Safety Monitoring and Evaluation in Late Phase Clinical Development:
An Application in OA Pain

José Pinheiro, *Janssen R&D*

Joint work with Camille Orman, Steven Wang, and Elena Polverejan – *Janssen R&D*

Janssen Research & Development

Outline

- Background: efficacy and safety in drug development
- Safety: signal detection vs. demonstration of lack of clinically significant harm
- Example: joint-related events in pain indication
- Concluding remarks

Background

- Traditional main goal of clinical drug development : establish clear efficacy under acceptable safety
- Regulatory approval paradigm has focused on formal statistical testing of efficacy, with safety evaluated in more exploratory fashion
- Usual requirements for drug approval:
  - (at least) two pivotal studies with primary efficacy endpoint statistically significant at 2.5% one-sided level
  - evidence of acceptable safety (as in lack of clinically worrisome signal) in reasonable amount of exposed patients data (indication-dependent, often in guidance doc)

Efficacy

- Generally drives study design (e.g., sample size) in late development (Ph 2 and 3): focus on statistical power to detect clinically relevant effect in primary endpoint
- Great regulatory concern about multiplicity issues: different doses, primary and secondary endpoints, subgroup analyses, etc
- Estimation also important for labeling purposes
- Efficacy analyses pre-specified in detail in study protocol and/or statistical analysis plan (SAP)
- Long history of interaction between industry and regulators to discuss acceptable/non-acceptable approaches

➔ Formalized, focused, well-understood

Safety

- Drives design in early development: first in human, dose escalation (SAD and MAD)
- In late development, mostly exploratory: tables and summary statistics for hundreds of safety endpoints, e.g., all adverse events (AEs), lab parameters, vital signs, biomarkers, etc
- Sponsors typically avoid including p-values for safety treatment comparisons: severe multiplicity
- Lack of regulatory concern about safety multiplicity issues: often no formal testing required
- Exceptions: QT prolongation, CV risk in diabetes

Safety (cont.)

- Challenge: rare safety signals (frequent ones detected early on, often even in pre-clinical phase)
- Sample sizes used in clinical development typically not large enough to detect rates of 1 per hundreds, or even thousands of patients
- Traditionally, strong safety signal detected during development and rare events post-approval and/or via meta-analyses (hopefully, but not always)
- Post-approval safety surveillance based on reported AE's – observational, denominator often unclear, hundreds of AE's, etc

Safety detection vs. demonstration

- Traditional paradigm has been safety monitoring: unsure where signal may show up, so look at hundreds of variables (but only descriptively)
- Demonstration of safety (or lack of clinically relevant harm) has more recent history: ICH E14 guidance on thorough QT studies from 2005; FDA guidance on CV risk for Diabetes drugs, 2008
  - Both specify acceptable upper limit for safety and require demonstration that risk associated with drug is below that limit (e.g., 1.8 hazard rate for Diabetes drugs at time of approval and 1.3 HR post approval)
  - Analysis done once at end of study or program

Challenges in safety evaluation

- How should safety monitoring be framed?
  - Statistically, like in hypothesis testing – if so, what is the null (no harm of drug)?
  - Should it be based just on clinical judgment (e.g., acceptable increase in event rate)? Statistical properties may still be needed, in any case
- How frequently should it be assessed? How widely (which variables to consider)? Operational concerns, false positive rate inflation (chasing noise), etc
- How to handle multiplicity problem? Patient vs. sponsor risk – who determines the right balance?

Case study: osteoarthritis pain

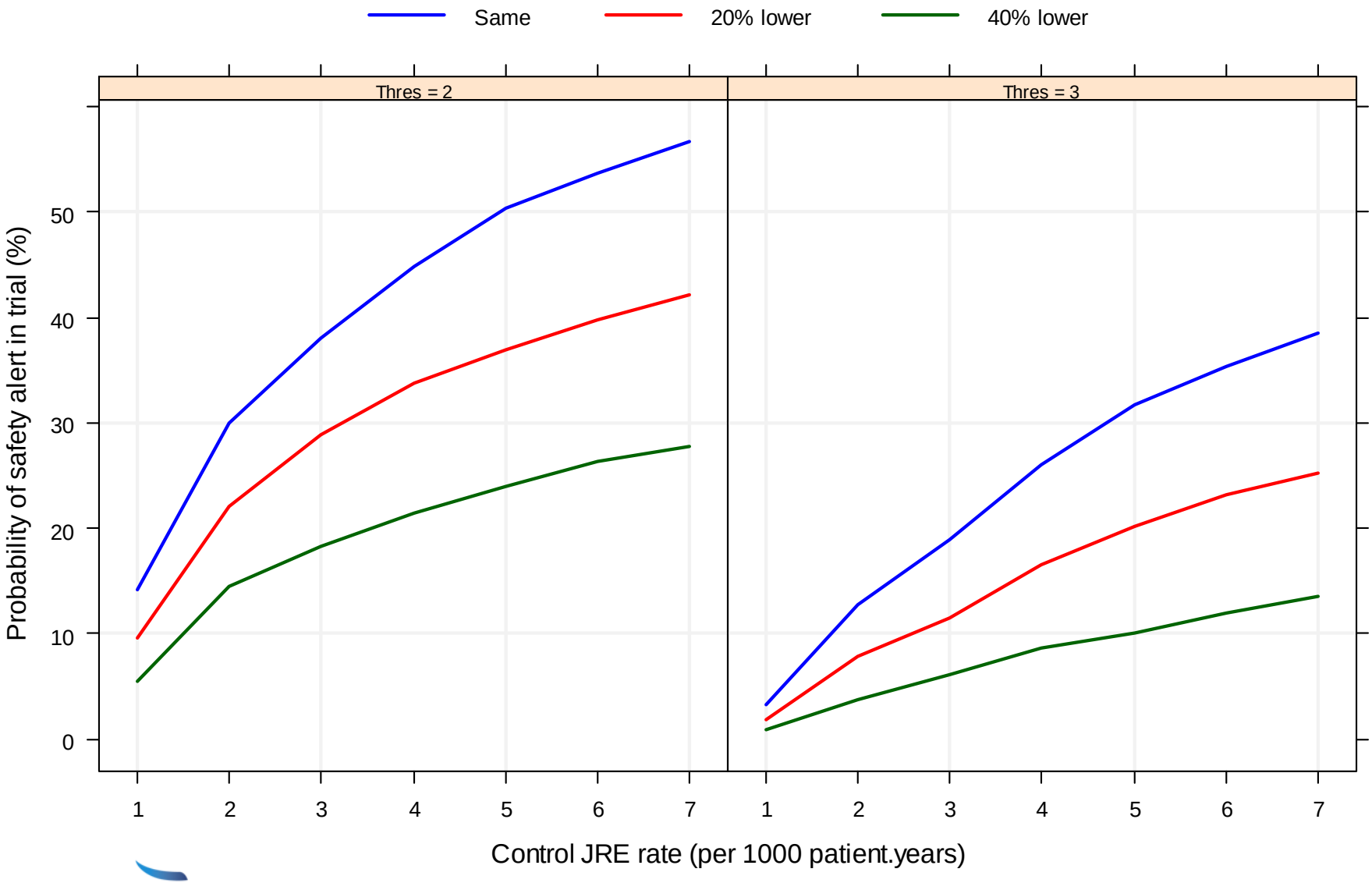
- Biologic (monoclonal antibody) with novel mechanism of action for severe chronic pain (patients not responding to conventional pain drugs)
- Several sponsors developing drugs with similar mode of action (same class), at different devel. phases
- Different pain indications (OA, lower back pain, cancer) – will focus on OA pain only
- Imbalance in number of patients undergoing joint replacement between drug and control groups led FDA to place whole class in clinical hold in Dec 2010
- Ongoing studies were stopped, with exception of cancer pain indication

Case study: OA pain (cont.)

- FDA advisory committee evaluated clinical hold: voted unanimously (21 vs. 0) to lift hold, but recommended putting in place safeguards to protect patients from potential risks
- Sponsors tasked with proposing safety monitoring rules to ensure patient safety:
  - early detection of harmful signal
  - focus on joint related events (JRE), not just joint replacement
- No guidance from FDA on rule's "operating characteristics" ➔ initial FDA feedback: safety monitoring **not** statistical, but clinical issue: rule to be based on simple threshold on difference in number of JRE (e.g., stop if difference ≥ 2)

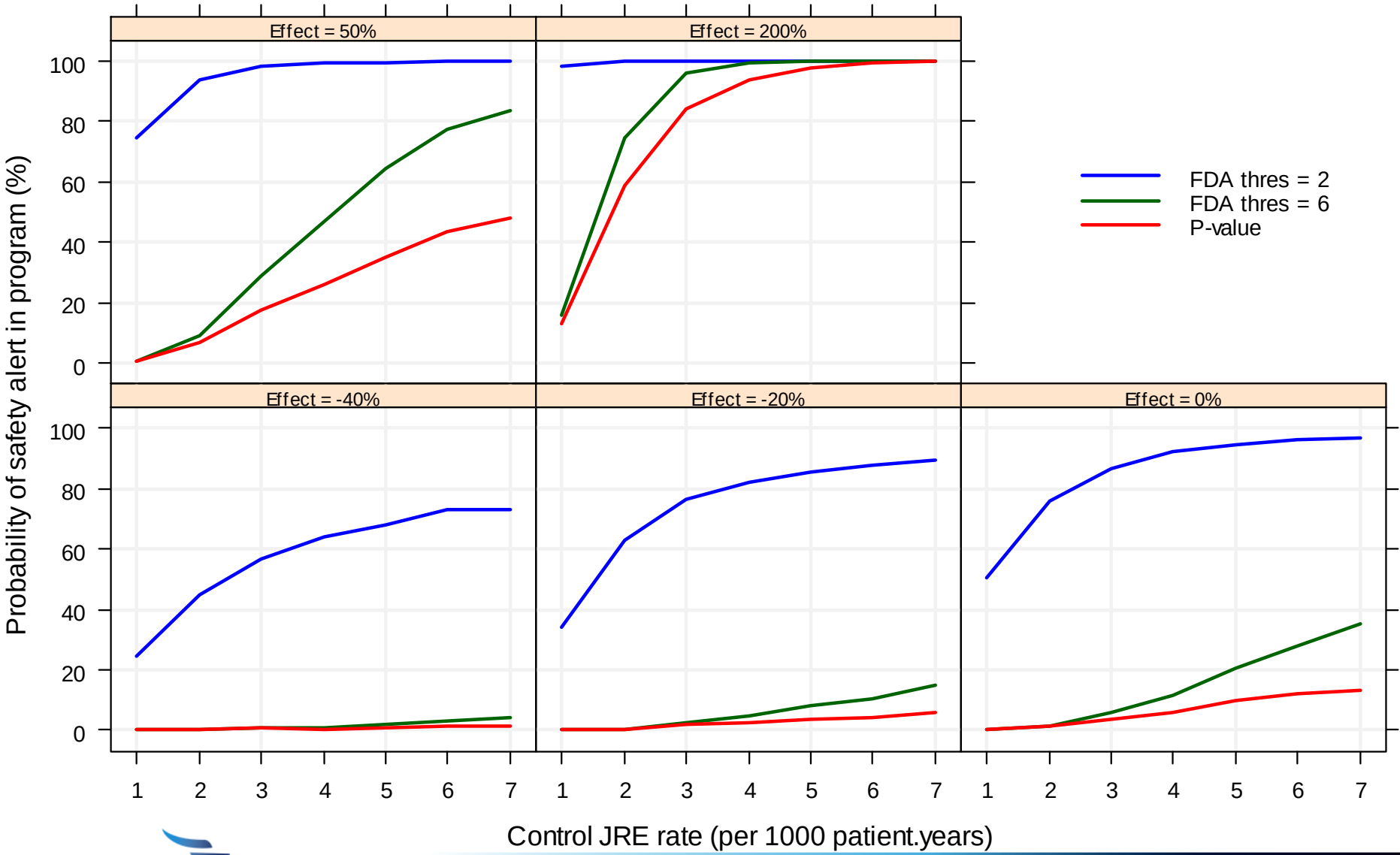
FDA alert rule – trial level

Combined arm, N = 200/arm



FDA alert rule – program level

Combined arm



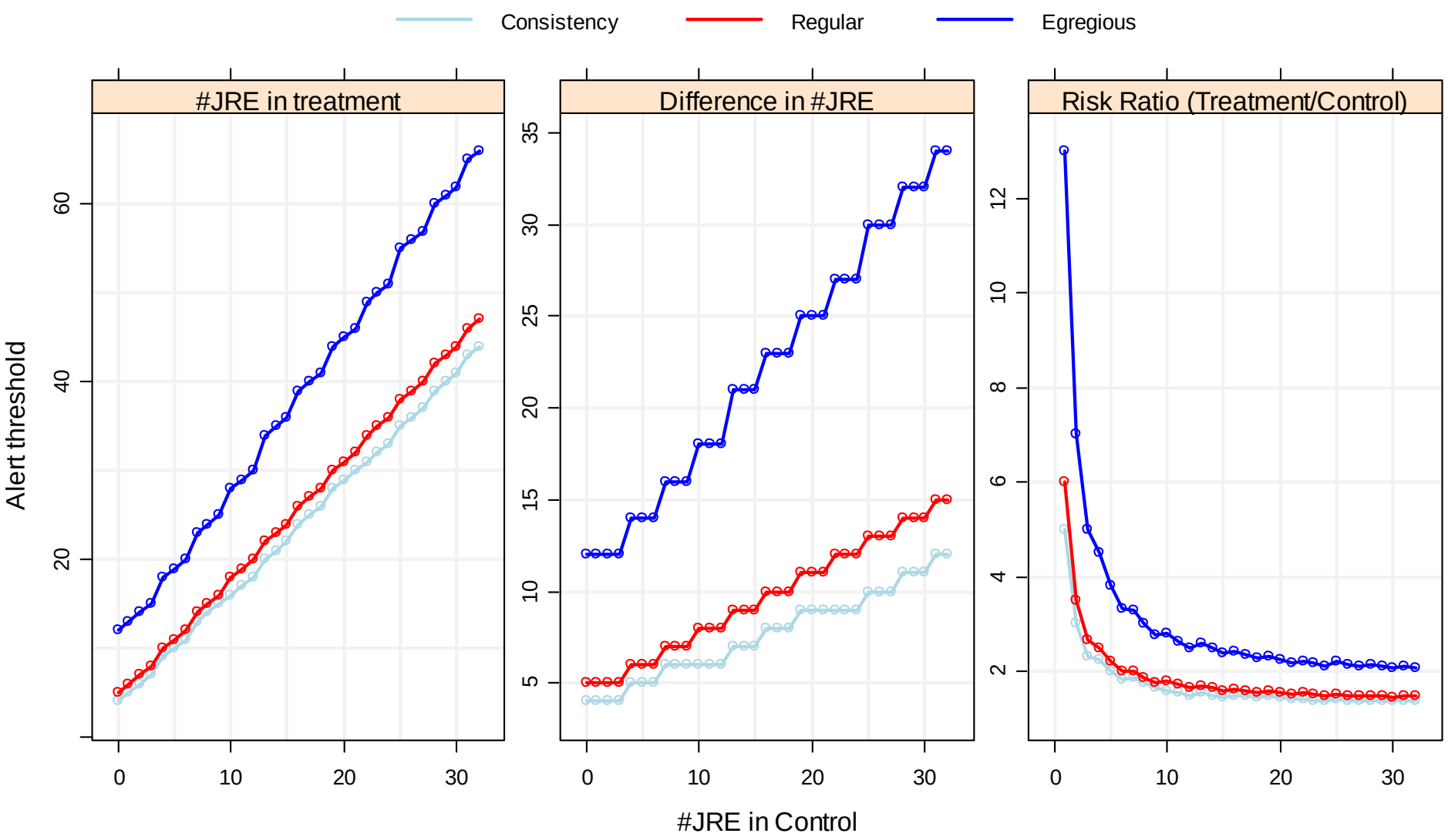
FDA Feedback on safety alert rule

- Cannot be based on hypothesis test (and p-values) – not a statistical problem, but a clinical one
- Agreed that simple threshold = 2 rule is not reasonable, but adamant that any rule needs to be based on difference in number of JRE (between treatment and control)
- No recommendations on desired sensitivity and specificity for rule: sponsor proposes and they react
- Valued simulation-based evaluation of rule operating characteristics and encouraged further discussions on alternative rules to be driven by it

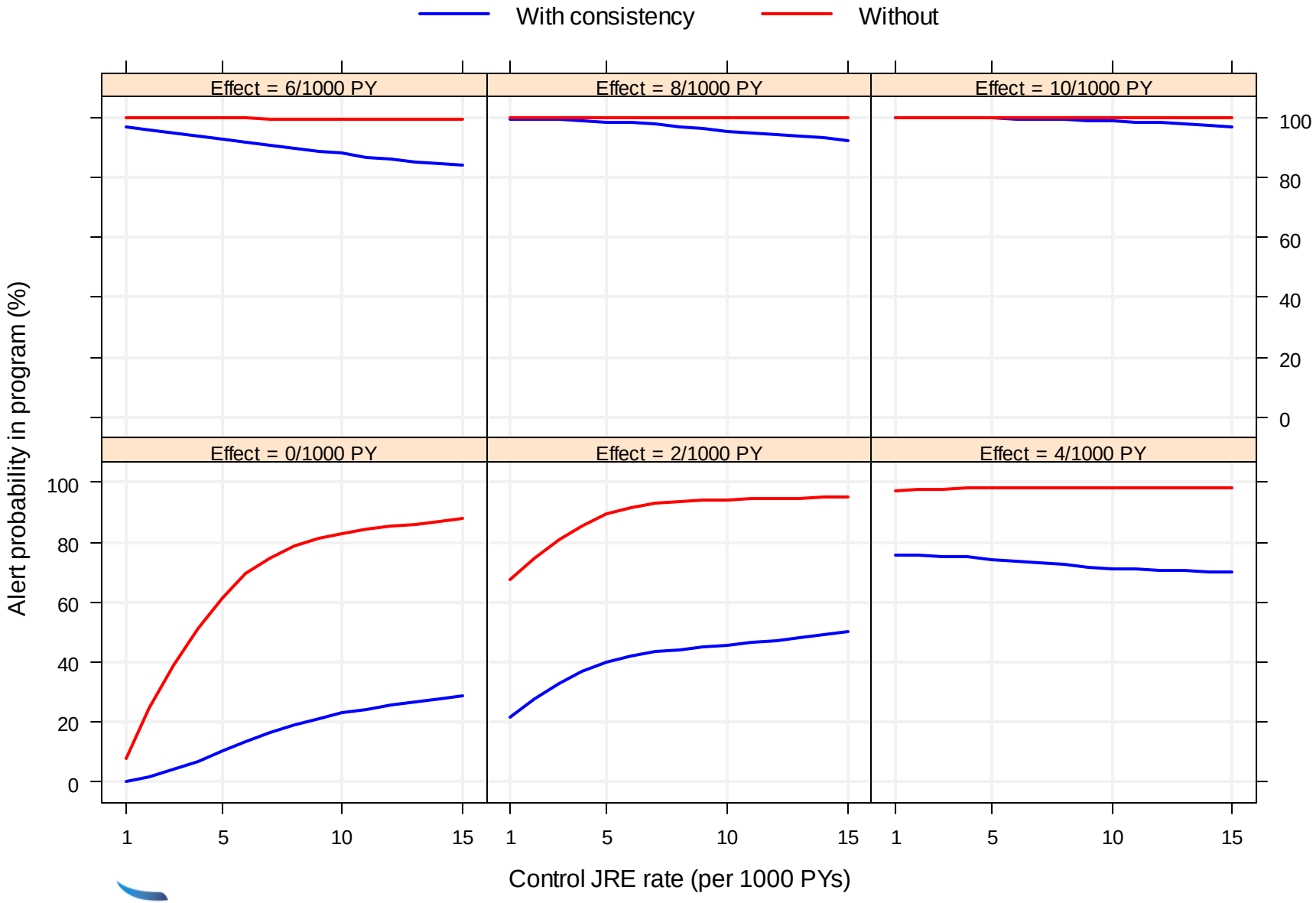
Revised rule: sliding thresholds

- Use step threshold function, increasing with number of JRE in control group (nC)
- Time-driven rule: evaluate every 4 weeks (operationally feasible)
- To reduce false positive at program level, required consistency in alert across trials: if triggered for one study, at least some safety trend seen in concomitant
- Included “egregious” alert that bypasses consistency rule and triggers alert by itself

Alert thresholds



Program alert prob., revised rule



Concluding remarks

- Formal safety detection rules and test are increasingly requested by regulatory agencies as part of development programs (pre-approval)
- Unlike efficacy, for which clear rules and guidance docs have been in place for decades, safety evaluation remains "work in progress" – part of broader discussion of acceptable benefit/risk of drugs
- Complexity of safety alert rules (multiplicity, threshold-based rules, etc) generally requires simulation for proper evaluation: need to quantify different risks (patients and sponsors)
- Needs balance needs between protecting patients vs. making investment in drug development too risky