**Interpreting the graphs**

Assuming a dataset where the proportion of false-recent results provides an estimate of the FRR, we have considered two simple statistics to illustrate the effect of the sample size.

First set of graphs: Probability of observing some *low* FRR

Description of graphs

- Inevitably, a *threshold* will have to be chosen, used to decide whether tests qualify for further investigation.
- In the first set of graphs (pages 1-6), the probability of observing a (raw estimate of the) FRR below (or equal to) the *threshold* is shown, as a function of the true FRR.
- The probability (read off the y-axis) depends on the:
    - sample size (separate graphs are shown for N = 50, 100, 200, 300, 500 and 1000 respectively),
    - the RITA's true FRR (read off the x-axis), and
    - the chosen *threshold* (blue line: 1%, red line: 5% and green line: 10%).
- For example: in a sample of N = 100, a RITA with a true FRR of 7% (x-axis) has a 30% chance (y-axis) of producing an estimate of the FRR less than or equal to 5% (red line).

Some notes

- While one could calculate the statistical power of a test about the FRR, given the hypothesis being tested and the significance level (type I error), the graphs do not strictly show the statistical power associated with a particular statistical test about the FRR - we are simply looking at the observed point/raw estimate of the FRR, and the chance of it being *low* (as defined by the *threshold*).
- In general, a RITA will have about a 50% chance of producing an estimate of the FRR below its *true* value (more accurate with increasing N). For example, a test with a true FRR of 5% will have about a 50% chance of providing an estimate of the FRR below 5%, and therefore the lines tend to pivot around approximately fixed points. As the sample size increases, the steepness of the lines (fixed around their pivots) increases.
- What is of interest is whether a test with a *truly* low FRR will fail to suggest a low FRR ("error A") or whether a test with a *truly* high FRR will happen to suggest a low FRR ("error B").
- Reading off the graphs, one can see how the risk of these errors is reduced with increasing sample size (dependent on choice of *threshold*).
- Given a sufficiently large sample, there are diminishing returns from increasing the sample size. For example, a RITA with a truly low FRR of 3% has an 80% chance of producing an estimate for the FRR of less than or equal to 5% at N = 50, 90% chance at N = 100, and close to 100% chance by N = 500.
- The choice of *threshold*, related to the willingness to risk "error" in discarding / continuing investigation of RITAs, is related to what is achievable with different sample sizes.

Second set of graphs: Confidence interval widths

Description of graphs

- A familiar measure of how much power a given sample size provides is the width of the confidence interval (CI) achieved. These widths relate to the power of statistical tests about whether the RITA's true FRR is below or above a chosen threshold.
- In the second set of graphs (pages 7-12), the width of the CI is shown.
- The width (upper (green) and lower (blue) limits of the 95% CI read off the y-axis) depends on the:
    - sample size (separate graphs are shown for N = 50, 100, 200, 300, 500 and 1000 respectively), and
    - the observed proportion of false-recent results (read off the x-axis).
- For example: in a sample of N = 100, if 7% (x-axis) of results are RITA-recent, the 95% CI for the FRR would span from 3% to 14% (y-axis).

Some notes

- As expected, the CI width decreases with increasing sample size (and smaller point estimates):
    - For example, for an observed FRR of 5%, the 95% CI for the FRR is 1%-15% for N = 50, narrowing to 3%-7% for N = 500.
- Again, given a sufficiently large sample, "diminishing returns" may be seen with larger samples.
- These graphs provide an indication of the increasing inferential power with increasing sample size. In the final assessments of the RITAs, narrow CIs would be of importance.