systematic Reviews of Medical Interventions Using Data from the PROSPERO Registry. *BMJ Open.* **7,** e012545. (2021) (Feb. 1, 2017).

293. Marshall, I. J. & Wallace, B. C. Toward Systematic Review Automation: A Practical Guide to Using Machine Learning Tools in Research Synthesis. *Systematic Reviews* **8,** 163. (2021) (July 11, 2019).

294. Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F. & Ng, L. Research Screener: A Machine Learning Tool to Semi-Automate Abstract Screening for Systematic Reviews. *Systematic Reviews* **10,** 93. (2021) (Apr. 1, 2021).

295. Gates, A., Guitard, S., Pillay, J., *et al.* Performance and Usability of Machine Learning for Screening in Systematic Reviews: A Comparative Evaluation of Three Tools. *Systematic Reviews* **8,** 278. (2021) (Nov. 15, 2019).

296. Hamel, C., Kelly, S. E., Thavorn, K., *et al.* An Evaluation of DistillerSR's Machine Learning-Based Prioritization Tool for Title/Abstract Screening – Impact on Reviewer-Relevant Outcomes. *BMC Medical Research Methodology* **20,** 256. (2021) (Oct. 15, 2020).

297. Lerner, I., Créquit, P., Ravaud, P. & Atal, I. Automatic Screening Using Word Embeddings Achieved High Sensitivity and Workload Reduction for Updating Living Network Meta-Analyses. *Journal of Clinical Epidemiology* **108,** 86–94 (Apr. 2019).

298. Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. Rayyan—a Web and Mobile App for Systematic Reviews. *Systematic Reviews* **5,** 210. (2021) (Dec. 5, 2016).

299. Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C. & Schmid, C. H. Semi-Automated Screening of Biomedical Citations for Systematic Reviews. *BMC Bioinformatics* **11,** 55. (2021) (Jan. 26, 2010).

300. Huang, K.-C., Chiang, I.-J., Xiao, F., *et al.* PICO Element Detection in Medical Text without Metadata: Are First Sentences Enough? *Journal of Biomedical Informatics* **46,** 940–946. (2021) (Oct. 1, 2013).

301. Cohen, A. M., Hersh, W. R., Peterson, K. & Yen, P.-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association* **13,** 206–219. (2021) (Mar. 1, 2006).

302. Van de Schoot, R., de Bruin, J., Schram, R., *et al.* An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews. *Nature Machine Intelligence* **3,** 125–133. (2021) (2 Feb. 2021).

303. Przybyła, P., Brockmeier, A. J., Kontonatsios, G., *et al.* Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods* **9,** 470–488. (2022) (2018).

304. Shemilt, I., Simon, A., Hollands, G. J., *et al.* Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* **5,** 31–49. (2022) (2014).

305. Olofsson, H., Brolund, A., Hellberg, C., *et al.* Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research Synthesis Methods* **8,** 275–280. (2022) (2017).

306. Ltd, V. H. I. *Covidence* `https://www.covidence.org/`. Covidence.

307. Langley, G. J., Moen, R. D., Nolan, K. M., *et al. The improvement guide: a practical approach to enhancing organizational performance.* (John Wiley and Sons, June 3, 2009).

308. Christoff, P. Running PDSA cycles. *Current Problems in Pediatric and Adolescent Health Care* **48,** 198–201 (Aug. 1, 2018).

309. Government of Canada, I. A. P. o. R. E. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018) – Chapter 2: Scope and Approach* (2022).

310. Qin, X., Liu, J., Wang, Y., *et al.* Natural Language Processing Was Effective in Assisting Rapid Title and Abstract Screening When Updating Systematic Reviews. *Journal of Clinical Epidemiology* **133,** 121–129 (May 2021).

311. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding* `http://arxiv.org/abs/1810.04805`. May 24, 2019. (2021).

312. Gu, Y., Tinn, R., Cheng, H., *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. Version 5. *ACM Transactions on Computing for Healthcare* **3,** 1–23. (2021) (Jan. 2022).

313. Nye, B., Li, J. J., Patel, R., *et al. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature* in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* ACL 2018 (Association for Computational Linguistics, Melbourne, Australia, July 2018), 197–207. (2021).

314. Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P. & Murad, M. H. Error Rates of Human Reviewers during Abstract Screening in Systematic Reviews. *PLOS ONE* **15,** e0227742. (2021) (Jan. 14, 2020).

315. Rodó, X., San-José, A., Kirchgatter, K. & López, L. Changing climate and the COVID-19 pandemic: more than just heads or tails. *Nature Medicine* **27,** 576–579. (2022) (Apr. 2021).

316. Nelson, C. Engineered pathogens: the opportunities, risks and challenges. *The Biochemist* **41,** 34–39. (2022) (June 1, 2019).

317. Jackson, S. S., Sumner, L. E., Garnier, C. H., *et al.* The accelerating pace of biotech democratization. *Nature Biotechnology* **37,** 1403–1408. (2022) (Dec. 2019).

318. Fink, S., Ruoff, F., Stahl, A., *et al.* Multiplexed Serum Antibody Screening Platform Using Virus Extracts from Endemic Coronaviridae and SARS-CoV-2. *ACS Infectious Diseases,* acsinfecdis.0c00725. (2022) (Mar. 16, 2021).

319. Deng, X., Achari, A., Federman, S., *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nature Microbiology* **5,** 443–454. (2022) (Mar. 2020).

320. Ackerman, C. M., Myhrvold, C., Thakku, S. G., *et al.* Massively multiplexed nucleic acid detection with Cas13. *Nature* **582,** 277–282. (2022) (June 2020).

321. Alavi, A., Bogu, G. K., Wang, M., *et al.* Real-time alerting system for COVID-19 and other stress events using wearable data. *Nature Medicine* **28,** 175–184. (2022) (Jan. 2022).

322. Shaw, Frederic. *A History of MMWR* in collab. with (Goodman), (Richard), (Lindegren), (Mary) & (Ward), (John). (2022).

323. Re3data.Org. GISAID. (2022) (2012).

324. *CEPI's collaborative task force to assess COVID-19 vaccines on emerging viral strains* (2022).

325. Hadfield, J., Megill, C., Bell, S. M., *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34,** 4121–4123. (2022) (Dec. 1, 2018).

326. Piwowar, H. A., Becich, M. J., Bilofsky, H., Crowley, R. S. & Workspace, o. b. o. t. c. D. S.
bibinitperiod I. C. Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. *PLOS Medicine* **5,** e183. (2022) (Sept. 2, 2008).

327. Kolesnikov, N., Hastings, E., Keays, M., *et al.* ArrayExpress update–simplifying data submissions. *Nucleic Acids Research* **43,** D1113–1116 (Database issue Jan. 2015).

328. Sutton, C. & Gong, L. *Popularity of arXiv.org within Computer Science* Oct. 14, 2017. (2022).

329. Watson, C. Many researchers say they'll share data — but don't. *Nature* **606,** 853–853. (2022) (June 21, 2022).

330. Jacobs, B. & Popma, J. Medical research, Big Data and the need for privacy by design. *Big Data & Society* **6,** 2053951718824352. (2022) (Jan. 1, 2019).

331. Byambasuren, O., Cardona, M., Bell, K., *et al.* Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada* **5,** 223–234. (2021) (Dec. 1, 2020).

332. Jahan, N., Brahma, A., Kumar, M. S., *et al.* Seroprevalence of IgG antibodies against SARS-CoV-2 in India, March 2020 to August 2021: a systematic review and meta-analysis. *International Journal of Infectious Diseases* **116,** 59–67. (2022) (Mar. 1, 2022).

333. Goldbeck-Wood, S. Evidence on peer review-scientific quality control or smokescreen? *BMJ (Clinical research ed.)* **318,** 44–45 (Jan. 2, 1999).

334. Andersen, M. Z., Fonnes, S. & Rosenberg, J. Time from submission to publication varied widely for biomedical journals: a systematic review. *Current Medical Research and Opinion* **37,** 985–993 (June 2021).

335. Peter Lombard [@peterlombard]. *Google sheet of seroprevalence studies by @DrJarlov* Twitter. (2022).

336. Eran Bendavid. *Personal Communication* E-mail. Aug. 10, 2010.

337. Douglas R. Lowy. *Update: COVID-19 Serology and Immunology Capacity Building* (National Cancer Institute, June 15, 2020), 22.

338. Institute, N. C. *Abstraction Process for the Serology Working Group* (2020).

339. National Cancer Institute. *COVID-19 SeroHub* COVID-19 Seroprevalence Studies Hub. (2022).

340. Comms, D. *SeroTracker* COVID-19 Immunity Task Force. (2022).

341. *RKI - Coronavirus SARS-CoV-2 - Results on SARS-CoV-2-seroprevalence in the general population - Update September 2022* (2022).

342. *Coronavirus (COVID-19) Infection Survey, antibody data, UK - Office for National Statistics* (2022).

343. Dunn, A. G. & Bourgeois, F. T. Is it time for computable evidence synthesis? *Journal of the American Medical Informatics Association* **27,** 972–975. (2022) (June 1, 2020).

344. Alper, B. S., Richardson, J. E., Lehmann, H. P. & Subbian, V. It is time for computable evidence synthesis: The COVID-19 Knowledge Accelerator initiative. *Journal of the American Medical Informatics Association* **27,** 1338–1339. (2022) (Aug. 1, 2020).

345. Convenes, C. *Cochrane Convenes: Preparing for and responding to global health emergencies | Cochrane Rehabilitation* (2022).

346. Wissel, B. D., Van Camp, P. J., Kouril, M., *et al.* An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time. *Journal of the American Medical Informatics Association: JAMIA* **27,** 1121–1125 (July 1, 2020).

347. Glaveski, S. Remote Work Should Be (Mostly) Asynchronous. *Harvard Business Review.* (2022) (Dec. 1, 2021).

348. Pérez-Gaxiola, G., Verdugo-Paiva, F., Rada, G. & Flórez, I. D. Assessment of Duplicate Evidence in Systematic Reviews of Imaging Findings of Children With COVID-19. *JAMA network open* **4,** e2032690 (Jan. 4, 2021).

349. Forum, M. H. *Context for the inventory* McMaster Health Forum. (2022).

350. Abbott, A., Cyranoski, D., Jones, N., *et al.* Metrics: Do metrics matter? *Nature* **465,** 860–862 (June 17, 2010).

351. Broad, W. J. The publishing game: getting more for less. *Science (New York, N.Y.)* **211,** 1137–1139 (Mar. 13, 1981).

352. Sarewitz, D. The pressure to publish pushes down quality. *Nature* **533,** 147 (May 12, 2016).

353. Kucharski, A. J., Funk, S. & Eggo, R. M. The COVID-19 response illustrates that traditional academic reward structures and metrics do not reflect crucial contributions to modern science. *PLoS biology* **18,** e3000913 (Oct. 2020).

354. Carpenter, C. R., Cone, D. C. & Sarli, C. C. Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine* **21,** 1160–1172 (Oct. 2014).

355. Page, M. J., McKenzie, J. E., Bossuyt, P. M., *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Revista Espanola De Cardiologia (English Ed.)* **74,** 790–799 (Sept. 2021).

356. Hershberg, E. *How Software in the Life Sciences Actually Works (And Doesn't Work)* New Science. (2022).

357. Menni, C., Valdes, A. M., Polidori, L., *et al.* Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: A prospective observational study from the ZOE COVID Study. *The Lancet* **399,** 1618–1624 (Apr. 2022).

358. Vogel, G. *Pandemic signals from the Sewer—What virus levels in wastewater tell us* Mar. 2022.

359. Buckingham-Jeffery, E., Isham, V. & House, T. Gaussian process approximations for fast inference from infectious disease data. *Mathematical Biosciences* **301,** 111–120 (July 2018).

360. *Studies* Imperial College London. (2022).

361. *COVID-19 Infection Survey - Office for National Statistics* (2022).

362. Campbell, J. R., Uppal, A., Oxlade, O., *et al.* Active testing of groups at increased risk of acquiring SARS-CoV-2 in Canada: costs and human resource needs. *CMAJ* **192,** E1146–E1155. (2022) (Oct. 5, 2020).

363. Operational considerations for respiratory virus surveillance in Europe - July 2022. *TECHNICAL REPORT,* 37.

364. De Lusignan, S. & Correa, A. Opportunities and challenges of a World Serum Bank. *The Lancet* **389,** 250–251. (2022) (Jan. 2017).

365. Mina, M. J., Metcalf, C. J. E., McDermott, A. B., *et al.* A Global lmmunological Observatory to meet a time of pandemics. *eLife* **9,** e58989. (2022) (June 8, 2020).

366. Niederberger, M. & Spranger, J. Delphi Technique in Health Sciences: A Map. *Frontiers in Public Health* **8.** (2022) (2020).

367. *PubPeer - Search publications and join the conversation.* (2022).

368. Fenton, A., Lello, J. & Bonsall, M. Pathogen responses to host immunity: the impact of time delays and memory on the evolution of virulence. *Proceedings of the Royal Society B: Biological Sciences* **273,** 2083–2090. (2022) (Aug. 22, 2006).

369. *The Apollo Program for Biodefense - Winning the Race Against Biological Threats* Bipartisan Commission on Biodefense. (2022).

370. Consortium, T. N. A. O. *A Global Nucleic Acid Observatory for Biodefense and Planetary Health* Aug. 5, 2021. (2022).

371. Duarte, N., Arora, R. K., Bennett, G., *et al.* Deploying wearable sensors for pandemic mitigation: A counterfactual modelling study of Canada's second COVID-19 wave. *PLOS Digital Health* **1,** e0000100. (2022) (Sept. 6, 2022).

372. Wohl, S., Lee, E. C., DiPrete, B. L. & Lessler, J. *Sample Size Calculations for Variant Surveillance in the Presence of Biological and Systematic Biases* Jan. 1, 2022. (2022).

373. Okhmatovskaia, A., Verma, A. D., Barbeau, B., *et al.* A simulation model of waterborne gastro-intestinal disease outbreaks: description and initial evaluation. *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2010,** 557–561 (Nov. 13, 2010).

374. Department of Economics {and} Social Affairs, U. N. D. o. E.
bibinitperiod S. A. P. D. *2019 Revision of World Population Prospects* (2021).

375. *COVID-19 Global Humanitarian Response Plan* (United Nations Office for the Coordination of Humanitarian Affairs, Mar. 25, 2020). (2021).

376. Ritchie, H., Mathieu, E., Rodés-Guirao, L., *et al. Coronavirus Pandemic (COVID-19)* Our World in Data. (2021).