Transmission of *Mycobacterium tuberculosis* (Mtb) in a primary care clinic in South Africa before and during the COVID-19 pandemic: Comparing molecular, clinical, environmental, and patient tracking data from two observational field studies

Statistical analysis plan (SAP)

SAP authors: Nicolas Banholzer[1*], Lukas Fenner[1]

Study authors: Nicolas Banholzer[1*], Keren Middelkoop[2], Juane Leukes[2], Kathrin Zürcher[1], Robin Wood[2], Matthias Egger[1], Lukas Fenner[1]

[1] Institute of Social and Preventive Medicine, University of Bern, Switzerland

[2] Desmund Tutu HIV Centre, University of Cape Town, South Africa

[*] Corresponding author: nicolas.banholzer@unibe.ch

Version number: 0.4

Version date: May 14, 2024

Contents

1 Introduction

This document describes the proposed presentation and analysis for the main paper(s) reporting results from a study estimating and comparing the transmission of *Mycobacterium tuberculosis (*Mtb*)* in a primary care clinic in Cape Town (South Africa) before and during the COVID-19 pandemic. Various infection control measures were introduced during the COVID-19 pandemic, including restricted patient flows, general mask-wearing, and improved ventilation. To assess the effects of these measures, the following data were collected during both study periods:

1. **Molecular data**: Concentration of *Mtb* genomes in bioaerosol samples.
2. **Clinical data**: Number of registered and diagnosed TB patients.
3. **Environmental data**: CO_2 and relative humidity.
4. **Patient tracking**: Tracking data to monitor person-time in the clinic.

Any deviations from the statistical analysis plan will be described and justified in the final report. The analysis will be carried out by identified, appropriately qualified and experienced statisticians, who will ensure the integrity of the data during their processing.

2 Background

In a previous study, we estimated the risk of *Mtb* transmission in a primary care clinic in South Africa between July and August 2019[1]. The study design, measurements, and setting were described in a protocol before the study commenced [2]. We repeated the study in the same clinic between October and November 2021 (during the COVID-19 pandemic). Multiple infection control measures were put in place during this time:

- **Restricted patient flows**: Patients received fixed appointments to avoid crowding and reduce waiting times at the clinic. They were asked to arrive at the earliest one hour before the appointment, otherwise, they were asked to wait outside the clinic. In addition, clinical attendees with respiratory symptoms were tested outside the clinic before entering.
- **Mask-wearing**: Compulsory mask-wearing for all clinic attendees and staff members, except young children.
- **Active ventilation**: Opening of all windows and doors throughout the day to promote natural ventilation.

2.1 Objectives of the study

The primary objective of the study is to assess and compare the risk of *Mtb* infection in the primary care clinic between the period before (2019) and during the COVID-19 pandemic (2021), and to assess the individual effects of the pandemic-related infection control measures.

2.2 Primary outcome

The concentration of *Mtb* genomes in bioaerosol samples expressed as the number of DNA copies per microliter per hour of sampling.

2.3 Secondary outcomes

1. Proportion of diagnosed TB patients among registered patients.
2. Proportion of presumptive TB patients among registered patients.

3. Outdoor air change rate (computed from indoor CO_2 levels and room occupancy).
4. Person-time in the clinic.

2.4 Variables

1. Molecular data (daily)
   (a) Number of *Mtb* genomes in bioaerosol samples (DNA copies per microliters per hour of sampling)
2. Clinical data (daily)
   (a) Number of registered patients.
   (b) Number of presumptive TB patients.
   (c) Number of newly diagnosed TB patients.
3. Environmental data (daily by minute)
   (a) CO_2 (parts per million [ppm]).
   (b) Relative humidity (%).
   (c) Temperature (°C).
4. Patient tracking data (daily by second)
   (a) Time and location of people in the clinic (observation ID, timestamp, x- and y-coordinates).
5. Building data
   (a) Width, breadth, height, and volume of the waiting and TB room of the clinic.
6. Secondary data
   (a) Estimated incidence/prevalence of TB [3].

2.5 Hypothesis framework

The null hypothesis will be that there is no true difference in the outcomes between the two studies. We will use a Bayesian framework to assess the null hypothesis and use the posterior distribution of the parameter modeling the difference in the outcome by study year to assess the direction and magnitude of the effects of the pandemic-related infection control measures that were implemented in the clinic.

2.6 Research hypotheses

Our hypothesized links between pandemic-related infection control measures and the concentration of *Mtb* in bioaerosol samples are illustrated with a direct acyclic graph (DAG) model (**Figure 1**). In the following, we develop our research hypotheses about the effects of infection control measures.

First, the number of *Mtb*-containing particles in the air depends on the generation of infectious particles. The total amount of particles generated depends on the person-time of infectious people in the clinic. Person-time in the clinic was aimed to be reduced by restricted patient flows. Therefore, this intervention should reduce the generated amount of infectious particles by clinic attendees.

Hypothesis 1: Restricted patient flows reduced the concentration of *Mtb* in bioaerosol samples.

In addition to reducing the number of inhaled particles from the air, mask-wearing reduces the number of infectious particles exhaled into the air.

Hypothesis 2: Mask-wearing reduced the concentration of *Mtb* in bioaerosol samples.

Second, the detection of *Mtb* in bioaerosols depends on the survival of *Mtb* over time, which depends on multiple environmental factors such as airflow, ventilation, CO_2, humidity, and UV radiation [4, 5, 6]. More active natural ventilation should accelerate the removal of infectious particles from the indoor air.

Hypothesis 3: Active natural ventilation reduced the concentration of *Mtb* in bioaerosol samples.

Ventilation may also affect other variables such as humidity, but the net effect on the generation and survival of infectious particles in the air is unclear [4]. We, therefore, do not develop a hypothesis for the effect of ventilation via humidity on the concentration of *Mtb* in bioaerosol samples. Nevertheless, we will adjust for a potential effect of relative humidity during analysis.

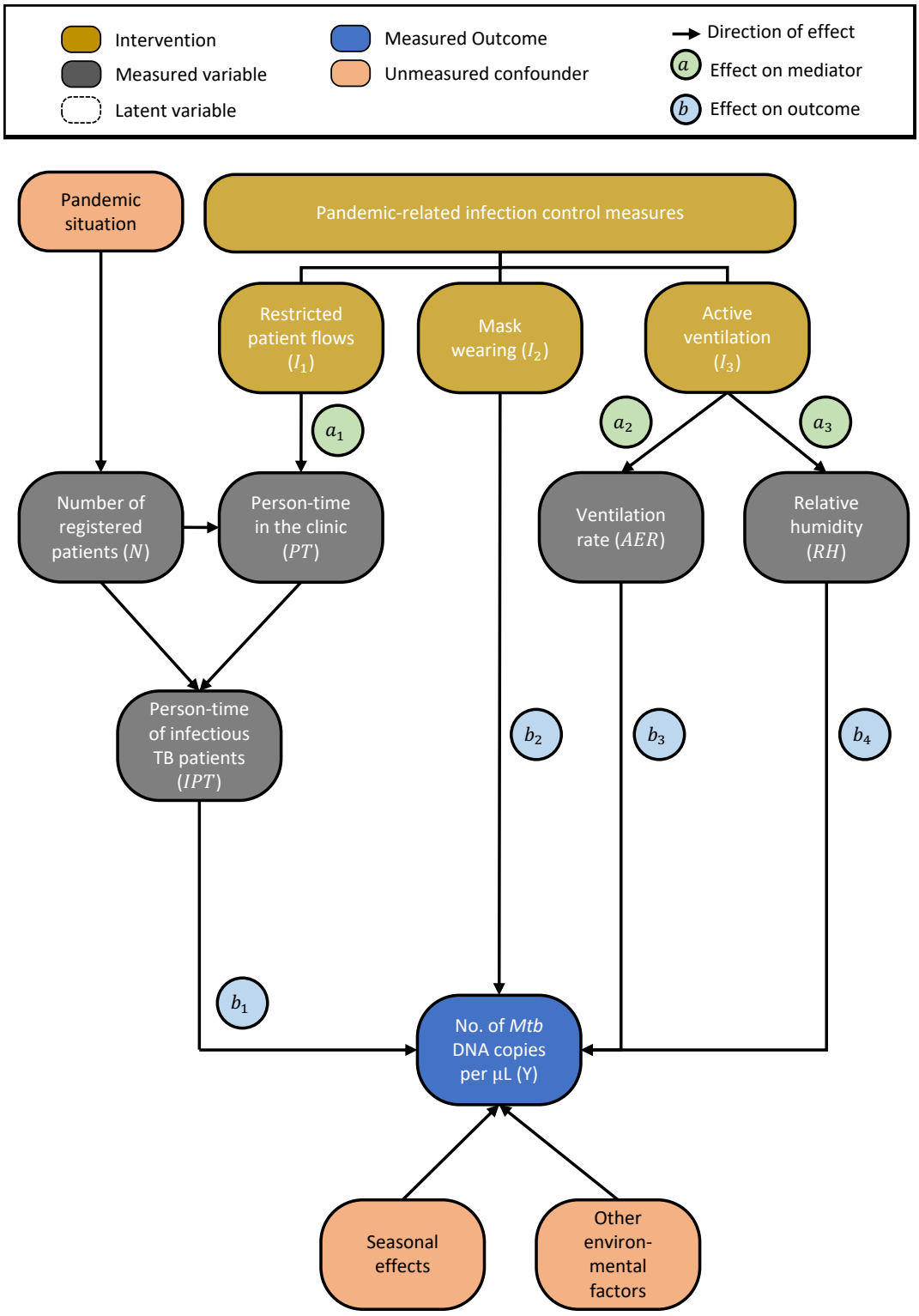
Our hypothesized effects are subject to potential unmeasured confounding and measurement error:

- Undiagnosed (subclinical) TB-infected patients also attend the clinic, may be infectious, and generate infectious *Mtb* particles [7].
- It is possible that *Mtb* transmission was lower during the COVID-19 pandemic (e.g. due to social distancing measures). However, it seems more likely that TB was under-diagnosed during the COVID-19 pandemic due to fewer people visiting the clinic [8, 9, 10]. People afraid of getting infected may have postponed their regular clinical visits, resulting in a smaller total number of clinical visits during the COVID-19 pandemic. We, therefore, adjust the effect of restricted patient flows on the person-time in the clinic for the daily number of registered patients.
- The generation of infectious particles depends also on the individual infectiousness of infected people [11]. More infectious individuals may generate more infectious particles per minute. The individual infectiousness of diagnosed TB patients is unknown. The removal of infectious particles depends not only on ventilation and humidity but also other, unmeasured environmental factors. For example, airflows may trap infectious particles in certain locations of the clinic [4]. Spatial variation in the concentration of infectious particles may introduce small measurement errors due to the fixed location of the bioaerosol sampling devices in each room of the clinic, yet the placement of the bioaerosol sampling devices was the same in both study years.
- The two studies that we compared were conducted between slightly different time periods, i.e. July to late August (before the pandemic) vs October to early November (during the pandemic), which possibly introduces seasonal effects. However, the weather conditions were comparable (only 1-2 months difference) and the infection control measures were maintained during the pandemic independent of the season.

3 Study populations

The target population includes all clinic attendees of the primary care clinic in South Africa between July and August 2019 (before the COVID-19 pandemic) and between October and November 2021. The primary outcome is the number of *Mtb* DNA copies per microliter per hour of sampling time. The sampling time was 2h in 2019 and 3h in 2021. Samples were collected twice per day, once in the morning and once in the afternoon, unless the clinic was temporarily closed. We have 39 observations (21 study days) in 2019 and 34 observations (19 study days) in 2021 when excluding observations where the primary outcome is missing. Considering also observations with missing

Figure 1: Directed acyclic graph (DAG) model showing the causal pathways from pandemic-related infection control measures (interventions) to the number of *Mtb* DNA copies per microliter per hour of sampling time (measured outcome), which is a proxy for the risk of infection (unmeasured outcome) in the clinic. Green-circled letters a denote effects of individual interventions on the mediators and blue-circled letters b denote effects of mediators on the outcome.



primary outcome, we have 41 observations in 2019 and 55 observations in 2021, corresponding to 21 and 28 study days, respectively.

4 Statistical analyses

4.1 Comparative analysis

We will employ descriptive statistics to summarize molecular, environmental, and patient tracking data. Categorical variables will be presented with the absolute number and as percentages. Continuous variables will be reported as the median and interquartile range (IQR) or mean and standard deviation (SD). We will create summary tables to describe all variables from molecular, environmental, and patient tracking data. We will create boxplots to compare the number of *Mtb* DNA copies per microliter in airborne samples, the number of diagnosed and registered patients attending the clinic, the number of person-time spent in the clinic, ventilation rates as computed from time-varying indoor CO_2 levels and room occupancy, and relative humidity.

4.2 Primary analysis

Based on the DAG model (**Figure 1**), we will use Bayesian linear regression models and mediation analysis to estimate the direct and indirect effects of the pandemic-related infection control measures I (binary variable for the study year 2021 vs 2019) on the log number of *Mtb* DNA copies per microliter per hour of sampling time in the waiting room of the clinic Y by daytime (morning and afternoon).

Direct effect of pandemic-related infection control measures

The direct effect of I on Y will be estimated with a Gaussian linear regression model

$$
\begin{aligned}
Y|\beta_0, \beta_1, \beta_2, \beta_3 &\sim \text{Normal}^+(\mu, \sigma) \\
\mu &= \beta_0 + \beta_1^{[D]} + \beta_2 \cdot I + \beta_3 \cdot N \\
\beta_0 &\sim \text{Student-t}_5\left(\mu_y, 2.5 s_y\right) \\
\beta_1^{[D]} &\sim \text{Normal}(0, \tau) \\
\tau &\sim \text{Student-t}_5^+(0, 1) \\
\beta_2 &\sim \text{Student-t}_5\left(0, 2.5\right) \\
\beta_3 &\sim \text{Student-t}_5\left(0, 2.5 \cdot \frac{s_y}{s_N}\right) \\
\sigma &\sim \text{Exponential}\left(\frac{1}{s_y}\right),
\end{aligned}
$$

where $\beta_0 + \beta_1^{[D]}$ is a day-specific varying intercept (non-centered parametrization) to consider intraday correlation, β_2 is the effect of I (differences between study years), and β_3 is the effect of N to adjust for the general impact of the pandemic situation on the number of clinic visits during the COVID-19 pandemic. Student-t_5^+ is a Student-t with 5 degrees of freedom, truncated at zero, μ_y and s_y are the empirical mean and standard deviation of the outcome, and s_N is the empirical standard deviations of the number of registered patients N.

Here and in the following, we will use weakly informative priors for all modeling parameters following recommendations on the choice of priors [12, 13, 14, 15, 16], unless stated otherwise. All priors represent our default but not definitive choices, and changes in the context of the data will be applied if necessary.

Indirect effects of pandemic-related infection control measures

We assume that mask-wearing I_2 affects Y directly, whereas restricted patient flows I_1 and active ventilation I_3 affect the outcome indirectly via the (infectious) person-time in the clinic PT (IPT), the ventilation rate AER, and relative humidity RH. The total indirect (mediated) effect of each intervention will be estimated using the product-of-coefficients method.

The indirect effect of I_1 on Y is

$$\beta_1^1 = a_1 \cdot b_1,$$

where a_1 is the effect of I_1 on PT (adjusted for N) and b_1 is the effect of IPT on Y (adjusted for I_2, AER, and RH).

The direct effect of I_2 on Y is

$$\beta_1^2 = b_2,$$

adjusted for IPT, AER, and RH.

The indirect effect of I_3 on Y is

$$\beta_1^3 = a_2 \cdot b_3 + a_3 \cdot b_4,$$

where a_2 is the effect of I_3 on AER, b_3 is the effect of AER on Y (adjusted for IPT, I_2, and RH), a_3 is the effect of I_3 on RH, and b_4 is the effect of RH on Y (adjusted for IPT, I_2, and AER).

IPT is the proportion of total person-time in the clinic PT that is contributed by diagnosed TB patients, which are considered infectious, and is computed as the proportion diagnosed TB patients among all registered patients times the person-time in the clinic PT. AER will be computed from time-varying indoor CO_2 levels and the number of people in the clinic. All variables will be computed or summarized by daytime. RH will be summarized with the mean.

4.3 Secondary analyses

4.3.1 Proportion of diagnosed TB patients among registered patients

The daily proportion of diagnosed TB patients among registered patients PD will be modeled with a Binomial logistic regression model

$$\begin{aligned}
PD, \beta_0, \beta_1, \beta_2 &\sim \text{Binomial-Logit}(\theta) \\
\theta &= \text{logit}(\mu) \\
\mu &= \beta_0 + \beta_1^{[D]} + \beta_2 \cdot I \\
\beta_0 &\sim \text{Student-t}_5(0, 2.5) \\
\beta_1^{[D]} &\sim \text{Normal}(0, \tau) \\
\tau &\sim \text{Student-t}_5^+(0, 1) \\
\beta_2 &\sim \text{Student-t}_5\,(0, 2.5)\ .
\end{aligned}$$

4.3.2 Proportion of presumptive TB patients among registered patients

The daily proportion of presumptive TB patients among registered patients PP will be modeled with a Binomial logistic regression model

$$\begin{aligned} PP, \beta_0, \beta_1, \beta_2 &\sim \text{Binomial-Logit}(\theta) \\ \theta &= \text{logit}(\mu) \\ \mu &= \beta_0 + \beta_1^{[D]} + \beta_2 \cdot I \\ \beta_0 &\sim \text{Student-t}_5(0, 2.5 \cdot s_{\text{PP}}) \\ \beta_1^{[D]} &\sim \text{Normal}(0, \tau) \\ \tau &\sim \text{Student-t}_5^{+}(0, 1) \\ \beta_2 &\sim \text{Student-t}_5\,(0, 2.5)\ . \end{aligned}$$

4.3.3 Outdoor air exchange rate

The outdoor air exchange rate AER will be computed daily from time-varying indoor CO_2 levels and the number of people in the clinic using a transient mass balance model [17]. The AER will be modeled with a Gaussian linear regression model

$$\begin{aligned} AER|\beta_0, \beta_1, \beta_2 &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \beta_0 + \beta_1^{[D]} + \beta_2 \cdot I \\ \beta_0 &\sim \text{Student-t}_5(\mu_{\text{ACH}}, 2.5 \cdot s_{\text{ACH}}) \\ \beta_1^{[D]} &\sim \text{Normal}(0, \tau) \\ \tau &\sim \text{Student-t}_5^{+}(0, 1) \\ \beta_2 &\sim \text{Student-t}_5\,(0, 2.5) \\ \sigma &\sim \text{Exponential}\left(\frac{1}{s_{\text{ACH}}}\right). \end{aligned}$$

4.3.4 Person-time in the clinic

The person-time in the clinic PT will be computed from the patient-tracking data and will be modeled with a linear regression model

$$\begin{aligned} PT|\beta_0, \beta_1, \beta_3 &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \beta_0 + \beta_1^{[D]} + \beta_2 \cdot I + \beta_3 \cdot N \\ \beta_0 &\sim \text{Student-t}_5(\mu_{\text{PT}}, 2.5 \cdot s_{\text{PT}}) \\ \beta_1^{[D]} &\sim \text{Normal}(0, \tau) \\ \tau &\sim \text{Student-t}_5^{+}(0, 1) \\ \beta_2 &\sim \ \text{Student-t}_5(0, 2.5) \\ \beta_3 &\sim \text{Student-t}_5\left(0, 2.5 \cdot \frac{s_{\text{PT}}}{s_{\text{x}}}\right) \\ \sigma &\sim \text{Exponential}\left(\frac{1}{s_{\text{PT}}}\right). \end{aligned}$$

The model will be adjusted for the number of registered patients N.

4.3.5 Subgroup analysis

The main analysis will be performed only for the waiting room of the clinic since tracking data was not collected in the TB room. As a subgroup analysis, we will perform a comparative analysis of the primary outcome in the TB room.

5 Supplementary analyses

5.1 Modeling transmission risks

We will model transmission risks using the Wells-Riley equation[18] as modified by Rudnick and Milton[19]

$$P = 1 - \exp\left(-\frac{fIqt}{n}\right),$$

where P is the risk of infection, f is the fraction of rebreathed air, I, is the number of infectious individuals in space, n is the total number of individuals in space, q is the generation rate of infectious quanta (doses of infectious particles), and t is the duration of exposure. The number of infectious individuals can be determined based on clinical data. The fraction of rebreathed air can be computed from CO_2 levels. The number of individuals in space and exposure times can be determined based on the patient tracking data. The unknown parameter q will be assumed based on reported estimates in the literature [20, 21, 22, 11, 23, 18]. The change in the generated infectious quanta due to mask wearing will also be informed by estimates from the literature [24, 25].

We will model the risk of infection based on the number of infectious people in space based on clinical data (presumptive and/or diagnosed TB patients) or reported estimates for the prevalence of TB by the World Health Organization[3].

6 Missing data

Missing data will be imputed using multivariate imputation by chained equations assuming missing at random. Each variable (number of registered patients, presumptive TB cases, diagnosed TB cases, person-time in the clinic, air change rate, relative humidity, and number of *Mtb* DNA copies per μL) will be imputed based on all other variables.

Analyses will be performed on each imputed dataset and the posterior draws of the modeled intervention effects will be pooled.

7 Statistical software employed

The statistical software R (version 4.2.1 or later) with RStudio will be used for all statistical analyses. Data pre-processing and comparative analyses will be performed in the `tidyverse` package (version 1.3.2 or later). Bayesian analyses will be performed with the probabilistic programming language Stan (version 2.21.0 or later), using the R interface provided by the `rstan` package (version 2.21.7 or later) and the `brms` package (version 2.18.0). The MICE package package (version 3.16.0) will be used for imputation.

References

[1] Zürcher K, Riou J, Morrow C, Ballif M, Koch A, Bertschinger S, et al. Estimating Tuberculosis transmission risks in a primary care clinic in South Africa: Modeling of environmental and clinical data. The Journal of Infectious Diseases. 2022;225(9):1642–1652. doi:10.1093/infdis/jiab534.

[2] Zürcher K, Morrow C, Riou J, Ballif M, Koch AS, Bertschinger S, et al. Novel approach to estimate Tuberculosis transmission in primary care clinics in sub-Saharan Africa: protocol of a prospective study. BMJ Open. 2020;10(8):e036214. doi:10.1136/bmjopen-2019-036214.

[3] World Health Organization. Global Tuberculosis Report 2022; 2022. Available from: `https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022` [last accessed 2023-05-15].

[4] Wang CC, Prather KA, Sznitman J, Jimenez JL, Lakdawala SS, Tufekci Z, et al. Airborne transmission of respiratory viruses. Science. 2021;373(6558):eabd9149. doi:10.1126/science.abd9149.

[5] Morawska L, Allen J, Bahnfleth W, Bluyssen PM, Boerstra A, Buonanno G, et al. A paradigm shift to combat indoor respiratory infection. Science. 2021;372(6543):689–691. doi:10.1126/science.abg2025.

[6] Haddrell A, Oswin H, Otero-Fernandez M, Robinson JF, Cogan T, Alexander R, et al. Ambient carbon dioxide concentration correlates with SARS-CoV-2 aerostability and infection risk. Nature Communications. 2024;15:3487. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11045827/`. doi:10.1038/s41467-024-47777-5.

[7] Patterson B, Bryden W, Call C, McKerry A, Leonard B, Seldon R, et al. Cough-independent production of viable Mycobacterium tuberculosis in bioaerosol. Tuberculosis. 2021;126:102038. doi:10.1016/j.tube.2020.102038.

[8] Soko RN, Burke RM, Feasey HRA, Sibande W, Nliwasa M, Henrion MYR, et al. Effects of Coronavirus disease pandemic on Tuberculosis notifications, Malawi. Emerging Infectious Diseases. 2021;27(7). doi:10.3201/eid2707.210557.

[9] Pillay Y, Pienaar S, Barron P, Zondi T. Impact of COVID-19 on routine primary healthcare services in South Africa. South African Medical Journal. 2021;111(8):714–719. doi:10.7196/SAMJ.2021.v111i8.15786.

[10] Uwishema O, Badri R, Onyeaka H, Okereke M, Akhtar S, Mhanna M, et al. Fighting Tuberculosis in Africa: The current situation amidst the COVID-19 pandemic. Disaster Medicine and Public Health Preparedness. 2022;16(6):1–3. doi:10.1017/dmp.2022.142.

[11] Escombe AR, Moore DAJ, Gilman RH, Pan W, Navincopa M, Ticona E, et al. The infectiousness of Tuberculosis patients coinfected with HIV. PLOS Medicine. 2008;5(9):e188. doi:10.1371/journal.pmed.0050188.

[12] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics. 2008;2(4):1360–1383. doi:10.1214/08-AOAS191.

[13] Gelman A. Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine. 2008;27(15):2865–2873. doi:10.1002/sim.3107.

[14] Gelman A, Hill J, Vehtari A. Regression and other stories. Cambridge University Press; 2020.

[15] Stan Development Team. Stan modeling language users guide (Version 2.21); 2022. Available from: `https://mc-stan.org/docs/2_21/reference-manual/effective-sample-size-section.html` [last accessed 2023-05-15].

[16] Gabry J, Goodrich B. Prior Distributions for rstanarm models; 2023. Available from: `https://cran.r-project.org/web/packages/rstanarm/vignettes/priors.html` [last accessed 2023-05-15].

[17] Batterman S. Review and Extension of CO2-Based Methods to Determine Ventilation Rates with Application to School Classrooms. International Journal of Environmental Research and Public Health. 2017;14(2):145. doi:10.3390/ijerph14020145.

[18] Riley RL, Mills CC, O'grady F, Sultan LU, Wittstadt F, Shivpuri DN. Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. American Review of Respiratory Disease. 1962;85:511–525. doi:10.1164/arrd.1962.85.4.511.

[19] Rudnick SN, Milton DK. Risk of indoor airborne infection transmission estimated from carbon dioxide concentration. Indoor Air. 2003;13(3):237–245. doi:10.1034/j.1600-0668.2003.00189.x.

[20] Mikszewski A, Stabile L, Buonanno G, Morawska L. The airborne contagiousness of respiratory viruses: A comparative analysis and implications for mitigation. Geoscience Frontiers. 2021;13(6):101285. doi:10.1016/j.gsf.2021.101285.

[21] Banholzer N, Schmutz R, Middelkoop K, Hella J, Egger M, Wood R, et al. Airborne transmission risks of tuberculosis and COVID-19 in schools in South Africa, Switzerland, and Tanzania: Modeling of environmental data. PLOS Global Public Health. 2024;4(1):e0002800. doi:10.1371/journal.pgph.0002800.

[22] Andrews JR, Morrow C, Walensky RP, Wood R. Integrating social contact and environmental data in evaluating Tuberculosis transmission in a South African township. The Journal of Infectious Diseases. 2014;210(4):597–603. doi:10.1093/infdis/jiu138.

[23] Nardell EA, Keegan J, Cheney SA, Etkind SC. Airborne infection: Theoretical limits of protection achievable by building ventilation. American Review of Respiratory Disease. 1991;144(2):302–306. doi:10.1164/ajrccm/144.2.302.

[24] McCreesh N, Karat AS, Baisley K, Diaconu K, Bozzani F, Govender I, et al. Modelling the effect of infection prevention and control measures on rate of Mycobacterium tuberculosis transmission to clinic attendees in primary health clinics in South Africa. BMJ Glob Health. 2021;6(10):e007124. doi:10.1136/bmjgh-2021-007124.

[25] Dharmadhikari AS, Mphahlele M, Stoltz A, Venter K, Mathebula R, Masotla T, et al. Surgical face masks worn by patients with multidrug-resistant tuberculosis: Impact on infectivity of air on a hospital ward. Am J Respir Crit Care Med. 2012;185(10):1104–1109. doi:10.1164/rccm.201107-1190OC.