Comparison of designs with interim analyses based on posterior and predicted probabilities

Francesca Michielin[1], Elias Laurin Meyer[2]

[1]Roche, Switzerland; [2]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria

Introduction

Often in oncology Ph1 studies, the response rate (ORR) is used to have a first flavor of efficacy of the new drug and depending on whether this is favourable or not compared to the standard of care (SoC), decisions to continue or stop development might be taken.

A lot of time is then spent to define gating strategies for final decision making and even more to early analyse the data, to understand whether there is an interesting signal.

When those interims are only used for internal decision making, they are not accounted for in the operating characteristics (OC), but what is the risk we are taking?

Here we run a set of simulations to see which are the differences between designs based on posterior probabilities and designs based on predicted probabilities.

Setting the scene for the simulations

For the simulations we considered a 40 patients trial, and a setting where the ORR for the SoC is 20%. A 15% improvement over this (thus: a 35% ORR) is considered exceptionally good, and a 10% improvement (thus: a 30% ORR) good enough.

At the final analysis (FA), if at least **13/40** responders are observed, a **GO** decision will be taken. With this rule, the posterior probability (PostProb) that the true response rate brings at least a 10% improvement over the SoC is above 60%, and a 15% improvement is at least 35% likely.

The **STOP** decision will be taken in case only **10** patients or less are responding out of **40** (Figure 1). This corresponds to a low posterior probability (either <30% or <15%) that the true response rate has respectively a 10% or 15% improvement over SoC.

Decision	ORR	Alternative rules
STOP	≤ **10/40** (25%)	PostProb(ORR>ORR SoC +10%) < 30% PostProb(ORR>ORR SoC +15%) < 15%
GO	≥ **13/40** (32%)	PostProb(ORR>ORR SoC+10%) > 60% PostProb(ORR>ORR SoC+15%) > 35%

Figure 1 – Summary of decision making rules at FA

For the interim analyses (IA) we explore different rules (see Figure 2), based either on posterior probabilities or predicted probabilities. In case of posterior probabilities, we simply repeat the very same rule applied at final. For predicted probabilities, we would recommend to GO or STOP at interim if the predicted probability of respectively a GO or a STOP is high (>80%).

This makes up 4 different scenarios, where the rule at FA is exactly the same, while the rules at interim might differ.

Figure 2 – Scenarios and decision making rules at IA

Actual decision making

Figure 3 shows how the actual decision making would look like following the above rules.
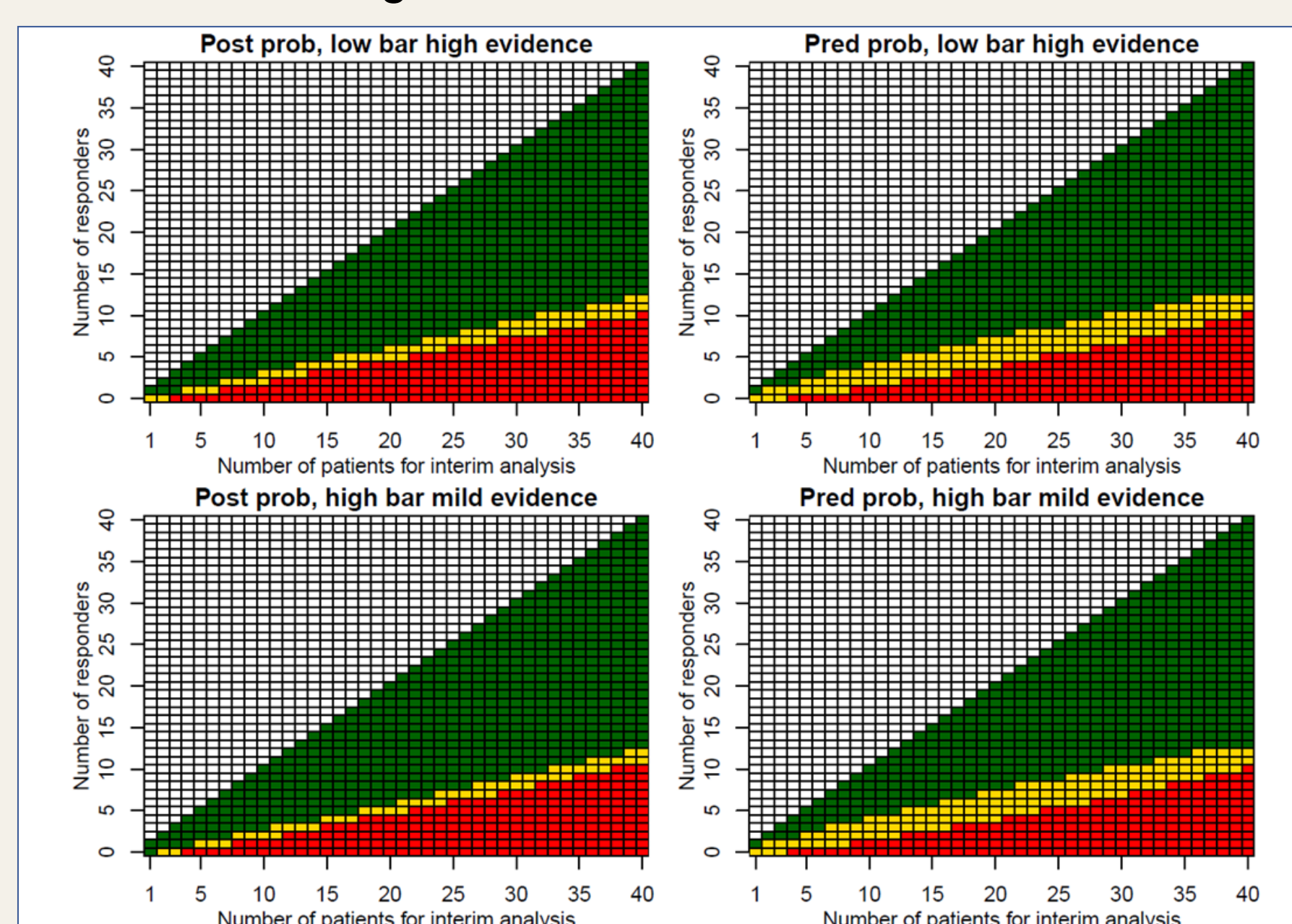
Figure 3 – Actual decision making in the various scenarios

- All 4 scenarios have the same decision rule at final (done on purpose!)
- Decision making based on predicted probability is more cautious than decision making based on posterior probability, and does not depend on the way we formulate the decision rule at final analysis as long as the actual rule at final analysis is the same
- Decision making at interim based on posterior probability applying the same rule as final will depend on the way we express the final rule (despite the actual rule at final analysis being the very same!)

Looking at decision making

Given our decision making rules we can now look at what decisions we would take for each simulated clinical trial if we were having an IA after each patient was enrolled.

Figure 4 shows decision making for the first 100 simulations (of 10000 simulations), taking a case where the true ORR was the same as the SoC as an example.
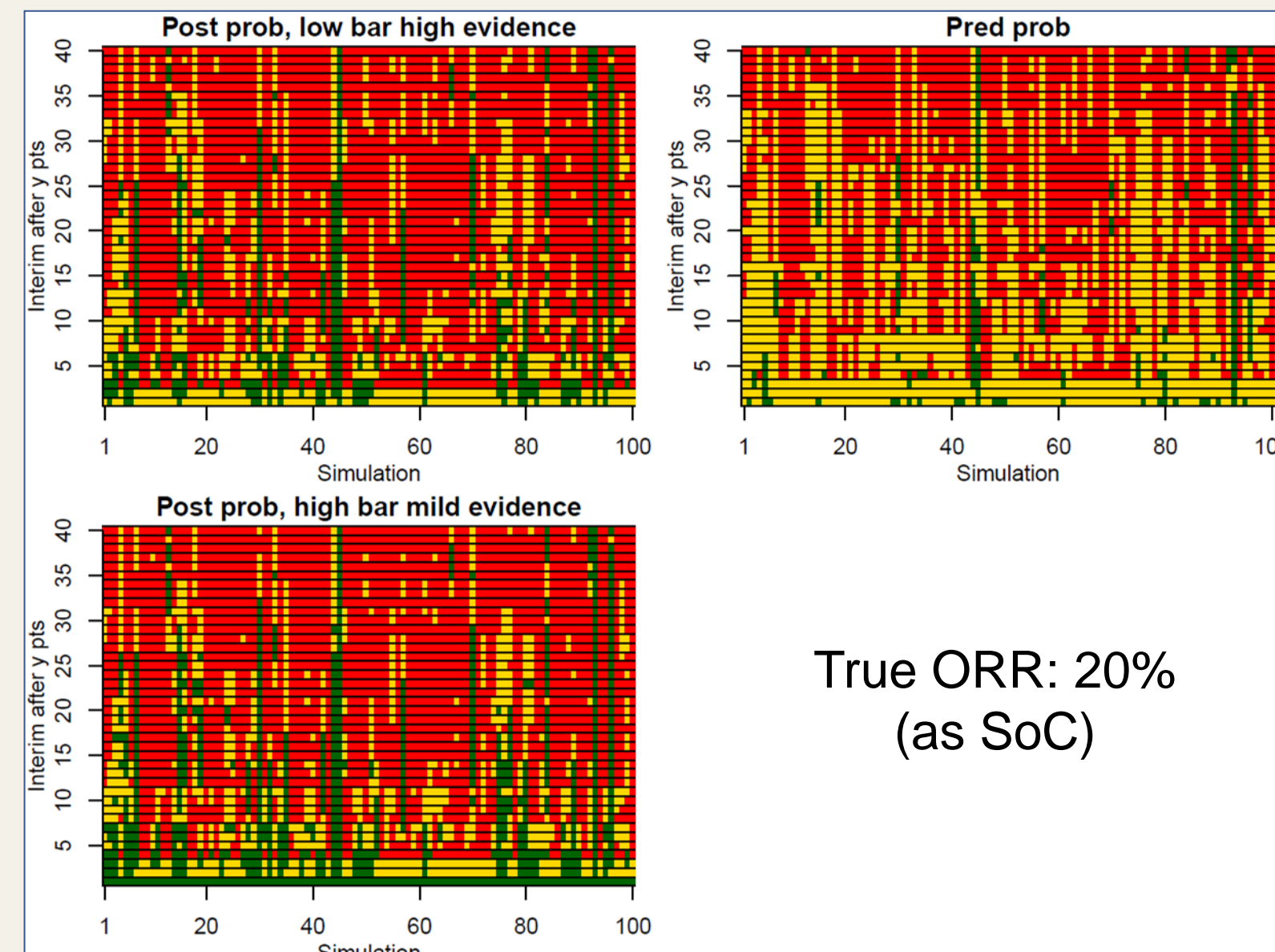
Figure 4 – Decision making for first 100 simulation

- In general decisions are not extremely coherent within each simulation run: depending on when we look at the data, our assessment of efficacy might vary
- Usually we do not easily switch between GO and STOP, but chance of giving different recommendations is high
- As expected from the rules, interim decisions based on predicted probabilities are more cautious than decisions based on posterior probabilities
- If rule is based on posterior probabilities, we'd declare more GO in the *high bar mild evidence* scenario compared to *low bar high evidence*

Looking at operating characteristics

How would the operating characteristics look like when we consider interim analyses? Specifically, we looked at:
- OC at each IA independently (Figure 5)
- OC thinking we will re-evaluate efficacy after each patient, but excluding the first few (IA are thus binding) – Figure 6
- OC when few binding IA only (i.e. 3 IA after 10 patients each, 2 IA after 20 and 30 patients, 1 IA only after 20 patients) – not shown

For illustration, we only report the OC when the true ORR was the same as for SoC, i.e. 20%.
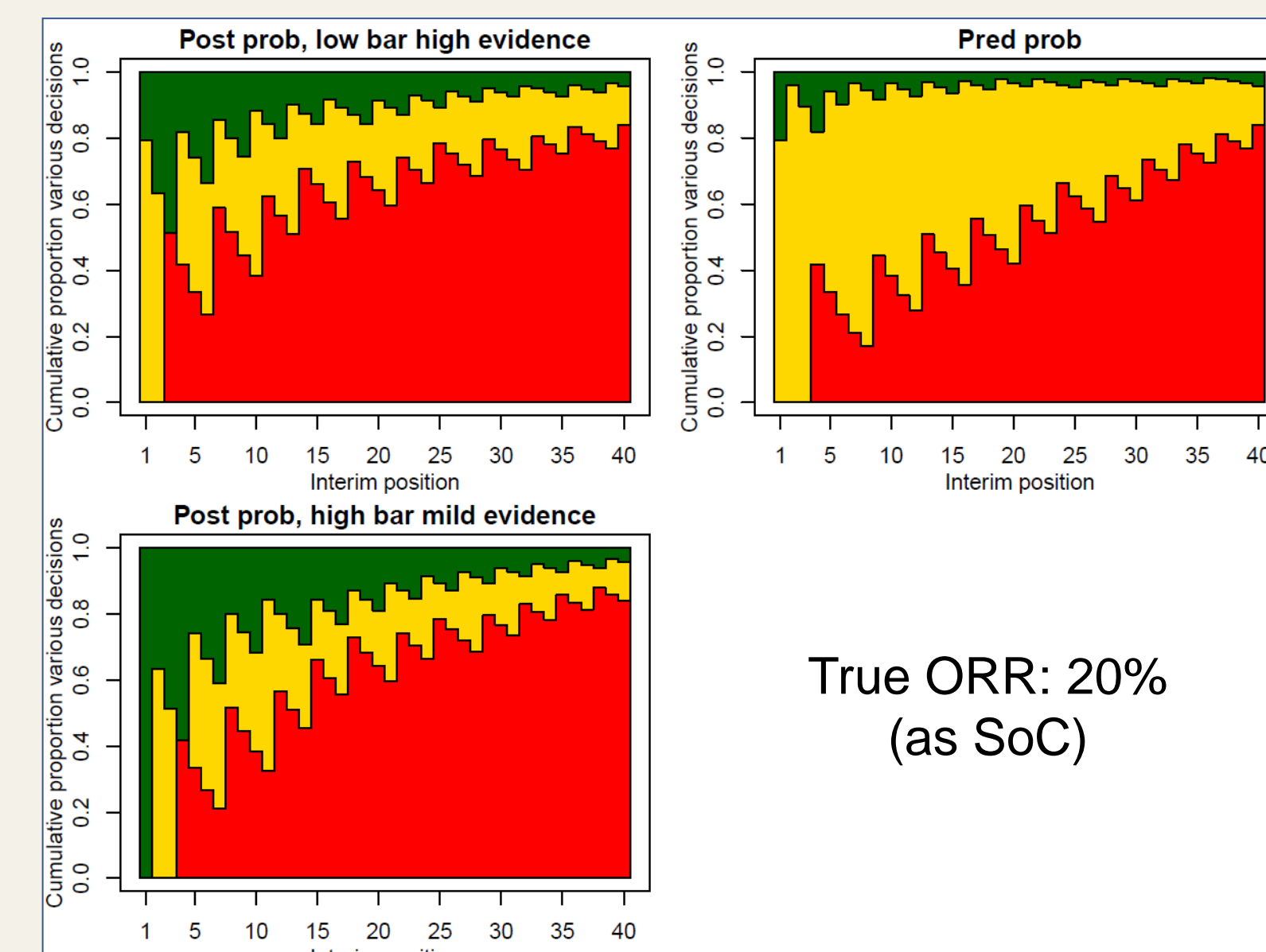
Figure 5 – OC at each IA

- With PostProb, the risk of false GO decisions is higher than with PredProb, especially early on.
- By using predicted probabilities we minimise the chance of false decisions, paying the price that in general we would be more hesitant to interpret the current efficacy as a clear GO or STOP signal.
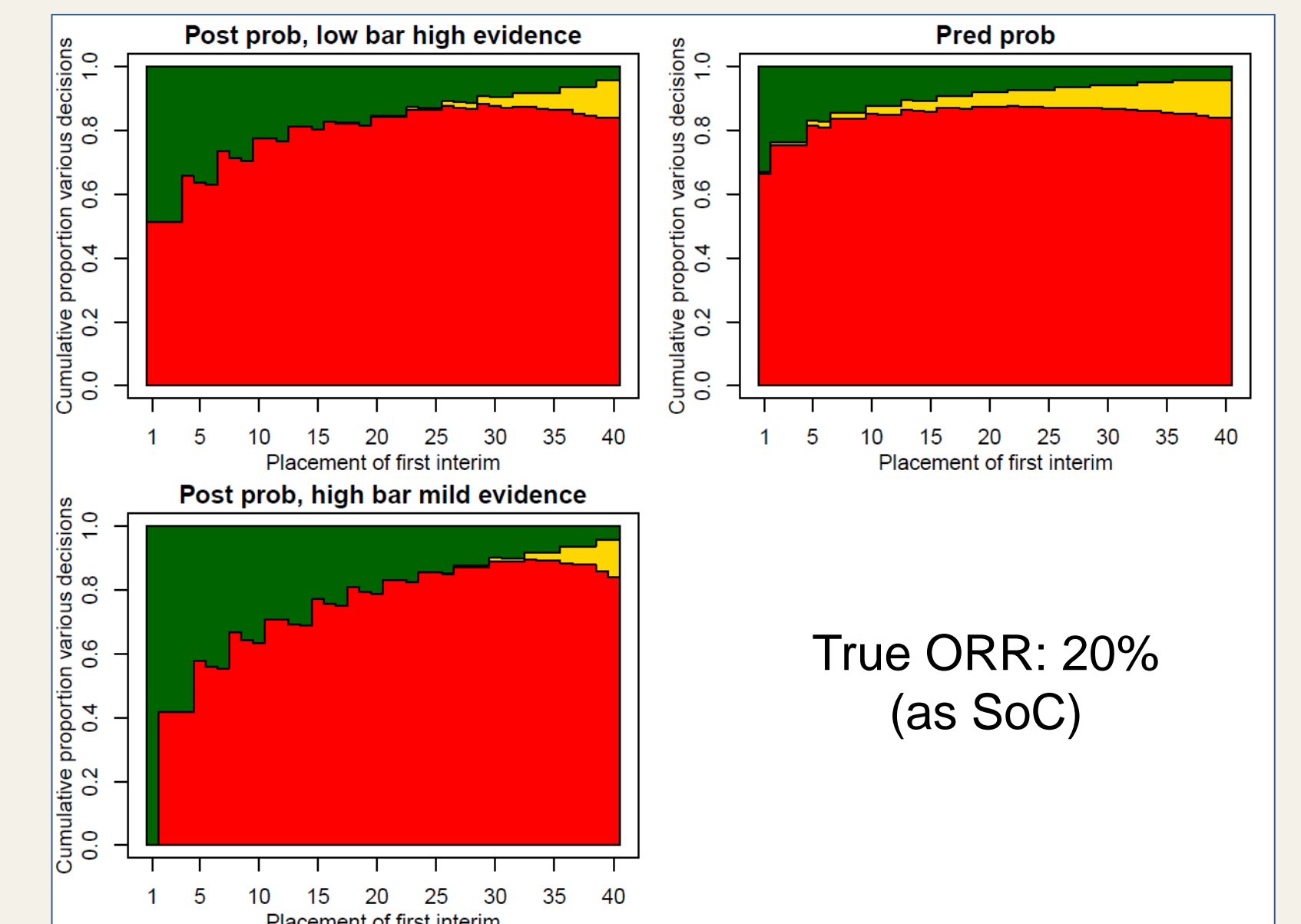
Figure 6 – OC for continuous assessment, moving time 1st IA

- OC might be very poor if we start with interim early on, especially if we use PostProb
- Designs that make use of the PredProb show a much lower erosion of the operating characteristics when including interim analyses
  - For instance, looking at simulations based on no improvement over SOC, those kind of designs slowly increase the chance of a false positive decisions when more assessments are included

Comparing the posterior probabilities scenarios

How would decision making vary if we choose a different threshold for the posterior probabilities rules (*still maintaining the same rule at final*)?
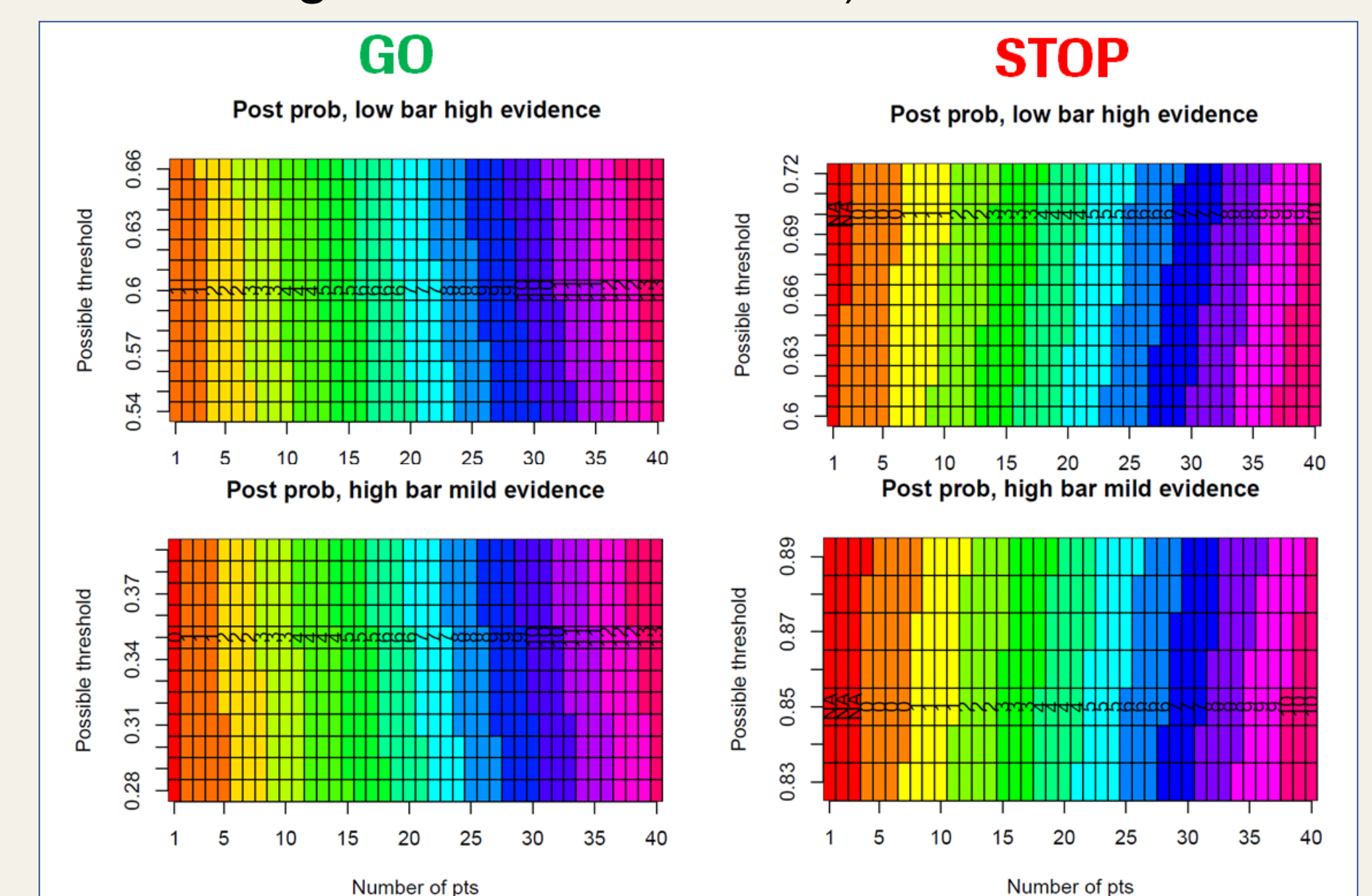
Figure 7 – Rules of PostProb scenarios with Δ thresholds

- Even if we maintain the same rule at final, the rule at interim might change widely depending on the specific threshold we set for the posterior probability – high variability of rules!
- While for many IA looking at the range of possible decisions we can conclude that we would GO more in the *high bar mild evidence* scenario, and STOP more in the *low bar high evidence* one, for some IA the range of possible recommendations overlaps.

Conclusion

- Designs where interim decision making is based on predicted probabilities are more 'cautious' than designs based on posterior probabilities when the level of required confidence is high (*e.g.* 80%)
- Designs where decision making at interim is based on predicted probabilities do not change properties as long as rule at FA does not change, while those based on posterior probabilities do change properties depending on the way we describe the rule at FA
- Not assessable *a priori* whether designs based on *high bar mild evidence* are preferable to *low bar high evidence*: rules might overlap
- Early assessments might bring high percentages of mistakes, this is minimised with predicted probabilities; still too early decision making is discouraged

Disclaimer: this work gives some food for thoughts, but the need to evaluate the operating characteristics of each new design remains!

Contact Information

Francesca Michielin
francesca.michielin@roche.com

Elias Laurin Meyer
elias.meyer@meduniwien.ac.at