1-1-2020

The Wisdom of a Crowd of Near-Best Fits: Drug-Resistant Tuberculosis in the United States

Ellie Mainou
Pennsylvania State University

Gwen Spencer
Convoy Inc.

Dylan Shepardson
Mount Holyoke College

Robert Dorit
Smith College, rdorit@smith.edu

Recommended Citation

Mainou, Ellie, Gwen Spencer, Dylan Shepardson, and Robert Dorit. 2020. "The Wisdom of a Crowd of Near-Best Fits". Letters in Biomathematics 7 (1), 15–35. https://lettersinbiomath.journals.publicknowledgeproject.org/index.php/lib/article/view/261.

LETTERS IN BIOMATHEMATICS
An International Journal

Intercollegiate Biomathematics Alliance

RESEARCH ARTICLE

OPEN ACCESS

The wisdom of a crowd of near-best fits: Drug-resistant tuberculosis in the United States

Ellie Mainou[a], Gwen Spencer[b], Dylan Shepardson[c], Robert Dorit[d]

[a]Department of Biology, Pennsylvania State University, University Park, PA; [b]Convoy, Inc., Seattle, WA; [c]Department of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA; [d]Department of Biological Sciences, 5 College Biomathematics Program, Smith College, Northampton, MA;

ABSTRACT

Antibiotic-resistant tuberculosis (TB) strains pose a major challenge to TB eradication. Existing US epidemiological models have not fully incorporated the impact of antibiotic-resistance. To develop a more realistic model of US TB dynamics, we formulated a compartmental model integrating single- and multi-drug resistance. We fit twenty-seven parameters to twenty-two years of historical data using a genetic algorithm to minimize a non-differentiable error function. Since counts for several compartments are not available, many parameter combinations achieve very low error. We demonstrate that a crowd of near-best fits can provide compelling new evidence about the ranges of key parameters. While available data is sparse and insufficient to produce point estimates, our crowd of near-best fits computes remarkably consistent predictions about TB prevalence. We believe that our crowd-based approach is applicable to a common problem in mathematical biological research, namely situations where data are sparse and reliable point estimates cannot be directly obtained.

ARTICLE HISTORY
Received November 27, 2018
Accepted December 18, 2019

KEYWORDS
model fitting, tuberculosis, disease dynamics, compartmental models, genetic algorithm

1 Introduction

The emergence of drug-resistant bacterial strains associated with a number of consequential human diseases poses a major threat to public health efforts over the next several decades. The World Health Organization (WHO) estimates that one third of the population may harbor TB (WHO, 2014). One major challenge to the global elimination of TB involves the increasing prevalence of drug-resistant strains. Even in countries where the near-elimination of TB once seemed possible, the presence of these new disease variants may be driving the rapid growth of infected and latently infected populations. Critically, the ability of researchers to reason in a data-driven way about how drug-resistance will impact future disease dynamics is undermined by a relative shortage of data. New strains have short histories, and incidence and prevalence estimates provided by public health agencies may only be tracking a subset of the relevant drug-resistant strains. Further, reliable estimates for important epidemiological parameters often do not exist in the literature, because of the rapid emergence of the new disease variants, local variability, and the inherent difficulty and expense of measuring these values directly.

In this paper, we develop a methodological approach meant to address these challenges, and use it to explore the potential impacts of drug-resistant strains on the future prevalence of tuberculosis in the United States. We expand on the prominent and widely-cited compartment model of Hill et al. (2012). We add new compartments to describe annual active infections by two drug-resistant strains of TB tracked since 1993 by the Centers for Disease Control and Prevention (CDC), and one strain that has not been tracked. These three new drug-resistant strains each also require a compartment to describe latent infections (that are also not CDC-tracked). In most cases, where Hill's model would have a single parameter (e.g. the proportion of immigrants that are latently infected with TB), our model now has four parameters (e.g. the proportion of immigrants that are latently infected with drug-susceptible TB, as well as the proportion latently infected with each of our three additional drug-resistant strains).

Our parameterization attempt of this complex model indicates that the relative sparsity of the available CDC data results in several challenging obstacles. A deterministic parameter-fitting algorithm returns parameter combinations that are highly dependent on an arbitrary choice of initial conditions. Because many widely different parameter combinations appear to lead

CONTACT Ellie Mainou ejm6007@psu.edu

to comparably low error function values, the conventional wisdom that the lowest-error parameter combination gives the best evidence about the true values of the parameters is problematic. On the other hand, a randomized sequential parameter-fitting method produces bimodal distributions for more than half of the parameters in our model. These behaviors expose the central hazard of our effort to incorporate drug-resistant strains: because we must rely on relatively sparse data to fit a large number of parameters, our error function exhibits a complex set of local minima. Further, sequential parameter-fitting methods become trapped at local minima on the boundary of the region of plausible parameter combinations, and fail to detect many better-quality local minima within the parameter range.

To conduct a more thorough survey of the local minima of our error function (which correspond to plausible parameterizations of our model), we implement a randomized genetic algorithm. Though we still find that runs of the genetic algorithm converge to a variety of local minima, focusing on a "crowd" of the lowest-error parameter combinations reveals remarkable consistency on several counts. First, predictions of future rates of TB in the United States appear quite stable across our crowd of low-error parameter combinations. Second, for several hard-to-directly-measure parameters, our crowd of low-error fits displays a striking degree of consensus about their values. In addition to the predictions we offer about future trajectories of drug-resistant TB and TB-specific epidemiological parameters, we argue that the crowd-based methodology we introduce here will provide useful directions for other mathematical modelers grappling with model parameterization in data-sparse settings.

2 Tuberculosis: Background and Prior Literature

Worldwide, in 2016, 4.1% of new TB cases were caused by multi-drug resistant (MDR) strains, defined as strains resistant to isoniazid and rifampin (WHO, 2016). Sharma et al. (2017) estimate that by 2040 up to 12.4% of cases in India, 32.4% in Russia and 5.7% in South Africa may be caused by MDR strains. Extensively drug-resistant (XDR) strains of *Mycobacterium tuberculosis* (Mtb) that do not respond to isoniazid, rifampin, at least one fluoroquinolone and at least one injectable second-line drug (WHO, 2016) may eventually constitute of 8.9% of cases in India, 9% in Russia and 8.5% in South Africa by 2040 (Sharma et al., 2017). In the US, both drug-sensitive and drug-resistant forms of TB are most often associated with individuals born outside of the US, and account for approximately two thirds of newly reported cases in the United States (Cain et al., 2008).

As is the case with most forms of drug resistance, the emergence of resistant TB strains is primarily the result of selection resulting from by the inappropriate use of antibiotics. The administering of a single drug or of combination regimens at incorrect dosages or intervals, the reduced compliance frequently seen with long-term therapeutic regimes, and the extensive distribution of ineffective or counterfeit antibiotics all select for drug-resistant strains, and provide the substrate for subsequent mutations that confer resistance to additional drugs. In the presence of antibiotics, resistant strains are at a selective advantage (Borrel and Gagneux, 2009). The acquisition of resistance due to inappropriate drug treatment is clinically supported by a study which found that isoniazid-resistant (H-resistant) strains acquired resistance to other drugs during treatment (Kim et al., 2008).

In the context of TB treatment, which previously required antibiotic regimes lasting several months, non-adherence has severe clinical and epidemiological implications. Non-adherent patients remain infected with Mtb bacteria, and specifically with resistant strains that survive the initial treatment (Dheda et al., 2014). In response to this challenge, WHO developed the Directly Observed Therapy Short course category (DOTS), in which patients are supervised by healthcare workers while taking the prescribed drugs (Srivastava et al., 2011). Critical to global TB control, the WHO director called DOTS the most important health breakthrough of the past decades (Garner and Volmink, 2003).

Several mathematical models have been developed to describe the dynamics of TB, each characterized by different assumptions. Basu et al. (2007) investigate the transmission of TB in the hospitals of South Africa by including three forms of TB: drug-susceptible (DS), multi-drug resistant (MDR) and extensively drug-resistant (XDR). They conclude that XDR TB will continue to be a serious public health problem unless important changes in public policy are implemented. They also acknowledge that certain strategies that prevent TB transmission in hospitals can alter this prediction. Cohen and Murray (2004) develop a model incorporating the effects of biological fitness. For this purpose, three types of Mtb bacteria are included in the model: DS, fit MDR, and unfit MDR. In this context, fitness is defined as the relative ability of bacterial strains to survive and infect new hosts. Given that antibiotic resistance incurs a fitness cost to the bacteria, the study aims to investigate the extent to which even less fit MDR TB strains can pose an epidemiological danger. They demonstrate that even lower fitness MDR TB strains can outperform drug-susceptible Mtb bacteria even in low antibiotic environments and thus pose an epidemiological risk. Subsequent models include DS, MDR and XDR TB strains and explore the effects of quarantining individuals with XDR TB, concluding that this practice may serve to reduce TB transmission (Bhunu, 2011). The model of Hill et al. (2012) divides the US population into US-born (USB) and foreign-born (FB) individuals, involves preferential mixing, and assumes that exogenous infections occur mainly within subpopulations. That model is deployed to investigate whether TB elimination, defined as the occurrence of at most one TB case per 100 million people (WHO, 2014) can be achieved in the US by 2100. The results of the analysis suggest that TB elimination is possible in the US-born population by 2100, but not in the foreign-born population. In fact, even if the transmission of TB is completely interrupted, the goal of eliminating TB in the foreign-born populations cannot be achieved in the proposed time frame (Hill et al., 2012).

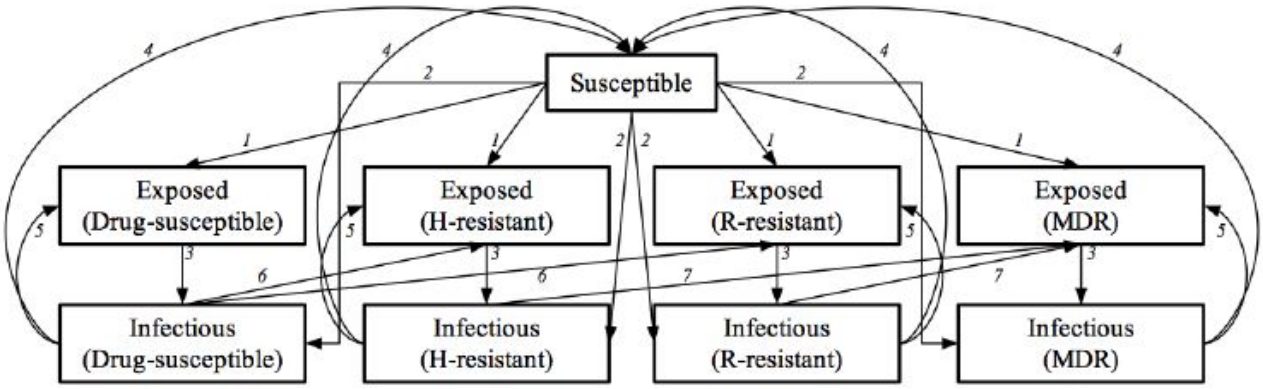

Figure 1: Schematic of the model. Susceptible individuals are infected and become either latently (1) or actively (2) infected. Latently infected individuals progress to active disease and become infectious (3). Infectious individuals receive treatment. If the treatment is successful, they become susceptible again (4). If treatment is unsuccessful, and resistance is not acquired, infectious individuals return to the corresponding latently infected compartment (5). If resistance is acquired, individuals who had active drug-susceptible (DS) TB enter the latently infected H-resistant or R-resistant compartment (6). Individuals who fail treatment for active H-resistant or R-resistant TB move into the MDR (latently infected) (7). For simplicity, cases resulting in death are not shown in this schematic (Justynski, 2016). The arrows represent the possible pathways linking the movement of individuals from one compartment to another.

We aim to develop and parameterize a mathematical model of TB transmission in the United States that explicitly incorporates the consequences of single- and multi-drug resistance, as well as the effect of the foreign-born population on overall TB prevalence.

3 Model

Our compartmental model builds on the model of Hill et al. (2012). We adhere closely to this widely-cited compartmental model of TB transmission dynamics, but we modify the model to include simple and realistic dynamics for the acquisition of drug-resistance. Our model divides the US population into the following categories: 1) susceptible, 2) latently infected, 3) actively infected and 4) TB deaths. Each infected category includes four sub-categories of strains with varying drug-resistance: Mtb strains that are 1) drug-susceptible (DS), 2) isoniazid (H)-resistant, 3) rifampin (R)-resistant and 4) multi-drug resistant (MDR). We exclude extensively drug-resistant (XDR) strains from our analysis due to their low frequency (CDC, 2007) and the added complexity they would impose to the model. The arrows in Figure 1 represent the possible pathways whereby an individual can move from one compartment to another.

In this model, susceptible individuals contacting actively infected individuals may themselves become infected. Infected individuals who develop symptoms within two years of infection move into the actively infected compartment; otherwise they migrate to the latent compartment. A proportion of latently infected individuals progress to active disease, and only actively infected individuals receive treatment. Depending on the efficacy of treatment, treated individuals will either be cured, returning them to the susceptible compartment, or else become asymptomatic (and thus no longer infectious), but will nonetheless remain latently infected. Individuals rejoining the latent category following treatment will either return to their original compartment (displaying the same resistance profile as their previous active infection) or they will acquire new drug resistance. Thus, actively infected patients who are unsuccessfully treated and survive will either become latently infected with drug-sensitive TB or will acquire resistance to H or R and move to the respective latent compartment. Similarly, patients with H-resistant (or R-resistant) TB can return to the latently H-resistant TB infected (or R-resistant) compartment, or migrate into the latently MDR TB infected category. Finally, unsuccessful treatment may result in the death of the patient.

The model also explicitly incorporates the impact of migration of foreign-born infected individuals, currently considered the main source of latent tuberculosis infection (LTBI) in the United States. While in principle—and to some extent, in practice—actively infected (symptomatic) people are not allowed entry into the US, some proportion of foreign-born individuals entering the US bear latent TB and can contribute to the latently-infected compartments according to their latent infection type. Finally, for simplicity, the schematic in Figure 1 does not depict natural births and deaths, although both are taken into consideration in the model.

This compartmental model is described by a coupled system of differential equations, where each equation represents the

change in the number of the individuals that belong to a specific compartment. The equations are presented below:

$$\frac{dS}{dt} = \rho N - qt_1\lambda\frac{SI_1}{N} - qt_2\lambda\frac{SI_2}{N} - qt_3\lambda\frac{SI_3}{N} - qt_4\lambda\frac{SI_4}{N} + z_1\phi_1 I_1 + z_2\phi_2 I_2 + z_3\phi_3 I_3 + z_4\phi_4 I_4 + (1-l)\alpha N - \mu_0 S \tag{1}$$

$$\frac{dE_1}{dt} = (1-p)qt_1\lambda\frac{SI_1}{N} - v_L E_1 + (1-y_1)(1-z_1)\phi_1 I_1 + l\alpha(1-r_2-r_3-r_4)N - \mu_0 E_1, \tag{2}$$

$$\frac{dI_1}{dt} = qt_1\lambda\frac{SI_1}{N} + qv_L E_1 - \phi_1 I_1 - \mu_0 I_1 - \mu I_1, \tag{3}$$

$$\frac{dE_2}{dt} = (1-p)qt_2\lambda\frac{SI_2}{N} - v_L E_2 + (1-y_2)(1-z_2)\phi_2 I_2 + \gamma(1-z_1)y_1\phi_1 I_1 + l\alpha r_2 N - \mu_0 E_2, \tag{4}$$

$$\frac{dI_2}{dt} = qt_2\lambda\frac{SI_2}{N} + v_L E_2 - \phi_2 I_2 - \mu_0 I_2 - \mu I_2, \tag{5}$$

$$\frac{dE_3}{dt} = (1-p)qt_3\lambda\frac{SI_3}{N} - v_L E_3 + (1-y_3)(1-z_3)\phi_3 I_3 + (1-\gamma)(1-z_1)y_1\phi_1 I_1 + l\alpha r_3 N - \mu_0 E_3, \tag{6}$$

$$\frac{dI_3}{dt} = qt_3\lambda\frac{SI_3}{N} + v_L E_3 - \phi_3 I_3 - \mu_0 I_3 - \mu I_3, \tag{7}$$

$$\frac{dE_4}{dt} = (1-p)qt_4\lambda\frac{SI_4}{N} - v_L E_4 + (1-z_2)y_2\phi_2 I_2 + (1-z_3)y_3\phi_3 I_3 + (1-z_4)\phi_4 I_4 + l\alpha r_4 N - \mu_0 E_4, \tag{8}$$

$$\frac{dI_4}{dt} = qt_4\lambda\frac{SI_4}{N} + v_L E_4 - \phi_4 I_4 - \mu_0 I_4 - \mu I_4, \tag{9}$$

$$\frac{dD}{dt} = \mu(I_1 + I_2 + I_3 + I_4), \tag{10}$$

$$\frac{dN}{dt} = \rho N + \alpha N - \mu(I_1 + I_2 + I_3 + I_4) - \mu_0 N, \tag{11}$$

where

S: the susceptible population,
E_1: the number of individuals latently infected with DS TB,
I_1: the number of individuals actively infected with DS TB,
E_2: the number of individuals latently infected with H-resistant TB,
I_2: the number of individuals actively infected with H-resistant TB,
E_3: the number of individuals latently infected with R-resistant TB,
I_3: the number of individuals actively infected with R-resistant TB,
E_4: the number of individuals latently infected with MDR TB,
I_4: the number of individuals actively infected with MDR TB,
D: number of deaths caused by TB,
N: total US population.

4 Parameter Fitting

Incorporating multiple drug-resistant TB strains to expand Hill's formulation for TB dynamics in the US results in a complex model. In addition to requiring a larger number of compartments and introducing many strain-specific transition rates between compartments, our model incorporates new dynamics describing how unsuccessful treatment contributes to the acquisition of antibiotic resistance.

The challenge of fitting this complex model is made all the more daunting by the relatively sparse available data concerning the prevalence of latent infections and by the dearth of direct experimental evidence about rates of resistance to commonly used antibiotics. As a result, deriving point estimates for all of the parameters values in our model seems an unlikely undertaking from the outset. Indeed, in the following subsections, we describe several difficulties we encountered in attempting to apply standard estimation techniques. In response, we revised our focus to concentrate on producing a collection (or "crowd") of informative parameter combinations.

4.1 Defining an Error Function

Our compartmental model involves a total of 30 parameters, only 3 of which have been previously determined with high precision. For the remaining twenty-seven parameters in our compartmental model, reliable point estimates are not available in the

Table 1: Parameter descriptions with documented ranges from existing literature.

Parameter	Description	Range
α_2	Proportion of initial LTBI cases that are H-resistant	$(0, 0.2)$
α_3	Proportion of initial LTBI cases that are R-resistant	$(0, 0.2)$
α_4	Proportion of initial LTBI cases that are MDR	$(0, 0.2)$
α	Immigration rate	0.00425
b	Proportion of initial active cases that are DS	$(0.5, 1)$
γ	Proportion H-resistance acquisition cases	$(0, 1)$
l	Proportion of immigrants that have LTBI	$(0, 0.3)$
λ	Effective Contact Rate	$(0, 30)$
μ	TB mortality rate	$(0, 0.5)$
μ_0	Mortality rate unrelated to TB	0.013
p	Proportion of exogenous infections that are acute	$(0, 0.3)$
ϕ_1	Rate of end of treatment (DS)	$(0.6, 0.9)$
ϕ_2	Rate of end of treatment (H-resistant)	$(0.5, 0.9)$
ϕ_3	Rate of end of treatment (R-resistant)	$(0.36, 0.5)$
ϕ_4	Rate of end of treatment (MDR)	$(0.3, 0.5)$
q	Proportion of active cases that have the potential to be infectious	$(0, 1)$
r_2	Proportion of immigrant H-resistant LTBI cases	$(0, 0.2)$
r_3	Proportion of immigrant R-resistant LTBI cases	$(0, 0.2)$
r_4	Proportion of immigrant MDR LTBI cases	$(0, 0.2)$
ρ	US birth rate	0.0179
t_1	Proportion of treatment time when individuals are infectious (DS)	$(0, 0.1)$
t_2	Proportion of treatment time when individuals are infectious (H-resistant)	$(0, 0.1)$
t_3	Proportion of treatment time when individuals are infectious (R-resistant)	$(0, 0.1)$
t_4	Proportion of treatment time when individuals are infectious (MDR)	$(0, 0.1)$
v_L	Progression rate from latent to active infection	$(0, 0.01)$
y_1	Proportion of failed treatments for DS TB that result in H- or R-resistance	$(0, 1)$
y_2	Proportion of failed treatments for H- or R-resistant TB that result in MDR	$(0, 1)$
z_1	Proportion of treatment courses for DS TB that are successful	$(0.6, 0.9)$
z_2	Proportion of treatment courses for H-resistant TB that are successful	$(0.5, 0.9)$
z_3	Proportion of treatment courses for R-resistant TB that are successful	$(0.5, 0.9)$
z_4	Proportion of treatment courses for MDR TB that are successful	$(0.1, 0.8)$

*Since q and λ always appear together in our model, we treat their product, $q\lambda$, as a single parameter with range $(0, 30)$.

Table 2: CDC data.

Year	D^i_{MDR}	D^i_{Active}	D^i_{HR}	D^i_{D}	Year	D^i_{MDR}	D^i_{Active}	D^i_{HR}	D^i_{D}
2014	91	9,421	680	NA*	2003	119	14,835	903	711
2013	96	9,565	653	555	2002	158	15,055	912	784
2012	86	9,941	683	510	2001	151	15,945	897	764
2011	127	10,510	753	536	2000	146	16,308	981	776
2010	105	11,161	699	569	1999	157	17,499	999	930
2009	114	11,523	762	529	1998	155	18,286	1,120	1,112
2008	107	12,895	835	590	1997	201	19,751	1,195	1,166
2007	124	13,282	798	554	1996	250	21,210	1,284	1,202
2006	124	13,728	845	644	1995	327	22,726	1,350	1,336
2005	125	14,063	842	648	1994	431	24,206	1,543	1,478
2004	128	14,499	872	662	1993	484	25,102	1,534	1,631

*For 2014 there was no data for deaths due to TB.

existing literature (see Table 1). Our task was therefore to estimate parameter values that produce reliable fits to TB incidence data obtained by the CDC for the years 1993–2014. The available annual data to which the model is fitted includes: 1) All active TB cases, 2) TB cases that are H-resistant, 3) TB cases that are MDR and 4) TB-related deaths.

We quantify the quality of a parameter fit to our model by minimizing the sum of the squared normalized errors. We define an error function using the following notation. Let

$$\begin{aligned}
f^i_{\text{Active}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) &:= \text{the model-predicted number of active cases of TB in the US for the year } i,\\
f^i_{\text{HR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) &:= \text{the model-predicted number of H-resistant cases of TB in the US for the year } i,\\
f^i_{\text{MDR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) &:= \text{the model-predicted number of MDR cases of TB in the US for the year } i,\\
f^i_{\text{D}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) &:= \text{the model-predicted number of deaths due to TB in the US for the year } i,
\end{aligned}$$

and D^i_{Active}, D^i_{HR}, D^i_{MDR}, D^i_{D} denote the respective data provided by the CDC (Table 2). The values of f^i are obtained as a function of our parameter values. Our compartmental model is implemented in R: for each set of parameter values, the model outputs the number of active, H-resistant and MDR TB cases, as well as deaths due to TB.

Next we define an error function, $E(\alpha_2, \alpha_3, \ldots, z_3, z_4)$. By minimizing $E(\alpha_2, \alpha_3, \ldots, z_3, z_4)$ we hope to parameterize our model in order to understand parameter values, and to predict future trajectories of TB infections by H-resistant, R-resistant, and MDR strains. When fitting simple models with strong assumptions about the distribution of observation errors, least squares minimization guarantees specific statistical properties. No such guarantees are possible when least-squares minimization is applied to complex model forms, including our compartment model. Despite the lack of theoretical guarantees for our setting, least squares is a standard and extensively applied measure of fit quality. Since the CDC data used to fit our model appear relatively smooth, outliers that distort the fit (due to the heavy penalty that the least squares form imposes on large deviations) seem unlikely. We therefore see no reason to suspect that minimizing the sum of absolute errors would be a more robust approach, and proceed to define a least-squares error function. Let

$$\begin{aligned}
E(\alpha_2, \alpha_3, \ldots, z_3, z_4) = \sum_{i=1993}^{2014} \Bigg[&\left(\frac{f^i_{\text{Active}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) - D^i_{\text{Active}}}{D^i_{\text{Active}}} \right)^2 + \left(\frac{f^i_{\text{HR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) - D^i_{\text{HR}}}{D^i_{\text{HR}}} \right)^2 \\
&+ \left(\frac{f^i_{\text{MDR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) - D^i_{\text{MDR}}}{D^i_{\text{MDR}}} \right)^2 + \left(\frac{f^i_{\text{D}}(\alpha_2, \alpha_3, \ldots, z_3, z_4) - D^i_{\text{D}}}{D^i_{\text{D}}} \right)^2 \Bigg]. \qquad (12)
\end{aligned}$$

We include normalization by annual CDC-observed populations in Equation (12) to avoid fits that predict large compartments precisely (namely, f^i_{Active}) at the expense of accuracy in compartments whose populations are orders of magnitude smaller (e.g. f^i_{MDR}). For example, the empirical CDC observations for 1993 document 25,102 total active TB cases, with only 484 active MDR-resistant cases (Table 2). Notice that a least-squares error function without normalization would consider the following two pairs of model predictions to be of equal quality:

- 25,102 active TB cases with 284 MDR cases, or
- 25,302 active cases with 484 MDR cases.

Notice that the first pair of predictions is grievously far from the the observed prevalence of MDR (under-estimating the empirical CDC observation by 41%). In contrast, the second pair of predictions is within 1% of the CDC's empirical observations for both populations. By dividing the discrepancy between the model prediction and the CDC value by the CDC value, we render compartments of different magnitudes comparable, so that a low aggregated error value is not obtained at the expense of extreme inaccuracy within smaller-sized compartments.

4.2 Minimizing Error: Challenges and Initial Attempts

The values of $f^i_{\text{Active}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, $f^i_{\text{HR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, $f^i_{\text{MDR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, and $f^i_{\text{D}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$ are determined as numerical outputs of our compartmental model. As a result, $E(\alpha_2, \alpha_3, \ldots, \ldots, z_3, z_4)$ is not differentiable. In general, parameter fitting can be complicated when the error function does not have an analytical form because any error-minimizing method applied must be derivative-free.Optimization methods for derivative-free settings are often referred to as *direct-search methods* or *zeroth-order methods*. While direct search methods are often heuristics (that is, many offer no global guarantees of optimality) in recent years the mathematical optimization community has recognized both that such methods can be effective in practice, and that they sometimes provide the only option for difficult optimization problems, particularly those involving the outputs of complex simulations (Kolda et al., 2003). Since our error function depends on model predictions that are produced in a highly-complex manner (based on Equations (1) through (11)), it seems quite difficult to assess whether certain levels of smoothness (or Lipshitz continuity) hold in neighborhoods of the minima. As such, we can not rely on the guarantees of more sophisticated modern derivative-free methods (Koziel et al., 2016).

Instead we will regard our error function as a *black-box function*: given a set of parameters, a black-box function returns only the numerical value of the function. Optimization methods for black-box functions typically rely on substantial numbers of function evaluations. Naive grid searches, are an example of a conceptually simple derivative-free black-box optimization method. In order to conduct a grid search, the plausible range for each unknown parameter is divided into uniform intervals. Each parameter now has some number of possible discrete levels: the resulting parameter combinations can be explicitly enumerated and the function value is then checked at every such parameter combination. Grid-search is effective when the number of parameters is small, but quickly becomes computationally unfeasible as the number of parameters grows. For example, even if we considered only 3 discrete values for each of our model's 27 unknown parameters, the number of error-function evaluations required to conduct a grid search would be a completely impractical 3^{27}.

Given that an enumeration-based grid-search technique was unfeasible for our high-dimensional parameter space even at coarse resolution, we first attempted to minimize Equation (12) with an algorithm that iteratively optimizes each of the 27 parameters. We divided the range of each parameter into 100 discrete possibilities. From a random 27-parameter combination (an initial condition), our iterative method proceeds through a permutation of the parameters. When a particular parameter is considered for updating, the current values of the other 26 parameters are fixed and this fixed 26-parameter combination is paired with each of the 100 values for the currently-considered parameter. The error function is evaluated at each of these hundred 27-parameter combinations: the best 27-parameter combination is adopted as the new best-so-far parameter combination. The algorithm proceeds to consider updating the next parameter from the permutation: fixing the remaining 26, and optimizing the 27th, etc. This method iterates repeatedly through a permutation of the 27 parameters until the 27-parameter combination stops changing. Over a run of the algorithm to convergence, this method traces a path of decreasing error-function values along a small subset of 100^{27} grid points in the plausible parameter space.

Unfortunately, we observed that the parameter combinations found at convergence by this iterative fitting method were heavily dependent on the order in which parameter-updates were attempted as well as on the initial conditions. This indicated the presence of many local minima for our error function, thus complicating the search for a global minimizer. We found that repeatedly re-randomizing the order of attempts to update the individual parameters did not alleviate these issues. While the existence of many local minima was a hazard we anticipated given the relative sparsity of the CDC data, we were further hindered by this troubling artifact of the sequential approach to error minimization.

Conducting many runs of the iterative method led to strong bimodal distributions for approximately half of the 27 unknown model parameters. For example, Figure 2a depicts the values of z_3 at convergence for 100 runs of the iterative error minimization approach. Recall that z_3 denotes the proportion of treatment courses for R-resistant TB that are successful. We considered this bimodal behavior troubling for two reasons. First, it is unclear what insight about z_3 can be gained from such a distribution. For example, reporting a typical measure of central tendency (mean, median) for the 100 values of the parameter z_3 from Figure 2a seems highly misleading. Indeed, closer inspection showed that comparably low-error parameter combinations appeared in both groups (the group near 0.5, and the group near 0.9). Second, the most frequent optimized values of the bimodal parameters corresponded to the absolute top and absolute bottom of the those parameters' known ranges—values that were hard-coded into the algorithm (from Table 3). In particular, it did not appear likely that all local minima of the error function just coincidentally happened to lie at the boundaries of the parameter space: removing the boundaries of the known parameter ranges caused the support of this strong bimodal behavior to shift outward to the new boundaries (Figure 2b). The sequential nature of our

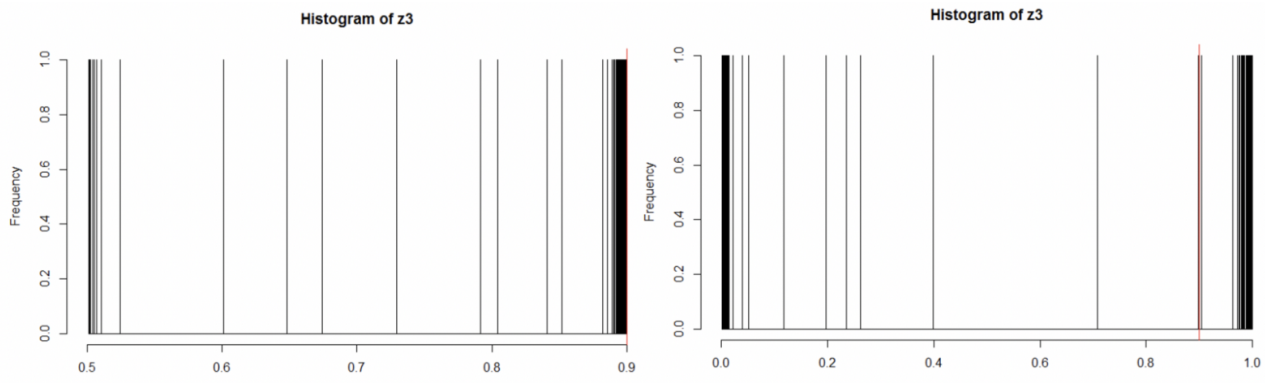

Figure 2: a) 100 runs of the iterative algorithm yield bimodal values. b) Allowing parameters to take on all possible values still produces bimodal values.

minimization method appeared to be heavily biased towards convergence to local minima at the boundaries of the parameter region. The "shift outwards" of the support of the bimodal distribution further suggested that the local minima discovered at region boundaries were not true local minima of the error-function but instead indicated a direction in which the error function was decreasing, but the search was truncated by the externally imposed parameter boundaries derived from the existing literature (Table 3). In response to this behavior, we shifted our focus towards an algorithmic framework that would fit all parameters simultaneously.

Before turning towards evolutionary/genetic methods, we briefly considered another standard class of methods for black-box optimization, referred to as *pattern search method*. Starting from an initial parameter combination, pattern search evaluates the black-box function at points identified by a "stencil" that is centered at the initial combination (Koziel et al., 2016). This stencil is typically a star with "exploratory" points extending in several directions. When these directions correspond to unit vectors of each dimension this is the classical "compass search" (Kolda et al., 2003) describes the long history and modern theoretical results related to this famous method). More generally, stencils may be based on collections of randomly-chosen vectors. If a point of the stencil gives a parameter combination with lower error than the current center of the stencil, then that point (aka, that parameter combination) becomes the new center of the stencil. Otherwise, the size of the stencil is reduced to check function values closer to the current center of the stencil. While pattern-search is a popular general method, because the ranges of our 27 unknown parameters vary substantially, it was unclear how to define the relative length of the "legs" of a stencil for pattern search in a natural way. For example, a stencil in which all legs have the same length would effectively be taking much larger steps away from the stencil center in some dimensions than in others. A major problem exposed by our iterative method was that allowing a single parameter to take a long step to the edge of its known range resulted in many other parameters immediately "snapping" to their hard-coded boundaries. Since we had already observed close dependence on initial conditions and the existence of many local minima with our iterative fitting method, we were wary of adopting a pattern-search approach where the outcome might depend critically on our definition of the stencil.

4.3 Minimizing Error: Turning to Genetic Algorithms

Genetic algorithms are known for being effective in cases of high-dimensional search spaces, non-linear objective functions, and complex collections of local minima. Randomized mutation steps can help genetic methods to "overcome premature stagnation" at poorer quality local minima (Koziel et al., 2016). Based on the early challenges we observed with our iterative fitting approach, genetic (or "evolutionary") minimization techniques appeared as a potentially fruitful approach given our setting.

To fit the model to the CDC data for the years 1993–2014 (Table 2) we implemented a genetic algorithm in R. In our genetic algorithm, an *individual* is defined as a vector of the 27 unknown parameter values for our compartment model (these values give the *profile* of the particular individual). For each *generation*, our genetic algorithm measures the fitness (that is, the error function $E(\alpha_2, \alpha_3, \ldots, \ldots, z_3, z_4)$ evaluated at an individual's vector of 27 parameter values) for each member of a population of 900 individuals. The most fit individuals (parameter combinations that achieve the lowest error) from each generation are selected to act as the progenitors of the next generation.[1]

Our population size of 900 was carefully chosen. Early tests suggested that smaller generation sizes (100, 400) were quite susceptible to drift. In this scenario, individuals in smaller populations may survive (or were lost to) the evolutionary process due to chance alone, resulting in limited exploration of the space of profiles and very limited evolution over consecutive generations.

[1]Details of breeding and mutation are discussed in Section 4.4.

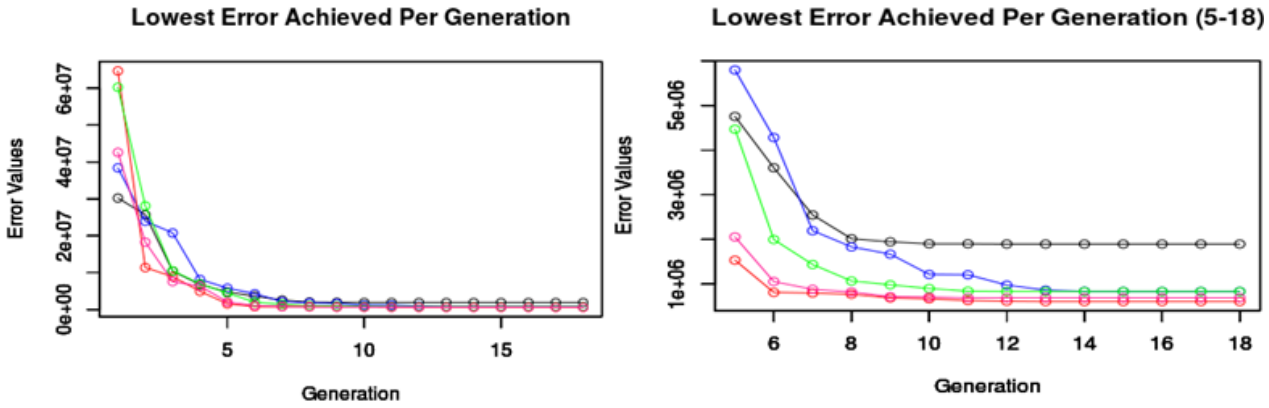

Figure 3: Runs of the genetic algorithm consistently discover low-error parameter combinations. a) The error of the fittest individual in each generation for five runs of the genetic algorithm. b) Same plot after initial period of fast evolutionary improvement. Truncated axes allow visual differentiation between values of error at convergence. Very similar final error levels could indicate that different runs of the genetic algorithm are effectively converging to a common optimal parameter combination. Closer investigation shows that this is not the case.

Repeated initial tests of the algorithm showed that a generation size of 900 counteracted the effects of drift and uncovered solutions of comparable quality to those found using substantially larger generation sizes.

4.4 A single run of the genetic algorithm

To initialize a run of the genetic algorithm, we randomly generate 900 individuals for the first generation. For each individual, a value for each of 27 parameters is chosen uniformly at random within a plausible range for the respective parameter (based on ranges in the existing literature, see Table 1). The quality (or *fitness*) of each parameter combination is quantified by the sum-of-squares error function described in Equation (12). We call individual A *fitter* than individual B if the error function in Equation (12) yields a smaller value for A than B.

After the fitness of each individual from generation n is quantified using Equation (12), the thirty fittest individuals are selected to produce the next generation. The next generation of 900 individuals is produced by *breeding* all possible ordered pairs of the thirty fittest individuals from the previous generation (with replication allowed within a breeding pair). Suppose that for generation n, each parent is a vector X_i, where $i \in [1, 30]$, of the form

$$X_i = \left(\alpha_2^{X_i}, \alpha_3^{X_i}, \alpha_4^{X_i}, b^{X_i}, \gamma^{X_i}, l^{X_i}, \mu^{X_i}, p^{X_i}, \phi_1^{X_i}, \phi_2^{X_i}, \phi_3^{X_i}, \phi_4^{X_i}, v_L^{X_i}, r_2^{X_i}, r_3^{X_i}, r_4^{X_i}, t_1^{X_i}, t_2^{X_i}, t_3^{X_i}, t_4^{X_i}, v_L^{X_i}, y_1^{X_i}, y_2^{X_i}, z_1^{X_i}, z_2^{X_i}, z_3^{X_i}, z_4^{X_i}\right)$$

Each member of the progeny is produced by 27 fair coin flips: each coin flip selects one of the two parental values for each parameter. For example, parents X_1 and X_2 may produce a random progeny like $(\alpha_2^{X_2}, \alpha_3^{X_1}, \alpha_4^{X_1}, b^{X_2}, \ldots, z_4^{X_1})$. Since we include self-breeding of X_i with X_i, the 30 fittest individuals from generation n are always are present in generation $n+1$. Since we breed all ordered pairs of the thirty fittest individuals in generation n, a pair of parents X_i and X_j will contribute two random progeny to the n + 1st generation.

In addition to the recombination described above, our model also incorporates mutation in the creation of generation n + 1. To avoid being trapped by local minima of the error function (Equation (12)), 100 individuals in each new generation are subjected to a mutation rate of 0.3. In essence, each of the 27 parameter values for an individual undergoing mutation have a 0.7 probability of remaining unchanged and a 0.3 probability that the existing value for that parameter will be replaced by a new value selected at random from within the parameter's range. As indicated, the 30 fittest individuals from the previous generation are present, unmutated, in the next generation.

This routine is repeated for each run of the genetic algorithm, and the resulting (lowest-error) individuals identified. During the first few generations, the fittest parameter combination quickly improves from one generation to the next. Eventually, evolutionary progress slows. Figure 3 depicts five typical runs of the genetic algorithm. Note that the lowest-error parameter combination discovered in the first generation (y-intercept of the left panel) is a random variable. Early runs of the genetic algorithm suggested that convergence to a single fittest individual (a single combination of parameter values that appears as the fittest individual in several subsequent generations) almost always occurred within 17 generations. We call this final fittest individual the *convergence individual* for a run of the genetic algorithm.

4.5 Typical Behavior of the Genetic Algorithm

Figure 3 suggests that repeated runs of the genetic algorithm may converge to a number of different low-error parameter combinations (that is, different algorithm runs may result in different convergence individuals). It is not clear from such a small set of algorithm runs whether these different "optimized" parameter combinations simply represent minor variations of each other (all effectively yielding the same parameterization of the model) or instead specify strongly different parameterizations of the model that could critically impact predictions about future TB prevalence. In this section we investigate such behavior more closely.

To better understand the behavior of the model and the resulting optimized parameter values, the genetic algorithm was run 1,000 times to convergence. Each run produced a final fittest individual (a vector of the 27 unknown parameter values that produces low error for Equation (12)). Panel a) of Figure 4 is a histogram of the final error values at convergence for each of the 1,000 algorithm runs. For context about the magnitude of these final error values, we note that a vector of predictions that falls within 8% of every one of the 87 CDC-observed values in Table 2 would yield an error value of 1.6×10^6. While some runs of the genetic algorithm converge to relatively high-error parameter combinations (at the right-hand side of Figure 4a)), the large majority of our genetic algorithm runs converge to a parameter combination with error below 1.6×10^6.

The worst fits at convergence in Figure 4a) have error values that are an order of magnitude larger than the best parameter combinations discovered among our 1,000 genetic algorithm runs. Naturally, we regard the parameter combinations yielding high-error values as less reliable than those resulting in the lowest error (at the left-hand side of the histogram). At the same time, we note that a number of different convergence individuals (that is, many runs of the genetic algorithm) achieve comparably low error at convergence.

In our experiments, the single best fit discovered among all 1,000 genetic algorithm runs has error value within a factor of two of the next 10 best algorithm runs, yet the parameter combinations differ substantially. Parameter values that minimize the error function are generally considered to provide the best proxy for true parameter values. Nevertheless, this conventional wisdom is sorely tested when many very different combinations of parameter values minimize the error function to a near-identical degree: it is not clear that any single near-optimal fit should be considered the most authoritative about the true parameter values.

We want to limit our attention to parameter combinations that give the highest quality fits to the empirical data. Across all 1,000 genetic algorithm runs, the error at convergence varied widely between 145,466 and 11,541,614 with a mean of 1,475,410 and a standard deviation of 1,112,003. By focusing on the 100 best runs of the genetic algorithm, we can substantially decrease the mean and variance of the error at convergence. For example, the 100th-best fit has an error of 620,069. We define a *crowd of near-best fits* in which this 100th-best fit is the worst member. The error value of this worst member matches the error of a set of predictions that is within $\approx 6.4\%$ for of every one of the 87 CDC-observed values in Table 2. The 100th-best fit has error within a factor of 4 of the best fit found among 1,000 algorithm runs, and has error lower than $\frac{1}{19}$ of the worst error at convergence among the 1,000 genetic algorithm runs. Limiting to a crowd of the 100 best fits discovered at convergence by genetic algorithm runs, the mean error of our crowd of fits drops to 519,661 (corresponding to predicting every one of the 87 CDC-observed values within 6%), with a standard deviation of 89,532. We judge that our crowd of near-best fits is composed of 100 high-quality fits to the empirical CDC observations.

4.6 The portion of the parameter space compatible with low error

The 100 genetic algorithm runs resulting in the smallest final error at convergence were used to produce histograms of each of the 27 unknown parameter values. These 100 best fits give a set of 100 vectors in the 27-dimensional parameter space which all produce low error. We argue that this *crowd of near-best fits* yields insight about the parameter space for our TB model.

Using our crowd of 100 low-error fits, we created histograms of the resulting optimized values for each parameter. For some parameters, (including α_2, γ, ϕ_1, ϕ_2, ϕ_4, r_2, r_3, r_4, t_1, t_2, t_4, y_1, y_2, z_1, z_2, z_3, and z_4) our fitted model does not appear to lead to new insight about the range of the parameter. For example, Figure 4b shows that the 100 best fits yield values of ϕ_1 that are dispersed across the previously-known range for ϕ_1 (Table 1).

In contrast, consider the two lower panels of Figure 4. Panel c) displays the concentrated histogram of v_L values that occurred among our crowd of 100 low-error fits. Similarly, panel d) displays a similar highly concentrated histogram of α_4 values from our crowd. In both panels, the horizontal axis displays the range of plausible values for that parameter based on the existing TB literature. In both cases, it appears that a large fraction of the ostensibly plausible parameter range is carefully avoided by our crowd of low-error fits.

We summarize the results of 27 histograms analogous to Figure 4, panels b–d in Table 3. Parameters α_3, α_4, b, l, μ, p, t_3, and v_L display remarkably narrow distribution in the values recommended by the 100 best fits with only a small number of outliers. The "Obtained Range" listed in Table 3 is a range that contains values of the respective parameter for at least 95 of our 100 low-error fits. Parameters for which our crowd of 100 fits suggests refining the known parameter range by more than 50% are highlighted. Several other parameter ranges are also narrowed by our crowd of 100 low-error fits, albeit less dramatically so (e.g. ϕ_3, the product $q\lambda$, and z_1).

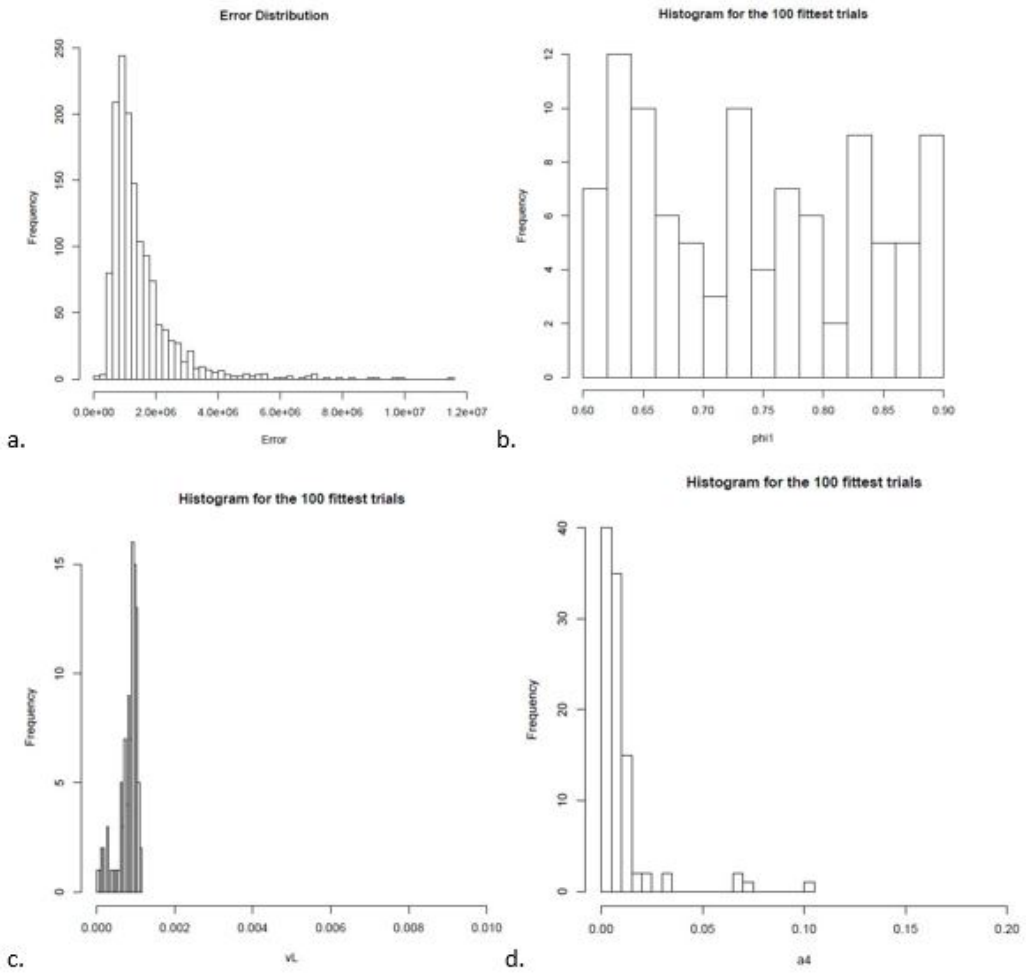

Figure 4: A crowd of near-best fits refines recommendations about the range of key parameters. a: Histogram of final error at convergence for 1,000 runs of the genetic algorithm. b: Example parameter (ϕ_1) where the crowd of 100 best fits does not recommend refining parameter range documented in existing literature (Table 1). c, d: Examples of parameters (v_L, α_4) where crowd of 100 best fits recommends substantially refining parameter range documented in existing literature. For b)-d), the horizontal axis displays the existing parameter range in the literature (Table 1).

Table 3: Obtained parameter ranges.

Parameter	Known Range	Obtained Range	Parameter	Known Range	Obtained Range
α_2	(0, 0.2)	(0.0006, 0.1663)	r_2	(0, 0.2)	(0.0017, 0.196)
α_3	(0, 0.2)	(0.00001, 0.0823)	r_3	(0, 0.2)	(0.0013, 0.1836)
α_4	(0, 0.2)	(0.0002, 0.0319)	r_4	(0, 0.2)	(0.0043, 0.1939)
b	(0.5, 1)	(0.5002, 0.6147)	t_1	(0, 0.1)	(0.000061, 0.0997)
γ	(0, 1)	(0.0859, 0.9944)	t_2	(0, 0.1)	(0.0012, 0.0996)
l	(0, 0.3)	(0.000095, 0.0201)	t_3	(0, 0.1)	(0.034, 0.0989)
μ	(0, 0.5)	(0.0018, 0.0605)	t_4	(0, 0.1)	(0.0028, 0.0982)
p	(0, 0.3)	(0.1833, 0.2992)	y_1	(0, 1)	(0.0034, 0.992)
ϕ_1	(0.6, 0.9)	(0.6004, 0.8968)	y_2	(0, 1)	(0.0149, 0.995)
ϕ_2	(0.5, 0.9)	(0.5013, 0.819)	z_1	(0.6, 0.9)	(0.7007, 0.8982)
ϕ_3	(0.3, 0.5)	(0.3002, 0.418)	z_2	(0.5, 0.9)	(0.501, 0.8996)
ϕ_4	(0.3, 0.5)	(0.30022, 0.4952)	z_3	(0.5, 0.9)	(0.5609, 0.8944)
$q\lambda$	(0, 30)	(8.0986, 29.613)	z_4	(0.1, 0.8)	(0.1636, 0.7978)
v_L	(0, 0.01)	(0.000049, 0.0011)			

We are keenly aware of the relative sparse data set we are using to optimize what is, after all, a relatively parameter-rich model. We are limited—as is often the case with infectious disease data—by the granularity of existing real world data. Nonetheless, we are gratified by the realization that even the sparse available data still allows us to narrow the range of the highlighted parameters from Table 3 well below previous estimates.

Five of the parameters for which the parameter range was considerably constrained (apart from μ and b) are directly linked to latent TB. For example, v_L, the progression rate from LTBI to active TB, is very difficult to observe clinically or to estimate from epidemiological surveillance data: the exact times at which individuals become infected with LTBI or progress to TB disease are almost never known, and estimates for the average progression rate from LTBI to active TB can vary dramatically from one population to another. Because the identification and treatment of latent cases is a necessary step towards TB elimination (Campbell et al., 2014), modeling approaches that refine our estimates of parameters governing the progression and transmission of LTBI are potentially very useful, especially when coupled with more intensive screening and treatment.

In order to visualize the impact of our crowd-based approach on the universe of possible parameter values we propose the notion of a parameter-space hypercube: a hyperdimensional volume whose edges represent possible values of each of the 27 unknown parameters in our model. We can, *a priori*, estimate the volume of this potential hypercube by multiplying the range lengths proposed in the existing literature for each of our 27 parameters. Thus, for instance, one edge of the cube represents the parameter b, whose range in the existing literature is $[0.5, 1]$. The length of this edge would be $1 - 0.5 = 0.5$. The volume of the parameter-space cube corresponding to Table 1 (that collects ranges from the existing literature) is the product:

$$\begin{aligned} V_1 = {} & (0.2-0)(0.2-0)(0.2-0)(1-0.5)(1-0)(0.3-0)(0.9-0.6)(0.9-0.3)(0.5-0.3) \\ & (30-0)(0.2-0)(0.2-0)(0.2-0)(0.1-0)(0.1-0)(0.1-0)(0.1-0)(0.01-0) \\ & (1-0)(1-0)(0.9-0.6)(0.9-0.5)(0.9-0.5)(0.8-0.1) = 2.091898 \times 10^{-14} \end{aligned}$$

The utility of our approach in refining—or, more precisely, reducing—the hypercube of potential parameter values can thus be estimated by calculating the volume of the parameter hypercube compatible with low-error fits to CDC data. As before, each edge in this new hypercube represents a parameter, but this time, the length of the edge is the length of the parameter range obtained by our optimization process (Table 3). While some edges of this parameter-space cube are no shorter than the range from the existing literature, others (such as v_L) are considerably reduced. The volume of the new parameter-space cube suggested by our analysis is $V_2 = 1.968407 \times 10^{-21}$. The ratio of the two hypercube volumes is thus the direct measure of the cumulative reduction in the parameter-space volume achieved by our genetic algorithm and crowd-of-fits approach.

$$\frac{V_2}{V_1} = \frac{1.96 \times 10^{-21}}{2.09 \times 10^{-14}} \approx \frac{1}{10^7}. \tag{13}$$

While the available CDC data may be too sparse to inform convincing point estimates of unknown parameters, our approach, when applied to the available CDC data presented in Table 2, nonetheless achieves a substantial reduction in the "plausible" parameter space. Our newly defined hypercube occupies $\approx 0.00001\%$ of the volume of the parameter space based on the values currently available in the existing literature.

5 A crowd of consistent TB projections through 2050

Table 3 reveals that in our crowd of 100 low-error fits, some unknown parameters have highly concentrated distributions, while other parameters remain relatively dispersed over the parameter range drawn from the existing literature. Yet, despite substantial uncertainty about the values of many parameters, we find that predictions of future TB prevalence (and of multi-drug resistant TB prevalence in particular) are strikingly and reassuringly consistent across the crowd of 100 low-error fits discovered by our genetic algorithm.

The black circles in Figure 5 depict the trajectories of the actual CDC data tracking total active TB cases, H-resistant TB cases, MDR resistant TB cases and annual TB deaths from 1993 to 2014 (see Table 2). The 100 green trajectories shown in Figure 4 are the predicted annual trajectories corresponding to each model parameterization from our crowd of 100 near-best fits. Despite substantial variation in the values of many model parameters, predictions from our crowd of 100 fits are remarkably consistent for the period 2015–2050. Our crowd of low-error fits tightly matches the historical data for total active TB cases and MDR cases, and subsequently exhibits very low variance in the prediction period (2015–2050).

For contrast, consider the left-hand panels of Figure 5 for H-resistant cases and Annual TB deaths. Some fits from our crowd appear very faithful to the historical data while other fits dramatically over- or underestimate these populations. We note that, since all fits in our crowd achieve quite low error according to Equation (12), it may be that some fits that have very high error for Annual TB Deaths are simultaneously very accurate for matching historical rates of H-resistant cases, and *vice versa*. There may be some trade-off between achieving low error for these two populations simultaneously—perhaps good fits for H-resistance

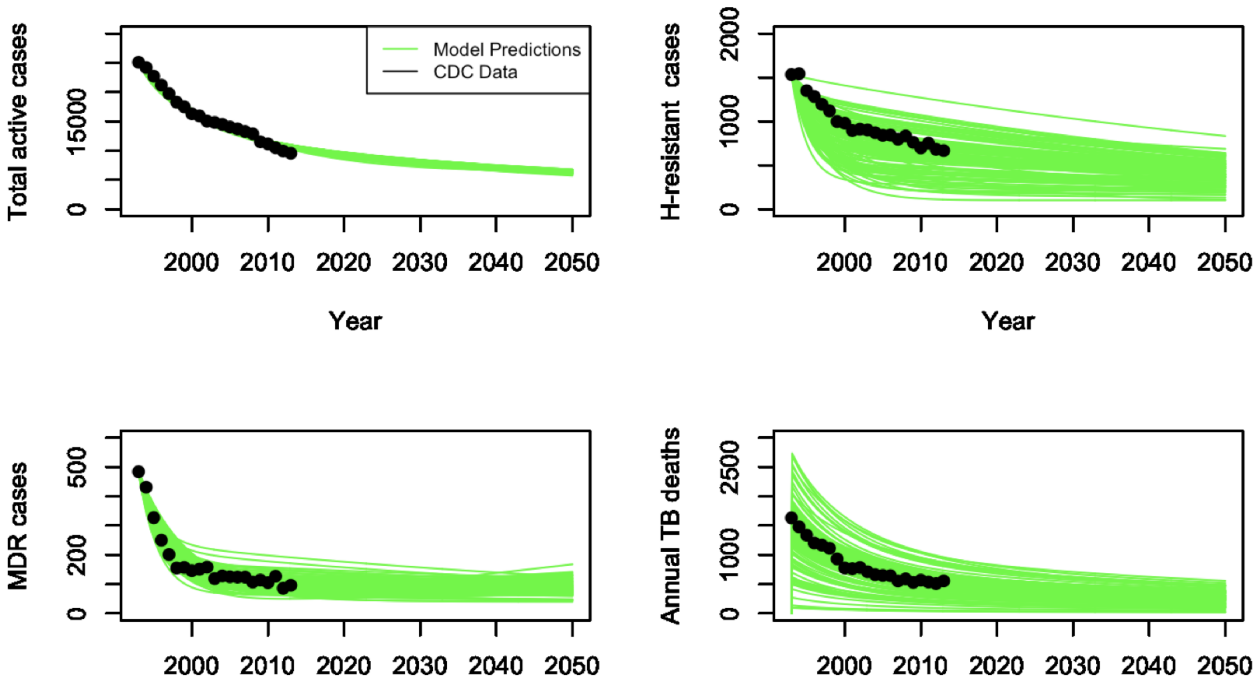

Figure 5: Plots of model fit including future predictions for the number of all active TB cases (*a*), H-resistant (*b*), and MDR cases (*c*), along with the annual number of TB deaths (*d*). The black circles indicate the CDC data and the green lines represent the trajectories predicted by parameter values derived from each of the 100 best fits.

Table 4: Summarizing a Crowd of 100 TB Projections.

Year	MDR Mean (Std.Dev.)	Total Active Mean (S.D.)	H-resistant Mean (S.D.)	Deaths Mean (S.D.)
2020	94.2 (23.7)	9468.2 (204.5)	559.5 (211.1)	402.8 (177.4)
2030	91.0 (21.7)	8168.7 (219.2)	502.9 (184.7)	341.3 (150.2)
2040	88.9 (21.7)	7202.2 (198.3)	454.7 (160.6)	299.8 (131.6)
2050	87.4 (24.0)	6413.3 (198.5)	412.9 (139.5)	267.5 (117.1)

demand certain parameter values that are somehow in conflict with achieving the best fits for Annual TB Deaths—but these subtleties are obscured by having a single aggregated error function.[2] Still, the variance of predictions across our crowd of 100 near-best fits appears to decrease by 2050 even for these relatively less-well-fit populations.

We provide summary statistics for our crowd of 100 near-best fits in Table 4. For the total number of annual active TB cases (the upper-left panel of Figure 5 and column 2 of Table 4), our crowd provides a consensus. While total TB cases will continue to drop until 2050 the apparent "leveling out" observed between 2000 and 2013 in the CDC data will continue. As a result, a substantial number of active cases (>6,000) are predicted in 2050. Our analysis discovers no low-error fit to the historical data in which annual active TB cases maintain a linear rate of decrease.

We also find incredible consistency in our crowd of predictions about annual MDR cases (lower left panel of Figure 5 and column 1 of Table 4). Our crowd agrees that decreases in the number of these hard-to-treat TB infections have effectively stagnated. At the same time, our crowd of near-best fits discovers almost no "plausible" parameterization of our model for which the number of MDR cases in the US grows by 2050 (and no parameterization for which this population grows substantially). Our crowd of near-best fits suggests that, provided that the current rates of immigrants entering the US with latent cases of various drug resistant TB strains is stable, the US is not in danger of a new epidemic of MDR tuberculosis.

6 Reflecting on Potential TB Interventions

A well-parameterized epidemiological model can help researchers gain insight about the effectiveness of potential policy interventions in reducing future disease burden. In this section, we explore TB projections from our crowd of near-best fits under

[2] In the conclusion, we mention a possible direction of future work that would explore beyond an aggregated error function.

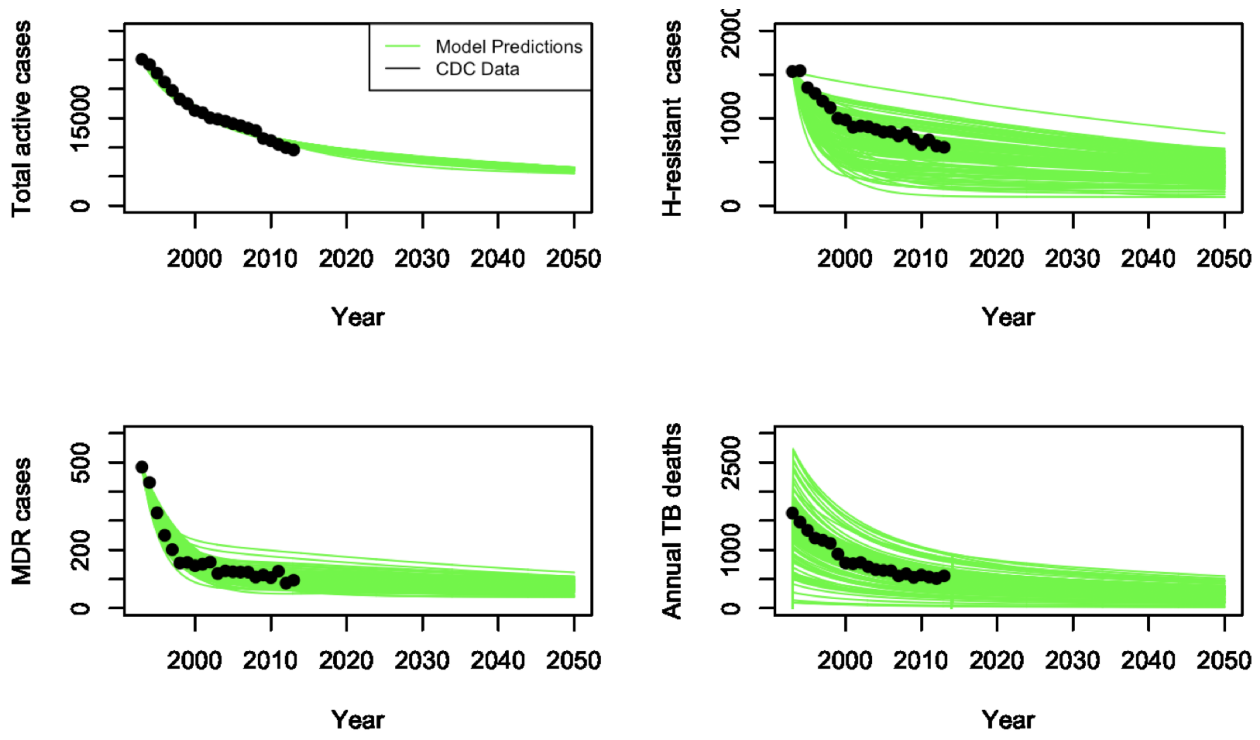

Figure 6: Plots of CDC data including future predictions for the number of all active TB cases (*a*), H-resistant (*b*), and MDR cases (*c*), along with the annual number of TB deaths (*d*) when l (the proportion of immigrants with LTBI) decreases by 50%. The black circles indicate the CDC data and the green lines represent the trajectories predicted by parameter values derived from each of the 100 best fits.

Table 5: Projections based on a 50% decrease in l (the proportion of immigrants with LTBI).

Year	MDR Mean (Std.Dev.)	Total Active Mean (S.D.)	H-resistant Mean (S.D.)	Deaths Mean (S.D.)
2020	93.1 (23.7)	9424.9 (231.3)	556.5 (211.0)	401.1 (176.6)
2030	87.4 (20.9)	8098.5 (261.9)	497.0 (185.0)	338.1 (148.9)
2040	82.5 (18.5)	7105.5 (233.9)	446.1 (161.1)	295.5 (129.9)
2050	78.9 (17.2)	6289.3 (194.1)	401.3 (139.9)	262.1 (115.1)

two stylized interventions.

First, we consider an intervention at the boundary of our model. An important driver of TB dynamics in the US is the influx of foreign-born individuals harboring TB infections. Investing in heightened screening and treatment of foreign-born individuals for LTBI could decrease the number of such cases. Suppose that this policy is implemented immediately and reduces the number of LTBI cases introduced through immigration by a factor of 2. In this scenario, for each fit in our crowd, the parameter l is reduced to half of its previous value (while the 26 other parameters are unchanged). We use this adjusted crowd of fits to produce projections of TB dynamics until 2050.

Figure 6 and Table 5 describe a crowd of predictions about the impact of this intensified immigrant screening and treatment intervention. As intuition would suggest, this intervention does decrease the value of each CDC-tracked compartment. Remarkably though, these decreases are very slight. Our crowd of fits consistently suggests a reduction in annual active TB cases and annual TB deaths of just a few percent. Economists suggest that widespread screening of immigrants for latent TB would carry high financial costs (Campbell et al., 2014; Cain et al., 2008; Verma et al., 2013); our crowd of fits provides little support for the cost-effectiveness of such measures.

Next we consider a different stylized intervention focused within the US. In this scenario, research investments in improving TB treatment regimens lead to reduced treatment times. In our model, decreasing treatment times is equivalent to increasing the transition rates ϕ_1, ϕ_2, ϕ_3, and ϕ_4 that describe the rate at which individuals exit the infectious compartments (I_i). Intuitively, fewer infectious individuals would be expected to produce fewer new exogenous infections. Figure 7 and Table 6 describe the predicted impacts of a 30% increase in ϕ parameters. While our crowd of fits don't suggest that such a change would allow TB elimination, we observe an almost immediate descent to a dramatically lower level of disease burden (Table 4). Specifically, by

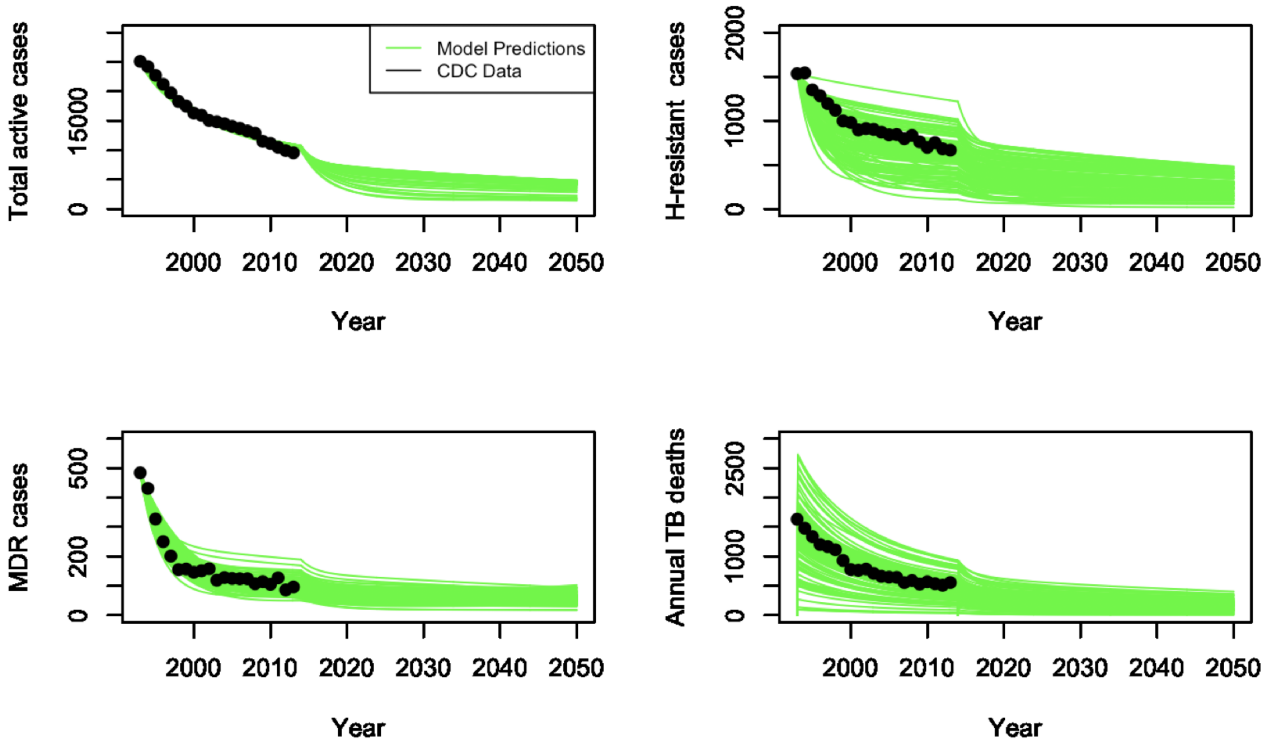

Figure 7: Plots of CDC data including future predictions for the number of all active TB cases (*a*), H-resistant (*b*) and MDR cases (*c*), along with the annual number of TB deaths (*d*) when ϕ_1, ϕ_2, ϕ_3, and ϕ_4 each increase by 30% (corresponding to reduced treatment times). The black circles indicate the CDC data and the green lines represent the trajectories predicted by parameter values derived from each of the 100 best fits.

Table 6: Projections based on a 30% increase in ϕ.

Year	MDR Mean (Std.Dev.)	Total Active Mean (S.D.)	H-resistant Mean (S.D.)	Deaths Mean (S.D.)
2020	69.6 (19.9)	6660.7 (839.1)	384.47 (165.5)	280.8 (127.8)
2030	63.13 (26.6)	5539.4 (960.6)	333.5 (149.3)	225.6 (105.9)
2040	60.2 (17.6)	4860.6 (847.1)	298.2 (129.9)	197.3 (92.8)
2050	57.9 (17.7)	4299.4 (729.3)	267.5 (112.7)	174.9 (82.0)

2020, the mean number of active cases falls to 6660, in contrast to 9468 if no such intervention was implemented. Our crowd of fits predicts that, decreasing treatment times by 30%—assuming this is medically feasible—would be a very effective policy. Short of major improvements in treatment, based on the structure of our model, a similar impact would be predicted for an intervention that simply reduces the exposure of infectious individuals to the general public by a similar fraction.

In the section below, we suggest how extensions of our approach (considering a crowd of low-error model parameterizations) could also be used to explore scenarios in which non-US latent TB demographics shift dramatically (e.g. if non-US rates of MDR prevalence were to grow substantially such that rates of latent MDR infection were much higher among the foreign-born US population than historic or current levels).

7 Sensitivity Analysis for the Error Function

Our crowd of 100 near-best fits exhibits wide dispersion for many parameters (see Table 3). One possible interpretation of this evidence is that our model predictions of $f^i_{\text{Active}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, $f^i_{\text{HR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, $f^i_{\text{MDR}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$, and $f^i_{\text{D}}(\alpha_2, \alpha_3, \ldots, z_3, z_4)$ (and consequently, our error function) are simply insensitive to some parameter values (at least within some neighborhood near low-error fits). A natural consequence of such insensitivity would be that the value of such parameters upon convergence of the genetic algorithm would effectively be selected uniformly at random from the known parameter range. Panel b) of Figure 4 certainly appears consistent with such an interpretation. To further investigate this behavior, we measured how modest scaling of the value of each parameter altered the quality of the model fit to the historical CDC data.

Table 7: Sensitivity analysis for the error function.

Scaled Param.	Mean Error Inflation From Scaling				Standard Deviation			
	0.7	0.9	1.1	1.3	0.7	0.9	1.1	1.3
α_2	1.0264	1.00158	1.0054	1.03763	0.05871	0.01423	0.01412	0.05458
α_3	0.97857	0.9918	1.00914	1.03042	0.038108	0.01156	0.01122	0.03472
α_4	1.0017	0.99716	1.0062	1.02887	0.03908	0.01487	0.01795	0.06536
b	6.15117	0.9916	1.9494	12.269	17.2059	0.62931	0.32181	11.9876
γ	1.00139	1.00046	0.99953	0.99861	0.001978	0.00065	0.00065	0.0019
l	0.981	0.993	1.0067	1.0211	0.01721	0.0059	0.00615	0.0191
μ	2.6202	1.07746	1.0756	1.99092	7.5455	0.2697	0.21785	3.53087
p	1935.66	175.755	4808.58	92879050.1	8438.26	891.495	31064.15	765673385.4
ϕ_1	137673328.2	4555.5004	110.23	931.5473	1096939828	30217.73	669.87	5519.19
ϕ_2	583.3503	1.12337	1.0664	1.2329	4484.821	0.60783	0.1303	0.4584
ϕ_3	24544.6	26.5628	6.005	60.959	1115507.519	90.015	7.503756	105.0462
ϕ_4	1.3262	1.00122	1.0166	1.059	0.90724	0.0289	0.01402	0.04597
$q\lambda$	1959.78	177.538	4814.15	85600819.4	8489.23	897.525	31073.27	704087630.8
r_2	0.9998	0.9999	1.00005	1.00017	0.0011	0.00038	0.0003	0.0011
r_3	0.9994	0.99981	1.0001	1.00057	0.002089	0.00067	0.00068	0.0019
r_4	1.00092	1.0002	0.9997	0.9993	0.00736	0.0022	0.00215	0.0059
t_1	1036.06	112.2003	3757.158	74799822.6	6086.234	689.79	25632.74	615263069.9
t_2	1.20956	1.04977	1.11854	247.27	0.50144	0.1364	0.5322	1616.31
t_3	47.5895	4.69193	22.05403	22996.3602	115.132	8.5292	114.92	167217.71
t_4	1.0297	1.00627	1.0038	1.1166	0.0402	0.0107	0.0168	0.36048
v_L	183.4333	5.46422	5.8159	188.6359	53.0857	1.36976	1.66199	59.2885
y_1	0.99927	0.9997	1.00024	1.00073	0.00173	0.00058	0.0005	0.00177
y_2	1.00388	1.00122	0.99884	0.99676	0.0044	0.00143	0.00138	0.00402
z_1	1.04657	1.01472	0.98604	0.96038	0.02137	0.00674	0.00639	0.01821
z_2	1.00318	1.00103	0.9989	0.99704	0.0021	0.00071	0.0007	0.00208
z_3	1.02641	1.00772	0.99334	0.98324	0.01503	0.00474	0.00465	0.0142
z_4	1.00051	1.0001	0.99982	0.999493	0.00048	0.00016	0.00015	0.00046

In a simpler setting, sensitivity analysis would be performed around a single specific optimized parameter combination. Instead, we have a crowd of 100 near-best fits: we perform sensitivity analysis around 100 different points in the 27-dimensional parameter space. Since we are uncertain whether some of these points give a more meaningful parameterization of our model than others, we summarize by reporting the mean and variance of the sensitivity across our crowd of near-best fits. We hope to gain insight about about the shape of the error function across a collection of 100 neighborhoods (one neighborhood centered at each parameter combination in our crowd of near-best fits).

For each of our 100 near-best fits, we performed a sensitivity analysis on the error function by scaling each parameter value obtained by 0.7, 0.9, 1.1, and 1.3 respectively, while keeping all other parameter values fixed. Each altered parameter combination was evaluated according to the error function (Equation (12)). This new error measurement was normalized by the error of the initial fit. Finally, for each parameter, we computed the mean and standard deviation of the normalized error over our crowd of 100 near-best fits. These values are reported in Table 7.

A number of factors must be considered in interpreting the values in Table 7. First, suppose that a near-best fit yields a parameter value near the boundary of the known literature range for that parameter (Table 3). In such cases, the scaling used to produce the values in Table 7 could result in a parameter value outside the range considered plausible in the existing TB literature. In turn, using biologically unrealistic values in our model could be a driver of some large error values after scaling. For example, t_1, the proportion of treatment time when individuals infected with drug-susceptible TB are infectious, occupies almost the entire range considered plausible in the existing literature across our 100 low-error fits (Table 3). Scaling each of the 100 optimized t_1 values by 1.3 gives a large average increase in the error function (though the large standard deviation given for scaling t_1 by 1.3 in Table 7 implies that the sizable mean may be driven by a smaller group of very large sensitivity measurements among the 100 low-error fits). We also underscore that some means of normalized error in Table 7 are less than 1. We interpret

this as an artifact of the genetic algorithm: while a value survives the evolutionary process by producing a low error, another nearby slightly better value may have been lost to the randomness of the evolutionary process, or perhaps never tested due to the granularity of our randomization procedure.

We describe an error function as being *sensitive to a parameter* if some mean normalized error that results from scaling that parameter by a value in $\{0.7, 0.9, 1.1, 1.3\}$ is more than 4 (that is the error function increases by a factor of 4 averaged across our 100 low-error fits). This threshold is intended to capture situations where a modest scaling of a parameter causes a substantial increase in the error across our crowd of 100 low-error fits (though the specific threshold choice of 4 is somewhat arbitrary). By this standard, the model appears to be particularly sensitive to the following parameters: b, p, ϕ_1, ϕ_2, ϕ_3, $q\lambda$, t_1, t_2, t_3, v_L. The biological meaning of these parameters may explain their effect on the model b represents the proportion of the initial cases that are drug-susceptible. Given that drug resistance plays an important role in TB dynamics, small deviations in values of b, can have a strong impact on the model. The parameter p is the proportion of exogenous infections that are acute (i.e. active) and $q\lambda$ is the progression rate from latent to active infections. Considering that active infections are contagious, a small change in their number can significantly impact transmission, accounting for the sensitivity of our results to values of the p parameter. $q\lambda$ represents the transmission of TB, t_1, t_2, and t_3 is the proportion of time when DS infected, H-resistant, or R-resistant infected individuals respectively can transmit the infection. Here again, the importance of these parameters in the sensitivity analysis comes as no surprise. Finally, parameters ϕ_1, ϕ_2, ϕ_3 are the rates at which infectious individuals (DS, H-resistant, and R-resistant respectively) terminate treatment. Given that inappropriate treatment, including the untimely termination of treatment, can lead to the acquisition of resistance, it is reasonable that changes in those parameters significantly impact the number of resistant cases.

For the parameters that are identified as sensitive based on Table 7, the standard deviation of the normalized error is frequently high, suggesting that the results are highly variable across our crowd of 100 near-best fits. Among the top 100 low-error fits some parameter sensitivity analyses led to sharp increases in the error, while others did not. This is to be expected, since it is unlikely that the error function behaves identically in the neighborhoods of 100 different local minima (or local near-minima). We note that the values reported in Table 7 depend strongly on the 100 low-error fits considered. Sensitivity analyses based on a smaller number or different crowd of fittest convergence individuals (final solutions delivered by genetic algorithm runs) could provide a different list of sensitive parameters.

We also note that in several cases where the quality of the model fit is highly sensitive to a parameter, the impact of scaling that parameter is strongly asymmetric. For example, t_2 may be scaled down or scaled *slightly upward* and still maintain low error across our entire crowd of 100 low-error fits. In contrast, when t_2 is scaled up by 1.3 at least some near-best fits in our crowd of 100 show a large inflation of the error function.

Before performing sensitivity analysis on our crowd of 100 near-best fits, it seemed natural to suspect that the error function would be highly sensitive to those parameters whose values appeared to be highly concentrated by the criteria of achieving a low-error fit in Table 3. Qualitatively, since achieving low error seems to require key parameters to occupy very specific and narrow ranges, it seems intuitive that the error function must be highly sensitive to such parameters. Conversely, it would seem intuitive that for any parameter for which the crowd of near-best fits seems ambivalent about its value (e.g. panel b) of Figure 4) the sensitivity of the error function to changes in that parameter value would also prove low.

Remarkably, Table 7 shows that the relationship between concentrated ranges and error-function sensitivity can be much more complex. A variety of different relationships are observed. Table 7 shows that the quality of model fit is highly sensitive to some parameters highlighted as having newly-narrowed ranges in Table 3 ($b, q\lambda$, and especially p). Proportional scaling of these key parameters[3] can dramatically increase the average error of the fit across our crowd of 100 near-best fits. Especially for p, the error function must be exceedingly "steep" on average in neighborhoods around near-best fits. In contrast, other parameters whose ranges are dramatically narrowed among our crowd of 100 near-best fits display quite low sensitivity under proportional scaling (α_3, α_4, l, and μ). That is, even though achieving a low-error fit forces these parameters to occupy a narrow range, within a small neighborhood around each of our 100 near-best fits, the error function is flat in all four directions.

At the same time, we find that the error function is actually *highly sensitive* to the values of several parameters whose resulting ranges in our model remained very dispersed ($\phi_1, \phi_2, \phi_3, q\lambda, t_1, t_2$, and t_3). While our crowd of 100 low-error fits made use of a wide variety of values for each of these parameters, Table 7 shows that the error function is on average quite "steep" in these directions near low-error fits. We conjecture that the combination of wide distribution of values with high sensitivity arises when *achieving low error requires that certain subsets of these seven parameters be correlated in highly specific ways*. Figure 8 provides a schematic in two synthetic dimensions to visualize how this could arise. For example, achieving a low-error fit to the CDC data may not force t_1 to occupy a particularly narrow range, but perhaps it does force the vector $(q\lambda, t_1, t_2, t_3)$ to occupy a very specific constrained volume within a 4-dimensional squashed parameter-value cube (again, see Figure 8 for a low-dimensional schematic of how this might arise). These four parameters control the rate of all new infections that arise through exposure (to any TB strain): while the CDC data may be too sparse to recommend narrow ranges for each individual parameter, the low-error fits in our crowd are discovering certain plausible tradeoffs among the values of these related parameters that result in

[3] This is a statement about proportional scaling of a single parameter *in the neighborhood* of a very low-error fit discovered by a run of the genetic algorithm.

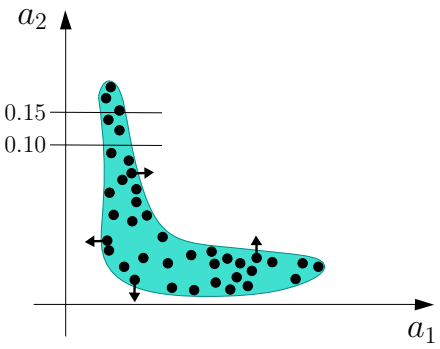

Figure 8: Constrained Low-Error Volumes in the Parameter Space Can Simultaneously Exhibit Wide Dispersion and High Sensitivity. Consider two synthetic parameters a_1 and a_2. Suppose that the green region of this 2-dimensional parameter space results in very low error, and that error is dramatically higher outside the green region. A crowd of low-error fits is depicted as black points. Parameters a_1 and a_2 both exhibit wide ranges across the crowd of fits, but many fits in the crowd are near the boundary of the low-error region such that scaling one parameter (while the other remains fixed) enters the high-error region.

predictions which closely replicate the populations observed in the historical CDC data. Under such an interpretation, the high sensitivity to proportionally scaling t_1 (Table 7) would indicate that moving away from a balanced combination of $(q\lambda, t_1, t_2, t_3)$ in the t_1 direction—with no coordinated adjustment of $(q\lambda, t_2, t_3)$—causes a steep increase in error in some neighborhoods of our crowd of low-error fits. While the distribution of t_1 obtained from our crowd of near-best fits is widely dispersed across the known range from the existing literature, our error function is far from ambivalent about the value of this parameter.

This "balanced combination" interpretation of our seven wide-range but high-sensitivity parameters suggests that new direct measurements of small subsets of these parameters could potentially force other parameters from the combination into more specific and narrow ranges (based on our low-error fit criteria). For example, in the synthetic schematic provided in Figure 8, if a new direct measurement revealed that $a_2 \in [0.10, 0.15]$ then achieving low error would require a_1 to come from a much narrower range than the crowd of fits initially suggested. That is, our crowd of near-best fits may not champion narrow ranges for these seven particular parameters individually, but our crowd does still encode a kind of structure that describes *combinations of smaller subsets of parameters* that are compatible with real rates of historical TB prevalence. The structure of these parameter combinations is hard to visualize because our crowd of near-best fits does not appear to have a "thin" projection into one or two of these seven dimensions (note that our "Obtained Ranges" in Table 3 are effectively projections of our crowd of near-best fits into one dimension). The absence of easily interpret-able low-dimensional projections for some subsets of crucially correlated parameters is likely a natural consequence of the complex system we model: several compartment populations are controlled by contributions from many other compartments (so that "balances" between related parameters/rates entering and leaving a compartment are frequently more than 2-way).

8 Conclusion

In epidemiological modeling, the emergence of new drug-resistant strains aggravates existing challenges related to data sparsity and the lack of reliable parameter estimates. Even for diseases with substantial global impacts, like tuberculosis, some data sparsity is unavoidable under current technology and regulatory policies (at present, the CDC has no mechanism in place to directly measure latently infected populations). In this paper, we extended a widely cited compartment model of TB dynamics in the US (Hill et al., 2012) to incorporate several drug-resistant strains. Our extended model results in both more compartments and more inter-compartment transition rates that demand strain-specific parameters.

Our initial attempt to fit our model based on 22 years of CDC data exposed that, in this data-sparse setting, many competing vectors of our 27 unknown parameters result in low values of our error function. When historical data is relatively sparse, it is not surprising that an error function can exhibit a complex set of local minima. For our TB study, we found that competing low-error parameter combinations could sometimes disagree dramatically about the values of important model parameters. This observation appeared to undermine the conventional wisdom that the lowest-error parameter combination gives the best indication about the true value of the parameters. This obstacle is not specific to disease modeling: a similar issue could arise in many data-sparse settings where local minima of similar quality form an elaborate landscape.

To move forward we introduced the notion of a *crowd of near-best fits*. We obtained a set of 100 low-error parameter combinations by repeatedly running a genetic algorithm to convergence. Selecting the best 100 runs out of 1,000 genetic algorithm runs, every fit in our crowd has an error value that matches or outperforms a set of predictions which has maximum deviation

of 6.4% across all 87 of the empirical quantities documented by the CDC. Next we explored whether our crowd of fits could provide insights analogous to the desired products of traditional model-fitting efforts.

Even though we did not acquire precise estimates for each compartment and each parameter, by considering a *crowd of near-best fits* we were still able to recommend dramatic restrictions of the possible ranges for seven key model parameters. Our seven dramatically restricted parameters include several that are difficult to measure through direct experiment. For example, we recommend newly narrowed ranges for the proportion of latent TB infections that are multi-drug resistant in the US, $\alpha_4 \in [0.0002, 0.0319]$, and for the rate of progression from latent to active infection, $v_L \in [0.000049, 0.0011]$. The parameter v_L plays an integral part in virtually every TB model, and modelers are heavily reliant on guesswork when estimating progression rate from latent to active TB in a population (Dowdy et al., 2015). Our goal is not necessarily to locate the very best set of parameters, but to find a collection of parameterizations of the model that fit the data. Assembling a collection of good fits can provide insight into epidemiological parameters that are known to be troublesome and hard to measure directly, but are nonetheless critical to TB dynamics (e.g. v_L).

Furthermore, we show that older predictions of declining TB prevalence that ignored drug-resistant strains (Hill et al., 2012) appear to be robust when these new types of infections are considered in our model. According to Hill et al., even if all TB transmission ceased instantly, there would still be a few cases even in 2100, mainly supplied by the foreign-born population. Indeed, our crowd of fits predict that despite the sharp decreases in recent TB cases, low prevalence is maintained in the United States through at least 2050 (Figure 5). Such predictions are consistent with other models (e.g. Shrestha 2017, Menzies 2018). This behavior is remarkably consistent across our large crowd of low-error model parameterizations. Our genetic-algorithm search discovers no low-error fits to the data that are consistent with rising future TB rates, despite the emergence and spread of recalcitrant MDR strains.

In the broader context of epidemiological modeling, our approach of using a genetic algorithm to generate a crowd of near-best fits offers an alternative method for parameterizing models to fit data. Modelers who report a particular set of parameters as the best fit for their data could consider studying their model's sensitivity to alternative parameterizations using this method of generating alternative near-best fits for comparison.

In Section 6 we explored predictions about how two stylized public health interventions would impact future TB prevalence. This exploration produced two modified versions of Figure 5. Figure 6 describes our crowd's consistent prediction that decreasing the number of latent TB cases introduced through immigration (e.g. through intensified border screening) would have only a mild effect on annual disease burden. On the other hand, in Figure 7 and the corresponding Table 6, our crowd of fits consistently predict that increasing the rate at which individuals leave the infectious compartment would have a substantial impact on future prevalence. Such an impact might be achieved as a result of investments in research that reduce treatment times, or by developing more effective isolation and treatment protocols for infected individuals.

We note that the value that can be extracted from sparse data is both model- and data-specific; we can not guarantee that the remarkable consistency revealed by our crowd-based analysis for drug-resistant TB will occur for other sparse data sets or alternative models. Still, we do hope that our crowd-based approach will motivate further exploration of how emergent infection-types can be explicitly incorporated both for TB and other infectious diseases. Indeed, it is our hope that the range of useful insights that could be accessed by our crowd-based model-fitting approach is wider than what we have been able to cover here. Towards that end we suggest several directions for future work.

Directions for Future Work. As in every computational study, we make a number of implementation choices that can be further explored. Our crowd of one hundred low-error fits provides us with an important signal about how to best parameterize our model. Nonetheless, it is important to note that different implementation choices might have resulted in a different set of best fits. Choosing a different error function, such as the least absolute deviation method might have yielded different results. Similarly, our use of a genetic algorithm required making somewhat arbitrary choices about generation size, mutation rate, recombination rate and breeding method. Here again, different choices might lead to slightly different outcomes. Our results likely exhibit some dependency on the choice of algorithm and on the specifics of its implementation. To defend against such artifacts, we are exploring the features of a crowd of best fits derived from the low-error fits generated across multiple optimization algorithms. Finally, our choice to consider a crowd size of the top one hundred near-best fits from the initial collection of 1000 genetic algorithm runs is somewhat arbitrary. Even though studying the 100 near-best fits ensures that we are examining fits with a very low error (from panel a) of Figure 4), a more specific numerical criterion could be used to determine membership in the crowd of best fits. Such a crowd might, for instance, be assembled by selecting only those fits whose error is at most 10% of the largest observed error at convergence. Alternatively, membership in this exclusive club of fits could be defined to include all fits discovered with an error no greater than twice the minimum error fit found among the 1000 genetic algorithm runs. Considering Figure 5, we might even imagine a 2-phase qualification process: to join the crowd of near-best fits, a fit must first achieve low error (according to a weighted error function), followed by a second triage that retains only those solutions that exhibit high-quality fits for each category of CDC data. This work has explored the sensitivity of our error function to changes in the values of the model parameters, but has not conducted a global sensitivity analysis for the 100 parameterized models that

we obtain (these models all have a common structure, but their parameter sets are drawn from our crowd of near-best fits). As with our predictions of future TB prevalence, sensitivity analysis could be conducted for each fitted model to explore whether the crowd of best fits consistently links a key outcome (e.g. annual TB Deaths in 2030) to the value of a single input (e.g. MDR cases in 2019). Such linkages could inform preventive public health efforts and could point to specific measurements that could substantially improve our prediction capacities.

The two stylized public health interventions simulated in this paper can also be extended to a broader range of potential policy suggestions. These might include varying the strain-specific rates of infection in both native and foreign-born populations, exploring the consequences of the increasing prevalence of single or multi-drug resistant strains of MTb around the globe on the prevalence and Tb-related mortality in the US, and investigating the effects of promiscuous transfer of resistance determinants across MTb strains.

Along similar lines, the role of extensively drug-resistant (XDR) strains on TB dynamics in the US deserves further scrutiny. XDR has become a major concern in TB control for countries with high TB prevalence. South Africa, for instance, has witnessed a 10-fold increase in XDR cases in the past decade, with 1500 reported XDR TB cases in 2012. This increase was predicted by Basu et al. (2007) who forecast an even larger increase in XDR cases if no additional interventions are implemented. Current evidence suggests that the majority of XDR cases arise from individual to individual transmission, rather than as a result of *de novo* resistant strains due to inappropriate treatment of MDR TB. This pattern was also seen in a social network analysis that investigated the connections among patients with XDR TB (Shah et al., 2017). It revealed that social networks create multiple opportunities for transmission, both in the community and in the clinic. As a result, direct transmission explained at least 59 of MDR TB cases. The relatively low success rate of treatment for XDR tuberculosis (<40%) and its high mortality rate (50–80%) (Shah et al., 2017) have increasingly attracted the attention of campaigns seeking to control and eradicate TB. The rising *in situ* prevalence of these resistant strains, coupled with the inexorable rise in international travel and cross-border movements, remind us that TB is a global challenge necessitating a concerted and data-driven global response.

Acknowledgements

The authors would like thank our collaborators, Chenyue Lu and Olivia Justynski, with whom this project began, as well as the Smith College Mathematics and Statistics department, and the Biomathematics Sciences Concentration for their support.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Basu, Sanjay, Jason Andrews, Eric M. Poolman, Neel R. Gandhi, N. Sarita Shah, Anthony Moll, Alison P. Galvani, and Gerald H. Friedland. 2007. "Prevention of nosocomial transmission of extensively drug-resistant tuberculosis in rural South African district hospitals: an epidemiological modeling study." *Lancet*, 370(9597):1500–7. 16, 34

Bhunu, C. P. 2011. "Mathematical analysis of a three-strain tuberculosis transmission model." *Applied Mathematical Modeling*, 35:4647–4660. 16

Borrell, S., and S. Gagneux. 2009. "Infectiousness, reproductive fitness and evolution of drug- resistant Mycobacterium tuberculosis."*International Journal of Tuberculosis and Lung Disease*, 13(12):1456–1466. 16

Cain, Kevin, Stephen R. Benoit, Carla A. Winston, and William R. MacKenzie. 2008. "Tuberculosis among foreign-born persons in the United States. " *The Journal of the American Medical Association*, 300(4):405–412. 16, 28

Campbell, Jonathon, Fawziah Marra, Victoria Cook, and James Johnston. 2014. "Screening immigrants for latent tuberculosis: do we have the resources?" *Canadian Medical Association Journal*, 186(4):246–247. 26, 28

Centers for Disease Control and Prevention. 2007. "Extensively Drug-Resistant Tuberculosis– United States, 1993-2006." *Morbidity and Mortality Weekly Report (MMWR)*. 56(11):250–253. Accessed November 15, 2018. www.cdc.gov/mmwr/preview/mmwrhtml/mm5611a3.htm 17

Cohen, Ted, and Megan Murray. 2004. "Modeling epidemics of multidrug-resistant tuberculosis of heterogeneous fitness." *Nature Medicine*, 10(10):1117–1121. 16

Dheda, Keertan, Tawanda Gumbo, Neel R. Gandhi, Megan Murray, Grant Theron, Zarir Udwadia, G. B. Migliori, and Robin Warren. 2014. "Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis." *Lancet Respiratory Medicine*, 2(4):321–338. 16

Dowdy, David, Christopher Dye, and Ted Cohen. 2015. "Data needs for evidence-based decisions: a tuberculosis modeler's 'wish list.'" *International Journal of Tuberculosis and Lung Disease*, 17(7):866–877. 33

Garner, Paul, and Jimmy Volmink. 2003. "Directly observed treatment for tuberculosis." *British Medical Journal*, 327(7419):823–824. 16

Hill, A. N., J. E. Becerra, and K. G. Castro. 2012. "Modelling tuberculosis trends in the USA." *Epidemiological Infections*, 140(10):1862-1872. doi:10.1017/S095026881100286X. 15, 16, 17, 32, 33

Justynski, Olivia. 2016. "A Four-Strain Model of Drug-Resistant Tuberculosis in the United States." (Unpublished undergraduate thesis). Mount Holyoke College, USA. 17

Kim, Yee Hyung, Gee Young Suh, Man Pyo Chung, Hojoong Kim, O Jung Kwon, Seong Yong Lim, Si Young Lim and Won-Jung Koh. 2008."Treatment of isoniazid-resistant pulmonary tuberculosis." *BMC Infectious Diseases*, 8(6). 16

Kolda, Tamara G., Robert Michael Lewis, and Virginia Torczon. 2003. "Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45: 385–482. 21, 22

Koziel, Slawomir (Eds.), Xin-She. Yang (Eds.), O. Kramer, David Echeverria Ciaurri, and Slawomir Koziel. 2016. "Computational Optimization, Methods and Algorithms." Springer Publishing Company, Incorporated. 21, 22

Shah, N. Sarita, Sara C. Auld, James C. M. Brust, Barun Mathema, Nazir Ismail, Pravi Moodley, …, and Neel R. Gandhi. 2011."Transmission of Extensively Drug-Resistant Tuberculosis in South Africa." *The New England Journal of Medicine*, 376(3):243–253. 34

Sharma, Aditya, Andrew Hill, Ekaterina Kurbatova, Martie van der Walt, Charlotte Kvasnovsky, Thelma E Tupasi, …, Peter Cegielski, and for the Global Preserving Effective TB Treatment Study Investigators. 2017. "Estimating the future burden of multidrug-resistant and extensively drug-resistant tuberculosis in India, the Philippines, Russia, and South Africa: a mathematical modelling study." *Lancet Infectious Diseases*, 17:707–715. 16

Srivastava, Shashikant, Jotam G. Pasipanodya, Claudia Meek, Richard Leff, and Tawanda Gumbo. 2011. "Multidrug-resistant tuberculosis not due to noncompliance but to between-patient pharmacokinetic variability." *The Journal of Infectious Diseases*, 204:1951–1959. 16

Verma, G., A. W. Chucj, and P. Jacobs. 2013. "Tuberculosis screening for long-term care: a cost-effectiveness analysis." *International Journal of Tuberculosis and Lung Disease*, 17(9):1170–1177. 28

World Health Organization. 2014. "WHO targets elimination of TB in over 30 countries." Accessed November 15, 2018. `http://www.who.int` 15, 16

World Health Organization. 2016. "Global tuberculosis report 2016." Accessed November 15, 2018. `http://www.who.int` 16