

# Small-Sample Behavior of Long-Memory Phase I Cancer Designs

Assaf Oron (with Peter Hoff)  
University of Washington, Seattle  
Presented at DAE 2009, University of Missouri,  
Columbia, MO  
**October 2009**  
[assaf@uw.edu](mailto:assaf@uw.edu)

## Acknowledgements

Current affiliation: MESA Air Pollution Study and Discover Center, Department of Environmental and Occupational Health Sciences, University of Washington.

- (spatial statistics and epidemiology, PI: Joel Kaufman)

This talk: basic concepts first appeared in Ch. 4 of a dissertation @UW Dept. of Statistics, 2007.

- Dissertation available on arxiv.org
- This is a far more refined version, based on an article draft and a 2009 Challenge Grant application (...)

Thanks to Nancy Flournoy, who invited me to talk here.

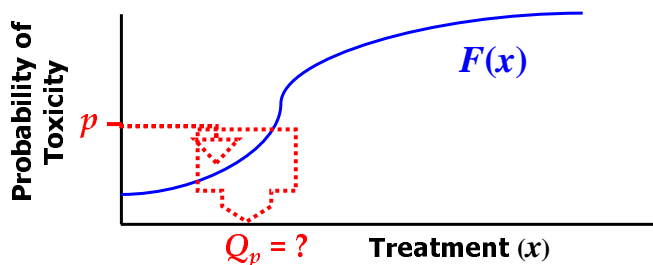
# Talk Outline

1. Overview
2. Demo: How Long-Memory Designs work (using a CRM example)
3. Problems with Long-Memory Designs
4. Workarounds and Potential Solutions

## Phase I Cancer (P1c): Framework

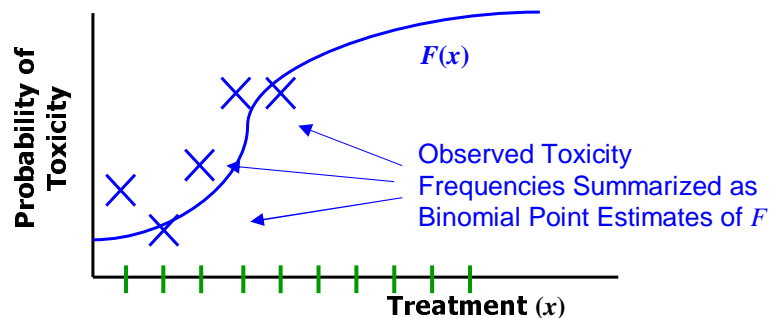
Toxicities dichotomized to “yes”/”no”. Probability for positive toxicity increases with increasing treatment ( $x$ ). **Response thresholds** assumed to have a smooth, increasing CDF  $F(x)$ . We know little else about  $F$ .

**Goal:** find the treatment that would yield a pre-specified toxicity probability (or frequency),  $p$ . Can be seen as a percentile of  $F$ :  $Q_p \equiv F^{-1}(p)$ ; the **target**



## Additional P1c Constraints

- $p$  is typically between 0.2 and 0.35;
- Discrete set of doses (typically ~4-10 levels); we want the one closest to target, a.k.a. the Maximum Tolerated Dose or MTD.
- Small sample size ( $n \approx 10-40$ );
- Trial subjects not a “random sample” by any means.



## Implications of the Constraints

If we spread trials over treatment levels, the Binomial point estimates will converge according to LLN/CLT.

- But this happens quite slowly, requiring hundreds of trials (overall) for reliable estimates.

Another idea: use likelihood-based regression via a parametric model-curve family,  $G(x;\theta)$  for  $F$ :

$$L(x, y; \theta) \propto \prod_{i=1}^n G(x_i; \theta)^{y_i} [1 - G(x_i; \theta)]^{1 - y_i}$$

- This has been the classical approach, and is still dominant.
- But we have poor knowledge of  $F$ ; so  $G$  is mis-specified, perhaps strongly. It is unclear whether pooling data to estimate a mis-specified model's parameters does more harm than good. This debate is somewhat tangential to my talk.

## Solutions: Adaptive Designs

Adaptive designs try to concentrate treatments around target.

Short-Memory designs: next treatment allocation based on recent responses and simple transition rules. Typically, observed toxicities trigger a de-escalation - and vice versa.

Long-Memory designs (LMP1c): at each point, allocation is based on all data via some estimation procedure.

- Since 1990, the bulk of published P1c statistical design-development work is about **long-memory** designs. Yet, nearly all P1c experiments are still run using **short-memory** designs (only 20 LMP1c experiments run in 1991-2006).

– *“...a startling paucity of translation of modern statistical methodology into the design of phase I cancer clinical trials”*  
(Rogatko et al., 2007) – i.e., **“practitioners should listen to us.”**

## Why Long Memory?

LMP1c proponents argue that using all available information for each allocation must be superior to using only part of the information (see e.g., O’Quigley and Zohar 2006, where the terms “Long-Memory”, “Short-Memory” were introduced).

- **Could there possibly be a downside to using all information for each allocation? This talk says yes.**

## Talk Outline

1. Overview
2. **Demo: How Long-Memory Designs work** (using a CRM example)
3. Problems with Long-Memory Designs
4. Workarounds and Potential Solutions

## Introducing CRM

By far the most popular (and most commonly discussed) LMP1c design, is a Bayesian method called Continual Reassessment Method (**CRM**; O'Quigley et al., 1990). CRM uses **a one-parameter model with a prior**.

A one-parameter family cannot be guaranteed to fit an unknown curve. So why is CRM so popular?

Rationale: no need for a fully specified, perfect model, because we are ultimately interested in only a single point.

- The point where  $F(x)$  crosses the horizontal line  $y=p$ . For this, one parameter might be enough.

## How CRM Works

Note: the parametric-model likelihood can be simplified to

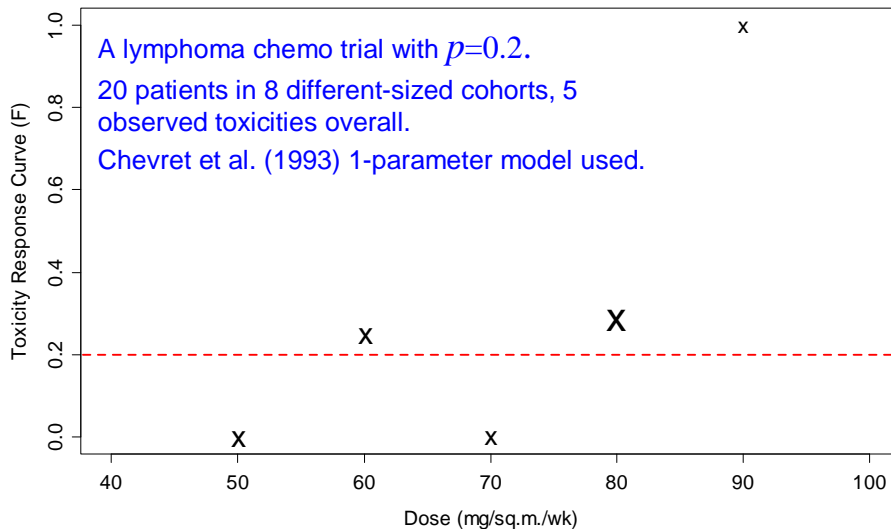
$$L(x, y; \theta) = \prod_{u=1}^m G(d_u; \theta)^{n_u \hat{F}_u} [1 - G(d_u; \theta)]^{n_u (1 - \hat{F}_u)}$$

That is, the available information is essentially concentrated at those doses  $d_u$  where we have observations (nothing P1c-specific here).

Fitting the model is equivalent to running a weighted-best-fit  $G$  curve through the point-estimates  $(d_u, \hat{F}_u)$

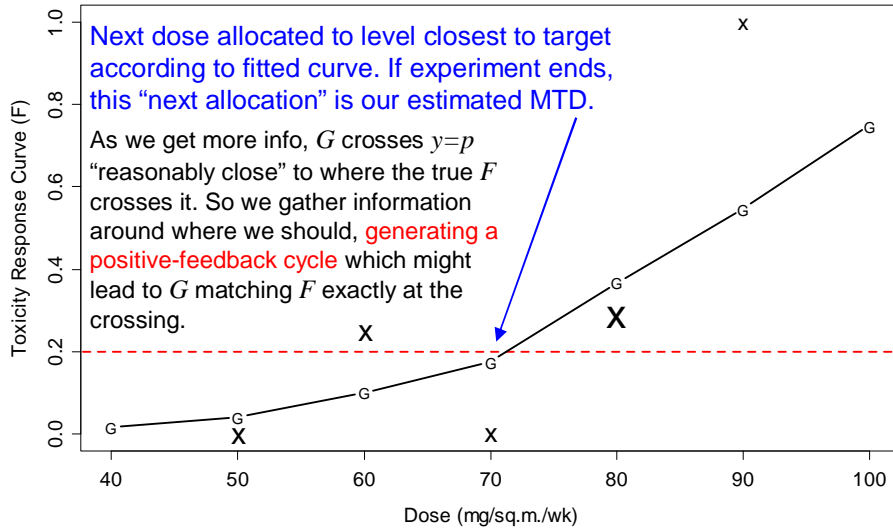
## How CRM Works: Flinn et al. (2000)

Flinn Et Al. (2000) Toxicity Frequencies and Model Fit



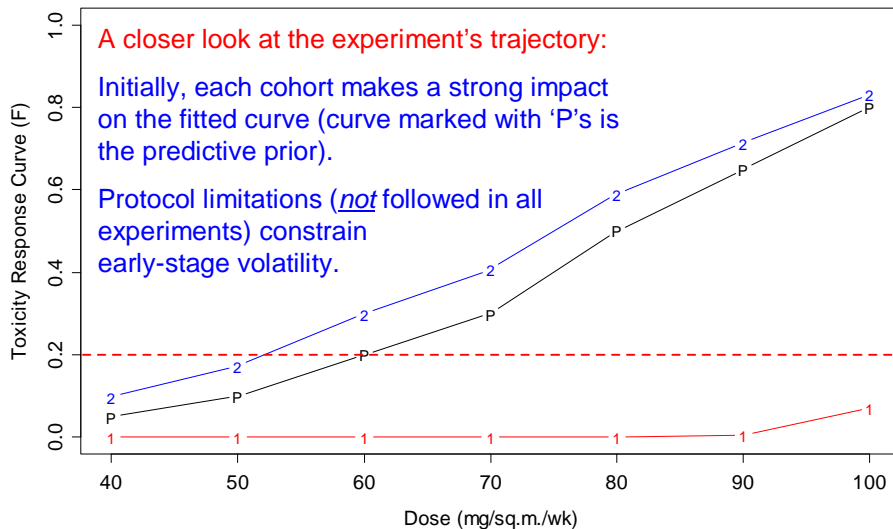
# How CRM Works: Flinn et al. (2000)

Flinn Et Al. (2000) Toxicity Frequencies and Model Fit



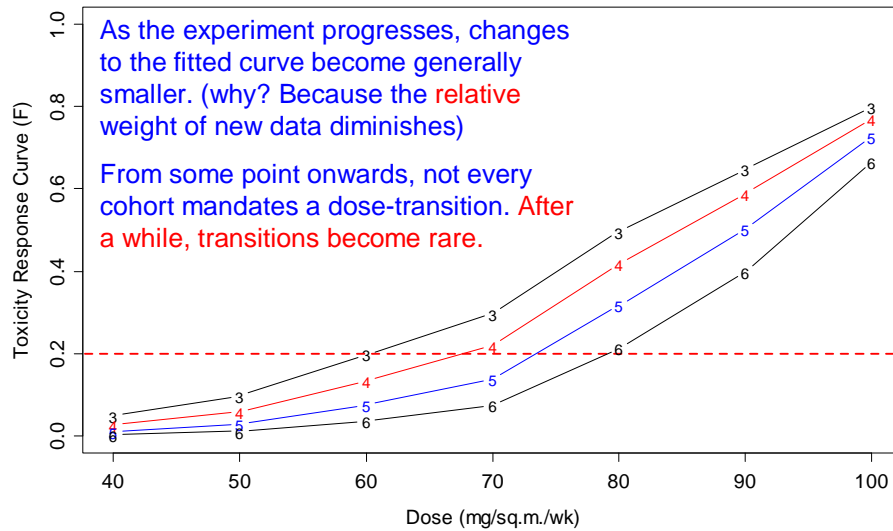
# How CRM Works: Flinn et al. (2000)

Flinn Et Al. (2000) Successive Model Fits



## How CRM Works: Flinn et al. (2000)

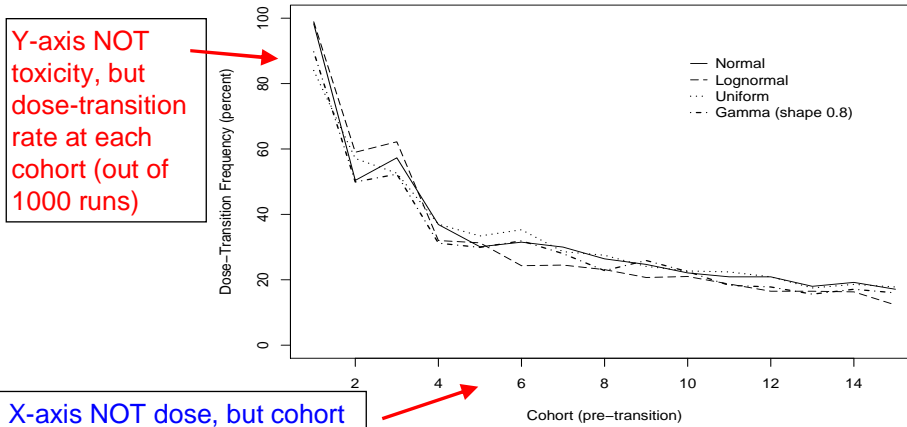
Flinn Et Al. (2000) Successive Model Fits



## Talk Outline

1. Overview
2. Demo: How Long-Memory Designs work (using a CRM example)
3. **Problems with Long-Memory Designs** (shown using 3 somewhat unusual figures)
4. Workarounds and Potential Solutions

## Long-Memory “Pseudo-Convergence”



Shown: simulation data (6 design levels, cohort size 2) with one CRM model and various true  $F$  curves. Dose-transition rate as function of cohort. Initially, dose-transitions are very frequent. Then they become rare, regardless of how well  $G$  fits  $F$ . This is very often mis-interpreted by researchers as “convergence”!

How do we know...

...it's not Convergence?

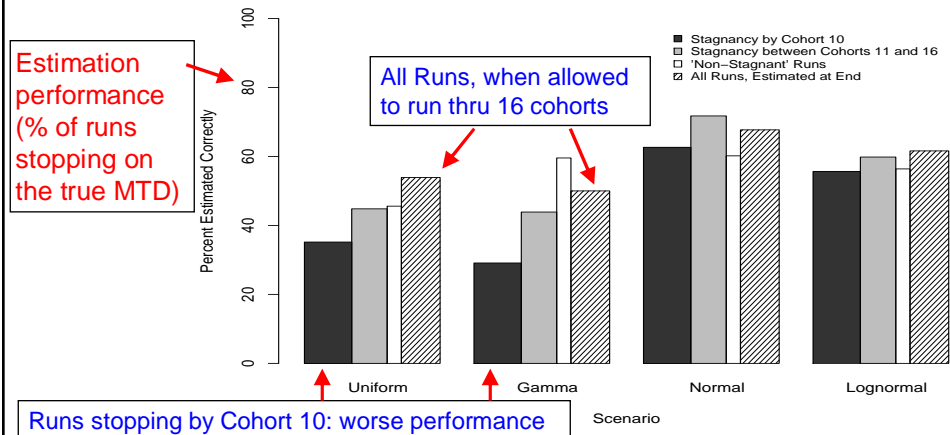
### Theoretical argument:

- Allocation convergence (when it occurs, e.g. the Shen and O’Quigley 1996 proof) is driven by the likelihood, which is driven by **Binomial point frequencies**. There is no way these have converged to their asymptotic values after only 10-20 observations, spread among several levels.

If this argument fails to convince, the question can be easily answered via **simulation**:

- Terminate runs when the same dose has been allocated  $t$  times in a row (this is in fact a stopping rule used sometimes). **If such behavior is indicative of convergence, this stopping rule’s performance should be very good.**

## How do we know it's not Convergence?



Same simulation as before (up to 16 cohorts of size 2 each, 1000 runs per ensemble). Runs terminated upon the 4<sup>th</sup> consecutive cohort at same dose – or after cohort 16 (if this doesn't happen by then).

- Note: “pseudo-convergence” behavior varies little between scenarios (recall previous figure), but **actual estimation performance varies a lot!**

## But perhaps this is the Best We Can Do?

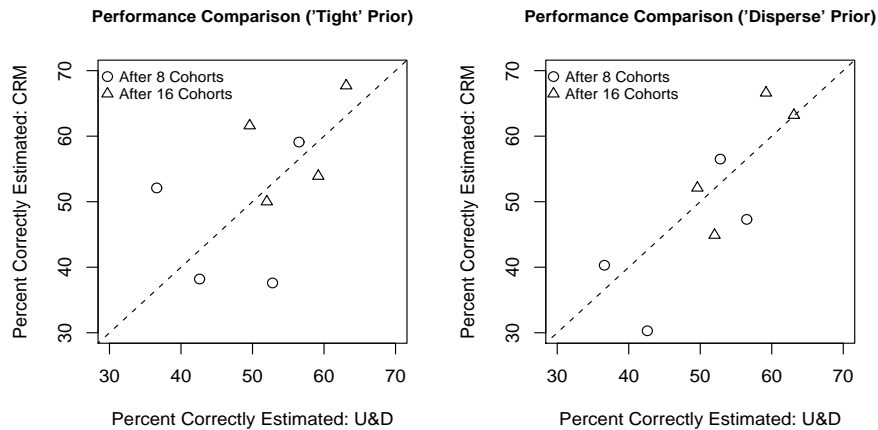
So far, **we have only compared CRM with itself**. Perhaps **it is still a drastic improvement** over Short-Memory designs?

- Perhaps. The most commonly used P1c design is a short-memory protocol called ‘3+3’. It has excellent toxicity control, but very poor and rather intractable estimation performance.

‘3+3’ was probably inspired by **Up-and-Down (U&D)**, a Short-Memory family of methods that has been around for two generations (Dixon and Mood, 1948).

- U&D has fixed transition rules based on recent responses. It converges quickly (geometrical rate) to **a stationary random walk around target**.
- U&D properties are much better understood, and are overall superior to ‘3+3’. **Most importantly: while U&D dose-transition decisions only use recent data, the final estimate – unlike ‘3+3’ – does use all data (via averaging or “isotonic regression” – Stylianou and Flournoy, 2002).**

## CRM vs. U&D Performance



Same simulation as before. U&D design has same cohort size as CRM (2).

- Overall performance roughly equivalent, but U&D more robust. CRM sensitive not only to scenario, but also to prior (see right frame vs. left; more about that soon).

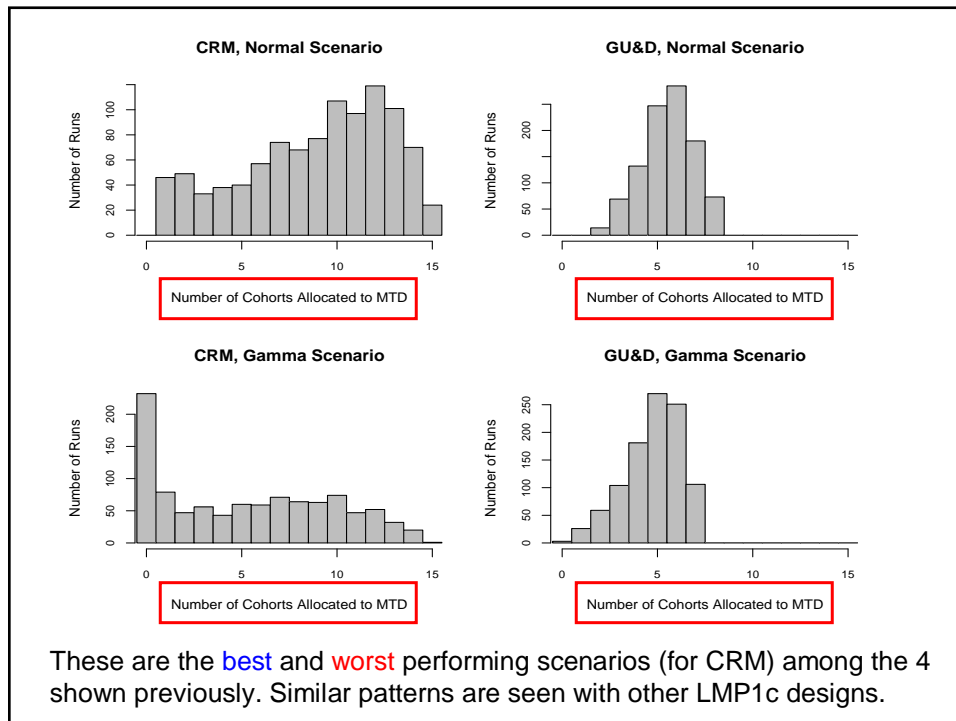
## LMP1c Run-to-Run Variability

If we put these 2 Long-Memory properties together...

- Settling early on a single dose, and staying there for a long while;
- Sometimes this is the right dose (i.e., the MTD) – but sometimes not;

...we conclude that Long-Memory designs should exhibit high run-to-run variability within the same scenario (because different runs settle on different doses).

- Therefore, we expect the total number of cohorts allocated to the correct MTD to vary greatly under Long-Memory designs, between runs under the same scenario. “% of trials at true MTD” is a commonly-quoted success criterion in P1c simulations, but is always reported in bulk. On the next slide, we observe its distribution broken down by run.



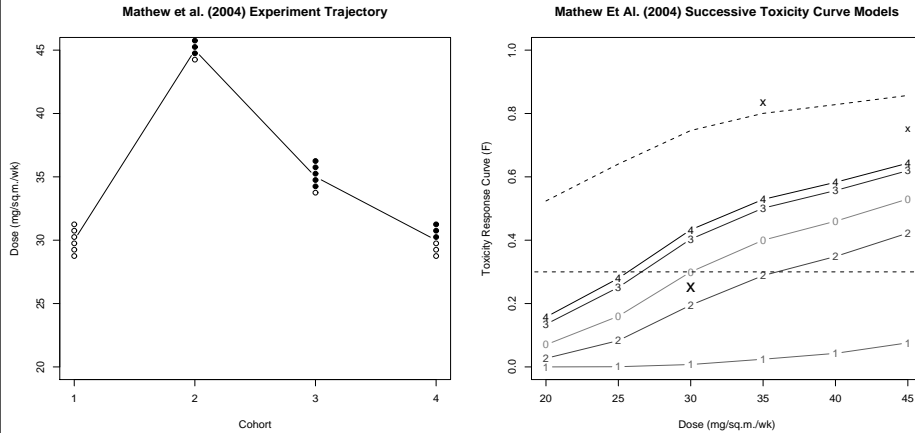
## LMP1c Sensitivity: Recap

Long-Memory’s “pseudo-convergence” property implies **a great sensitivity to early cohorts**. If these cohorts are well-behaved, we will probably do fine. If not, things might go badly. This property is shared by all LMP1c designs, including nonparametric ones.

**Parametric Bayesian designs** like CRM, are also...

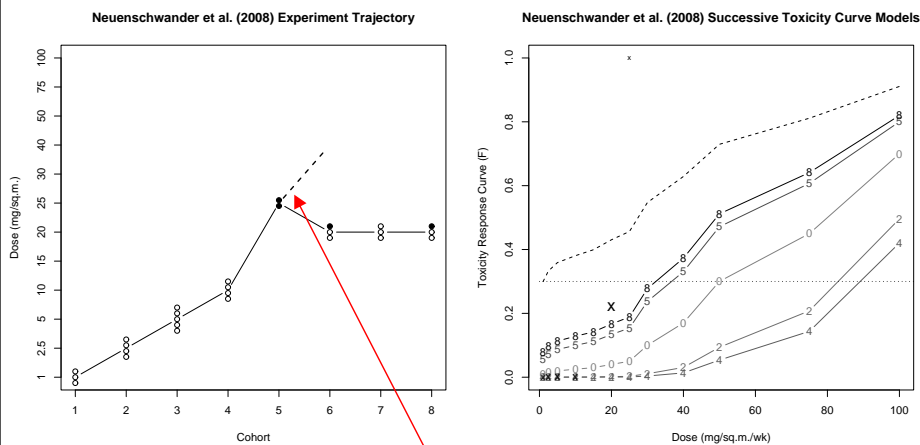
- ...sensitive to prior-predictive weight: if true MTD is up/down-weighted in the prior, performance will be good/poor;
- ...sensitive to model constraints, which inevitably affect and limit model-fitting response.
- For Bayesian designs, when all 3 “stars” (early cohorts, predictive-prior weight, model fit) are aligned, we can get spectacular performance. Otherwise...

## Are CRM 'Disasters' seen in Practice?



Some protocol issues here (dose jumps allowed, huge cohorts). However, the main problem is that **cohort 1 data are at odds with all the rest** – and because of the long memory, the effects linger. **G being so shallow** doesn't help either.

## Are CRM 'Disasters' seen in Practice?



Left, dashed line: original "counter-intuitive" CRM allocation decision after cohort 5. At that point, **model and design were both replaced**.

- Authors' choice of prior toxicity rates constrained  $G$  to be **very shallow up to a dose of ~30**; hence the recommended jump in spite of toxicities.

## Talk Outline

1. Overview
2. Demo: How Long-Memory Designs work (using a CRM example)
3. Problems with Long-Memory Designs
4. **Workarounds and Potential Solutions**

## Long-Memory Workarounds

### Outside CRM:

The second-most-popular Bayesian P1c design is Escalation With Overdose Control (EWOC, Babb et al. 1998).

- Two parameters provide a much better local fit (the discrete dose design seems to mandate 2 parameters);
- Model flexibility comes at the price of a more sluggish early-stage response (prior is “heavier”);
- A (low) posterior-percentile decision rule alleviates some CRM issues - but not the inherent early-cohort sensitivity.

Other LMP1c subtypes have been suggested; like EWOC, they resolve some issues but not the early-cohort sensitivity.

## Long-Memory Workarounds (2)

### Within CRM:

- The shape of  $G$  is directly controlled by prior probabilities at  $d_u$ . Seasoned CRM designers apparently “engineer” these probabilities not according to scientific knowledge, but rather in order to produce an initial behavior similar to ‘3+3’. Thus, a plausible early stage gives way to eventual “convergence” – and then victory can be declared.
- However, these recipes have not been presented and publicized at the same level as the original methods sans recipes. **Someone just reading the literature has no inkling that they exist, and may eventually run into trouble.**
- The original LMP1c rationale of “letting the data play the lead role”, is to a large degree sacrificed in return for some measure of robustness.

## How to Resolve these Issues?

First, recall some **good** things about LMP1c:

- Can accommodate variable cohort sizes and any target percentile (the best-in-class U&D designs cannot);
- **Does have the ability to eventually use all information - although this, as we've seen, should be done with great care.**

Early in the experiment it is probably safer to rely on a Short-Memory design ('3+3', U&D, etc.).

- Two-stage designs that switch from Short- to Long-Memory after the 1<sup>st</sup> observed toxicity have been suggested (e.g. Storer, 2001). From what we have seen here, this transition point is too early and too abrupt.
- **At JSM '07 and in my dissertation (Ch. 5), a hybrid design was suggested, with the Short-Memory allocation overridden by Long-Memory only when supported by evidence.** Simulation studies show it provides moderate performance improvement while retaining Short-Memory's robustness. This design seems to work best with a nonparametric “interval-based” Long-Memory component (Ivanova et al. 2007), rather than with CRM.

## Finally:

It is quite possible, that for many P1c applications the best choice will simply be a **Short-Memory design, brought up-to-date with recent theory**. Since these designs have received scant method-development attention since the early 1970's, there is plenty of room for improvement there.

More fundamentally, the statistical community should communicate effectively to practitioners **the limitations** of what can be achieved with 20-odd binary observations on a non-representative sample. **This can only help the field.**

- As a didactic example, consider weather forecasts: not so long ago, popular culture toyed with the idea that humans would eventually control their weather. Now, the public has come to accept that even short-term forecasts have very limited abilities.

## References

- Babb et al., *Stat. Med.* 17 (1998), 1103-1120
- Dixon and Mood, *JASA* 43 (1948), 109-126
- Flinn et al., *Ann. Oncology* 11 (2000), 691-695
- Ivanova et al., *J Stat. Plan. Infer.* 137 (2007), 2316-2327
- Mathew et al., *J Clin. Oncology* 22 (2004), 3323-3329
- Neuenschwander et al., *Stat. Med.* 27 (2008), 2420-2439
- Oron, *Ph.D. Dissertation* (2007, arxiv.org)
- O'Quigley and Zohar., *Brit. J Cancer* 94 (2006), 609-613
- O'Quigley et al., *Biometrics* 46 (1990), 33-48
- Rogatko et al., *J Clin. Oncology* 25 (2007), 4982-4986
- Shen and O'Quigley, *Biometrika* 83 (1996), 395-405
- Storer, *Stat. Med.* 20 (2001), 2399-2408
- Stylianou and Flournoy, *Biometrics* 58 (2002), 171-177