How many (multiple) imputations do I need for an important analysis?

Tim P. Morris, MRC Clinical Trials Unit at UCL
@timpmorris.bsky.social
Andrew Atkinson, James R. Carpenter

PSI 2025 Conference

Smarter Studies
Global Impact
Better Health

Multiple imputation

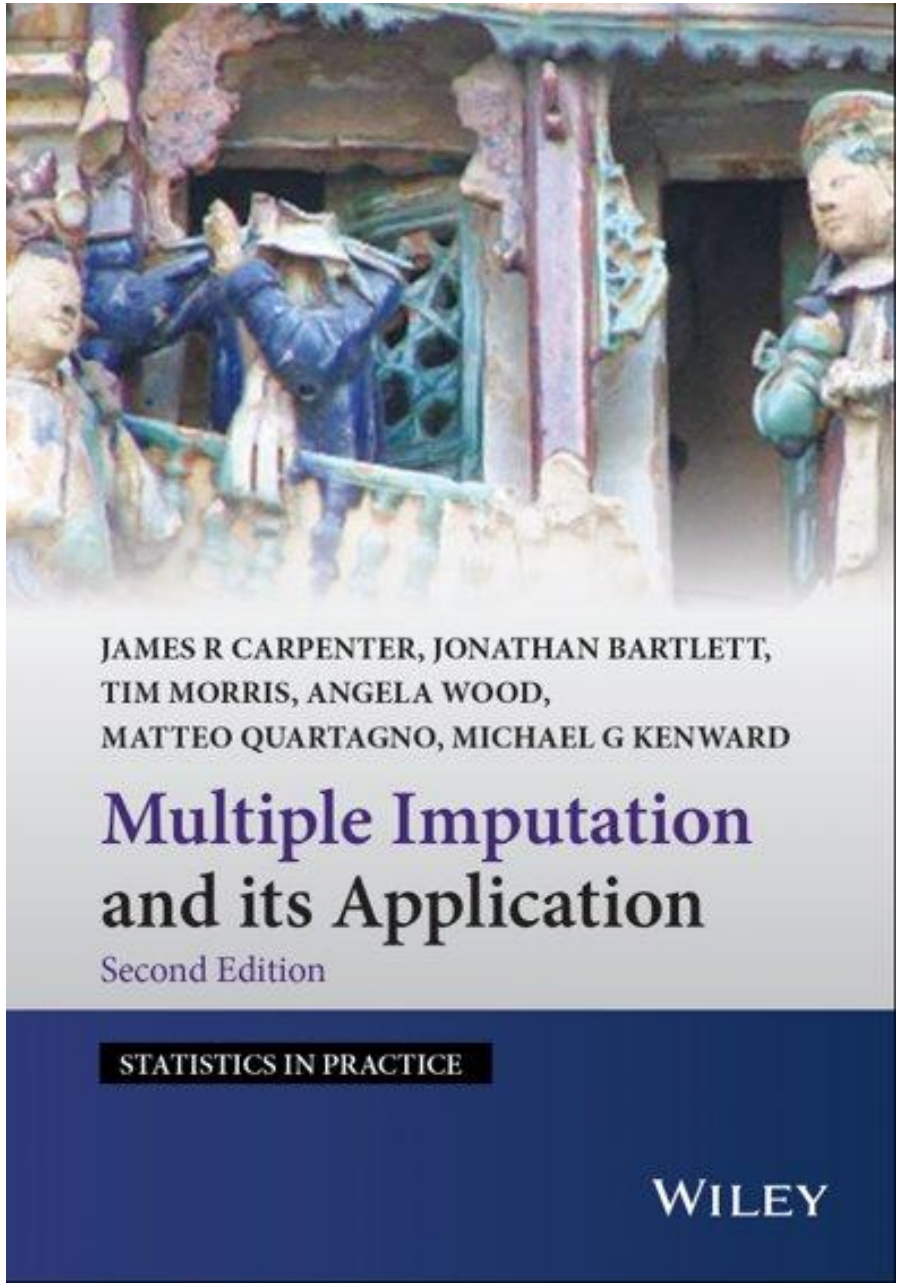
	Step	Details
1.	**Impute**	Repeatedly draw (simulate) the missing values from the posterior predictive distribution of an imputation model, to produce K imputed datasets.
2.	**Analyse**	Conduct the same analysis on each imputed dataset.
3.	**Combine**	Bring together the K sets of results using 'Rubin's rules' to produce a combined inference.

The dreaded question

The validity, versatility and pitfalls of multiple imputation are well-established and documented (see exhibit to right 😉)

A common question about multiple imputation is, ‘How many imputations do I need?’

Note: step 1 (**Impute**) involved simulation.



A range of (valid) answers

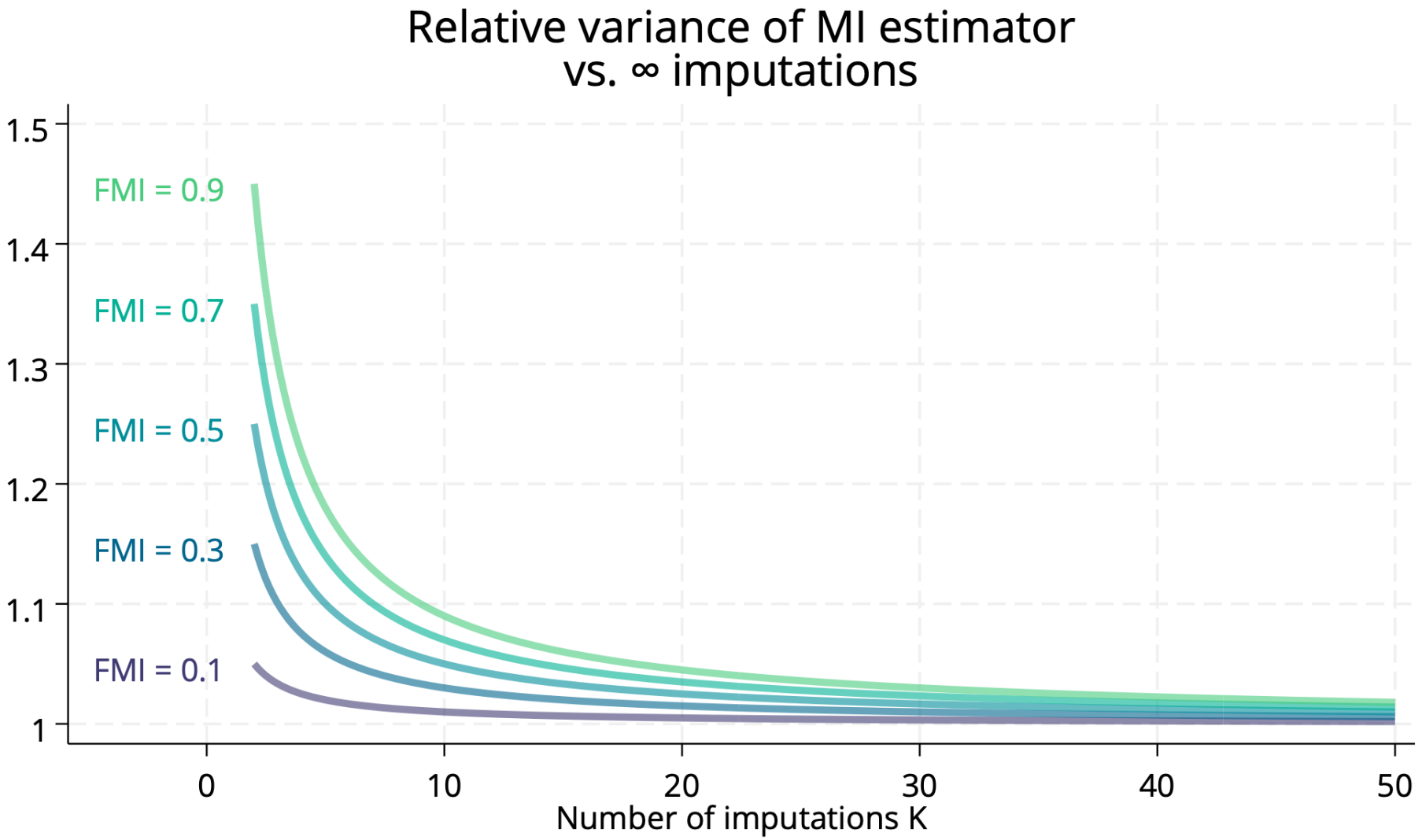
If multiple imputation is valid at all, $K = 2$ is valid!

As many as you can!

The problem with '$K = 2$ is valid' is
it's like saying 'if a trial with $n = 2{,}000$ is valid,
so is the same design with $n = 2$'.
It misses a rather important point.

Approach	Description	Reference
Efficiency	Consider relative efficiency of K imputations (vs. ∞) and realise that 5–10 is typically enough.	Rubin (1987)
'Linear' rule	Choose K equal to the number of incomplete cases (up to 50%).	Bodner (2008)
'Quadratic' rule	Produce some pilot imputations, estimate the fraction of missing information, and decide how many more you need to make your SE reproducible (note: two stages).	Von Hippel (2020)
Direct MCSE	**If we are comfortable with using 'pilot' imputations, we can work directly with MCSE!**	**Carpenter et al. (2023)**

Relative variance (Rubin, 1987)

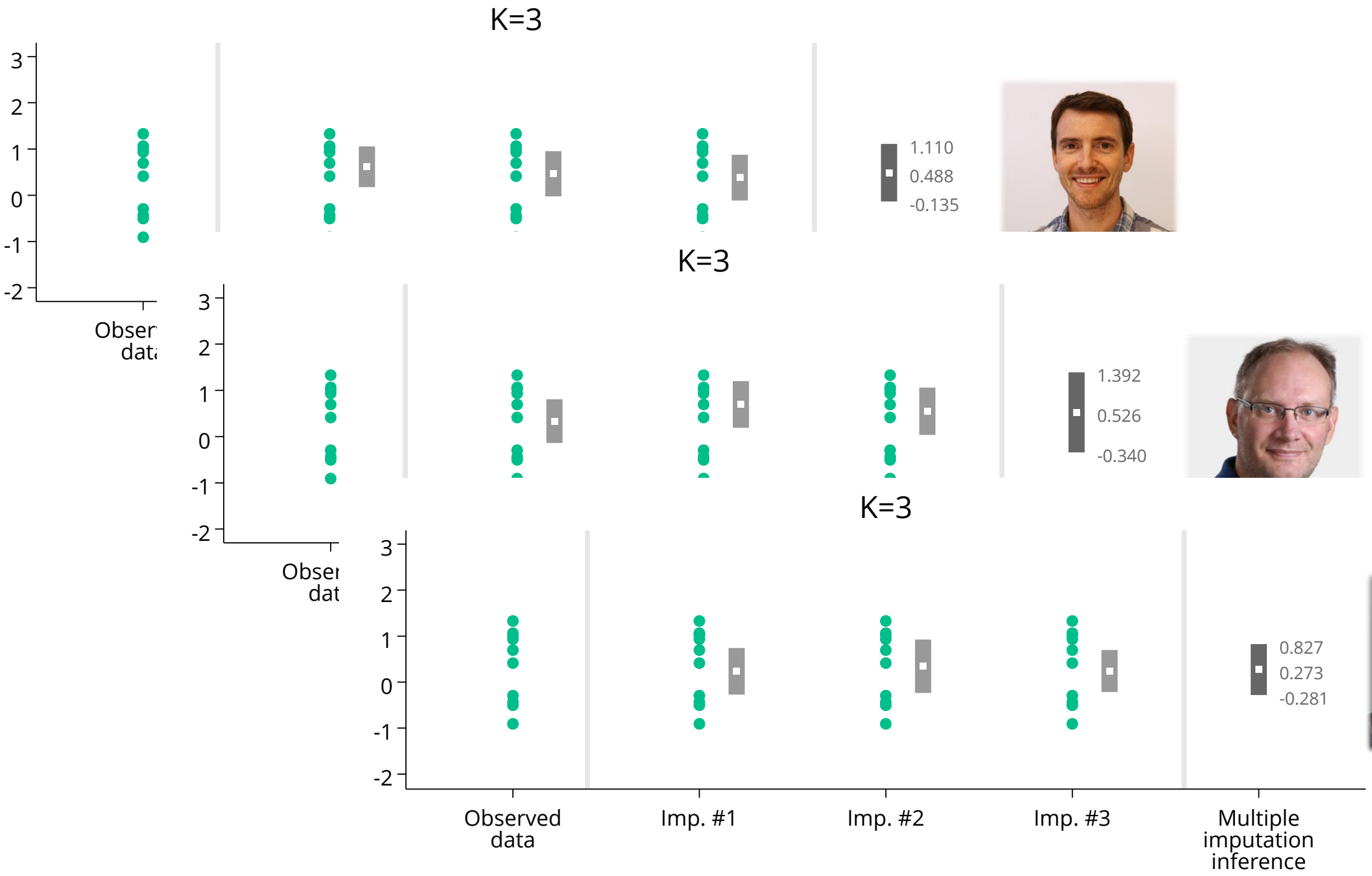


Monte Carlo SE

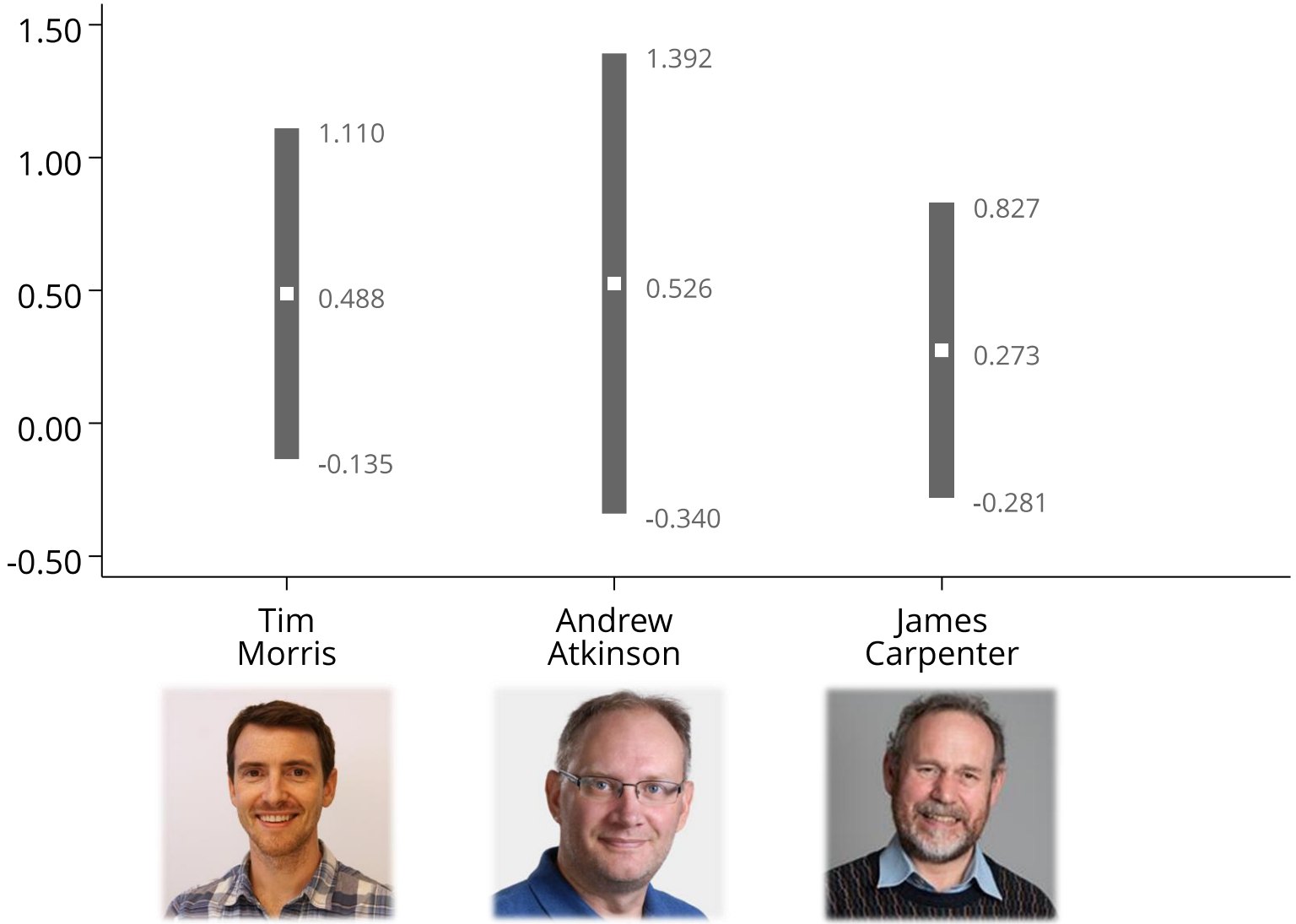
This is all well and good, but…

Recall that step 1 of multiple imputation was “Repeatedly **draw** (**simulate**) the missing values”

What if I had used a different seed and produced different imputations…
WHAT THEN!!


K=3
3
2
1
0
-1
-2
1.110
0.488
-0.135
K=3
3
2
1
0
-1
-2
1.392
0.526
-0.340
K=3
3
2
1
0
-1
-2
0.827
0.273
-0.281
Observed data
Imp. #1
Imp. #2
Imp. #3
Multiple imputation inference

Our MI inference with $K = 3$



Monte Carlo error

The bad news is that multiple imputation inference has this unwelcome source of variation.

The good news is:

1. We can estimate Monte Carlo error using jackknife (Efron & Gong, 1983) – don't have to do more imputations
2. Monte Carlo error reduces as K increases, and 'more imputations pls' is better than 'collect more data'
3. Monte Carlo SE is a standard error, so we have a rough idea of how fast it reduces (locally linear in $\sqrt{1/K}$)

We can get MCSE for any statistic

```
Multiple-imputation estimates                   Imputations       =          5
Linear regression                               Number of obs     =         15

DF adjustment:   Small sample                            max      =       6.56
                                                F(   0,     .)    =          .
Within VCE type:          OLS                   Prob > F          =          .

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   .4497113   .2618279     1.72   0.132    -.1779574     1.07738
             |   .0625612    .045419     0.28   0.064     .1433758    .2001906
------------------------------------------------------------------------------
Note: Values displayed beneath estimates are Monte Carlo error estimates.
```

The proposal

1. Begin with a pilot number of imputations, e.g. $K = 20$
2. Estimate MCSE for statistics of interest, e.g. p-value
3. Project how many more imputations would reduce MCSE to an acceptable value
4. Add imputations

Notes:

- This does not need any manual input!
- You can repeat steps 2–4

Does it work?

I've used two types of simulation study to evaluate this proposal:

1. A 'standard' simulation study (which I'll show you)
2. Repeated imputation within a specific dataset

Simulation study

Aim is to evaluate accuracy of proposed method for choosing number of imputations

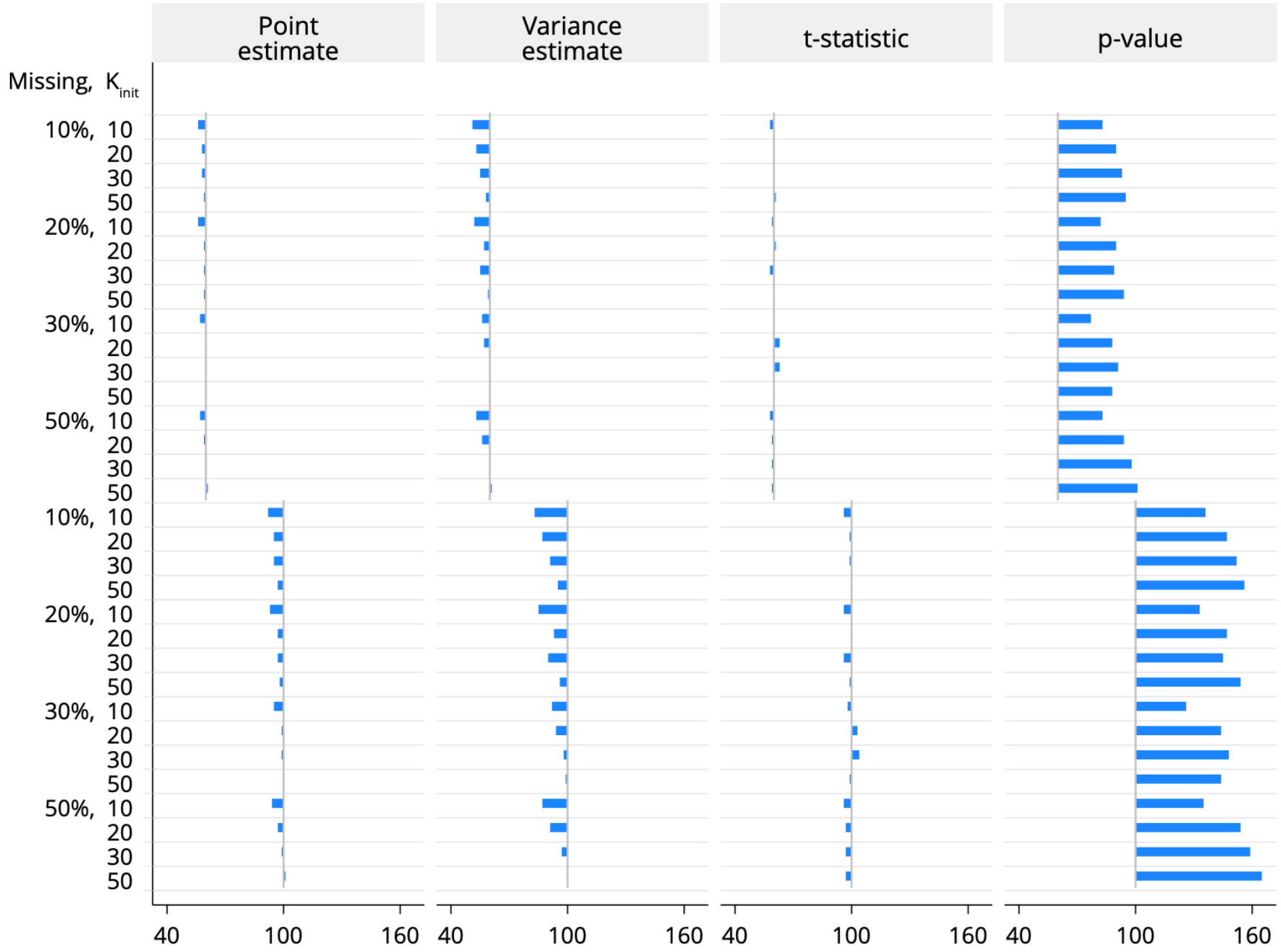
Data-generating mechanism takes some parameters from a real trial, but the setup is relatively simple:

- 300 participants, 150 per arm.
- Quantitative outcome and covariate, correlation around 0.5
- Data missing-at-random given arm and covariate, with varying proportions of missingness.

Estimand/target: in the trial is $E(Y^1 - Y^0)$ **but** for this simulation study, the 'target' is the true number of required imputations (either 60 or 100)

Results

Blue lines show the projected number of imputations compared with the actual number needed (60 or 100)



Some conclusions

- Because this is such a familiar idea from simulation studies, it seems obviously right!
- I would of course be happy to hear what further reassurance you would want.
- Recall that I am proposing you do this for important analyses where you would be mortified by high Monte Carlo error.
- Do not do it every time you use multiple imputation.

References and acknowledgements

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Bodner TE (2008) What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4), 651–675.

von Hippel PT (2020) How many imputations do you need? [...]. *Sociological Methods & Research*, 49(3), 699–718.

Carpenter JR, Bartlett JW, Morris TP, Wood AM, Quartagno M, Kenward MG. (2023) *Multiple Imputation and its Application, 2nd edition*. Wiley.

Efron B, Gong G (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.

Acknowledgements

Jonathan Bartlett
Matteo Quartagno
Tobias Muetze

Pls. note: My acknowledgement does not imply their endorsement!