

Synthetic Model Combination: A new machine learning method for pharmacometric model ensembling

Authors:

Alexander Chan: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK.

Richard Peck: Pharma Research and Development (pRED), Roche Innovation Center, Basel, Switzerland; and Department of Pharmacology & Therapeutics, University of Liverpool, Liverpool, UK.

Megan Gibbs: Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Gaithersburg, Maryland, USA.

Mihaela van der Schaar: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK.

Corresponding Author:

Alexander Chan; Centre for Mathematical Sciences, Wilberforce Rd, Cambridge, CB3 0WA; +44 (0) 1223 760425; alexjchan@maths.cam.ac.uk.

Keywords:

Precision medicine, Individualization, Model evaluation

Conflict of Interest Statement:

No competing interests declared for this work.

Funding Information:

No funding was received for this work.

Abstract

When aiming to make predictions over targets in the pharmacological setting, a data-focussed approach aims to learn models based on a collection of labelled examples. Unfortunately, data sharing is not always possible, and this can result in many different models trained on disparate populations, leading to the natural question of how best to use and combine them when making a new prediction. Previous work has focused on global model selection or ensembling, with the result of a single final model across the feature space. Machine learning models perform notoriously poorly on data outside their training domain however due to a problem known as covariate shift, and so we argue that when ensembling models the weightings for individual instances must reflect their respective domains - in other words models that are more likely to have seen information on that instance should have more attention paid to them. We introduce a method for such an instance-wise ensembling of models called Synthetic Model Combination (SMC), including a novel representation learning step for handling sparse high-dimensional domains. We demonstrate the use of SMC on an example with dosing predictions for Vancomycin, although emphasise the applicability of the method to any scenario involving the use of multiple models.

Introduction

The ability of a model to correctly represent a population is of necessity limited by how well the data used to build it represents the population. Given the enormous variability within

and between human populations for pharmacokinetics and pharmacodynamics, this is a significant challenge for the use of models and their ability to make useful predictions about effects in humans. This problem is further increased by the variability within diseases which are generally not homogenous entities where all patients have exactly the same disease processes.

When attempting to apply one or more previously developed models to make predictions about new patients or populations it is a very challenging task for us to know how well the individual models should perform, making the task of choosing the most appropriate model (or ensemble) difficult. This is compounded by the problem that the provided models could perform poorly for two main reasons: Firstly, the model itself may not have been flexible enough to properly capture the underlying true function present in the data; and secondly, in the area that they are making a prediction there may not have been sufficient training data used for the model to have been able to learn appropriately - i.e. the model is extrapolating (potentially unreasonably) to cover a new feature point (Figure 1).

Various methods have been developed to address this problem. At the simplest level there are typically multiple models developed each using different datasets that are claimed to better represent the population or at least a specific subset of the population. However, the reference populations used to build each model are usually only small subsets of the whole. Methods such as Bayesian model averaging are used to combine multiple models to try to capture the value of each - however they usually assume the model's performance is independent of the populations, ignoring their training domain. Data and model repositories have been proposed to allow the development of more definitive models of diseases and of

the PKPD of therapies [1]. There has been some success developing publicly accessible repositories of disease models that make the data and scope of the models more transparent, for example the DDMoRe model repository [2] and several academic institutions, charities and pharma companies make some trials data available through the non-profit organisation Vivli [3]. However, the extent of the data accessible is limited by some participants and phase 1 studies or PK data are excluded by some pharma participants. Sharing data brings its own challenges, not the least being that there may not be permission to use the data for anything other than the specific purpose for which it was first developed.

Synthetic Model Combination (SMC) is a new machine learning method that leverages and combines multiple models in an effective manner. Unlike existing methods, it focuses on building ensembles in an instance-wise manner before any additional data has been collected, that is to say that for each new test point over which a prediction needs to be made it constructs a new ensemble. This effectively means that SMC is able to select models for each test case (patient) that it thinks will be most effective for the given case, based primarily on whether the case is likely to have been well represented in the training domain of each model. In this paper we introduce and describe SMC, illustrate its use through an example of multiple pharmacometric models and try to stimulate ideas about other possible applications relevant to pharmacometrics, clinical pharmacology and drug development and use.

Background and Related Work

Synthetic Model Combination can be used in any situation where there are multiple alternative models. For the purposes of illustrating the method we chose to consider the common situation of multiple population PKPD models. In the case of the antibiotic vancomycin this is clinically important as population PKPD models are commonly used to guide selection of doses to achieve a target AUC and maximise the chances of effective therapy without nephrotoxicity and it is important to identify how best to use the multiple possible models [4]. We emphasise that this specific example is an illustration for the purposes of explaining the methodology. We anticipate there are many other situations, including other drugs and other cases where multiple models are available for which SMC may be even more useful.

With multiple models our goal is to combine them in the most appropriate way, taking the form of constructing ensembles. Here we differentiate between what we firstly describe as Naive Ensembling (NE), where multiple models have been trained (through bootstrapping or on different datasets) in order to reduce the expected bias or get an estimate of distributional uncertainty. Performance-based Model Averaging (PBMA) on the other hand works by selecting models with higher weights based on an estimate of the performance of the model; a practical and common approach being Bayesian Model Averaging (BMA) [5] Given an appropriate (usually uniform) prior, we calculate the posterior probability that a given model is the optimal one - and once this is obtained the models can be marginalised out during test time predictions, creating an ensemble weighted by each model's posterior probability. The posterior being intractable, the probability is approximated using the Bayesian Information Criterion (BIC) [6] - which requires a likelihood estimate over some validation set and is estimated as:

$$p(\mathcal{M}_i|\mathcal{D}) = \exp\left(-\frac{1}{2}\text{BIC}(\mathcal{M}_i)\right) / \sum_{i=1}^N \exp\left(-\frac{1}{2}\text{BIC}(\mathcal{M}_i)\right)$$

With this it is important to note the subtle difference in setup to the problem we are trying to work with. In all cases, it is assumed that there is some ordering for the models that holds across the feature space and so a global ensemble is produced with a fixed weighting \hat{w} such that $w(x) = \hat{w} \forall x \in \mathcal{X}$. This causes failure cases when there is variation in the models across the feature space, since it is a key point that BMA is not a model combination method [7]. This being an important distinction and one of the main reasons BMA has been shown to perform badly under covariate shifted tasks [8] - that is to say tasks where the testing distribution differs from the training distribution, a scenario that is well known to affect the quality of a model's predictions [9]. That being said, it can be extended by considering the set of models being averaged to be every possible combination of the provided models [10], although this becomes even more computationally infeasible.

This has led to a family of ensemble methods that calculate their weights slightly differently, replacing the $\text{BIC}(\mathcal{M}_i)$ term in the above equation with other measures of the "quality" of the model, such as AIC, log likelihood, or negative sum of squared errors - all of which are explored in the work of Uster et al. [11]. This appears to be the extent of the exploration in model averaging in the clinical pharmacological setting. We have summarised the properties of these methods in Table 1.

Traditionally PBMA models would not be used for instance-wise predictions, since in a typical supervised learning setting each patient would only have a single set of covariates

and outcome associated with them and so it would not be possible not be able to get an estimate of the performance for a given individual that would be different from the population as a whole. This reflects the central problem with global ensembles that run on the assumption that the measure of 'goodness' of each individual model holds the same across the feature space. That is to say that each model will be just as effective at predicting for old diabetics as it would for infants, however this is unlikely as models are often trained in different subpopulations and it is expected that they might all react differently to a drug. Global ensembles implicitly assume that this covariate shift is not the case and as such suffer when it is - they apply the same ensemble of models to every single new test point regardless of what data the model was trained on.

However, in PopPK settings, we sometimes wish to predict a patient's AUC having already observed one or multiple observations - in this case we would be able to use these few observations to get a performance estimate and thus weight models via PBMA. Despite this, we will likely still only have at most single digit observations for a patient and so there is a risk that we don't have enough signal to fit appropriate models and may potentially overfit. We still may want to make predictions without any observations, and as noted in [11], this is an area that PBMA does not handle and simply reduces to a naive ensemble.

Methods

Synthetic Model Combination

Unlike the previously mentioned methods, even without any observations at all, SMC does not search for a global ensemble - rather it asks the question; for a given individual x_i , what do we think is the best ensemble? This could naturally vary quite considerably from individual to individual, especially if the models were trained on data from relatively disparate populations

We explain at a high level the method here but refer the interested reader to [12] for more in depth detail. The method can be broken down into essentially three main steps.

The first step in SMC is to use the demographic information reported alongside published models to produce a density estimate such that we can sample from each model's effective support.

This aims to create a crude estimate of the region on which the model was trained on in the original feature space - for example, the general distribution of heights, weights, sex etc..

We will often expect this feature space to be high dimensional and the information available to be not hugely detailed, limiting our ability to use this original space to make meaningful predictions.

Given the flexibility in the form of what we allow the information to take, SMC must remain relatively agnostic to this step. A common example of the type of information we expect will simply be example feature samples, and in this case a simple kernel density estimate [13] or other density estimation method could be employed. On the other hand, when models are published, authors will often also provide demographic information on the patients that

were involved in the study, such as the mean and variance of each covariate recorded. In this case we may simply want to approximate the density using a Gaussian and moment-matching for example.

The second step is to learn a representation space for the individual features which will be maximally informative for considering which models will be effective at making accurate predictions on an individual. The principal aim is to lower the relevant dimensionality of the data such that density modelling is effective in the learnt space - but this can also be effective in bending and compressing the space such that regions of model training data are moved closer together or further apart based on whether they produce useful and transferable models. In cases when the dimension of covariates is already low, this step is not always

Learning the representation takes the form of a Variational / Regular Autoencoder [14] with additional auxiliary losses. This is trained on the features of the testing set as well as samples from the densities for individual models that were generated in the first step. Choosing the latent dimension to be low results in learning a representation space that compresses the useful information in the features and aims to move training regions that are transferable closer together.

In the final step, we remodel the original densities in the feature space now in the representation space, so that we can calculate ensemble weights for individuals based on their density under each model. Given model densities in the feature space $p_j^x(x)$, we construct a corresponding density in the representation space $p_j^z(z)$ - this can be achieved

simply by sampling from $p_j^x(z)$, passing through the encoder f_θ and modelling the new density with a kernel density estimate.

From here, we calculate weights as the relative density a feature representation has under the densities in the new space:

$$w_i(x) = \frac{p_i^z(f_\theta(x)) + \gamma}{\sum_{j=1}^N p_j^z(f_\theta(x)) + \gamma}$$

with a regularisation hyperparameter γ chosen to be very small such that an outlier's weights are not dominated by the closest model.

This step simply weights models by an individual's density in the new representation space, meaning that models that are more likely to have seen features similar to the individual (or ones with transferable features) will play a bigger part in the ultimate prediction for the given individual.

The quantity can be used to inform the confidence of any prediction made by SMC.

Particularly low values will indicate that the feature had low density under all the domains and as such it may be likely that none of the models were accurate. We note as well that assuming a hierarchical generative model for the test data where one of the models training data distributions is selected and then sampled from - this can be interpreted as the posterior probability that a test instance was sampled from a model's domain and is thus well represented by it.

Interacting with PBMA

The key thing to note is that unlike PBMA methods, SMC does not need any sort of observation in order to individualise the weights of the ensembles. This means that it can be applied in areas that PBMA cannot, such as making direct a priori predictions about a patient's AUC before any observations have been observed. However, when PBMA can be applied, since both it and SMC fundamentally use different signals to generate their weights, they can reasonably be combined at the same time in order to achieve the benefits of both models - i.e., the weights from both methods can be calculated individually and then combined in order to produce a final weighting. Interestingly, in the case of BMA we can potentially see this as a case of SMC learning an appropriate prior distribution for the weights that is then updated based on the performances of the models on the new observations, allowing for a natural integration into the current framework that allows for the best of both methods.

Results

A Case Study in Vancomycin

For vancomycin, the latest consensus dosing guidelines from the Infectious Diseases Society of America [15] recommend adjusting dose to achieve a target area under the curve (AUC). Many population PK models have been developed to adjust the doses on an individual patient basis taking into account important patient covariates.

We base our experiment around those of [11] who themselves consider a model averaging approach through the application of model averaging. We use simulated patients provided by the authors to evaluate the effectiveness of SMC in the accuracy of predicting the AUC across a number of settings when a number $\in \{0 \text{ (A priori)}, 1, 2, 3\}$ of concentration measurements are taken in a 48-hour period. Ultimately, we have six models, each from a separate subpopulation {extremely obese [16], critically ill post heart surgery [17], trauma patients [18], intensive care patients [19], septic [20], hospitalised patients [21]}, as well as a variety of demographic information for each. In our experiments we focus on the age, height, weight, and creatinine clearance levels as have been shown to be strongly associated with drug response [11] and are provided for each model.

We use the exact same test simulations as the original authors in order to more accurately explore the impact of SMC on the predictions made - we refer the interested reader to their paper for exact details of how the simulations were produced. In summary though, covariates were sampled from a global population before AUC observations being sampled for 1000 patients from each of the PopPK models used for a total cohort of 6000 simulated patients.

In Table 2 we report the relative root-mean-square error RMSE of the predictions - the lower, the better. We can see that SMC consistently performs competitively, especially when combined with PBMA, although it does not appear to be outperforming the competition in any significant sense. However, we note that the simulation setup here is not based on the underlying assumption that we make. I.e., when simulating patients based on the model of Adane et al. [16] for clinically obese patients, the current simulations still

generate covariates from a normal population, and actually only a small minority of the patients would be considered obese. Consequently, in order to evaluate the performance of SMC in what we consider a more realistic setting we develop a method to subsample the original simulations in order to obtain a population for each model that more accurately reflects the population on which each model was developed.

In order to select a smaller sample of 1000 patients, we first modelled the density of each of the patient populations based on the demographic statistics provided in each of the original papers. Then for each of the 6000 simulated patients we evaluated the likelihood that their covariates came from each model and selected the model with the highest likelihood. If this selected model matched the model from which the AUC observations were simulated, then the patient was kept and otherwise discarded. This mimics a rejection sampling method for the covariates from the original model demographics using the sampling method of [11] as the base distribution. This results in a population where each model only simulated data for patients whose covariates were likely under their reported demographic information.

In Table 3 we report the relative root-mean-square error RMSE again of the predictions. In this situation it becomes clear that SMC can take advantage of the setting where our assumptions more appropriately line up with the simulations. Still, in the question of only using SMC or PBMA, we can see that they both seem to perform roughly equivalently - except in the 'A Priori' setting where PBMA cannot be applied properly and as such SMC significantly improves upon it. What is clearer however, is that when both are used in combination they can each take advantage of their different properties and outperform the other methods individually.

Discussion

Existing methods for ensembling multiple models all promote models based on an estimate of their global performance - i.e., identifying which models on average work best. This can be a problem if our estimate of performance does not coincide with the particular population that we ultimately want to make predictions about, given models often perform poorly on data outside of the domain on which they were trained. SMC aims to overcome this problem by specifically modelling the regions of the space where it believes individual models should perform well. It works by learning a special representation of the feature space that we can map new test cases to, before modelling in that space which regions each model works on. With this, SMC constructs a unique ensemble for each test case, using only the models it feels appropriate - allowing it to ignore models that may perform well in general but not be very effective for a specific case. As a result, we expect then for SMC to bring some benefit where the goal is to predict a target when given a number of models, each of which have been developed in different, heterogeneous, populations and it is not clear which models are better than others.

It should be pointed out as well that in cases where you have no labelled data it can be hard - or impossible - to obtain such an estimate of global performance, meaning that often this approach may not even be possible in the first place, as is pointed out in [11]. However, when it is possible, it does not interfere with SMC in that you can calculate weights according to both methods and then combine them as you like.

The key take-away being that by introducing SMC, you do not need to give up the benefits of global ensembling for identifying good average models.

To explain and illustrate SMC and how it differs from existing model ensembling methods we used an example of multiple population PK models. However, SMC is not limited to use only with population PK models. It can be used in any situation where multiple models have been developed to address the same issue, for example multiple disease progression models of the same disease. It is likely there are other situations too and we hope to inspire ideas for other applications.

In order to identify good models on the individual level, SMC models the regions of the feature space for which models should be able to produce good predictions based on a level of epistemic uncertainty. This epistemic uncertainty could in turn be used in a feedback system for identifying regions of the space for which we have no good models. This would allow for future targeted data collection, allowing practitioners to identify subpopulations that require more information - not wasting resources collecting information on patients which we can already predict well for.

As we show in Figure 3, the ML community has developed a range of methods for what to do in the cases of different amounts of information. SMC is not the definitive answer on how to use the knowledge of model training domains, but it is the first, and highlights an important consideration that practitioners should be aware of when making their own predictions. We hope that this work will inspire future investigation, particularly in the task of adapting models to new populations we have no existing data on.

In general, we aim for the key takeaway for practitioners to be the following: *If the individual models seem unlikely to transfer well across populations then incorporating an aspect of SMC into ensemble predictions is unlikely to damage predictive power in the worst case and will most likely improve predictions.*

To conclude, in this paper we have introduced the framework of Synthetic Model Combination to the clinical pharmacological and pharmacometric community - an instance-wise approach to ensembling models in order to make predictions with only models that have seen similar individuals during their training phase. We demonstrated how it can be applied in the averaging of PopPK models with the real case study of estimating the effectiveness of vancomycin precision dosing, and the impact that could have in terms of the appropriate treatment of patients.

Study Highlights

What is the current knowledge on the topic? Model averaging population pharmacokinetic models is known to improve the predictive accuracy when informing optimal dose selection.

What question did this study address? Appropriately averaging models is challenging, and current methods ignore important information about the demographics of which population a model is based on, a concept this study leverages for improved performance.

What does this study add to our knowledge? Incorporating demographic information into model averaging methods allows us to improve the quality of predictions while maintaining the gains of current methods.

How might this change drug discovery, development, and/or therapeutics? The algorithms described in the study may increase the accuracy of precision dose calculations among other targets.

Acknowledgements

AC would like to acknowledge and thank Microsoft Research for its support through its PhD Scholarship Program with the EPSRC. This work was additionally supported by the Office of Naval Research (ONR) and the NSF (Grant number: 1722516).

Author Contributions

AC helped conceive the research problem, designed and completed the experiments, and wrote the manuscript.

RP helped write the manuscript as well as guide the research direction and experiments.

MG helped write the manuscript as well as guide the research direction.

MVS conceived the research problem and oversaw the project.

References

1. Powell JR, Cook J, Wang Y, Peck RW, Weiner D. Drug Dosing recommendations for all patients: A roadmap for change. *Clin Pharm Ther.* 2021;109:65-72.

2. The Drug Disease Model Resources Foundation (DDMoRe).
<https://www.ddmore.foundation>. Accessed 12th August 2022
3. Vivli - Center for Global Clinical Research Data. <https://vivli.org>. Accessed 16th August 2022.
4. Broeker A, Nardecchia M, Klinker KP, Derendorf H, Day RO, Marriott DJ, Carland JE, Stocker SL, Wicha SG. Towards precision dosing of vancomycin: a systematic evaluation of pharmacometric models for Bayesian forecasting. *Clinical Microbiology and Infection*. 2019 Oct 1;25(10):1286-e1.
5. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*. 1997 Mar 1;92(437):179-91.
6. Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2012 Mar;4(2):199-203.
7. Minka TP. Bayesian model averaging is not model combination. Available electronically at <http://www.stat.cmu.edu/minka/papers/bma.html>. 2000 Dec:1-2.
8. Izmailov P, Nicholson P, Lotfi S, Wilson AG. Dangers of Bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*. 2021 Dec 6;34:3309-22.
9. Kouw WM, Loog M. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*. 2019 Oct 7;43(3):766-85.
10. Kim HC, Ghahramani Z. Bayesian classifier combination. In *Artificial Intelligence and Statistics* 2012 Mar 21 (pp. 619-627). PMLR.
11. Uster DW, Stocker SL, Carland JE, Brett J, Marriott DJ, Day RO, Wicha SG. A Model averaging/selection approach improves the predictive performance of model-

- informed precision dosing: vancomycin as a case study. *Clinical Pharmacology & Therapeutics*. 2021 Jan;109(1):~~175~~83.
12. Chan AJ, van der Schaar M. Synthetic Model Combination: An Instance-wise Approach to Unsupervised Ensemble Learning. *Advances in Neural Information Processing Systems*. 2022.
 13. Terrell GR, Scott DW. Variable kernel density estimation. *The Annals of Statistics*. 1992 Sep 1:1236-65.
 14. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013 Dec 20.
 15. Infectious Diseases Society of America. <https://www.idsociety.org/practice-guideline/vancomycin/>. Accessed 13th September 2022
 16. Adane ED, Herald M, Koura F. Pharmacokinetics of Vancomycin in Extremely Obese Patients with Suspected or Confirmed *S taphylococcus aureus* Infections. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. 2015 Feb;35(2):127-39.
 17. Mangin O, Urien S, Mainardi JL, Fagon JY, Faisy C. Vancomycin pharmacokinetic and pharmacodynamic models for critically ill patients with post-sternotomy mediastinitis. *Clinical pharmacokinetics*. 2014 Sep;53(9):849-61.
 18. Medellín-Garibay SE, Ortiz-Martín B, Rueda-Naharro A, García B, Romano-Moreno S, Barcia E. Pharmacokinetics of vancomycin and dosing recommendations for trauma patients. *Journal of Antimicrobial Chemotherapy*. 2016 Feb 1;71(2):471-9.
 19. Revilla N, Martín-Suárez A, Pérez MP, González FM, Fernández de Gatta MD. Vancomycin dosing assessment in intensive care unit patients based on a population

- pharmacokinetic/pharmacodynamic simulation. *British journal of clinical pharmacology*. 2010 Aug;70(2):201-12.
20. Roberts JA, Taccone FS, Udy AA, Vincent JL, Jacobs F, Lipman J. Vancomycin dosing in critically ill patients: robust methods for improved continuous-infusion regimens. *Antimicrobial agents and chemotherapy*. 2011 Jun;55(6):2704-9.
21. Thomson AH, Staatz CE, Tobin CM, Gall M, Lovering AM. Development and evaluation of vancomycin dosage guidelines designed to achieve new target concentrations. *Journal of antimicrobial chemotherapy*. 2009 May 1;63(5):1050-7.

Figure Legends

Figure 1: a) Instance-wise Ensembles. Here we represent the density of the training features for three separate models - M1, M2, and M3. Given new test points A, B, and C, we need to construct predictions from these models. A is well represented by both M2 and M3 while B only has significant density under M3. C looks like none of the models will be able to make confident predictions.

b) Different models are useful for new patients. PopPK models are often trained on certain demographic groups given the studies that are designed for data collection. For a new patient who doesn't necessarily fit into one of the existing demographics, different models may be more or less relevant and accurate. Naive ensembles ignore this fact and always incorporate evenly the predictions of each model, SMC on the other hand aims to up-weight the models that would appear to be more relevant.

Figure 2: a) Synthetic Model Combination Training Algorithm. Algorithm outlining the main steps in training for SMC.

b) Inference Flow Diagram. As a new case comes in the first step is to calculate all of the individual models’ predictions, using NONMEM for example. Then, like with any model averaging algorithm the weights must be calculated. Performance-based model averaging methods have a set of weights independent of the new case, whereas SMC maps the new case features to a latent space that is then used to calculate individual weights for that case.

Figure 3: Methods based on varying information. A selection of methods from the spectrum of information available to a practitioner. SMC lies quite far towards the little information end, aiming to only take use of some demographic information from each of the models and not require any labelled training points.

Table 1: Comparison of modern model averaging/ensembling methods.

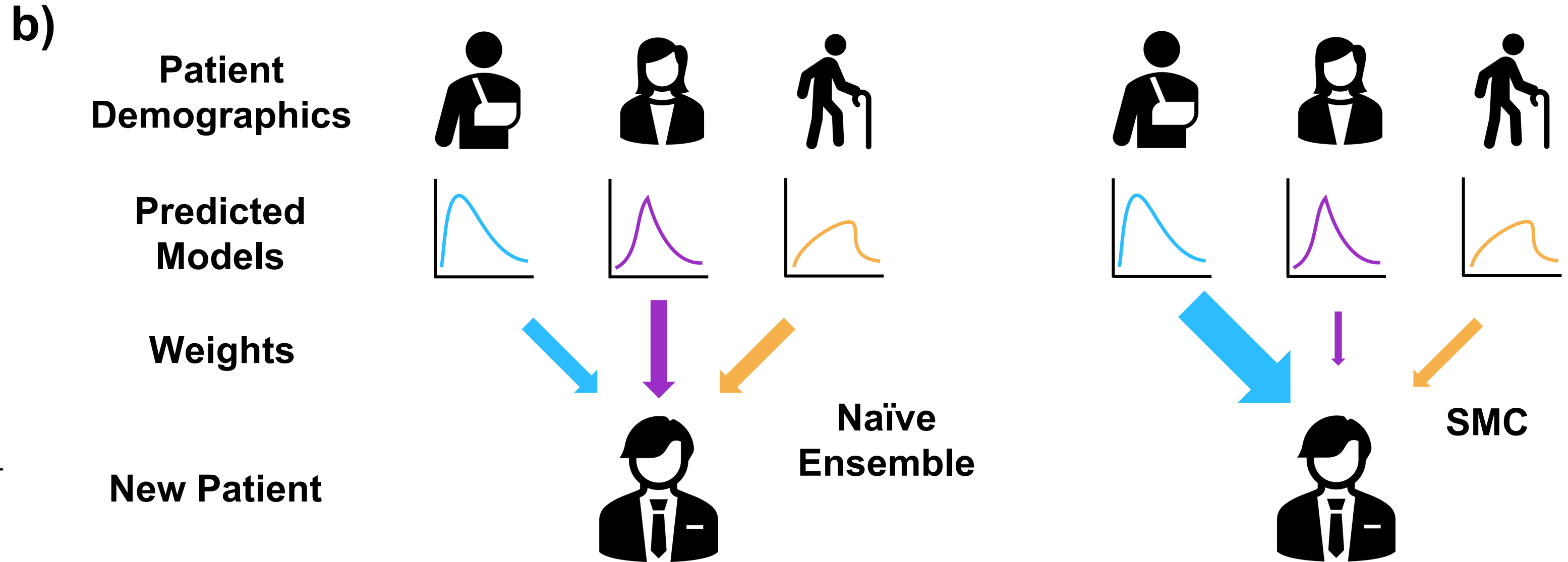
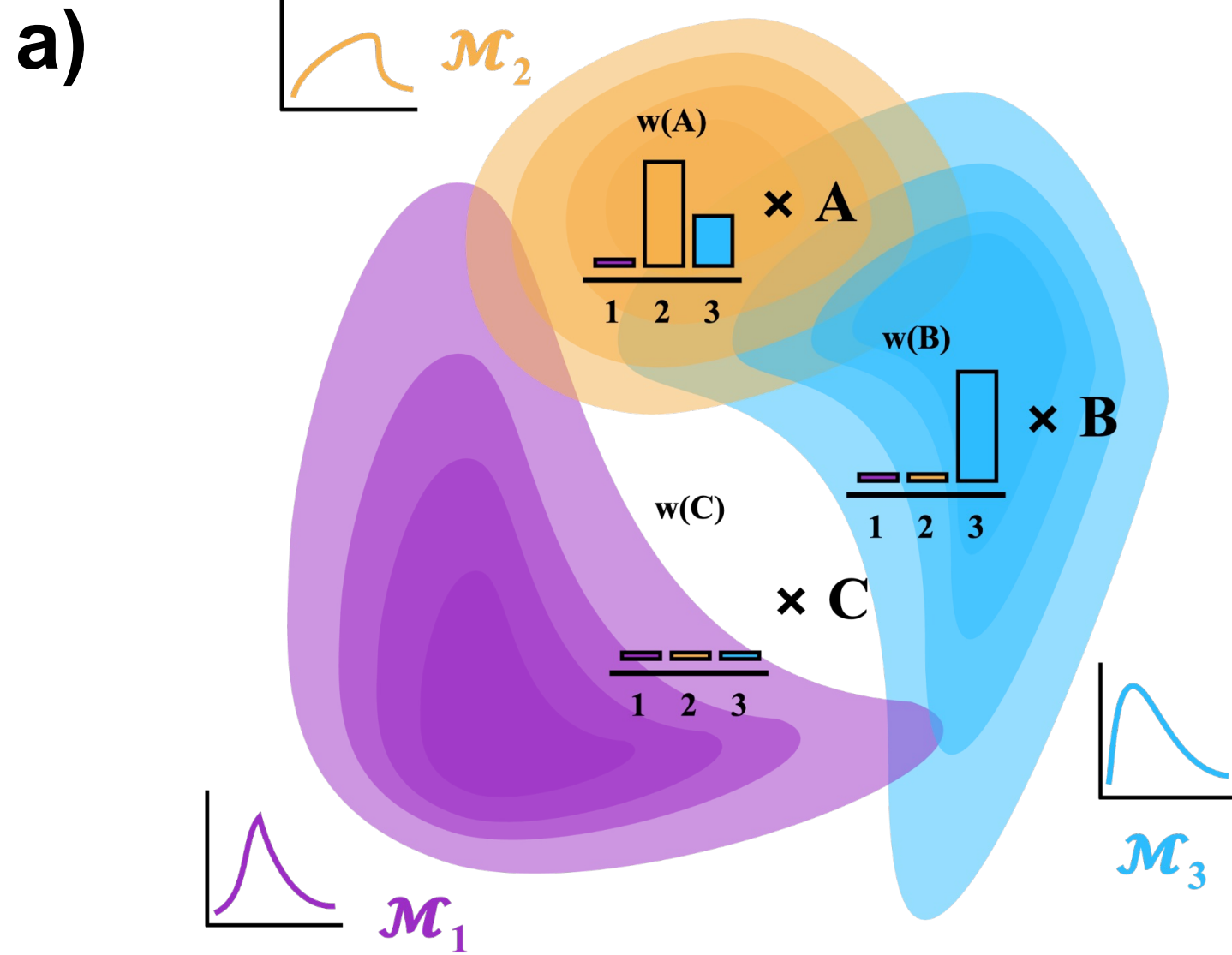
Table 2: RMSE for AUC predictions from models for a number of concentration measurements over a 48-hour period on the full cohort of 6000 simulated patients.

Table 3: RMSE for AUC predictions from models for a number of concentration measurements over a 48-hour period on 1000 subsampled patients.

Method	Instance-wise?	Require new data?	Weights
Naive Ensemble	No	No	$w = 1/N$
PBMA	No	Yes	$w_i \propto BIC(\mathcal{M}_i)$
Model Selection	No	Yes	N/A
Synthetic Model Combination	Yes	No	$w_i(x) \propto p_i^Z(f_{\theta}(x))$

	A Priori	One	Two	Three
Adane et al.	52.4 \pm 0.3	32.9 \pm 0.2	32.7 \pm 0.3	27.4 \pm 0.3
Mangin et al.	57.4 \pm 0.2	27.8 \pm 0.1	25.5 \pm 0.3	17.7 \pm 0.1
Medellin-G et al.	51.4 \pm 0.2	25.7 \pm 0.1	22.8 \pm 0.1	17.3 \pm 0.1
Revilla et al.	35.1 \pm 0.1	20.8 \pm 0.1	18.5 \pm 0.1	15.4 \pm 0.1
Roberts et al.	31.7 \pm 0.1	18.7 \pm 0.1	18.1 \pm 0.1	14.7 \pm 0.1
Thomson et al.	34.5 \pm 0.1	22.9 \pm 0.1	20.6 \pm 0.1	16.8 \pm 0.1
Ensemble	38.4 \pm 0.1	21.9 \pm 0.1	20.0 \pm 0.1	16.1 \pm 0.1
PBMA	38.4 \pm 0.1	19.0 \pm 0.1	17.4 \pm 0.1	13.9 \pm 0.1
SMC	36.4 \pm 0.1	21.0 \pm 0.1	19.3 \pm 0.1	15.4 \pm 0.1
SMC + PBMA	36.0 \pm 0.1	19.0 \pm 0.1	17.6 \pm 0.1	14.1 \pm 0.1

	A Priori	One	Two	Three
Adane et al.	80.0 \pm 0.7	47.3 \pm 0.4	40.9 \pm 0.6	32.5 \pm 0.3
Mangin et al.	83.3 \pm 0.6	36.2 \pm 0.3	33.0 \pm 0.3	21.4 \pm 0.2
Medellin-G et al.	76.5 \pm 0.6	33.5 \pm 0.2	28.5 \pm 0.2	20.9 \pm 0.2
Revilla et al.	32.7 \pm 0.4	21.6 \pm 0.2	20.0 \pm 0.2	16.7 \pm 0.2
Roberts et al.	35.8 \pm 0.3	20.6 \pm 0.1	20.0 \pm 0.2	16.1 \pm 0.1
Thomson et al.	48.2 \pm 0.4	30.2 \pm 0.2	25.8 \pm 0.2	20.9 \pm 0.1
Naive Ensemble	55.7 \pm 0.4	28.5 \pm 0.2	24.9 \pm 0.2	19.3 \pm 0.2
PBMA	55.7 \pm 0.4	20.8 \pm 0.2	19.3 \pm 0.2	15.2 \pm 0.2
SMC	41.8 \pm 0.5	22.1 \pm 0.2	19.6 \pm 0.2	15.2 \pm 0.2
SMC + PBMA	47.5 \pm 0.5	20.5 \pm 0.1	18.7 \pm 0.2	14.6 \pm 0.2



- a) Algorithm 1: Synthetic Model Combination**
- Result:** Test predictions using mapping from data to model weights
- Input:** $\{(\mathcal{M}_j, \mathcal{I}_j)\}_{j=1}^N$ and \mathcal{D}_T ;
1. Use information to produce density models;
 2. Sample data from models and combine with test data;
 3. Learn representation space;
 4. Re-model densities in new space;
 5. Calculate weights in new space;
 6. Make predictions $\{\hat{y}_i\}_{i=1}^M$ over test set;
- Return:** $\{\hat{y}_i\}_{i=1}^M$

