PSI Journal Club

27th October 2010

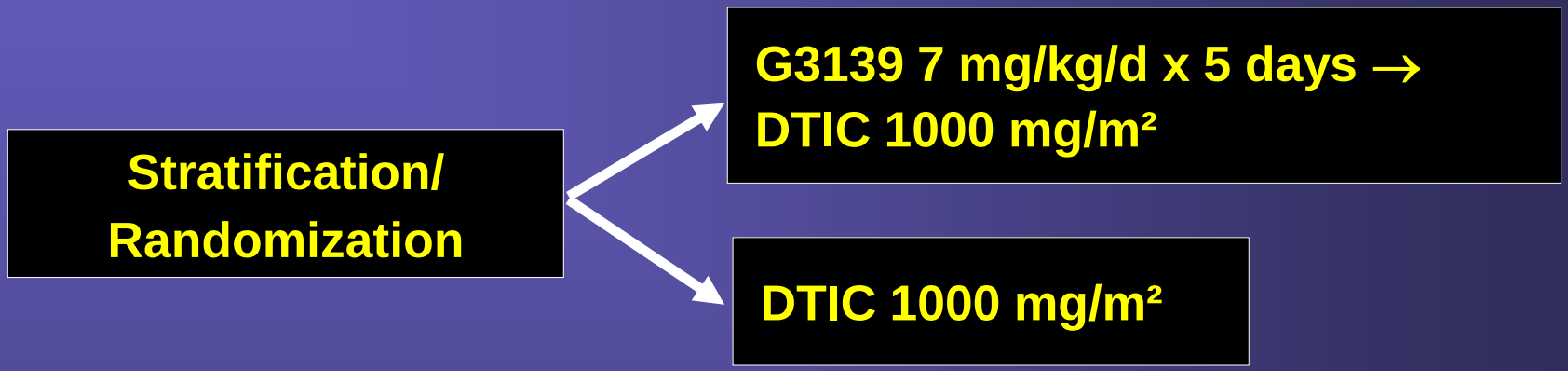
On assessing the presence of evaluation-time bias in PFS in randomized trials

Richard Kay, Jane Wu and Janet Wittes

Presentation Outline

- Background to the Case Study
- Progression Free Survival Results
- Evaluation-time bias
  - Reviewer simulations
  - Alternate assumption simulations
- Sensitivity analyses
- Conclusions

GM301* Study Design



- Sample size: N=771 (386 G+DTIC/ 385 DTIC)
- Cycles every 21 days (maximum of 8)
- Minimum follow-up: 24 months
- Primary endpoint: overall survival
- Secondary endpoints: response rate, progression free survival (PFS)

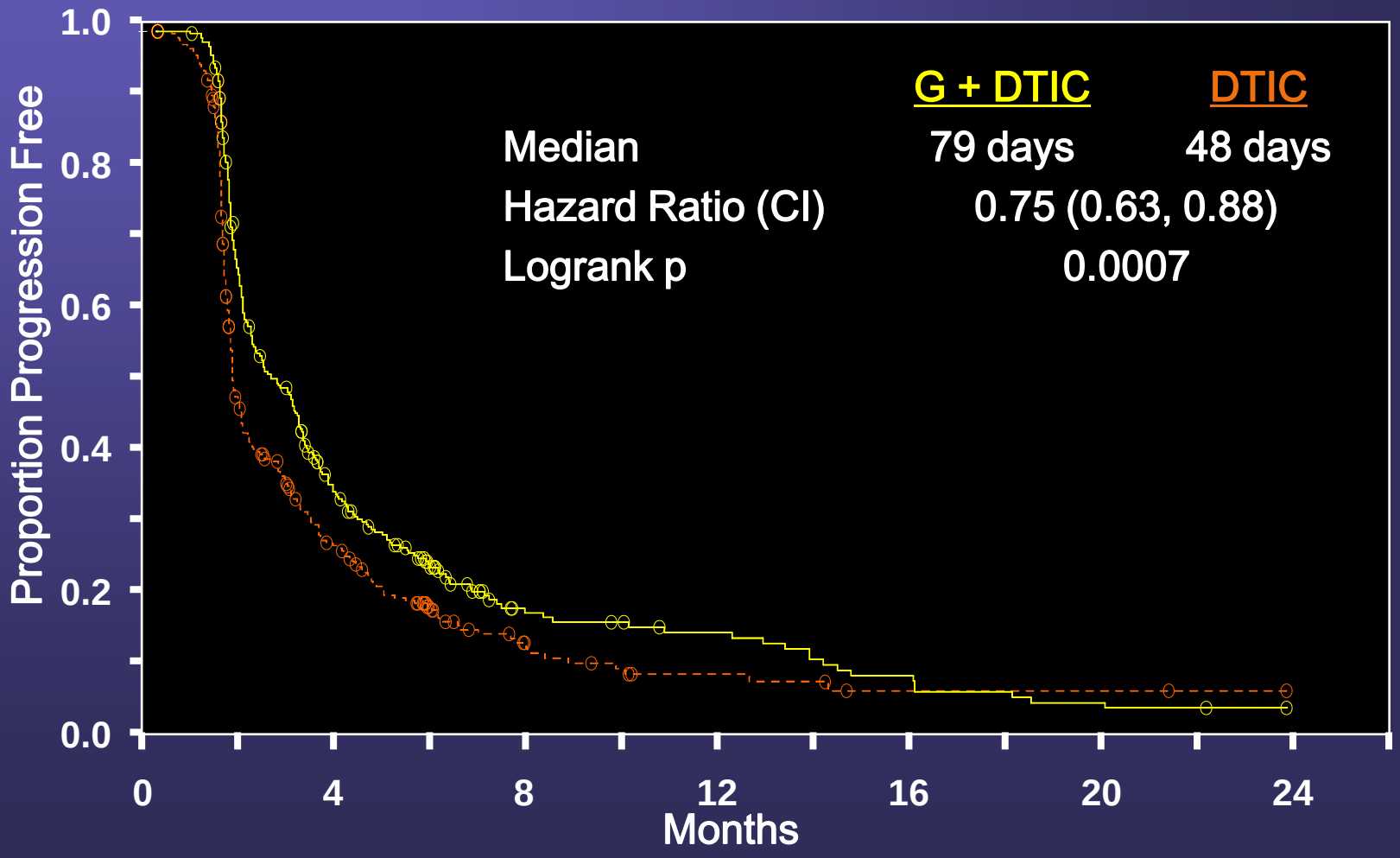
*Bedikian et al., J Clin Oncol, 2006.

Background

- Scans at baseline and every 2 cycles
- Patients could be examined at other time points, usually due to disease progression
- Progression could be declared outside of scheduled assessment times
- NDA filed in 2004 – based on ‘at least 6 months of follow-up’ for all patients
- Concerns* expressed by reviewer regarding presence of evaluation-time bias for PFS

* www.fda.gov/ohrms/dockets

Progression Free Survival 24 months



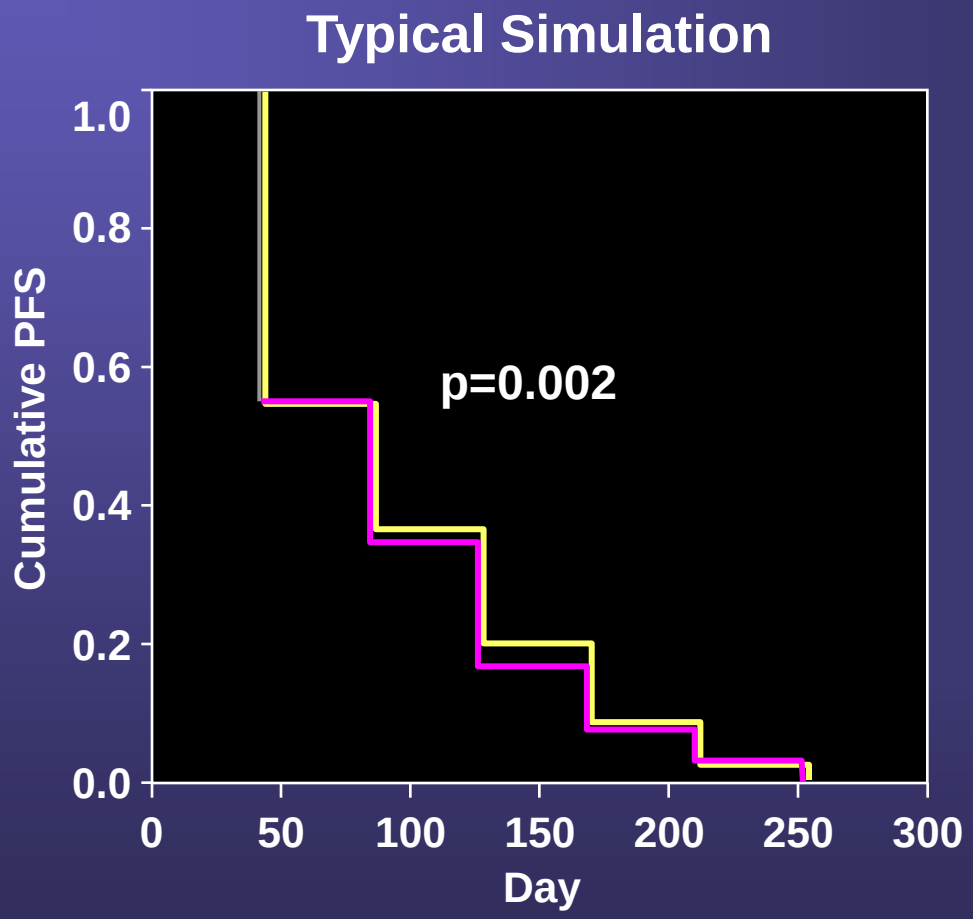
Reviewer Concerns

- Reviewer noted - time to first assessment differed between two treatment groups (median = 48 days on Genasense + DTIC, = 43 days on DTIC alone, logrank $p<0.0001$) – similar differences seen for assessments 2 and 3
- Treatment schedules were different (Genasense + DTIC given over 5 days, DTIC alone is a one hour infusion)
- Reviewer argued that assessments were consequently delayed in the Genasense + DTIC arm causing bias
- Sponsor argued that differences in assessment times were caused by patients progressing more rapidly in the DTIC alone arm and visiting their physicians earlier to report symptoms of progression – trial design had assessment times the same in both treatment groups

Reviewer Simulations

- Reviewer undertook simulation study (Model 1) to evaluate whether differences in assessment timing could explain difference in PFS
- Model 1 assumptions:
  - Distribution of PFS exponential, median PFS = 50 days in both groups
  - First assessment day 44 for Genasense + DTIC; day 42 for DTIC control
  - Subsequent assessments at days 86, 128, … for Genasense + DTIC, days 84, 126, … for DTIC control
- Based on N = 300 per treatment group and 5000 simulations the proportion of simulations giving $p<0.05$ was *98%*

Model 1 Simulations



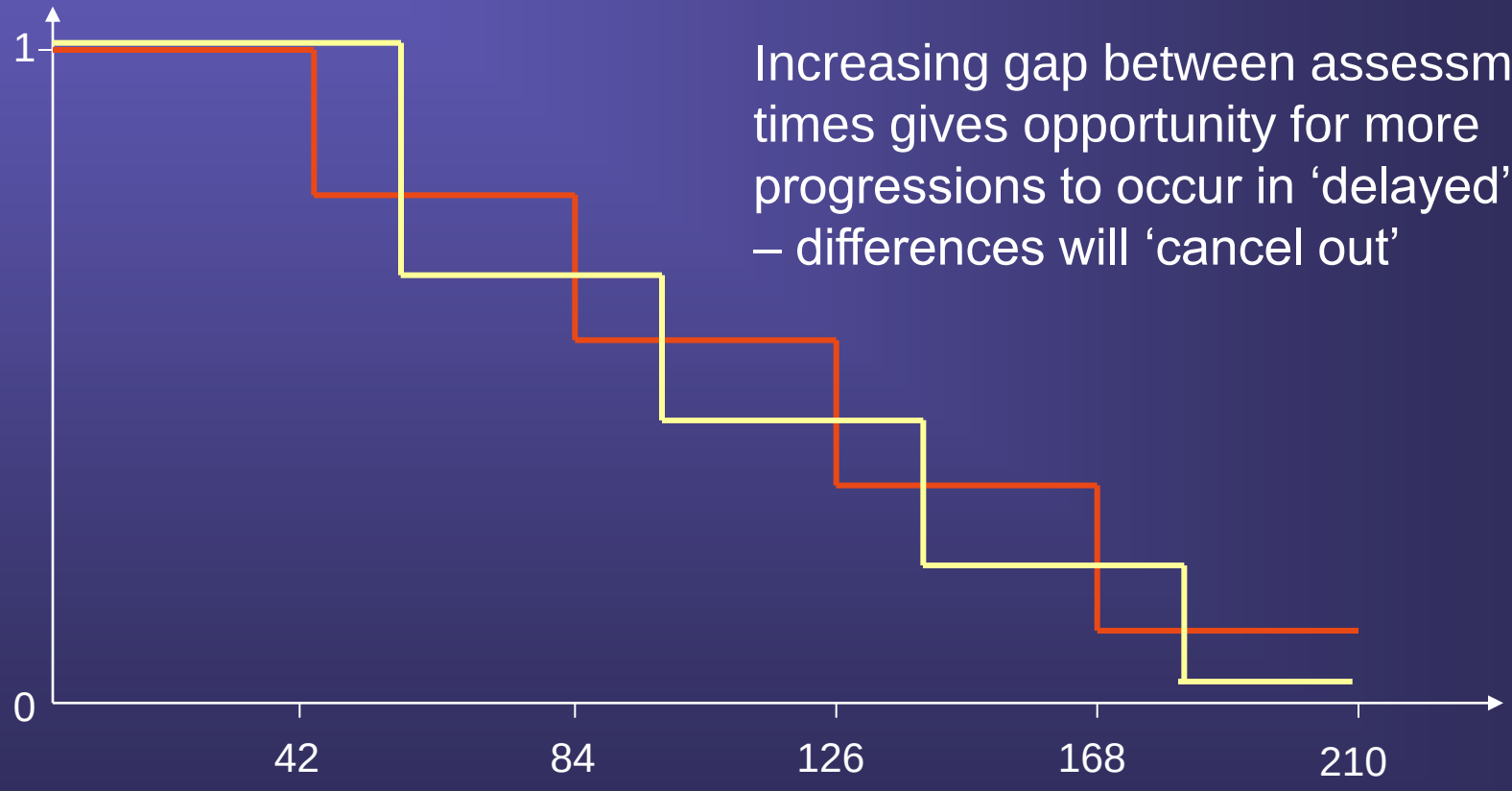
Model 1 Simulations

- PFS curves almost identical – highly significant differences result from properties of the logrank test
- Simulations display bizarre properties –bias *increases* as gap between assessment times *decreases*
  - 21 day gap, false positive rate = 63%
  - 2 day gap, false positive rate = 98%
  - 1/10th day gap, false positive rate = 99%

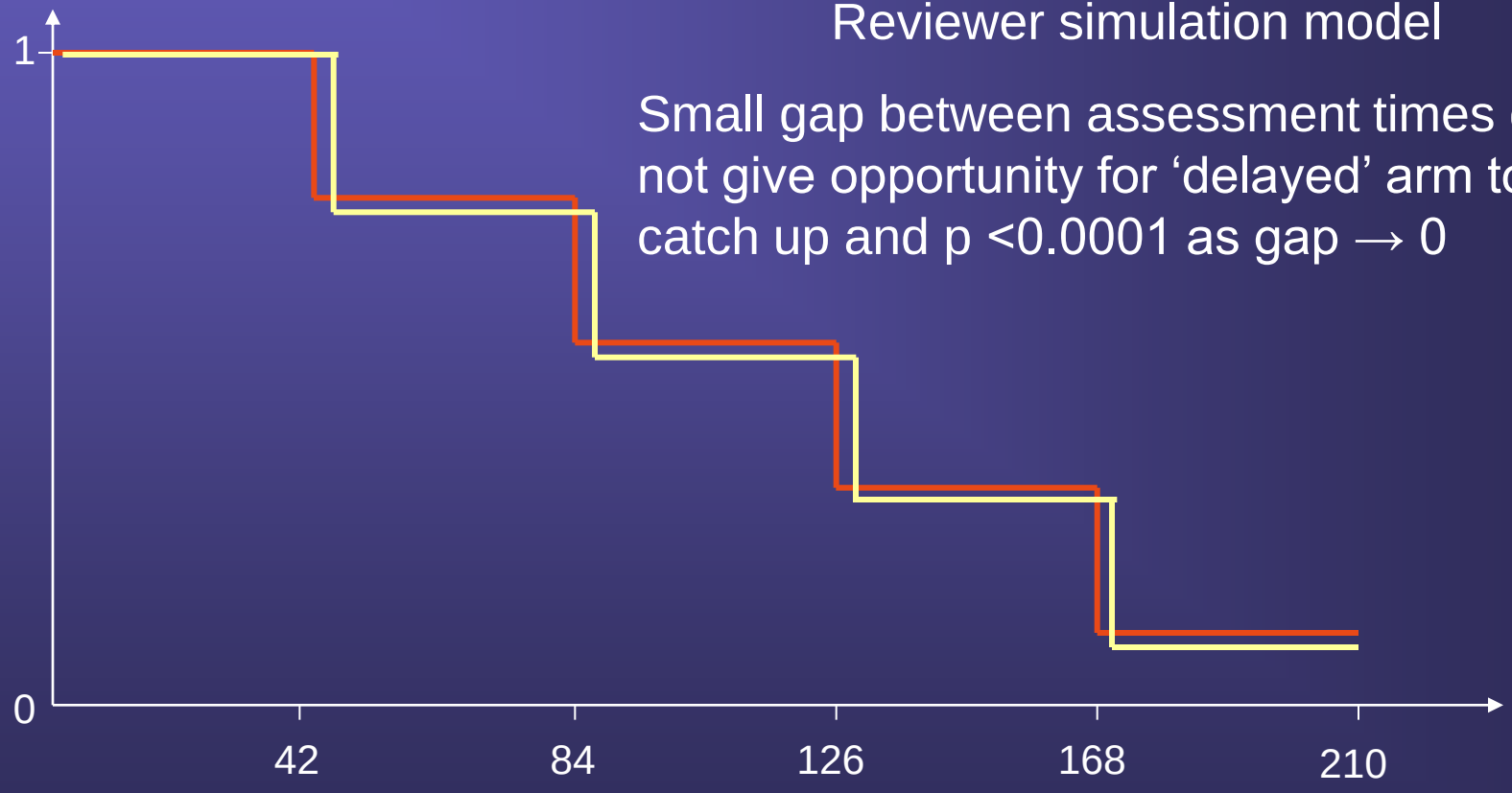
Model 1 Simulations

Reviewer simulation model

Increasing gap between assessment times gives opportunity for more progressions to occur in ‘delayed’ arm – differences will ‘cancel out’



Model 1 Simulations

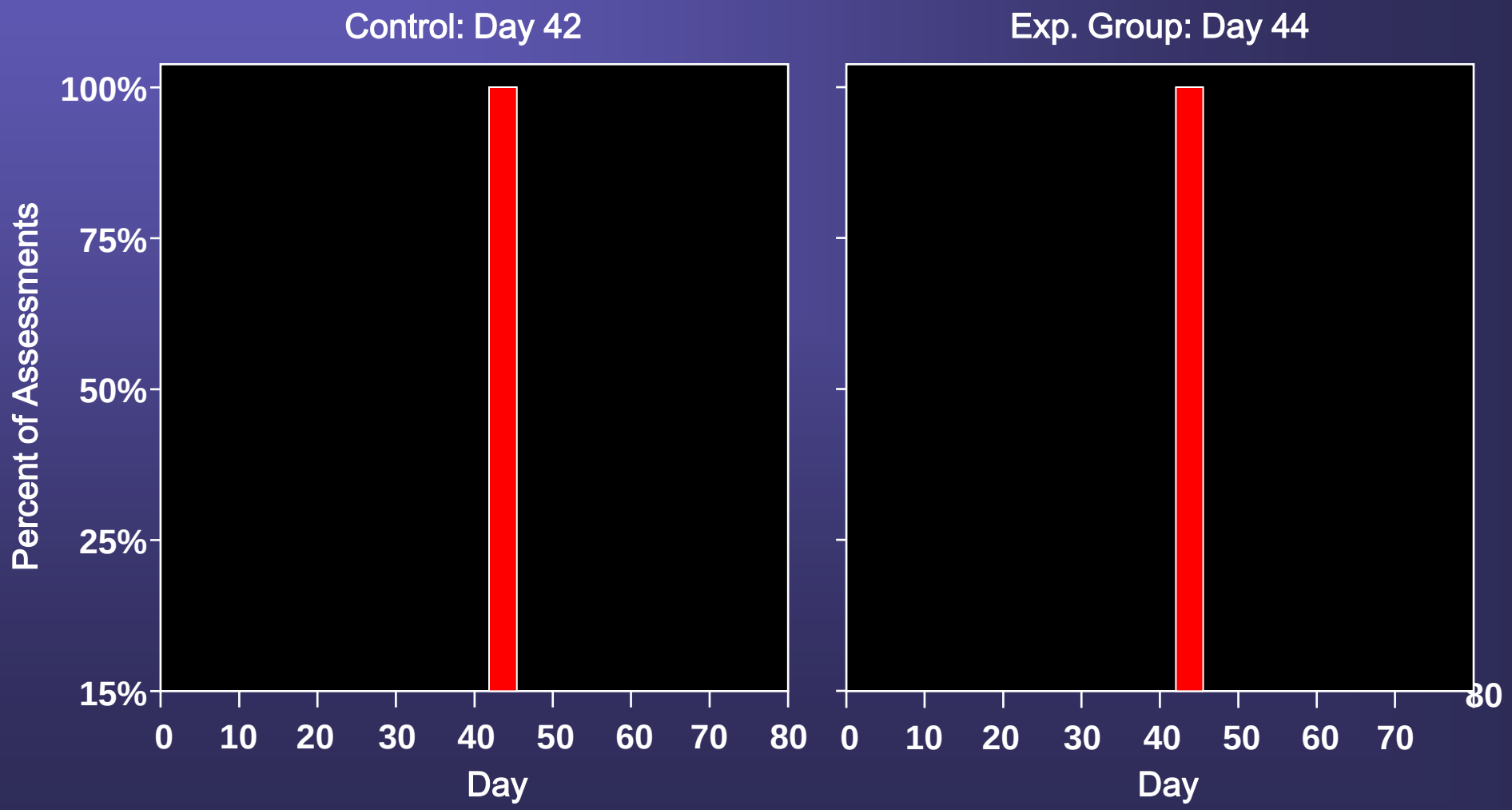


Reviewer simulation model

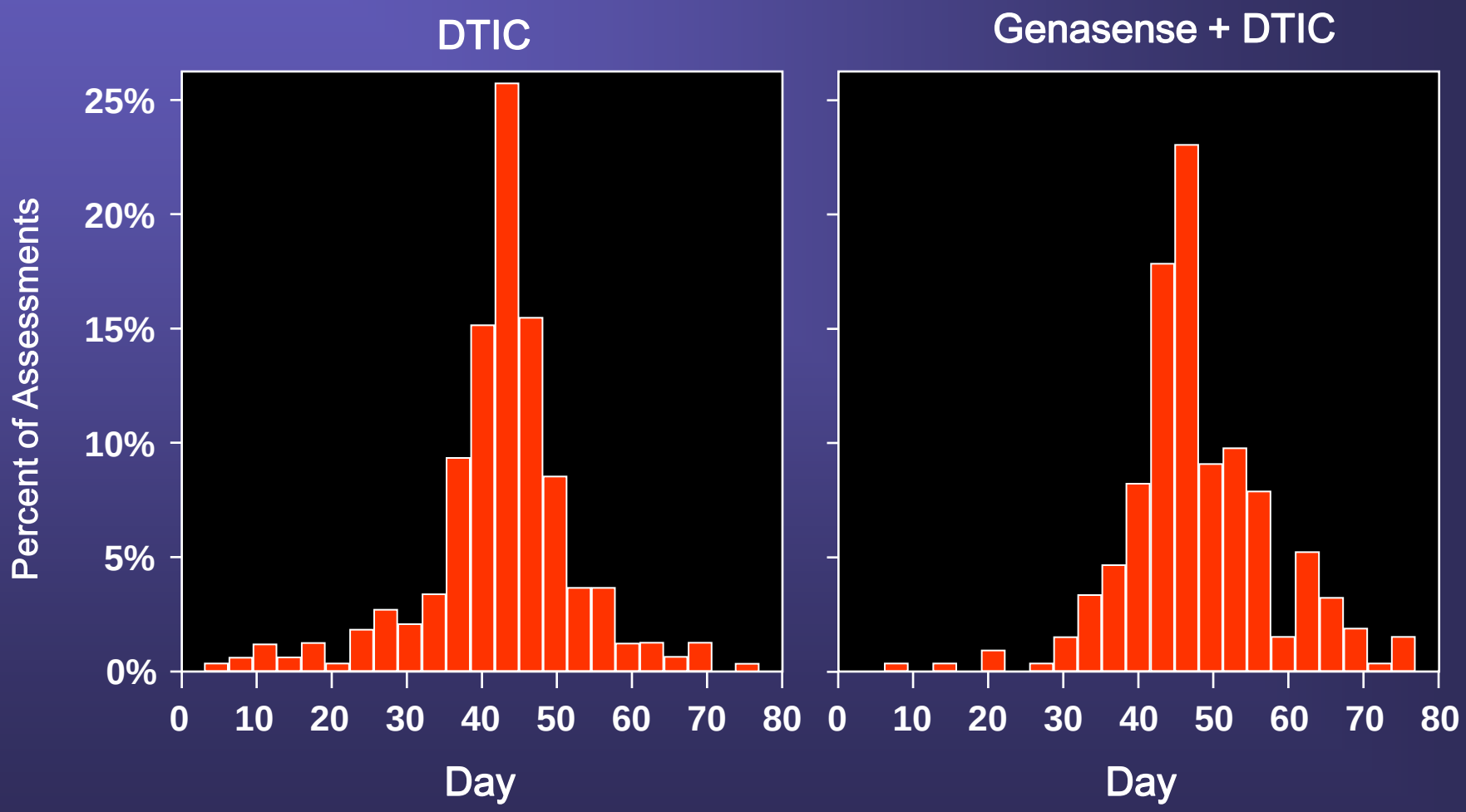
Small gap between assessment times does not give opportunity for ‘delayed’ arm to catch up and $p < 0.0001$ as gap $\rightarrow 0$

Assumptions for Time to First Assessment

(Model 1 Assumption: All Assessments Occur on Same Day; Standard Deviation=0)

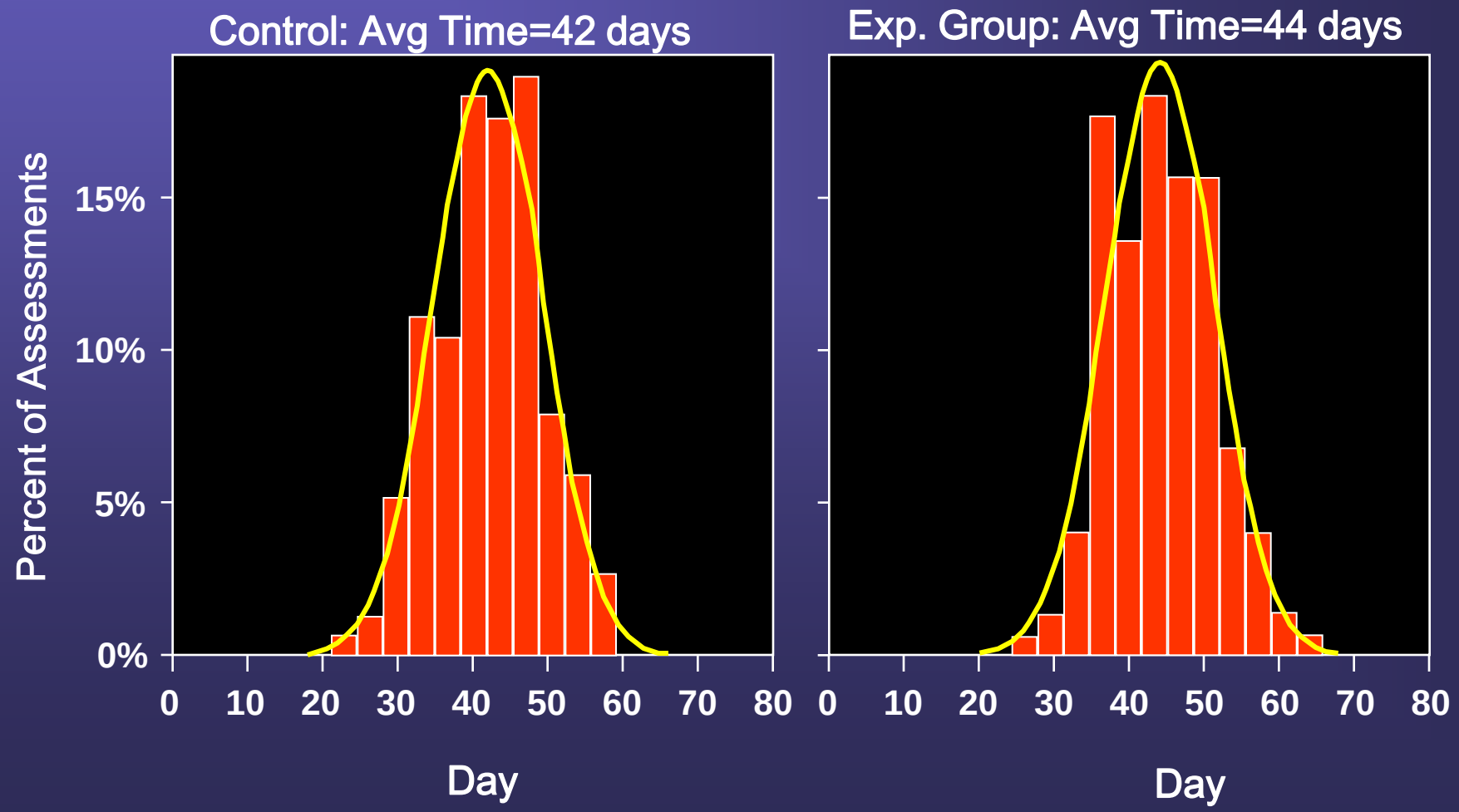


Study GM301- Actual Distribution of Time to First Assessment



Alternative Simulation of Time to First Assessment

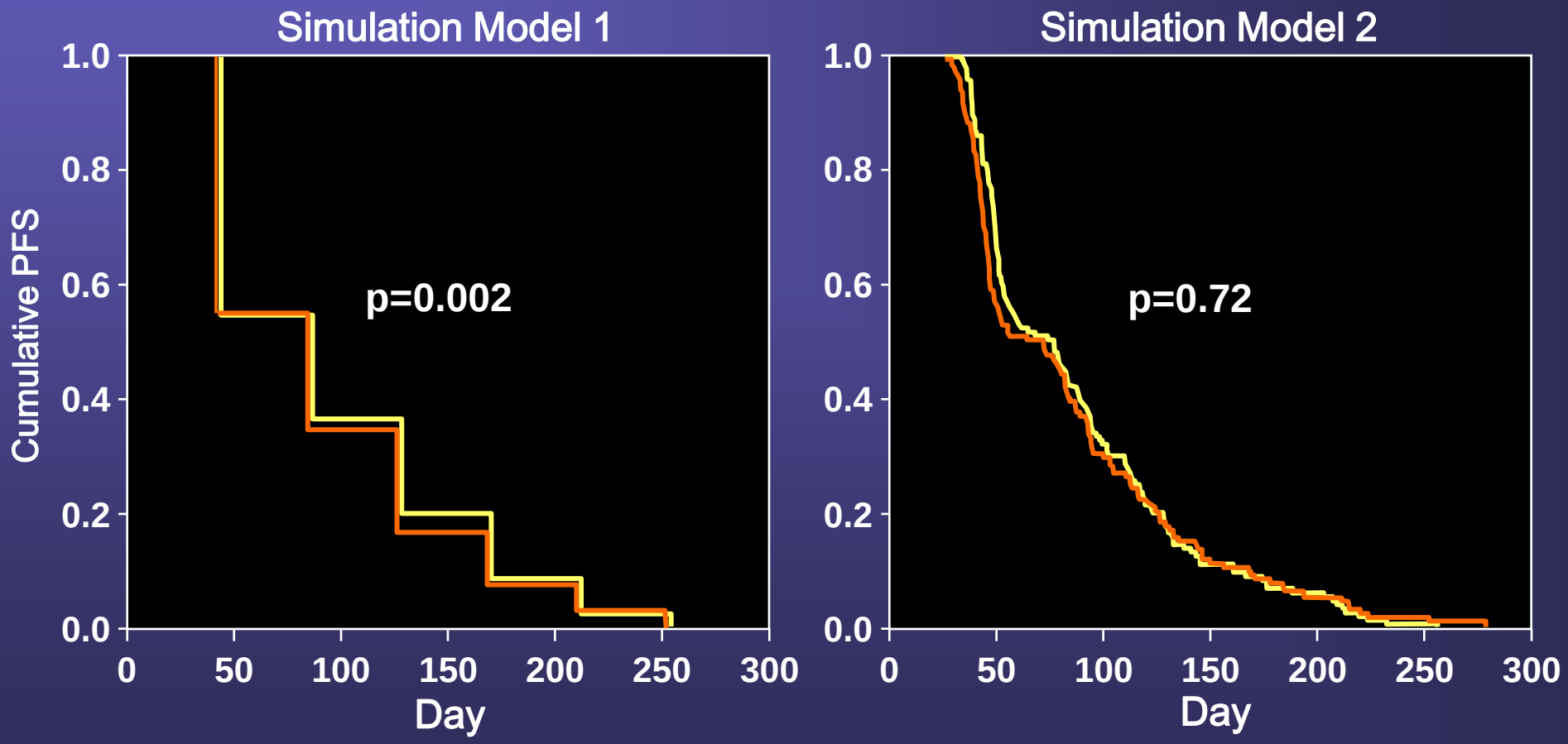
(Model 2 Assumptions: Normal Distribution; Standard Deviation = 10 days)



Model 2 Simulations-Realistic Assumptions

- Alternative simulation study (Model 2):
  - Distribution of PFS exponential, median PFS = 50 days in both groups
  - First assessment day 44 for Genasense + DTIC; day 42 for DTIC control on average; normal distribution, standard deviation = 10 days
  - Subsequent assessments at days 86, 128, … for Genasense + DTIC and days 84, 126, … for DTIC control on average; normal with sd = 10 days
- Based on N = 300 per treatment group and 5000 simulations showed that the proportion of simulations giving $p<0.05$ was *5.7%* - bias almost eliminated

PFS Simulation Models



Simulations

- Simulations based on inappropriate assumptions can be very misleading
- Reviewer simulations (Model 1) presented at 2004 ODAC meeting*
  - *'The simulation results suggested that the chance of falsely inferring treatment differences in PFS could be very large indeed even for slightly different assessment schedules between the two treatment groups'*
  - *'Difference in assessment intervals may explain observed PFS effect'*
- Undermined the positive PFS results for Genasense – major influence in negative ODAC vote

* www.fda.gov/ohrms/dockets

Simulations

- Differences in the scheduling of assessment times between the two treatment arms in GM301 could not have accounted for the observed differences in PFS
- HR = 0.75, p = 0.0007 at 24 months minimum follow-up
- HR = 0.73, p = 0.0003 at 6 months minimum follow-up

Sensitivity Analyses

- Sensitivity analyses are the correct way to address concerns about *evaluation-time bias* resulting from possible differential assessment time strategies– several undertaken for GM301
- PFS re-analysed by classifying each progression according to cycle in which it was observed – statistical significance retained (HR = 0.84, p = 0.048)
- Other sensitivity analyses support robustness of treatment effect

Conclusions

- Reviewer simulations (Model 1) of *evaluation-time bias* are flawed and fail to recognise the behaviour of the logrank test
- Alternative simulations with realistic assumptions (Model 2) show bias minimal
- In specific settings sensitivity analyses correct way to assess possible evaluation-time bias
- Assessment asymmetry could not have accounted for the positive result for PFS with Genasense in GM301