VIEWPOINT

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1671 Published online 16 January 2015 in Wiley Online Library

Viewpoint: model selection uncertainty, pre-specification, and model averaging

Björn Bornkamp*

Scientific progress in all empirical sciences relies on selecting models and performing inferences from selected models. Standard statistical properties (e.g., repeated sampling coverage probability of confidence intervals) cannot be guaranteed after a model selection. This viewpoint reviews this dilemma, puts the role that pre-specification can play into perspective and illustrates model averaging as a way to relax the problem of model selection uncertainty.

Keywords: dose-response, modelling

1. INTRODUCTION

Most available theory for statistical inference assumes that the statistical model is fixed and not selected based on the data. In practice, however, estimates and confidence intervals are often calculated based on this premise even if a model selection has been performed to select structural assumptions such as the covariates, residual distribution, or functional form in a regression analysis. The distribution of estimators post model selection can differ substantially from the distribution when the model is fixed, typically leading to too narrow confidence intervals (type I error rate inflation) and potentially bias. Reviews and illustrations of the problem can be found among many others in [1–5] or in Chapter 7 of [6]. A point similarly critical from a practical perspective is that model selection is not 'stable' [7]. Small changes in the data might lead to a different model being selected and potentially a complete change in the ultimate conclusions, leading to non-robust decision-making as the conclusions produced by an unstable procedure are hard to replicate on new data sets.

One approach to deal with model selection uncertainty is to pre-specify the statistical model prior to the data analysis. In pharmaceutical development, pre-specification of the analysis method is mandated through the ICH E9 guidance [8], although the document also leaves some limited room for data-based changes when these are pre-specified (see Sections 5.1, 5.5). Note however that in a strict statistical sense, pre-specification only makes inference 'valid', if the full statistical model is pre-specified. A pre-specified model selection procedure also suffers from the problems mentioned above. On the other hand, pre-specifying a complete statistical model can also be a source of bias if the specified model is not adequate. This might have a similarly (or even stronger) negative impact on the repeated sampling properties, such as bias but also coverage of confidence intervals. Because of this trade-off, pre-specification can only be a partial pragmatic solution to the problem of model selection uncertainty.

In addition, it is clear that scientific progress in all empirical sciences relies on selecting *non pre-specified* statistical models and performing inferences from these models. In pharmaceutical sciences, for example, model selection is unavoidable in the exploratory stage of development: less is known about the compound and it is difficult to pre-specify one statistical model at the design stage. So important decisions, for example, about the treatment regimen used in confirmatory trials or potential patient subgroups have to be made based on inferences from selected models, with all associated problems.

Nevertheless, there are ways to deal with model selection uncertainty better than just ignoring the issue, and in the past few years, a number of proposals have been made. In what follows, I will concentrate on the relatively simple situation, where the enumeration of all potentially plausible models is possible at the design stage and will consider only two approaches: Bayesian model averaging and bootstrap aggregating (bagging). Both are rather differently motivated but are very generally applicable 'recipes' to deal with model selection uncertainty.

If one a priori thinks that a set of models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ is plausible, also, the posterior distribution in a Bayesian paradigm will be supported on the same set of models. Picking one model and ignoring the rest of the posterior uncertainty does not seem adequate, unless required by the decision context. So ultimately, inference for the quantity of interest would use all models, weighted according to their posterior probability. This acknowledgement of model uncertainty is also the main reason why Bayesian inference, depending on the particular model used, often includes an implicit 'multiplicity' adjustment [9]. In addition, averaging over multiple models will be more stable than just picking the one best model. Raftery and Zheng [10] provide a list of references of simulation studies to illustrate the generally good

Novartis Pharma AG, 4002 Basel, Switzerland

**Correspondence to: Björn Bornkamp, Novartis Pharma AG, Basel, Switzerland. E-mail: bjoern.bornkamp@novartis.com*

performance of model averaging. Several forms of non-Bayesian (or approximately Bayesian) model averaging have also been proposed, for example, based on using weighted inference (e.g., weighted averages for quantities of interest) with weights

$$w_k \propto \exp(-0.5\mathrm{IC}_k), \quad (1)$$

where IC_k is an information criterion (such as AIC or BIC) calculated for model $\mathcal{M}_k$ (see also [6,11]).

An alternative method to acknowledge model uncertainty is bagging, originating from the machine learning literature, which can also be seen as a form of model averaging (see, for example, [12] or Chapter 8 in [13] for an introduction). For bagging, one bootstraps the data set and applies model selection to each bootstrap data set and finally uses the averaged bootstrap estimate to obtain one final estimate. Bagging will result in more stability as illustrated in [12] based on simulations and heuristic arguments. By randomly perturbing the original data and averaging the estimate of the different models, inference will be less driven by particular aspects of the observed data set and focus more on the main features of the data set, which will be represented in most of the resampled bootstrap data sets. In a way, one 'smoothens' model selection. Another useful side-product of bagging is that one also obtains bootstrap confidence intervals for the quantities of interest [14]. A difference of bagging to Bayesian model averaging is that one augments or perturbs the original data set with more variability to obtain a more stable estimation, whereas for Bayesian model averaging, the weights come from the uncertainty in the actually observed data.

A concern that is often brought up with respect to model averaging methods is that they are more difficult to interpret, as there is no longer one model providing inferences. This is an important point. However, in the end, inference always focuses on particular, interpretable quantities, which are not necessarily a parameter in the model, for example, the treatment effect for a particular patient group or a target dose of interest in a dose–response analysis. Estimates and confidence intervals for these quantities can always be extracted from a model averaging procedure as well. When it comes to variable selection, the decision context might sometimes require that only a small number of variables/models should be reported. This could be carried out by just reporting the models/variables with the most posterior probability or bootstrap probability to be selected, which comes at the cost of a reduced acknowledgment of model selection uncertainty.

Both approaches are quite generally applicable recipes to acknowledge model selection uncertainty. The problem of adequately acknowledging model selection appears in many areas (another challenging area that can be re-casted as model selection problem is, for example, subgroup selection). For the purpose of the presentation, here, we will just present an example related to dose–response modeling.

1.1. Example: dose–response analysis

Here, we use data from the dose–finding study NCT00900146 obtained from `clinicaltrials.gov`. In this study, the experimental treatment was administered as an add-on treatment to Metformin for patients with type 2 diabetes. In the design, there were four active doses, 5, 15, 50, and 150 mg, and a placebo group. The randomization was imbalanced with 175 patients on the placebo group and roughly 90 patients in the other groups. The primary endpoint was the change from baseline in HbA1c after 4 months of treatment. I simulated a data set that had the same patient numbers, empirical means, and standard deviations as given on the website. Pre-specifying one dose–response model at the design stage is difficult in these situations, which is why, for example, the Multiple Comparison Procedures and Modeling (MCP-Mod) methodology ([15,16]) uses a pre-specified candidate set of dose–response models and then either selects or averages over the candidate models once data are available.

Assume that y_i, the response for patient i, is generated as $y_i \sim N(\mu(x_i), \sigma^2)$, where x_i is the dose administered for patient i and $\mu(x)$ the dose–response model. In addition, assume that a priori, it was considered that the mean function is given either by a power model $\mu(x) = \beta_0 + \beta_1(x/150)^{\exp(\alpha)}$ or by a quadratic model $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. For the power model, the parameter $\alpha \in \mathbb{R}$ determines the shape of the dose–response curve, which can range from concave ($\alpha < 0$), linear ($\alpha = 0$), to convex ($\alpha > 0$) monotonic functions. Non-monotone dose–response relationships are rare but can sometimes not be ruled out at the trial design stage. Assume that in this situation, there was a clear indication that the quadratic model should be included in the candidate set of models.

In Figure 1(i), one can see the data set with both model fits and pointwise confidence intervals. It can be seen that both models lead to rather different conclusions regarding the dose–response curve. The power model suggests that the full treatment effect is reached already at a 15-mg dose, and the effect size is ~0.15 over placebo. The quadratic model, however, suggests that the maximum effect is reached at 75-mg dose with a larger effect of ~0.22. The fit of the quadratic model is quite discomforting, as it predicts a maximum effect larger than the effect at any observed dose and at a location where there is no data point. On the other hand, there seems to be a slight hint of a non-monotonicity in the data, hence, it is also not clear (from the data alone) whether the power model is more adequate.

When calculating the BIC of the models, one obtains 1157.031 for the power model and 1156.934 for the quadratic model, which means, in a formal model selection paradigm, one would go ahead with the quadratic model (the AIC leads to the same conclusions as both models considered here have the same number of parameters). However, a slight change in any of the observations can lead to a situation where the BIC (or AIC) would prefer the power model with entirely different conclusions about the dose–response curve and the maximum achievable effect size. Clearly, it seems wrong to discard one model over the other in this situation, in particular, because they differ quite dramatically when it comes to the quantities one is ultimately interested in.

Calculating model weights according to the formula (1), one ends up with a weight of 51% for the quadratic model and 49% for the power model. In Figure 1(ii), one can observe the model averaged dose–response curve using the BIC weighted model averaging and bagging. Bagging was performed here by drawing 2000 bootstrap resamples from the data set (stratified by dose) and then for each resample, fitting both models to the data and letting the model with minimum BIC predict the dose–response curve. The black line corresponds to the median of the bootstrap estimates per dose and the confidence limits to the 2.5% and 97.5% quantiles. Confidence intervals for BIC model averaging were calculated using a normal approximation with the variance obtained from the law of the total variance. In this case, both averaging procedures lead to rather similar results, and both fits are intuitively more satisfying compared with either of the fits alone.

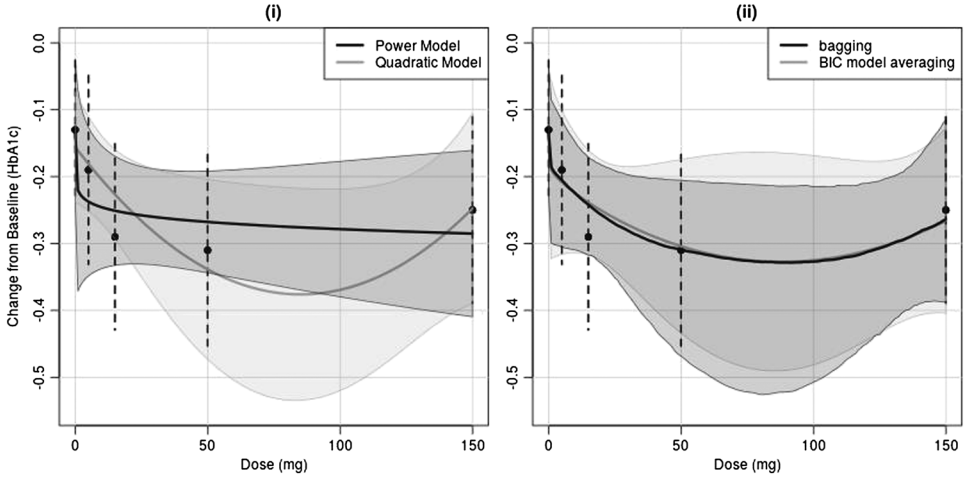

Figure 1. Dose–response curve. Dots indicate the group point estimates, and vertical lines indicate the corresponding 95% confidence intervals for an ANOVA model. In (i), the model fits are displayed, and in (ii), the model averaging model fits are displayed, both with pointwise 95% confidence intervals.

The situation presented here is clearly specific to the studied data set and an extreme example; however, similar situations are quite likely to occur in practice. The signal to noise ratio in this endpoint was small in this example, but this does not affect the point being discussed here on whether to use model selection or model averaging.

2. CONCLUSIONS

Statistical inference after model selection should take into account model selection uncertainty to be able to obtain more honest (i.e., usually wider) confidence intervals and more stable decision-making compared with the naive approach of performing model selection and ignoring the resulting uncertainty. This is independent of whether an analysis was pre-specified or not.

Bayesian model averaging and bagging are two fairly general methods to acknowledge model selection uncertainty in performing statistical inference. Both methods, however, require that one is able to enumerate the models of scientific plausibility only based on the previous data. In the situation, when model building is even more data-driven, one last resort are cross-validatory techniques to at least evaluate predictive abilities and correct optimism of the models, as outlined, for example, in Chapter 7 of [13].

REFERENCES

[1] Chatfield C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* 1995; **158**:419–466.
[2] Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B* 1995; **57**:45–97.
[3] Leeb H, Pötscher BM. Model selection and inference: facts and fiction. *Econometric Theory* 2005; **21**:21–59.
[4] Leeb H, Pötscher BM. Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 2006; **34**:2554–2591.
[5] Leeb H, Pötscher BM. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 2008; **24**:338–376.
[6] Claeskens G, Hjort NL. *Model selection and model averaging*. Cambridge University Press: Cambridge, 2008.
[7] Breiman L. Heuristics of instability and stabilization in model selection. *Annals of Statistics* 1996; **24**:2350–2383.
[8] ICH. E9, statistical principles for clinical trials, 1998. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf [Accessed 6th January 2015].
[9] Berger J. Multiplicity control and model prior probabilities, 2012. Presentation at CBMS: Model Uncertainty and Multiplicity, Available at: http://cbms-mum.soe.ucsc.edu/lecture4.pdf [Accessed on 6th January 2015].
[10] Raftery A, Zheng Y. Discussion: performance of Bayesian model averaging. *Journal of the American Statistical Association* 2003; **98**:931–938.
[11] Burnham KP, Anderson DR. *Model selection and multimodel inference: a practical information theoretic approach* (2nd edn). Springer: New York, 2002.
[12] Breiman L. Bagging predictors. *Machine Learning* 1996; **24**:123–140.
[13] Hastie T, Tibshirani R, Friedman J. *Elements of statistical learning* (2nd edn). Springer: New York, 2009.
[14] Efron B. Estimation and accuracy after model selection. *Journal of the American Statistical Association* 2014; **00**:00–00.
[15] Pinheiro JC, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine* 2014; **33**:1646–1661.
[16] European Medicines Agency. Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of phase II dose finding studies under model uncertainty, 2014. Available at: http://goo.gl/imT7IT [Accessed on 6th January 2015].