DISCUSSION AT HEI PANEL ON REPRODUCIBLE RESEARCH

Richard L Smith

Department of Statistics and Operations Research and Department of Biostatistics, University of North Carolina at Chapel Hill

and

Statistical and Applied Mathematical Sciences Institute

HEI Annual Conference

Chicago, April 30, 2018

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

samsi
NSF Duke NCSU UNC

OVERVIEW

- I. Some thoughts on the recent EPA Proposed Rule: Strengthening Transparency in Regulatory Science
- II. Technical discussion: Reliability of Inference from Perturbed Datasets
- *The usual disclaimer: what I say here are my own personal views and not the official position of my employer or any organization I work with*

I. Some thoughts on the recent EPA Proposed Rule: Strengthening Transparency in Regulatory Science

- Intent is “to strengthen the transparency of EPA regulatory science” by “ensur[ing] that the data underlying [scientific studies] are publicly available in manner sufficient for independent validation….. in a fashion that is consistent with law, protects privacy, confidentiality, confidential business information, and is sensitive to national and homeland security.”
- What’s wrong with that?

Ambiguities in the wording

- “Data (where necessary, data would be made available subject to access and use restrictions)”
  - How does this apply to Medicare data? Is this already “available” within the terms of the rule?
- “Where data is controlled by third parties, EPA shall work with those parties to endeavor to make the data available…”
  - What if the “third party” is a university that cites IRB and human subjects policies to deny EPA’s request?
- “EPA shall implement the provisions…in a manner that minimizes costs”
  - But there will still be costs – how will these be budgeted?
- “The [EPA] Administrator may grant an exemption if … it is not feasible to ensure … data … is publicly available…”
  - What does “feasible” mean in practice? How would this provision apply to a university with its IRB protocols?

We clearly need more information, but my initial advice would be not to rush to judgment

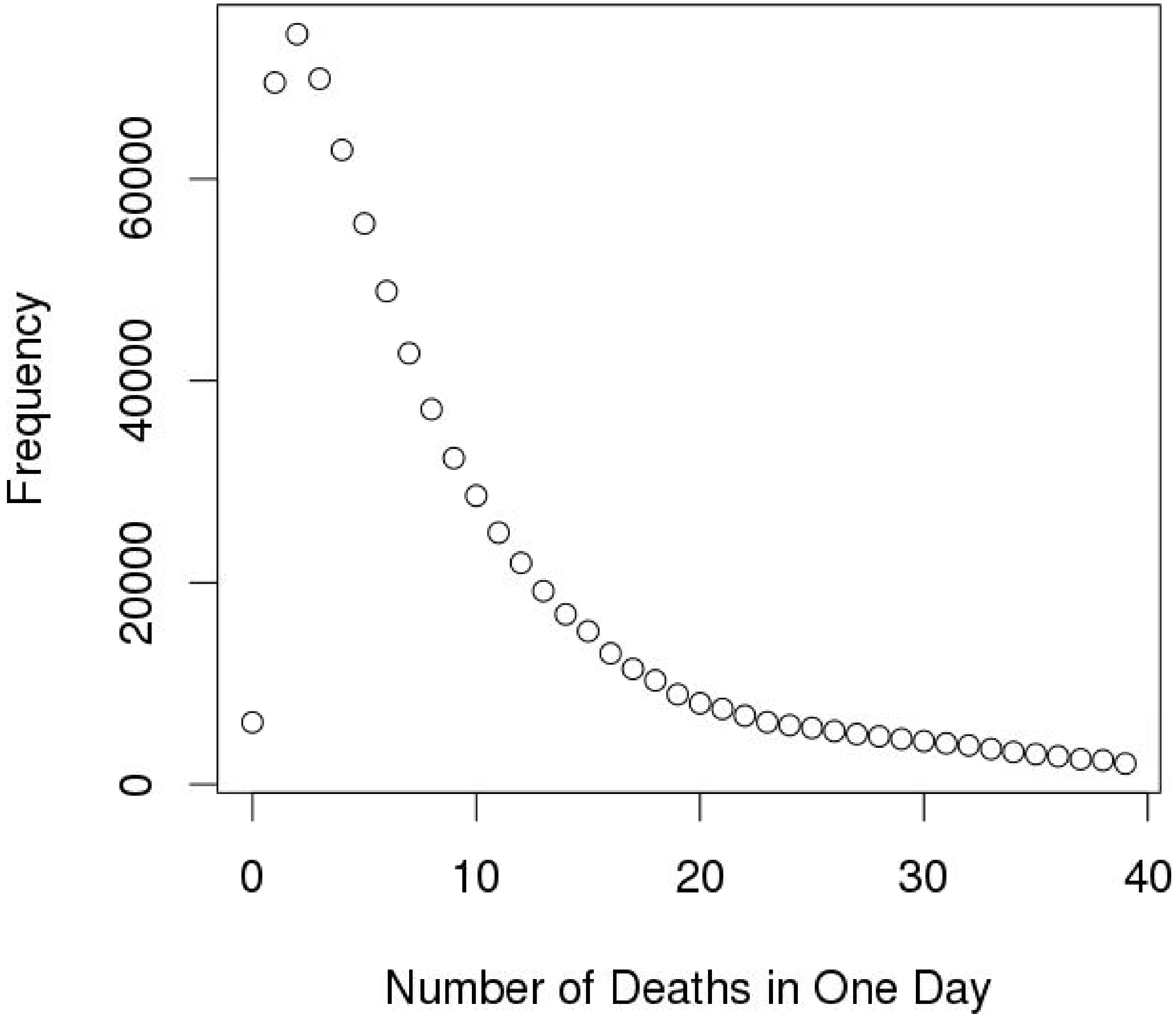
II. Technical discussion: Reliability of Inference from Perturbed Datasets

- This example is based on Medicare data
- One way to preserve confidentiality is to perturb the data before making it available
- The objective of this small example is to examine how a modest perturbation will affect the inferences of interest

Data administered by Gradient Corporation with financial support provided by a consortium of trade organizations, including the American Petroleum Institute, the American Forest & Paper Association, the American Wood Council, and The National Council for Air and Stream Improvement; the ExxonMobil Corporation also contributed funding.

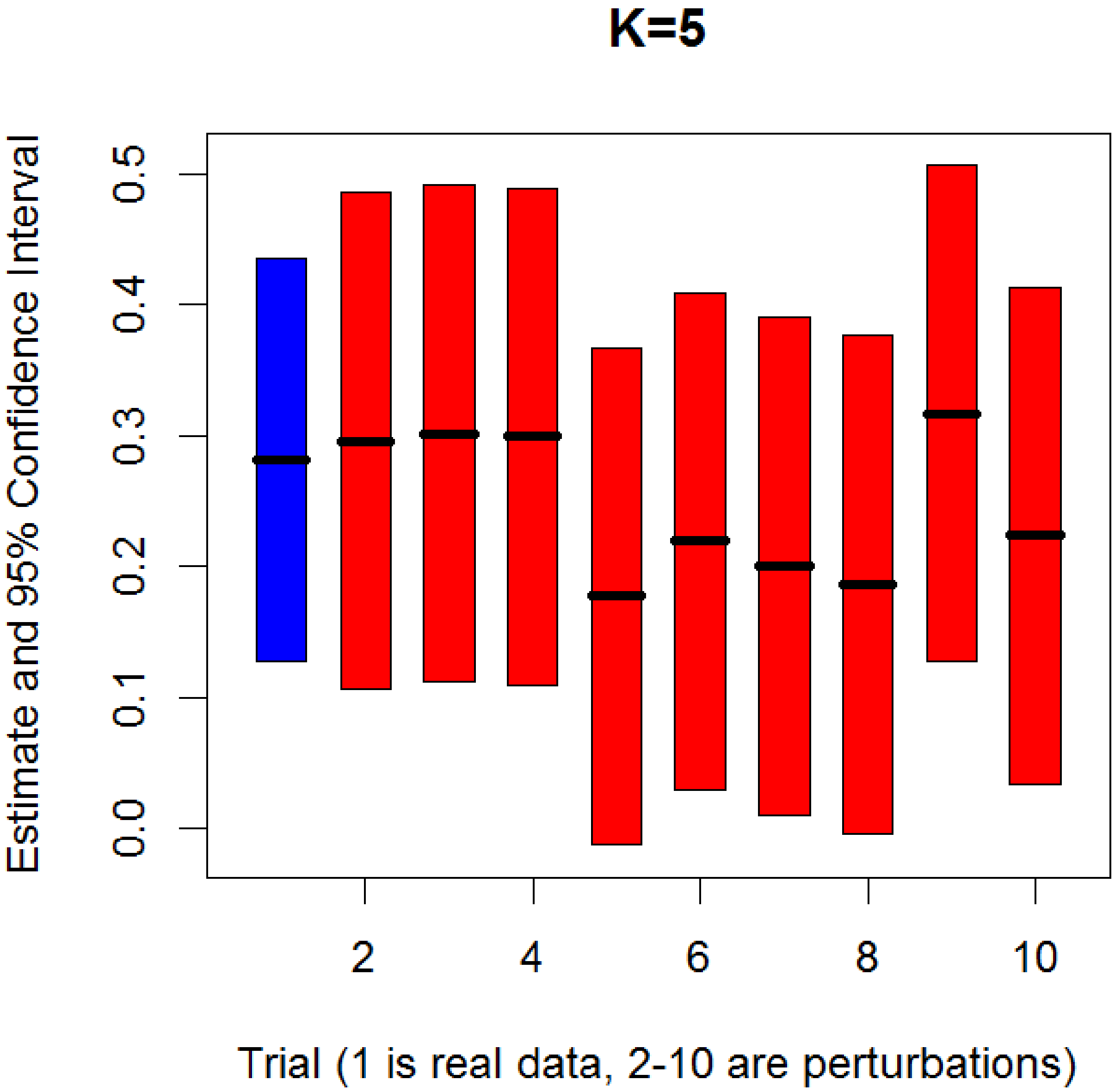
- 13 years of daily data from 399 counties
- Count of over-65 deaths in each county in each day, with $PM_{2.5}$ and meteorological variables
- Estimated a linear dose-response curve for each county over the range 0-35 μg/m^3 (expressed as percent rise per 10 μg/m^3), then combined over all counties using a weighted least squares (WLS) approach
- Total 779,317 days; 7,908,669 deaths; max number of deaths/day is 259

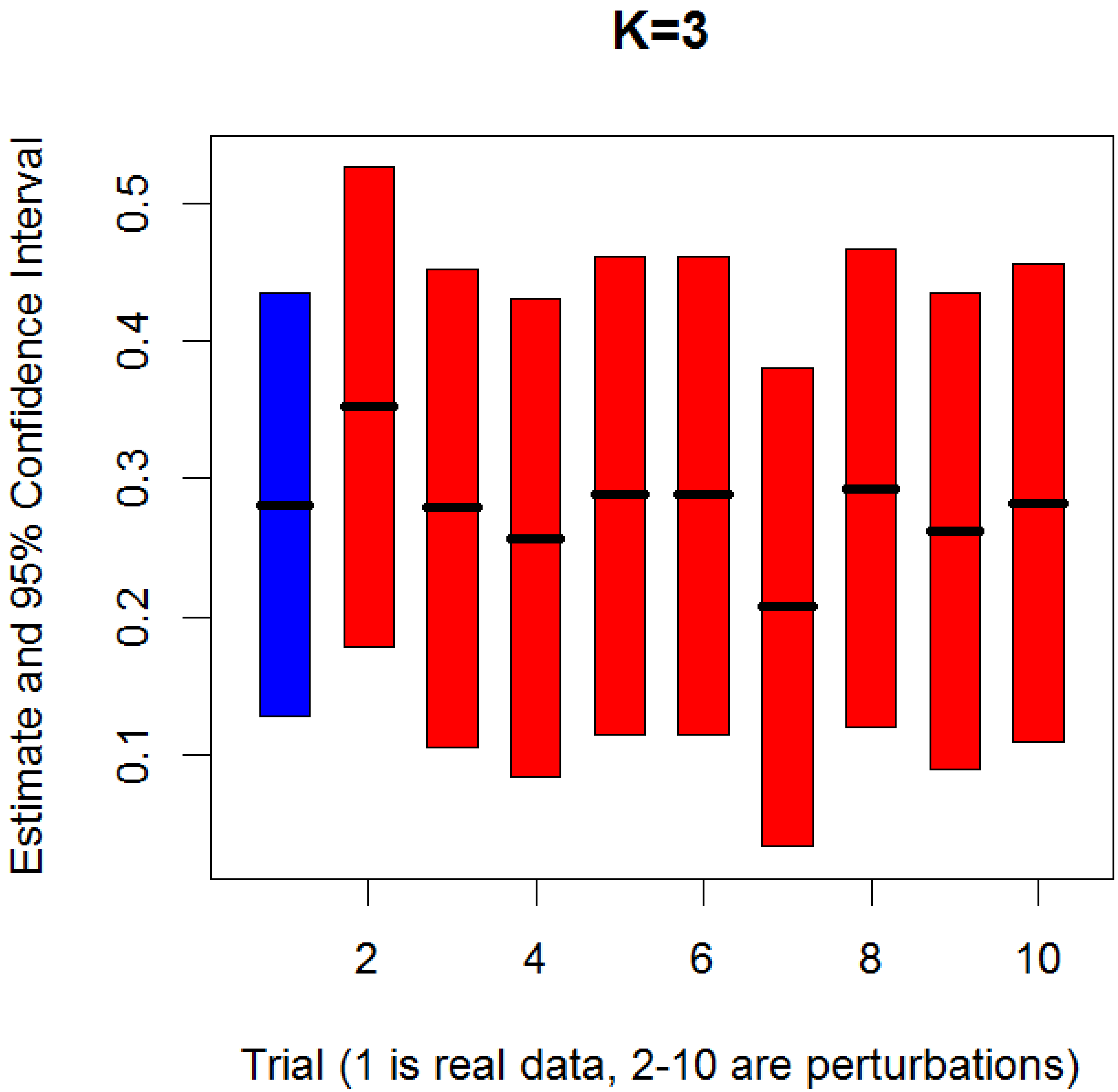
Frequency of Mortality Counts

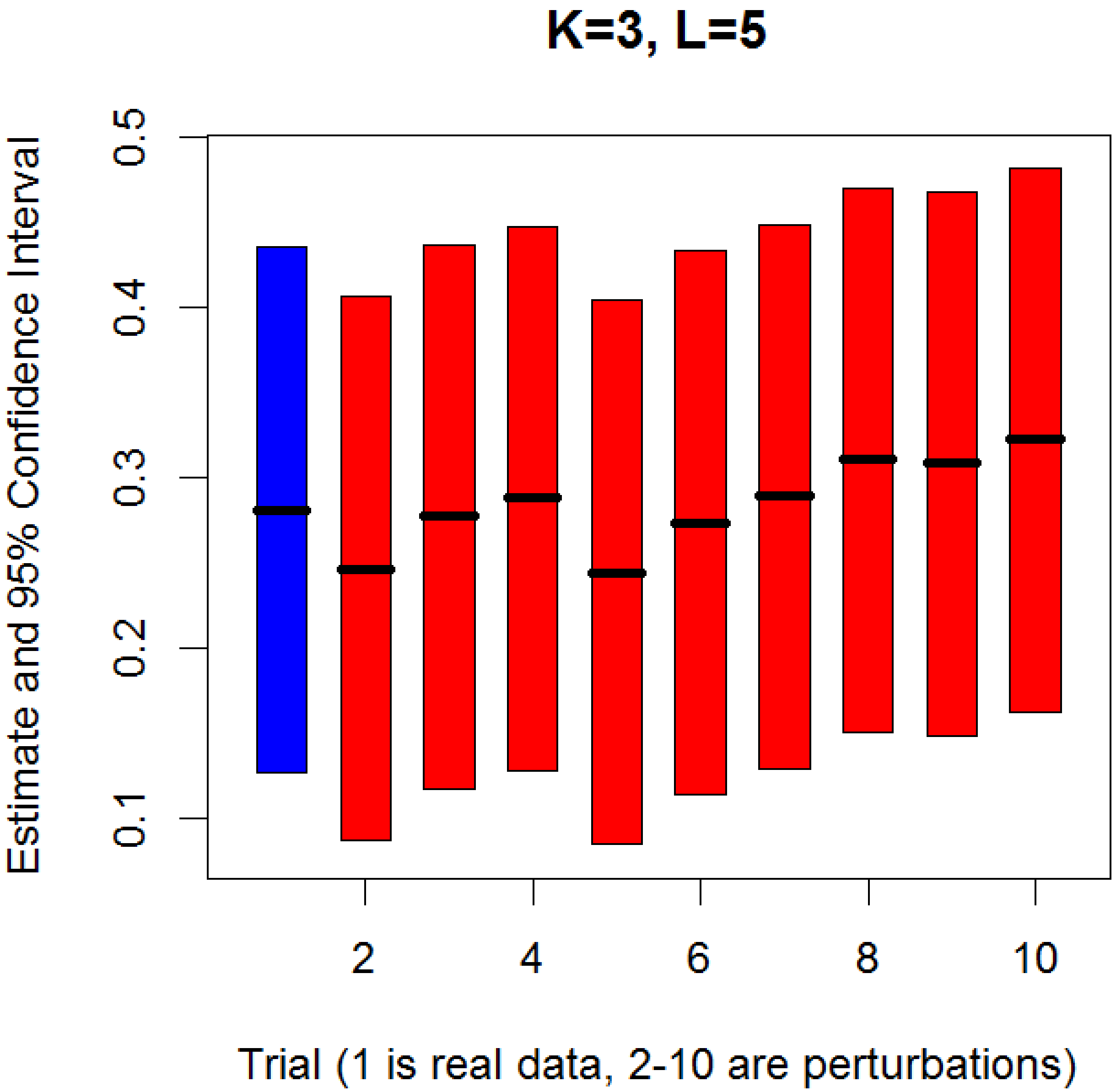


Proposed Perturbations

- For each county and each day, randomly add or subtract up to K to each mortality count, preserving non-negativity
- Variant: only do this for days with a count of ≤L
- First experiment: K=5, no limit on L
- Second experiment: K=3, no limit on L
- Third experiment: K=3, L=5
- Fourth experiment: K=1, L=3


K=5
Estimate and 95% Confidence Interval
0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10
Trial (1 is real data, 2-10 are perturbations)

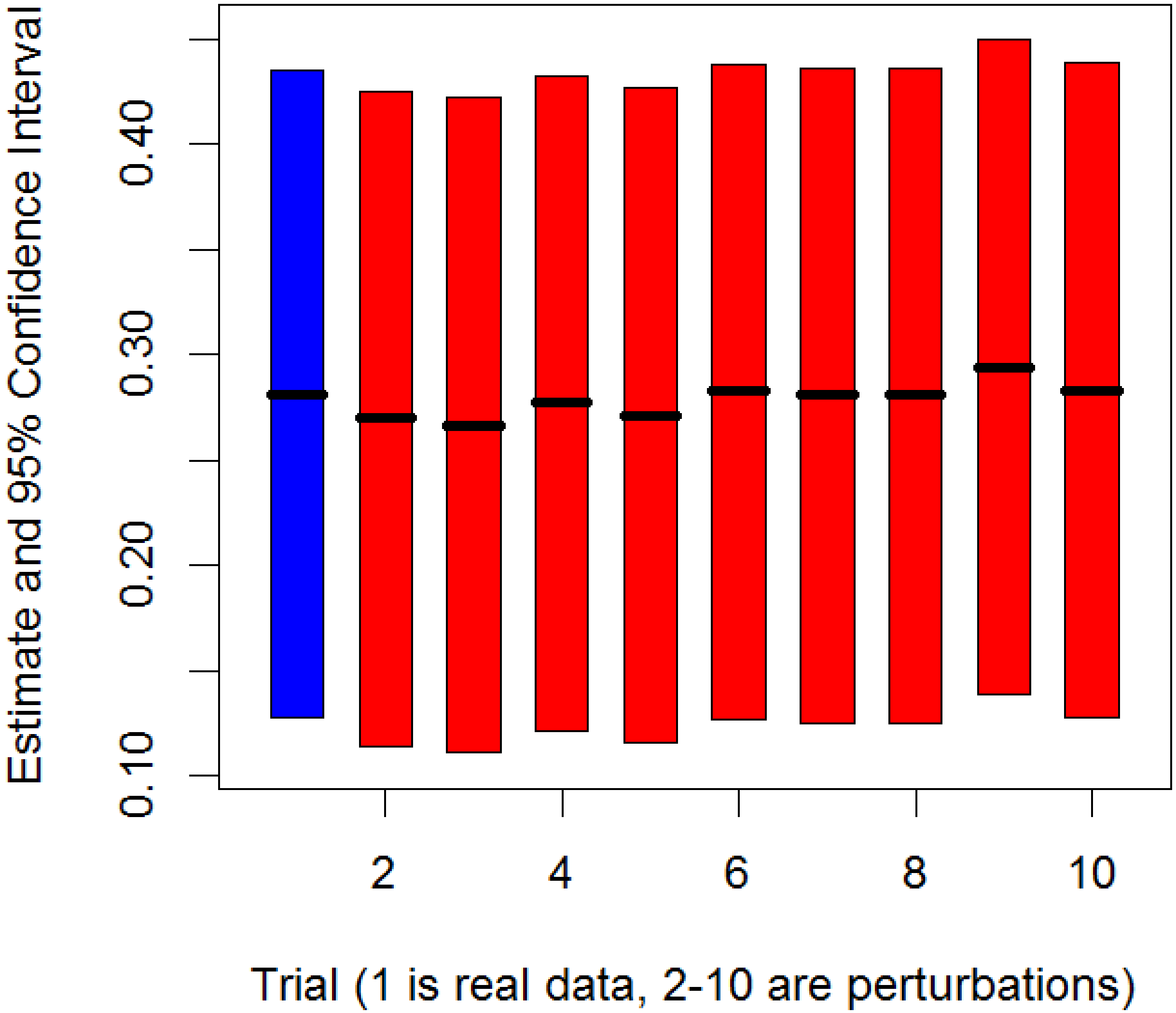

K=3
Estimate and 95% Confidence Interval
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10
Trial (1 is real data, 2-10 are perturbations)

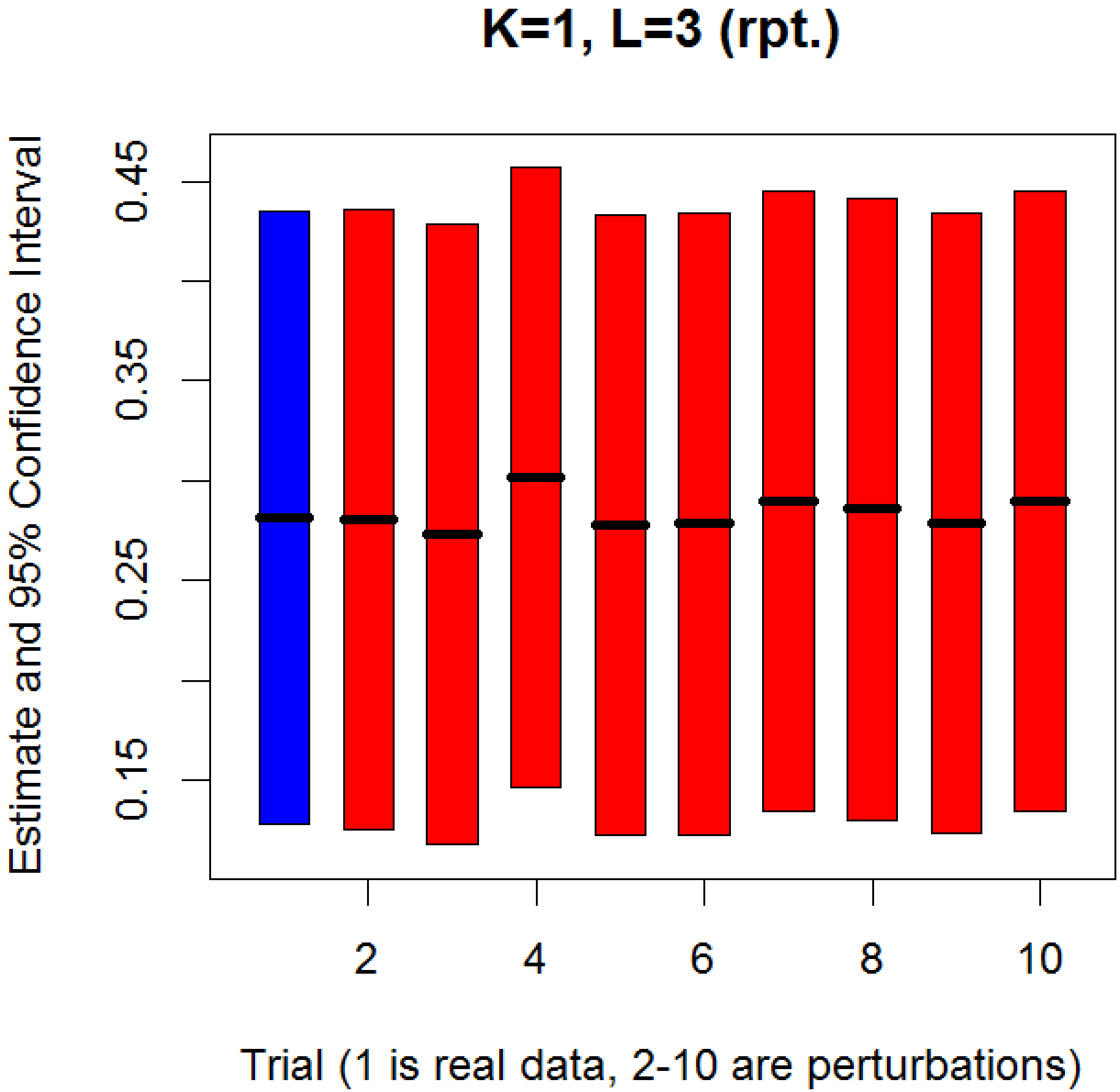


K=3, L=5

Estimate and 95% Confidence Interval
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10
Trial (1 is real data, 2-10 are perturbations)

K=1, L=3


Estimate and 95% Confidence Interval
0.10
0.20
0.30
0.40
2
4
6
8
10
Trial (1 is real data, 2-10 are perturbations)


K=1, L=3 (rpt.)
Estimate and 95% Confidence Interval
0.15
0.25
0.35
0.45
2
4
6
8
10
Trial (1 is real data, 2-10 are perturbations)

Conclusions

- The effect of the perturbation on the final result is small but not negligible
- The relative effect of the perturbation could be much larger for problems of genuine interest, e.g. nonlinear dose-response curve over 0-12 μg/m^3
- The exact form of the perturbation *does* matter, e.g. results including L seem better
- The results used by EPA for regulation would presumably still be those based on the true data – would this meet the provisions of the new rule?
- It would still be necessary to work with the relevant agency (in this case, CMS) to implement this kind of data publication in practice, and it’s not clear whether they’d allow that

CMS = Centers for Medicare & Medicaid Services