Model Averaging in Viral Dynamic Models

Antonio Gonçalves, France Mentré, Annabelle Lemenuel-Diot, Jérémie Guedj

Model Averaging in Viral Dynamic Models

Antonio Gonçalves[1], France Mentré[1], Annabelle Lemenuel-Diot[2] and Jérémie Guedj[1]

[1]Université de Paris, IAME, INSERM, F-75018 Paris, France

[2]Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel

Corresponding author: Antonio Gonçalves, antonio.goncalves@inserm.fr

IAME INSERM U1137,

16 rue Henri Huchard 75018, Paris, France

Tel: +33 1 57 27 75 39

Abstract

The paucity of experimental data makes both inference and prediction particularly challenging in viral dynamic models. In presence of several candidate models, a common strategy is model selection (MS), in which models are fitted to the data but only results obtained with the "best model" are presented. However, this approach ignores model uncertainty, which may lead to inaccurate predictions. When several models provide a good fit to the data, another approach is model averaging (MA) that weights the predictions of each model according to its consistency to the data.

Here we evaluated by simulations in a nonlinear mixed-effect model framework the performances of MS and MA in two realistic cases of acute viral infection: i) inference in presence of poorly identifiable parameters, namely initial viral inoculum and eclipse phase duration ii) uncertainty on the mechanisms of action of the immune response.

MS was associated in some scenarios with a large rate of false selection. This led to a coverage rate lower than the nominal coverage rate of 0.95 in the majority of cases and below 0.50 in some scenarios. In contrast, MA provided better estimation of parameter uncertainty, with coverage rates ranging from 0.72 to 0.98 and mostly comprised within the nominal coverage rate. Finally, MA provided similar predictions than those obtained with MS.

In conclusion, parameter estimates obtained with MS should be taken with caution, especially when several models well describe the data. In this situation, MA has better performances and could be performed to account for model uncertainty.

Introduction

Since 1995 and the two seminal papers providing an estimate of the half-life of HIV particles in blood (1,2), the use of viral dynamic models has considerably expanded. Applications have been summarized in a recent issue of Immunological Reviews (3), showing their relevance for understanding the host-pathogen interactions in both chronic and acute infections (4–6). In the last decade, parameter estimation of these models has increasingly relied on nonlinear mixed effect models (NLMEM), a statistical approach that improves both precision and accuracy of estimates by explicitly taking into account the between-subjects variability in the model (7,8). This is particularly true in the case of antiviral drug development where NLMEM have become central to support optimal treatment strategies in presence of a large variability in the response (9,10).

Although inference has been greatly facilitated by the use of NLMEM, viral dynamic models remain often characterized by a lack of theoretical or practical identifiability (7,8). In fact the availability of powerful algorithms for inference has mechanically led to the development of increasingly complex models, questioning the reliability of viral kinetic parameter estimates. In order to improve identifiability of these models, a commonly used strategy is to fix parameters to plausible values and then to check the impact of these choices by conducting sensitivity analyses. For instance in acute viral infection, one can fix the initial viral inoculum or the eclipse phase duration, two parameters that can hardly be estimated using only viral load data (11,12). Data fitting can also be used to evaluate the plausibility of different biological assumptions. In that case the usual approach is model selection (MS), where a predefined set of candidate models are fitted to the data and the model providing the best fit to the data (based on Akaike or Bayesian Information Criteria) is selected and carried forward in the analysis. In both contexts, these approaches, by focusing the predictions on a single

model, ignore the model uncertainty and may lead to wrong predictions (13) and potentially inaccurate biological conclusions (14–16).

In this paper, we propose to use model averaging (MA) as an alternative approach to MS in viral dynamics. MA is a conceptually simple approach, where the uncertainty related to each candidate model is taken into account and predictions associated to each model are weighted based on their consistency with the data (17,18). Through an extensive simulation study, we compare parameter estimates and predictive performances of model averaging versus model selection. We discuss the benefits and limits of model averaging compared to model selection. Simulations are inspired from recent works in Zika and Ebola virus dynamics (19,20) representing two typical settings encountered in viral dynamic modeling: i) a set of parameters are fixed to arbitrary values of a given biological model to ensure identifiability; ii) model selection relies on the comparison of fitting criterion of a set of pre-defined different biological models.

Material and methods

Model selection and model averaging

Model for the observations. Let Y_{ijm} denote the j^{th} log viral load measurement of subject i at time j, and suppose that *m*=1,..,M candidate models can be used to simulate the data. The model for the observations is defined as:

$$Y_{ijm} = log_{10}\left[V_m\left(t_j, \theta_{im}\right)\right] + e_{ijm} \quad (1)$$

where V_m is the viral load prediction function given by model *m*, θ_{im} is the vector of individual parameters under model *m*, t_j the time of viral load measurement, assumed to be similar for all patients and all models, and e_{ijm} the residual error. Individual parameters θ_{im} are log-normally distributed and depend on the vector of fixed effects μ_m and the vector of random effects $\eta_{im} \sim \mathcal{N}(0,\Omega_m)$ with $\theta_{im} = \mu_m \times e^{\eta_{im}}$. The variance-covariance matrix Ω_m is assumed to be diagonal. Residuals errors are assumed to be independent and normally distributed $e_{ijm} \sim \mathcal{N}(0, \sigma_m^2)$. Each biological model *m* is therefore associated with a set of population parameters, Ψ_m, of dimension $p_m = \dim(\mu_m, \Omega_m, \sigma_m)$.

Inference and model selection. For each candidate model, one can estimate the parameters using maximum likelihood estimates, providing, for each model, an estimate of the population parameters, noted $\hat{\Psi}_m$. One can also provide the confidence intervals of the parameters of interest under each model. This can be done using the asymptotic approximation where the density function of the estimated parameter $\hat{\Psi}_m$, $p(\hat{\Psi}_m)$, is assumed to be Gaussian with a variance-covariance matrix given by the inverse of the Fisher Information Matrix (FIM^{-1}).

Then, the most common approach is to select the model that best describes the data. This can be done using various criteria that rely on penalizations of the log-likelihood (LogL), such as

the Akaike information criteria (AIC), the consistent Akaike (CAIC) or the Bayesian information criteria (BIC) (21–24). In line with previous analysis (25,26), we relied on AIC given by $AIC_m = -2LogL(\hat{\Psi}_m) + 2p_m$. The analysis then focuses on the results (i.e., parameter estimates, confidence intervals, and predictions) obtained with the "best" model, i.e., the model associated with the lowest AIC among the m candidate models, noted AIC_{min}, with parameter estimates noted $\hat{\Psi}_{MS}$.

Model averaging. As explained above, MS is limited in the sense that it ignores the uncertainty associated with each model and only focuses on a post hoc selected model (14). Alternatively, one can use model averaging (MA) to take into account the fact that several candidate models may provide a reasonable fit to the data . In this approach a weight is attributed to each candidate model, w_m, proportional to AIC, such as $w_m = \frac{e^{\frac{-\Delta AIC_m}{2}}}{\sum_{m=1}^{M} e^{\frac{-\Delta AIC_m}{2}}}$ where $\Delta AIC_m = AIC_m - AIC_{min}$ (14,17,18). In that case the MA estimator of Ψ_m is given by $\hat{\Psi}_{MA}$, with a density function given by $p(\hat{\Psi}_{MA}) = \sum_{m=1}^{M} w_m \, p(\hat{\Psi}_m)$. Another approach could be to consider only the models that are responsible for the majority of the weight (0.9 or 0.8), and equally average them (see Discussion section).

Viral dynamic settings

Our objective is to compare model selection and model averaging in two typical contexts of viral dynamic models. In the first setting, we focus on the issue arising from using model selection when some parameters of the model cannot be identified and are fixed to arbitrary values. In the second setting, we focus on the issue arising from using model selection when several different biological models can be proposed to fit the data.

Setting I: viral dynamic models in presence of poorly identifiable parameters. We here focus on the standard target cell limited (TCL) model given by:

$$\frac{dT}{dt} = -\beta TV \quad (2)$$

$$\frac{dI_1}{dt} = \beta TV - kI_1 \quad (3)$$

$$\frac{dI_2}{dt} = kI_1 - \delta I_2 \quad (4)$$

$$\frac{dV}{dt} = \pi I_2 - cV \quad (5)$$

$$T_{t=0} = T_0;\ I_{1,t=0} = 0;\ I_{2,t=0} = 0;\ V_{t=0} = V_0$$

where, T are the target cells, I_1 the infected cells in eclipse phase, I_2 the productive infected cells and V the viral load in plasma. The model depends on the following disease parameters: β the infectivity rate constant, k the eclipse rate, δ the infected cell elimination rate, π the viral production rate constant, c the clearance of free virus, T_0 the initial number of target cells and V_0 the initial viral load. For the ease of interpretation and fitting, we reparametrized the model as $R_0 = \frac{\beta\pi T_0}{c\delta}$, the basic reproductive ratio, instead of β, where R_0 represents the number of secondary infection caused by one infected cells when the target cells are abundant. For the sake of simplificty we focused here on a simple, exponentially distributed, duration for the eclipse phase, but more complex models can be considered (27).

Not all parameters of the TCL model can be uniquely identified when only the viral load data are available (28–30) and this issue is not circumvented when parameters are estimated using NLMEM. This can be shown by analyzing the expected standard errors obtained with the approximated Fisher Information Matrix (http://www.pfim.biostat.fr/; see more details in (30)). Table I provides the expected standard errors obtained with 30 individuals sampled 3 days from day 3 up to day 18 post infection using typical parameter values close to those found during Zika infection in nonhuman primates (19). Although being theoretically identifiable, several parameters are associated with a very large expected standard error. This

can be corrected by fixing V_0 and k to some arbitrary values, leading to expected relative standard errors lower than 30% for all parameter values. We here aim to evaluate the impact of the choice of k and V_0 on parameter estimates.

We also conducted a sensitivity analysis on the parameters variability of setting I (k=4 d^{-1} and V_0=10^{-4} copies.mL^{-1}). We reported the influence of lower and higher variabilities (ω=0.1 and ω=1, respectively) on the performances of MS and MA.

Setting II: viral dynamic models including the immune response. In order to evaluate the impact of testing different biological assumptions in parameter estimates and predictions, we considered 4 additional models integrating the role of innate or adaptive immune response in the control of viral replication and inspired from the models used to describe Ebola infection in nonhuman primates (20). These models extend the TCL model with an additional compartment, noted F. This compartment is not observed and can therefore represent any biological entity involved in viral clearance, such as cytokine, macrophages, T-cell or antibodies (5,32,33). We assumed that F is produced at a rate q proportional to the number of productively infected cells, I_2, and is eliminated at a rate d_F (20). Thus F could either i) increase the number of refractory infected cells (refractory model, R), ii) decrease the viral production (production inhibition model, PI), iii) increase the clearance of productive infected cells (cytotoxic model, C) or iv) increase the clearance of the virus (virus-killing model, V). In all models, the effects of F followed an E_{max} relationship with ϕ the maximal effect of F and θ the sensitivity parameter. Table II displays the four model equations.

Simulations and parameter estimation

For each setting, the simulation procedure and parameter estimation under both MS and MA are described below.

Parameter values. In the first setting, we aimed to evaluate the impact of fixing the two poorly identifiable parameters, k and V_0, in the target cell limited model. For that purpose we defined a set of M=9 candidate models with values for V_0 and k equal to $V_0 = 10^{-5}$; 10^{-4} or 10^{-3} copies.mL^{-1} and k = 1; 4 or 20 d^{-1}. The other parameter values are given in Table III. Figure 1A shows that viral load is biphasic with a peak close to 10 $\log_{10}$ copies.mL^{-1} in all 9 scenarios, but the time to peak depends on k and V_0 to a lesser extent.

In the second setting V_0 and k were fixed to 10^{-4} copies mL^{-1} and 4 d^{-1}, respectively. To ensure a fair comparison between the models, parameters were chosen to predict a similar contribution of the immune response to viral control, as measured by the area under the curve of the $\log_{10}$ viremia from 0 to 20 days (AUC). Thus, in all four models, the parameter values were such that AUC=100 $\log_{10}$ copies.days.mL^{-1} while assuming $\phi^* = 0$ would lead to AUC 120 $\log_{10}$ copies.days.mL^{-1} (i.e., the absence of an immune system would lead to a 20% increase in AUC) (Figure 1B). The values of the TCL model were chosen to lead to a similar AUC=100 $\log_{10}$ copies.days.mL^{-1}.

We assumed that R_0, δ, π, θ and ϕ were the estimated fixed effects. Those parameters, with the exception of θ, where associated to an intermediate between-subject variability, ω, equal to 0.3. Other parameters were assumed to be known with values given in Table III.

Data simulation. For each model we simulated S=300 datasets of N=30 individuals using the same population parameter values, Ψ_m^*, given in Table III. Therefore $S \times M = 2700$ datasets were simulated in the setting I, and 1500 datasets were simulated in the setting II. We assumed measurements were made at days 3, 6, 9, 12, 15 and 18, with a lower limit of

quantification (LLOQ) of 1 $\log_{10}$ copies.mL^{-1} (20), and a measurement error term, σ, equal to 0.7 $\log_{10}$ copies.mL^{-1}.

Parameter estimation. Each of the *s*=1,…, S dataset, was fitted using the M candidate models of each setting. The set of parameter estimates obtained on the dataset *s* using the model *m*, namely R_0, δ, π, θ and ϕ and their corresponding between-subjects variabilities if specified, was noted without loss of generality $\widehat{\Psi}_m^s$. Parameter estimates were obtained by maximization of the likelihood using the SAEM algorithm implemented in the MONOLIX software (version 2018, release 2). We used k_1=800 and k_2=200 iterations for the exploratory and smoothing phases, respectively. We used the asymptotic approximation to derive the probability density function of $\widehat{\Psi}_m^s$, noted $p(\widehat{\Psi}_m^s)$, assumed to be Gaussian with a variance-covariance matrix given by the inverse of the Fisher Information Matrix (FIM^{-1}). The FIM was computed by stochastic approximation with at least 100 and up to 800 iterations. Of note, among the M models used to fit the data, only one is the true model (noted TM), i.e., the model used to generate the data, and we note $\widehat{\Psi}_{TM}^s$ the parameter estimates obtained by fitting the dataset *s* with TM.

For MA, 95% confidence intervals of $\widehat{\Psi}_{MA}^s$ was then calculated by sampling 10,000 values in the mixture distribution $p\left(\widehat{\Psi}_{MA}^s\right) = \sum_{m=1}^{M} w_m^s \, p\left(\widehat{\Psi}_m^s\right)$ and computing the associated 2.5th and 97.5th percentiles (14,18,34).

Performances of model averaging and model selection for estimation

Model selection. For each scenario, we reported the distribution weight of each candidate model as well as the proportion of simulations where each candidate model was selected (based on AIC, see above).

Parameter estimates and comparison with true parameter value. For each scenario, we reported the coverage rate obtained for each parameter with estimator based on MS, MA or TM, defined as the proportion of simulated datasets for which the true value of the parameter was contained in the 95% confidence interval of the estimated parameter. The coverage rates were compared with the prediction interval of a Binomial distribution with p=0.95 and S=300, i.e., [0.923; 0.973] and were reported for parameters R_0, π and δ in setting I and R_0 and δ in setting II.

Performances of model averaging and model selection for prediction

Finally, we aimed to evaluate MA in the context of prediction, i.e., the capability to anticipate the effect of a change in the experimental setting. We focused on the prediction of the impact of an antiviral treatment limiting the viral production π with efficacy ε ($0<\varepsilon<1$) on the predicted proportion of patients with undetectable viral load (10 copies mL^{-1}) at a given time point. We assumed that treatment was initiated at time t=6 and lasted until t=20 days, which coincides with the end of the follow-up. We considered 3 levels of efficacy on decreasing viral production with a factor $1-\varepsilon$, namely ε = 0.90, 0.95 and 0.99, and we focused on the prediction at t=20. For each model and each value of ε, Monte-Carlo simulations were used to the expected proportion of patients below the limit of detection noted $P_m^{*,\varepsilon}(\%) = P[V_m(t=20,\Psi_m^*,\varepsilon)<10]$.

Following what has been done above, one can calculate, for each simulated dataset, the estimate that would be given by model selection, given by $P_{MS}^{s,\varepsilon} = P\left[V_{MS}\left(t=20,\widehat{\Psi}_{MS}^{s},\varepsilon\right)<10\right]$ or by model averaging $P_{MA}^{s,\varepsilon} = \sum_{m=1}^{M} w_m^s \; P\left[V_m\left(t=20,\widehat{\Psi}_m^{s},\varepsilon\right)<10\right]$. Likewise for the sake of comparison, one can also calculate the probability obtained by fitting the data under

the true model, $P_{TM}^{s,\varepsilon} = P\left[V_m\left(t = 20, \hat{\Psi}_{TM}^{s}, \varepsilon\right) < 10\right]$. These values were summarized by calculating the bias and root mean square error (RMSE), given by $\frac{1}{S}\sum(P_{MA}^{s,\varepsilon} - P_m^{*,\varepsilon})$ and $\sqrt{\frac{1}{S}\sum\left(P_{MA}^{s,\varepsilon} - P_m^{*,\varepsilon}\right)^2}$ in the case of MA (similar applies to calculate the bias and RMSE in the case of MS or TM). For the sake of graphical representation, proportions of patients below the limit of detection were presented as percentages and biases and RMSE were therefore expressed in percentages.

Results

Setting I

The first setting focused on the comparison between model averaging and model selection when parameters of the model (e.g., the eclipse phase, *k*, and the initial inoculum V_0) cannot be identified and are fixed to arbitrary values (see Table III).

Overall, the true set of parameter values was selected up to 62% of the simulations. The two parameters did not have the same rate of selection, with the correct values for *k* and V_0 being selected up to 71% and 96%, respectively (Figure 2A). Although the true model was not systematically associated with the lowest AIC, it was associated in all scenarios with the largest weight among the candidate models with a median value comprised between 0.32 and 0.55 (Figure 2B). In all cases considered, at least two models had a weight greater than 0.20.

We next evaluated the impact of these results on parameter estimates and coverage rates. The estimation of R_0 using model selection was associated with a poor coverage rate between 0.46 and 0.63 (Figure 3). Results for the loss rate of infected cells, δ, were better with a coverage rate ranging from 0.53 and 0.94, and was comprised in the nominal 0.95 coverage

rate in 3 out of 9 scenarios. For the viral production π, MS showed coverage rates ranging from 0.68 and 0.96, and was comprised in the nominal 0.95 coverage rate in 5 out of 9 scenarios. Model averaging largely improved the coverage rates for all parameters and gave results close to those obtained with the true model. The coverage rates was between 0.91 and 0.98 for R_0, between 0.72 and 0.95 for δ and between 0.78 and 0.98 for π. Further the coverage rates were comprised in the nominal 0.95 coverage rate in 7 out 9 scenarios for R_0 and 5 out of 9 scenarios for δ and π. All confidence intervals can be found in supplemental figures S1, S2 and S3.

Lastly we explored the effect of simulating with less (ω=0.1) or more (ω=1) inter-subjects variability. In both cases, MS provided subnominal coverage rates but MA corrected them (see Supplemental Figure S4). Eventually, we observed poorer coverage rates with ω=0.1 and improved with ω=1 compared to ω=0.3.

Setting II

In the second setting, we assessed the properties of parameter estimates when several biological models can be proposed. We focused on models characterizing the effect of the immune response, considering that the immune response compartment could alternatively make cells refractory to infection, limit the production of virus, increase the elimination of infected cells or increase the elimination of free virions (see Table III).

Unlike what was found in the previous setting, the chance of selecting the true model was largely dependent on the model considered. In fact these chances were equal to 97% for the refractory model but this percentage could decrease to 58% with the cytotoxic model (Figure 4A). Conversely, the models were also associated with a large rate of false selection with rates

ranging from 3 to 10% for the refractory model, and up to 20% for the production inhibition model. In the case of the target cell limited model, the chances of correctly selecting it were equal to 88% and the rate of false selection were ranging from 1 to 19%. The median weight associated to the true model ranged from 0.43 to 0.99 (Figure 4B).

Accordingly, MS provided satisfactory coverage rates for target cell limited and refractory model (Figure 5); however it failed to achieve the nominal coverage rate in all other models, with values ranging from 0.62 to 0.91 for R_0 and from 0.50 to 0.89 for δ. This could be improved by taking into account model uncertainty and using model averaging. Indeed the coverage rates ranged from 0.86 to 0.99 for both parameters in all models considered. In fact, MA had even better performances than the true model in some cases, which achieved subnominal coverage rate in 3 of the 5 considered scenarios (Figure 5). All confidence intervals can be found in supplemental figures S5 and S6.

Finally, we compared the predictive performances of MS, MA and the true model. For that purpose, we predicted the effect of a putative antiviral treatment on the proportion of patients having undetectable viremia at end of follow up (day 20). Here as well the performances obtained using model selection and model averaging were compared. In all cases, the percentage of undetectable viral loads at end of treatment was accurately predicted for both MS and MA, with no more than 4% of bias in all cases considered (Figure 6). In term of precision of estimation, the results were also largely similar in most scenarios, with RMSE ranging from 0.4 to 30.5% in all cases. In one case, namely ε=0.95, we found that MA outperformed the results obtained by MS. Here as well, the results obtained by model averaging were largely comparable with those obtained with the true model.

Discussion

The objective of this study was to compare the estimation and the predictive performances of model selection and model averaging in the context of viral dynamic models. We explored two frequent issues encountered when developing viral dynamic models with uncertainty related either to (I) unidentifiable parameters or (II) the presence of several candidate biological models. In the two settings MS provided poor coverage rates of typical parameters. This stems from the fact that MS neglects model uncertainty and focuses on one single "best model", leading to overconfidence in the parameter estimates. This can be corrected under certain conditions using MA, which provided better coverage rates and achieved the nominal coverage rate in most scenarios studied. MA can also be relevant to predict the effect of intervention, such as the percentage of patients that would achieve undetectable viral loads during treatment. Thus extending results found in other contexts, in particular dose finding studies (25,35).

By offering a simple framework to take into account model uncertainty, MA accounts for the fact that in many situations several biological models are plausible. Our study shows the limitations of reporting only the best model. For instance, in the case of the target cell limited model, we found that the chance to conclude wrongly to an immune response controlling the infection was equal to 11%. In the case of the refractory model, which has been proposed as a driving force in several acute infection (33,36), our results were more reassuring, with a rate of false rejection of only 3%. This risk was larger with other models integrating an immune response, with rates of false rejection greater than 60% in some cases. By weighing the predictions of alternative models, MA avoids the caveat of MS. As advocated in other contexts (37,38), MA can be used to more transparently discuss model uncertainty and to stimulate new data acquisition (13).

Although MA offers a simple alternative to MS, it also presents the defects of its virtue. As MA weighs the models according to their information criterion, using MA is relevant only if one model does not largely outperform the others. The weight value leading to "outperformance" is arbitrary, and depends also on the number of candidate models. Accordingly, one may question the need to use weights when making predictions. As suggested by a reviewer, we conducted a simulation where all models having a weight greater than a given threshold (0.1 or 0.2) were considered as equally likely in the prediction, and this approach provided results close to those obtained with MA (Supplemental Figure S7).

MA still requires to make important assumptions that need to be kept in mind. First we used the asymptotic Gaussian approximations to calculate the standard errors. This assumption may not hold for all models, depending on their complexity and data paucity, as can been seen in some cases of Figures 3 & 5. Other approaches have been proposed in the context of NLMEM to calculate the standard error more precisely, such as bootstrapping (38), sampling importance resampling (39) or Hamiltonian Monte-Carlo methods (HMC) (40). Future work will be needed to evaluate in which contexts these methods, which are computationally demanding, are beneficial. Second in our simulations, we assumed that there was a true model and that it was part of the candidate models. Although there is no "true model" in real data, we made this hypothesis to stress that MA should be performed only with biologically relevant models. Likewise, MA should not be used to "blindly" average predictions of any models and modelers should, prior to the analysis, develop other models at hand and, if possible, discuss and perform new experiments to discriminate between them (13). In that perspective, using MA to calculate CI is meaningful only if parameter have the same interpretation across the candidate models. This is the case for half-life or viral production rates but is less evident for derived parameters such the basic reproductive number R_0 (39,40). Finally, MA does not substitute to a proper analysis of parameter identifiability. In fact the

differences between MA and MS may simply reveal a poor practical identifiability, i.e., the fact that data available are not sufficient to precisely estimate parameters (8) and/or that the biological question is wrongly formulated (13). This is also what we observed here, with the wrong selection of models being in part due to the fact that the models had a poor practical identifiability, at least for some parameters. In order to be performant, MA requires that only a limited number of models are tested. It is only when a reasonable number of models remain that MA can be relevant, as an alternative to Bayesian approaches, that may be tedious in particular a non-linear mixed effect framework.

Acknowledgments

Antonio Gonçalves was funded by a grant from Roche Pharmaceutical Research and Early Development. The authors declare that there is no conflict of interest regarding the publication of this article. The authors also would like to acknowledge Hervé Le Nagard and Lionel de la Tribouille for the use of CATIBioMed calculus facility.

Supplemental information

A readily usable R code to compute weights and confidence intervals in MA is provided in supplementary material. The example is based on digitized Zika data of Best at al. *PNAS* 2017 and illustrates model averaging in the context of poorly identifiable parameters (setting I).

References

1. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature. 1995 Jan 12;373(6510):123–6.

2. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, et al. Viral dynamics in human immunodeficiency virus type 1 infection. Nature. 1995 Jan 12;373(6510):117–22.

3. Perelson AS, Ribeiro RM. Introduction to modeling viral infections and immunity. Immunological Reviews. 2018;285(1):5–8.

4. Perelson AS. Modelling viral and immune system dynamics. Nat Rev Immunol. 2002 Jan;2(1):28–36.

5. Best K, Perelson AS. Mathematical modeling of within-host Zika virus dynamics. Immunological Reviews. 2018;285(1):81–96.

6. Ciupe SM. Modeling the dynamics of hepatitis B infection, immunity, and drug therapy. Immunological Reviews. 2018;285(1):38–54.

7. Lavielle M, Mentré F. Estimation of population pharmacokinetic parameters of saquinavir in HIV patients with the MONOLIX software. J Pharmacokinet Pharmacodyn. 2007 Apr;34(2):229–49.

8. Guedj J, Thiébaut R, Commenges D. Practical identifiability of HIV dynamics models. Bulletin of Mathematical Biology. 2007 Oct 25;69(8):2493–513.

9. Snoeck E, Chanu P, Lavielle M, Jacqmin P, Jonsson EN, Jorga K, et al. A comprehensive Hepatitis C viral kinetic model explaining cure. Clinical Pharmacology & Therapeutics. 2010 Jun;87(6):706–13.

10. Nguyen T, Guedj J. HCV kinetic models and their implications in drug development: HCV kinetic models and their implications. CPT: Pharmacometrics & Systems Pharmacology. 2015 Apr;4(4):231–42.

11. Handel A, Longini IM, Antia R. Towards a quantitative understanding of the within-host dynamics of influenza A infections. Journal of The Royal Society Interface. 2010 Jan 6;7(42):35–47.

12. Smith AM, Adler FR, Ribeiro RM, Gutenkunst RN, McAuley JL, McCullers JA, et al. Kinetics of coinfection with influenza A virus and streptococcus pneumoniae. Grenfell BT, editor. PLoS Pathogens. 2013 Mar 21;9(3):e1003238.

13. Ganusov VV. Strong inference in mathematical modeling: a method for robust science in the twenty-first century. Frontiers in Microbiology [Internet]. 2016 Jul 22 [cited 2019 Apr 4];7. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01131/abstract

14. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics. 1997 Jun;53(2):603.

15. Boulesteix A-L. Ten simple rules for reducing overoptimistic reporting in methodological computational research. Lewitter F, editor. PLOS Computational Biology. 2015 Apr 23;11(4):e1004191.

16. Kirk PDW, Babtie AC, Stumpf MPH. Systems biology (un)certainties. Science. 2015 Oct 23;350(6259):386–8.

17. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2. ed., [4. printing]. New York, NY: Springer; 2010. 488 p.

18. Claeskens G, Hjort NL. Model selection and model averaging [Internet]. Cambridge: Cambridge University Press; 2008 [cited 2019 Sep 30]. Available from: http://ebooks.cambridge.org/ref/id/CBO9780511790485

19. Best K, Guedj J, Madelain V, de Lamballerie X, Lim S-Y, Osuna CE, et al. Zika plasma viral dynamics in nonhuman primates provides insights into early infection and antiviral strategies. Proceedings of the National Academy of Sciences. 2017 Aug 15;114(33):8847–52.

20. Madelain V, Baize S, Jacquot F, Reynard S, Fizet A, Barron S, et al. Ebola viral dynamics in nonhuman primates provides insights into virus immuno-pathogenesis and antiviral strategies. Nature Communications. 2018 Dec;9(1).

21. Bertrand J, Comets E, Mentré F. Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. Journal of Biopharmaceutical Statistics. 2008 Nov 7;18(6):1084–102.

22. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987 Sep;52(3):345–70.

23. Anderson DR, Burnham KP. Understanding information criteria for selection among capture-recapture or ring recovery models. Bird Study. 1999 Jan;46(sup1):S14–21.

24. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications: The Bayesian information criterion. WIREs Comp Stat. 2012 Mar;4(2):199–203.

25. Buatois S, Ueckert S, Frey N, Retout S, Mentré F. Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. AAPS J. 2018 29;20(3):56.

26. Aoki Y, Röshammar D, Hamrén B, Hooker AC. Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. J Pharmacokinet Pharmacodyn. 2017 Dec;44(6):581–97.

27. Kakizoe Y, Nakaoka S, Beauchemin CAA, Morita S, Mori H, Igarashi T, et al. A method to determine the duration of the eclipse phase for in vitro infection with a highly pathogenic SHIV strain. Sci Rep. 2015 Sep;5(1):10371.

28. Xia X, Moog CH. Identifiability of nonlinear systems with application to HIV/AIDS models. IEEE Transactions on Automatic Control. 2003 Feb;48(2):330–6.

29. Wu H, Zhu H, Miao H, Perelson AS. Parameter identifiability and estimation of HIV/AIDS dynamic models. Bulletin of Mathematical Biology. 2008 Apr;70(3):785–99.

30. Miao H, Dykes C, Demeter LM, Cavenaugh J, Park SY, Perelson AS, et al. Modeling and estimation of kinetic parameters and replicative fitness of HIV-1 from flow-cytometry-based growth competition experiments. Bulletin of Mathematical Biology. 2008 Aug;70(6):1749–71.

31. Dumont C, Lestini G, Le Nagard H, Mentré F, Comets E, Nguyen TT, et al. PFIM 4.0, an extended R program for design evaluation and optimization in nonlinear mixed-effect models. Computer Methods and Programs in Biomedicine. 2018 Mar;156:217–29.

32. Baccam P, Beauchemin C, Macken CA, Hayden FG, Perelson AS. Kinetics of influenza A virus infection in humans. J Virol. 2006 Aug;80(15):7590–9.

33. Pawelek KA, Huynh GT, Quinlivan M, Cullinane A, Rong L, Perelson AS. Modeling within-host dynamics of influenza virus infection including immune responses. PLoS Comput Biol [Internet]. 2012 Jun 28 [cited 2019 Jun 26];8(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3386161/

34. Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. Statist Med. 2014 May 10;33(10):1646–61.

35. Schorning K, Bornkamp B, Bretz F, Dette H. Model selection versus model averaging in dose finding studies: K. SCHORNING *ET AL.* Statistics in Medicine. 2016 Sep 30;35(22):4021–40.

36. Saenz RA, Quinlivan M, Elton D, MacRae S, Blunden AS, Mumford JA, et al. Dynamics of influenza virus infection and pathology. J Virol. 2010 Apr;84(8):3974–83.

37. Hoeting JA, Adrian E. Raftery, Madigan D. Bayesian model averaging: a tutorial. Statist Sci. 1999 Nov;14(4):382–417.

38. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2. ed., [4. printing]. New York, NY: Springer; 2010. 488 p.

39. Lloyd AL. The dependence of viral parameter estimates on the assumed viral life cycle: limitations of studies of viral load data. Proceedings: Biological Sciences. 2001;268(1469):847–54.

40. Ribeiro RM, Qin L, Chavez LL, Li D, Self SG, Perelson AS. Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. Journal of Virology. 2010 Jun 15;84(12):6096–102.

List of Tables

Table I: **Expected standard error (SE) of the fixed effect parameters using a target cell limited model when the estimated parameters include or do not include the initial inoculum, V_0, and the eclipse rate, k.** Expected standard errors were calculated using PFIM software and for a study design including 30 subjects sampled every 3 days from day 3 to day 18.

Parameter (units)	Estimation of R_0, δ, V_0, k and π			Estimation restricted to R_0, δ and π		
	Estimate	SE	Relative SE(%)	Estimate	SE	Relative SE(%)
R_0	12	62.0	516%	12	0.84	7.0%
δ (d^{-1})	1	0.10	10%	1	0.063	6.3%
π (copies.cell^{-1}.d^{-1})	6000	3625	604%	6000	1446	24.1%
V_0 (copies.mL^{-1})	10^{-4}	29.7	743%	10^{-4} (fixed)	-	-
k (d^{-1})	4	38.9	971%	4 (fixed)	-	-
c (d^{-1})	20 (fixed)	-	-	20 (fixed)	-	-
T_0 (cells.mL^{-1})	10^8 (fixed)	-	-	10^8 (fixed)	-	-

Table II: Differential equations system of immune response models. At t=0 we have $T_{t=0} = T_0$; $I_{1,t=0} = 0$; $I_{2,t=0} = 0$ and $V_{t=0} = V_0$.

	Target cell limited	**Refractory**	**Production inhibition**	**Cytotoxic**	**Virus-killing**
$\frac{dT}{dt} =$	$-\beta TV$	$-\beta TV - \frac{\phi TF}{F+\theta}$	$-\beta TV$	$-\beta TV$	$-\beta TV$
$\frac{dI_1}{dt} =$	$\beta TV - kI_1$	$\beta TV - kI_1$	$\beta TV - kI_1$	$\beta TV - kI_1$	$\beta TV - kI_1$
$\frac{dI_2}{dt} =$	$kI_1 - \delta I_2$	$kI_1 - \delta I_2$	$kI_1 - \delta I_2$	$kI_1 - \delta I_2 - \frac{\phi I_2 F}{F+\theta}$	$kI_1 - \delta I_2$
$\frac{dV}{dt} =$	$\pi I_2 - cV$	$\pi I_2 - cV$	$\pi\left(1 - \frac{\phi F}{F+\theta}\right) I_2 - cV$	$\pi I_2 - cV$	$\pi I_2 - cV - \frac{\phi FV}{F+\theta}$
$\frac{dF}{dt} =$	$qI_2 - d_F F$	$qI_2 - d_F F$	$qI_2 - d_F F$	$qI_2 - d_F F$	$qI_2 - d_F F$

Table III: Parameter values used for simulations

Parameter (units)	Setting I	Setting II				
	Target cell limited	Target cell limited	Refractory	Production Inhibition	Cytotoxic	Virus-killing
R_0 [a]	12	12				
δ [a] (d^{-1})	1	1				
π [a] (copie.cell^{-1}.d^{-1})	6000	250	6000			
ϕ [a]	-	-	1	0.99	0.9	36.5
θ	-	-	2200	325000	3	0.001
V_0 (copies.mL^{-1})	{10^{-5}; 10^{-4}; 10^{-3}}	10^{-4}				
k (d^{-1})	{1; 4; 20}	4				
c (d^{-1})	20	20				
T_0 (cells.mL^{-1})	10^8	10^8				
q (d^{-1})	1	1				
d_F (d^{-1})	0.4	0.4				
SD of the additive error	0.7	0.7				

[a]: parameters for which inter-individual variability ω=0.3

Legend to Figures

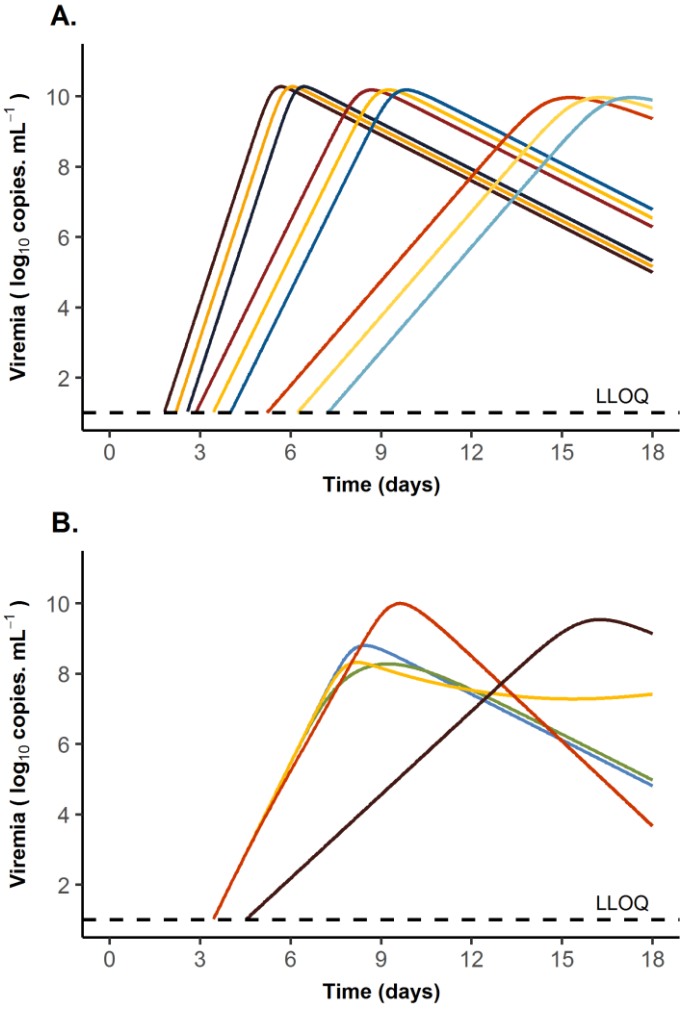
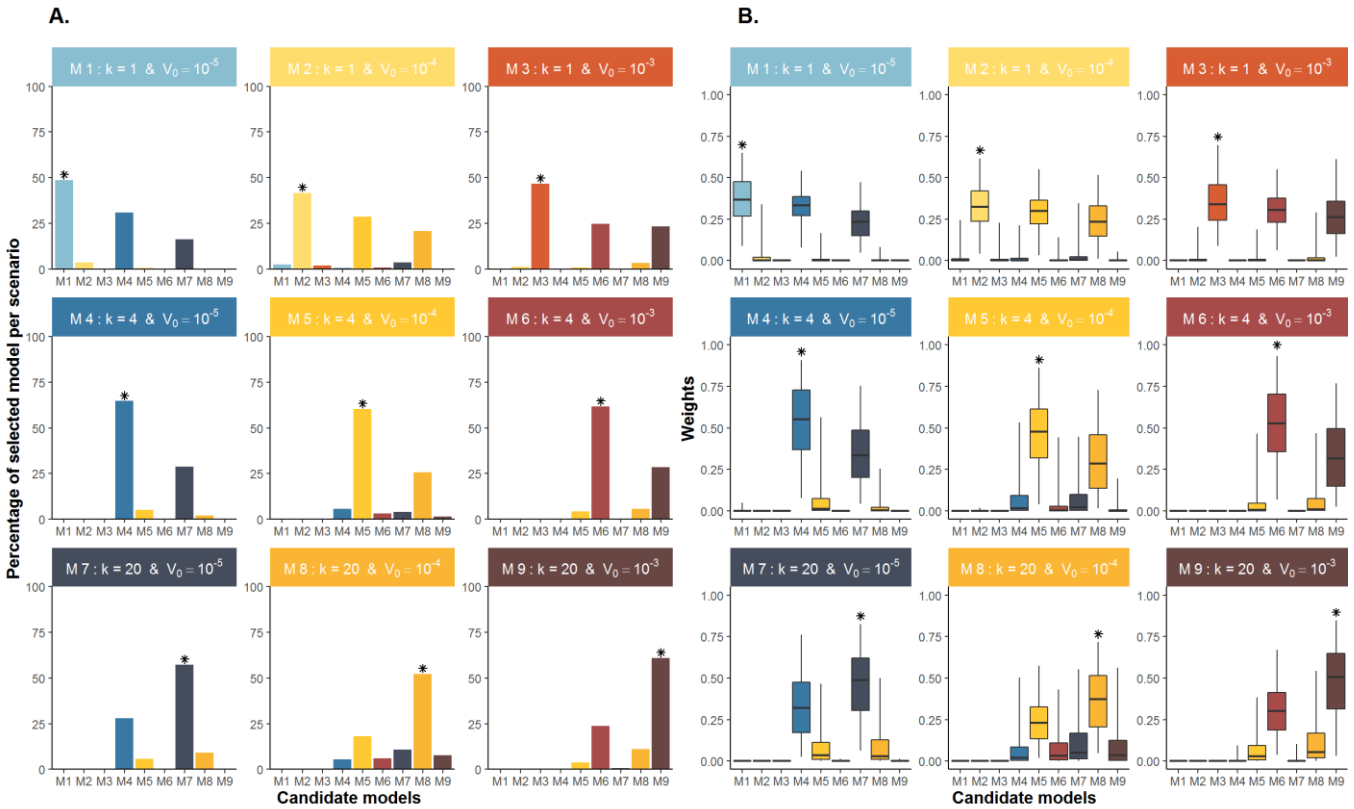


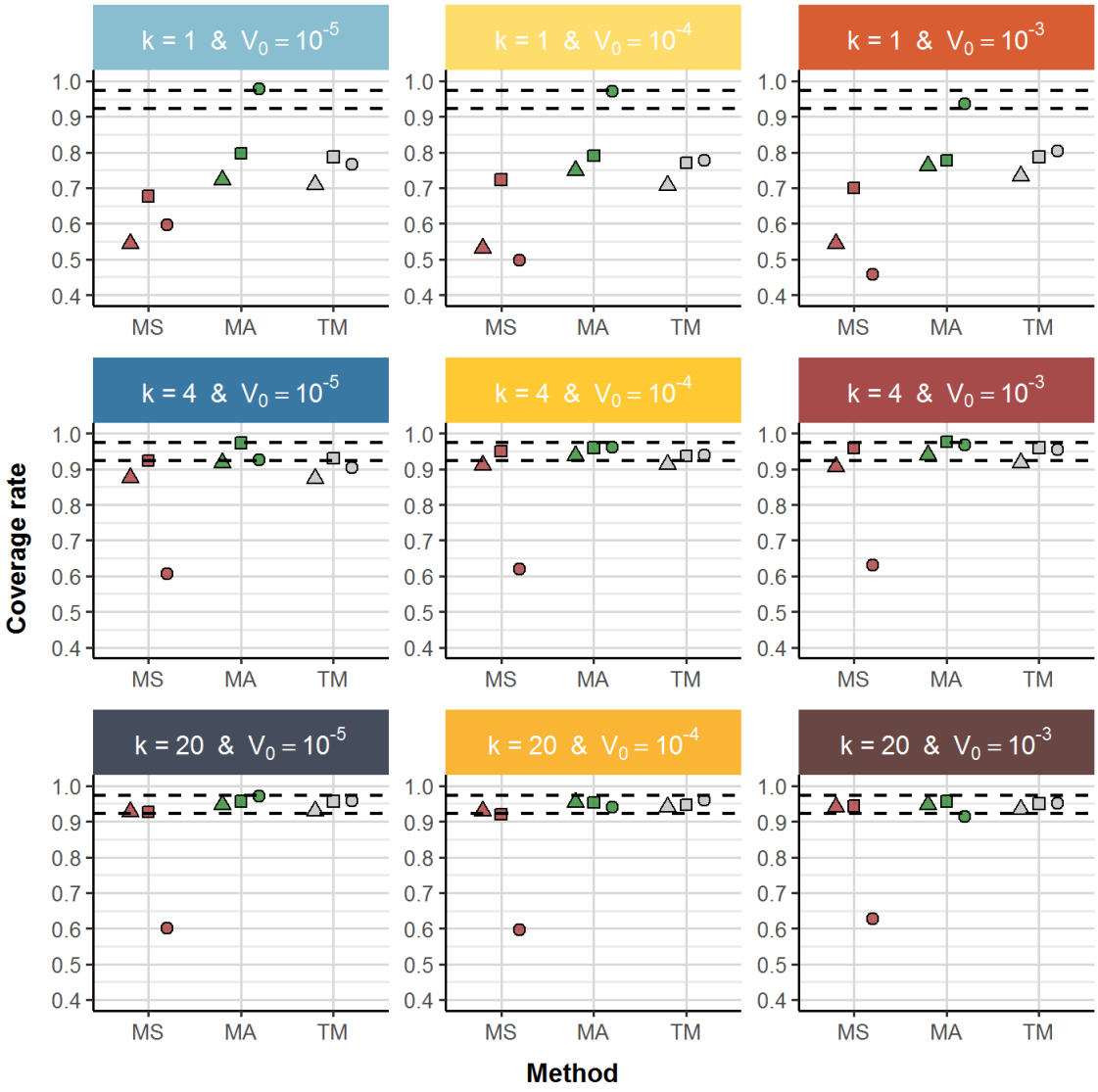
Figure 1. Viral kinetics profiles obtained with the population parameters for each candidate model. (A) and (B) correspond to the simulation settings I and II, respectively. In Panel A, curves are regrouped by 3. At left, the first 3 curves correspond to models with k = 20 d^{-1}; center, models with k = 4 d^{-1} and right, models with k = 1 d^{-1}. Within each group, red curves correspond to models with $V_0 = 10^{-5}$ copies.mL^{-1}, yellow curves to models with $V_0 = 10^{-5}$ copies.mL^{-1} and blue curves to models with $V_0 = 10^{-5}$ copies.mL^{-1}.



503

504 **Figure 2. Setting I.** (A) For each scenario, the percentage of simulations where each
505 candidate model was selected using AIC. Title of the facet indicates the true model. (B)
506 Boxplots of weights (whiskers from the 2.5th to the 97.5th percentile) associated with each
507 candidate model using AIC values. The asterisk denotes the true model in each scenario.

508



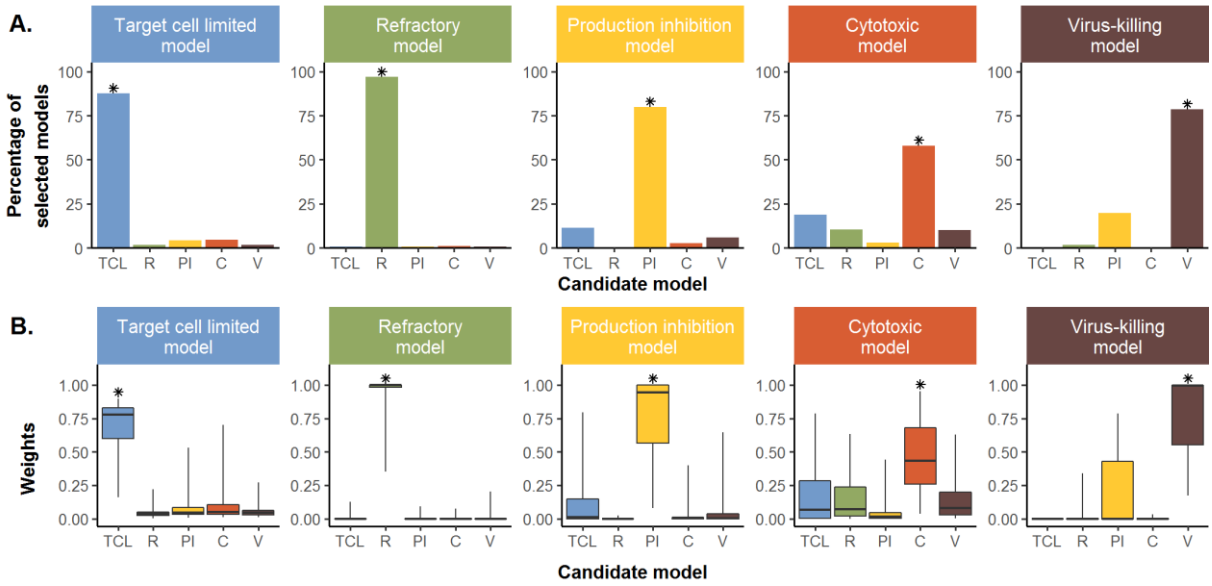
509

510 **<u>Figure 3.</u> Coverage rate of R_0 and δ in setting I.** Coverage rate of the parameters R_0 (dots),

511 π (squares) and δ (triangles) for each scenario using model selection, model averaging or the

512 true model. Dashed lines represents the prediction interval around 0.95.

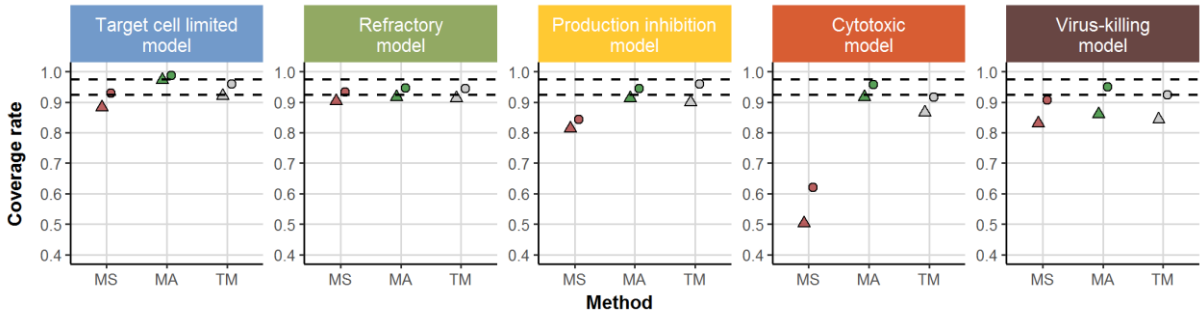
513



514

515 **Figure 4. Setting II.** (A) For each scenario, the percentage of simulations where each
516 candidate model was selected using AIC. Title of the facet indicates the true model. (B)
517 Boxplots of weights (whiskers from the 2.5th to the 97.5th percentile) associated with each
518 candidate model using AIC values. The asterisk denotes the true model in each scenario.

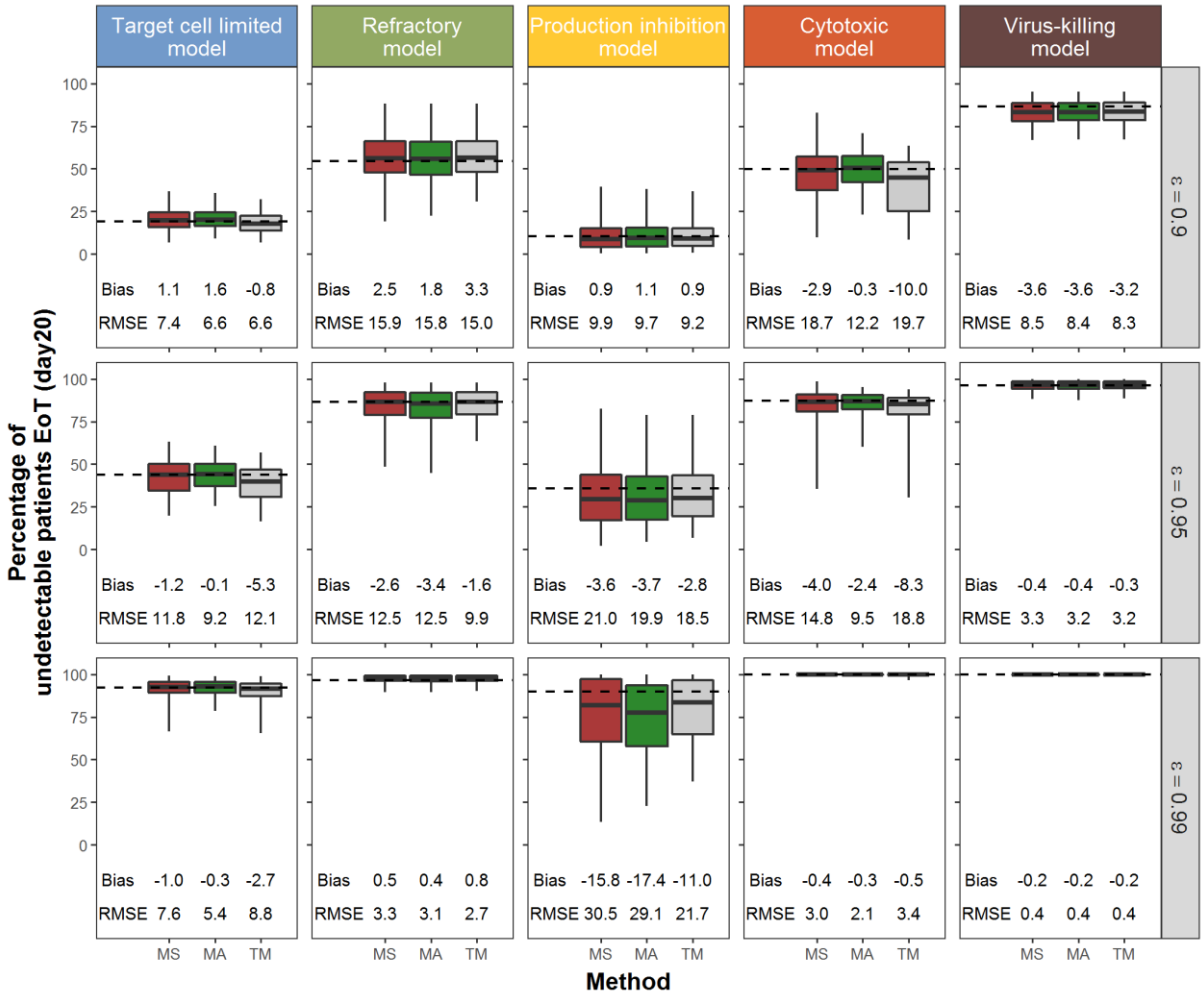
519



520

521 **Figure 5. Coverage rate of R_0 and δ in setting II.** Coverage rate of the parameters R_0 (dots)
522 and δ (triangles) for each scenario using model selection, model averaging or the true model.
523 Dashed lines represents the prediction interval around 0.95.

524



<u>Figure 6.</u> Distribution of the expected proportion of patients below the limit of detection at day 20 using model selection (red), model averaging (green) or the true model (grey). For each scenario, whiskers represent the 2.5th to the 97.5th percentile and each row corresponds to a different value of the treatment effect, noted ε.