EPSRC ‘New Horizons’ funding opportunities

Analysis of the influence of review anonymisation interventions

Summary

There is some evidence that the review anonymisation used in EPSRC’s New Horizons funding opportunities was associated with a reduction in bias related to applicants’ ethnicities. Analysis indicated a trend toward reducing bias in relation to other personal characteristics of applicants, but differences were not statistically significant.

Introduction

EPSRC has run two New Horizons (NH) pilot opportunities: the first in 2020 followed by a second instance, incorporating several process changes, in 2021.

New Horizons was intended to:

- support speculative, high-risk research ideas that could potentially offer high reward
- trial ways to minimise the bureaucracy of the application process, particularly for the applicant
- deliver a faster process compared to current standard funding routes, whilst maintaining robust decision making that is proportionate to the level of risk

While reduction or elimination of bias related to the personal characteristics of applicants was not a stated aim of either opportunity, the use of anonymised applications might have been expected to have that effect. Note that information that we collect to support equalities monitoring are not made available to assessors even when anonymisation is lifted.

In both rounds of NH the initial assessment process was to some extent double anonymised – that is, the applicants did not know who the reviewers were, and the reviewers were not told who the applicants were.

Round 1 started with double anonymised written peer review of full proposals, followed by a panel meeting at which applicant information was shared with panel members only in the final stages. Round 2 involved a two-stage process of outline and then full proposals, with both stages reviewed by expert panels. In Round 2 the outline stage was double anonymised while the full proposal stage was not: panel members were aware of the membership of teams submitting full proposals.

Round 1 attracted 1203 applicants as Principal Investigator, Round 2 770.

Because both the processes and the subject areas of the two rounds differed substantially it is not particularly useful to compare them directly with each other. It is however possible to compare each of them with a selected comparator process, in this case EPSRC Responsive Mode schemes in the physical and mathematical sciences (Round 1) and engineering and ICT (Round 2.)

This analysis focuses on

- the award rates of applicants within groups defined by age, disability status, ethnicity and sex
- differences between award rates across those groups
- differences between those measures in NH and comparator opportunities

No comparison is made of the composition of NH applicants and those in the comparator activities.

Findings

Figure 1 shows award rates by characteristic of the Principal Investigator (the terminology used to describe the lead applicant at the time the opportunities were run) of the application for NH rounds 1 and 2. There are two subplots for ethnicity: one showing a five-way grouping and another using a binary classification. This chart cannot be used to infer reliably anything about differences in award rates across groups. Some groups have few members, meaning that even large rate differences may not be meaningful. Analysis shown later will address the question of the significance of findings.

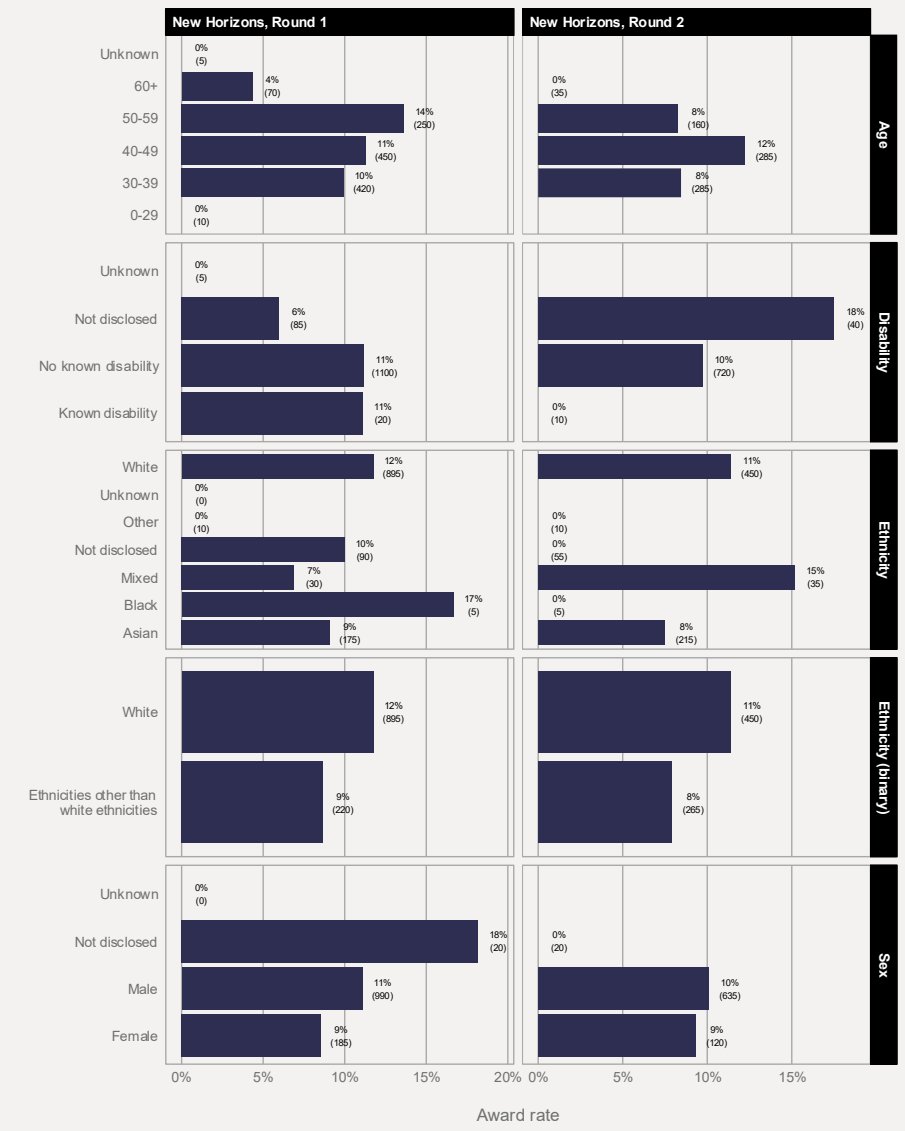

Figure 1 New Horizons award rates by Principal Investigator age group, disability status, ethnicity and sex. The total number of applicants as Principal Investigator was 1203 in Round 1 and 770 in Round 2. Individual group totals are shown on the chart in parentheses and are rounded to the nearest 5. Award rates are rounded to the nearest 1%.

For both Rounds 1 and 2, Male applicants had higher award rates than did Female applicants, and White applicants had higher award rates than did applicants declaring ethnicities other than White. Note that the 'White' group may include applicants from white ethnic minority backgrounds.

The age and disability data cannot be interpreted so straightforwardly. Age data is analysed further in later charts and contains some groups with very few members. Interpretation of disability data is complicated by the high level of missing data and a low level of applicants declaring disabilities.

Differences in award rates across binary groups are a more direct indicator of bias than are the award rates themselves. Figure 2 shows these differences, for NH Rounds 1 and 2, and their respective comparators.

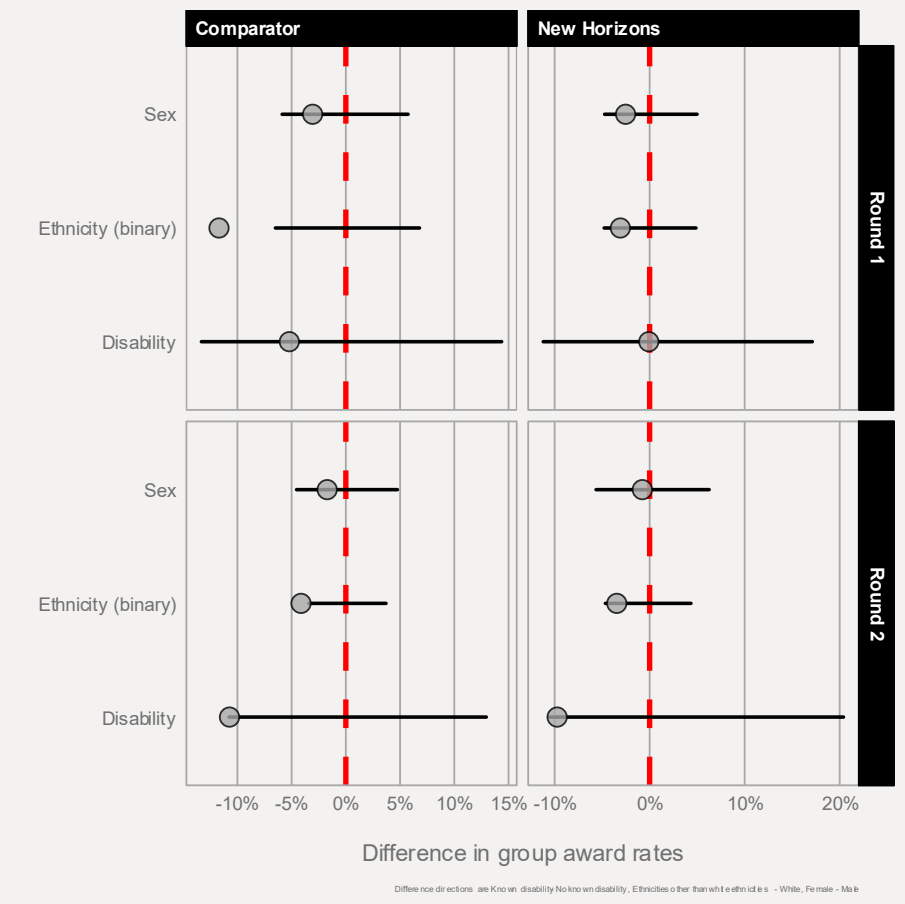

Figure 2 Comparison of award rate differences across binary PI disability status, ethnicity and sex categories. No difference (0 percentage points) is highlighted with a red, dashed, line. Ranges may not be symmetric around 0 as the underlying distribution of outcomes may not be symmetric. Ranges covering 95% of the distribution are shown.

It is likely that no difference between groups will ever be exactly zero. Intuitively we know that, even if there is no bias, we can expect to see some noise in the data. Using the method described in the annex, Figure 2 shows, as horizontal lines, ranges that indicate the level of variation that we might expect to see if there was in fact no association between the characteristic and the funding decision. Where we have less data, for example with disability status, these ranges are larger, indicating greater uncertainty.

If a circle marking a measured value falls outside the indicated range, we might reasonably conclude that it is a sign of something unusual and that there is therefore an association of some kind between that characteristic and decision outcomes.

All the NH rate differences fall within the range of expected values. In a formal sense, this suggests that the observed differences in award rates, while sometimes large, are compatible with a belief that there is no consistent bias associated with those characteristics.

While the award rate differences relating to disability status and sex in the comparator opportunities are also within expected ranges, those relating to (binary) ethnicity are not. This is evidence of an ethnicity-related bias in those comparator opportunities, with that bias favouring White applicants.

Does this mean that NH was unbiased, and therefore that the NH process was preferable? Not necessarily.

While the observed rate differences in both rounds of NH fall within expected ranges, they are still negative and favour the majority group (‘No known disability’, ‘White’, ‘Male’). The same pattern is seen in all the data. Taking broader evidence into account, it is reasonable to believe that the NH opportunities retained some bias in favour of White applicants. Evidence relating to possible sex or disability biases is mixed and no conclusions can safely be drawn.

It is possible to compare the differences between the inter-group differences in NH and comparator opportunities to understand whether the two processes showed the same level of bias. For example, we might determine how large the difference in award rates between Female and Male applicants was in NH Round 1 and subtract from it the difference in award rates between Female and Male applicants in the relevant comparator opportunity.

This ‘difference in difference’ analysis is shown in Figure 3. If the difference-in-differences is zero, the level of bias in each opportunity was the same. Figure 3 also includes expected ranges for these differences-in-differences, and the same interpretation of the results can be applied: if the point is inside the expected range, the variation from zero is within the range we might expect to see if there was no difference in bias across the two opportunities (NH, comparator.)

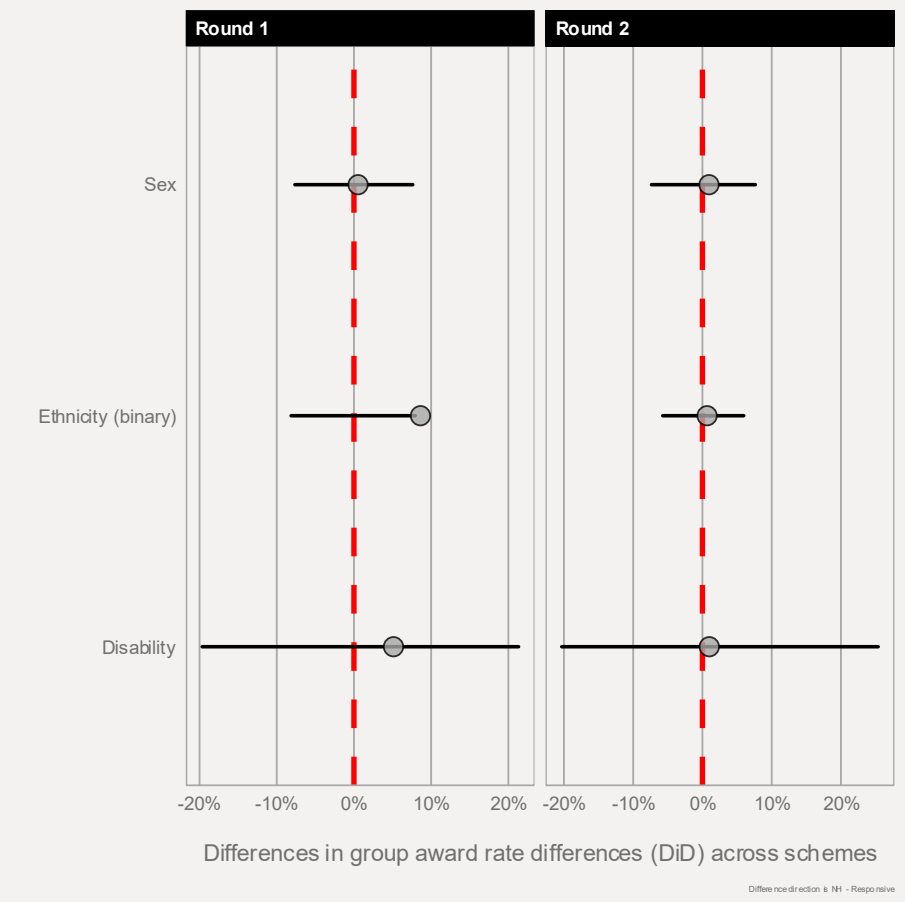

Figure 3 Difference in binary differences between New Horizons opportunities and their comparators, by round.

Five of the six differences-in-difference were both rather small and well within expected ranges (though all were positive, indicating a tendency to see less bias in NH.) But in Round 1 the difference-in-difference was strongly positive in relation to binary ethnicity. The simple interpretation of this is that in NH Round 1, the ethnicity bias was significantly (in a statistical sense) less than that seen in its comparator (though note that a bias in favour of White applicants was still seen in both opportunities: less bias is not the same as no bias.)

Although NH passes one formal statistical test, we still cannot be sure that NH overall was in some sense less biased as a process than was the more traditional responsive mode process against which it is being compared here. We might expect more of a difference in relation to more characteristics if this was the case. Larger datasets will be needed to explore these questions further.

Data on applicants' ages needs to be treated slightly differently as it features five, not two, groups. Figure 4 shows the award rate by PI age group and, again as horizontal grey lines, the range of award rates we might expect to see were there no association between success probability and applicant age. There are very few applicants in the '0 -29' age category in NH1, resulting in very

broad expected ranges for that group. The observed award rate of 0% falls within the expected range, which is anything between 0% and nearly 40%. There were no applicants in this category in NH2.

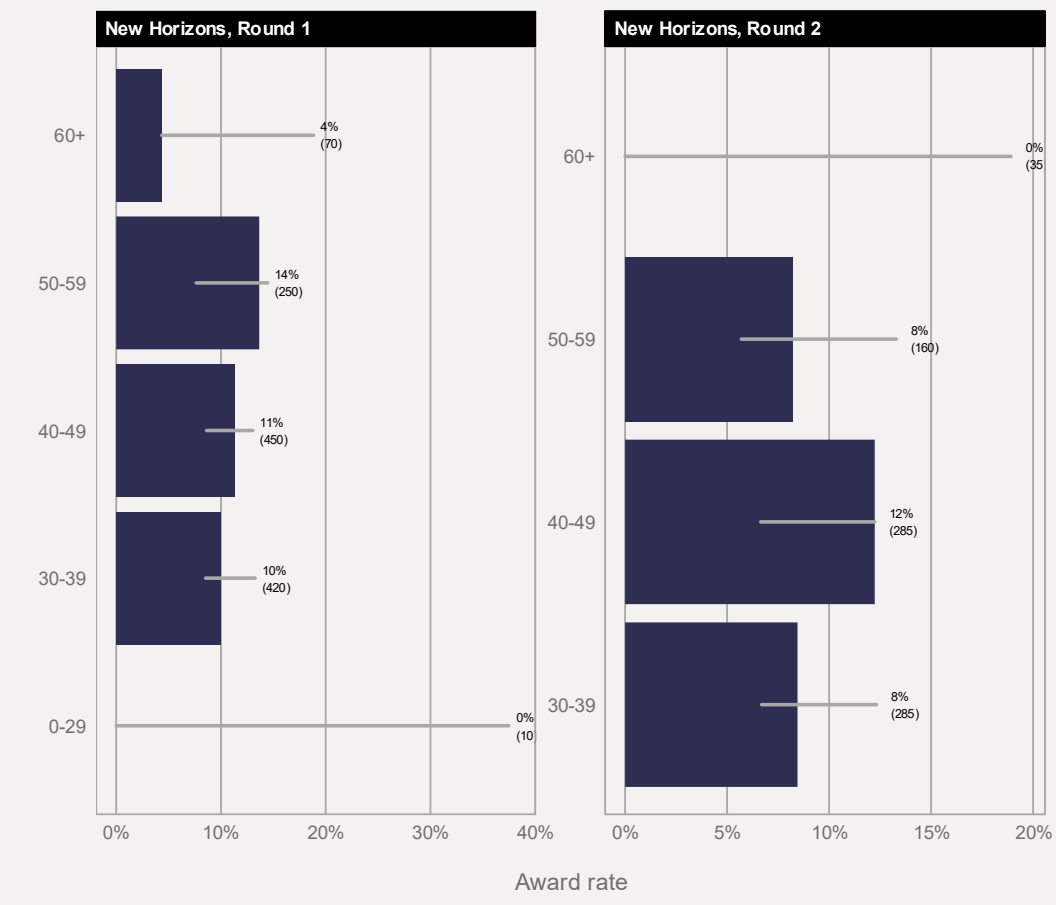

Figure 4 Award rates by applicant age category, including 95% expected ranges. Individual group totals are shown on the chart in parentheses and are rounded to the nearest 5. Award rate differences are rounded to the nearest 1% (so that a stated difference of '11%' indicates an 11 percentage point difference.)

In both rounds of New Horizons, the observed award rates for all groups are within the range expected. There is also no clear sign of an age-related gradient.

Figure 5 shows the differences between every possible pair of age ranges' award rates seen in each round, again with intervals showing the expected range of differences for each, shown as grey horizontal lines.

The intervals are a formal guide to the significance of differences. The expected ranges shown reflect the numbers in each group (fewer members leads to wider ranges) so the paucity of data in

some groups is accommodated by the analysis in the form of wider expected ranges. This is particularly apparent in comparisons involving the 0-29 age group in NH1.

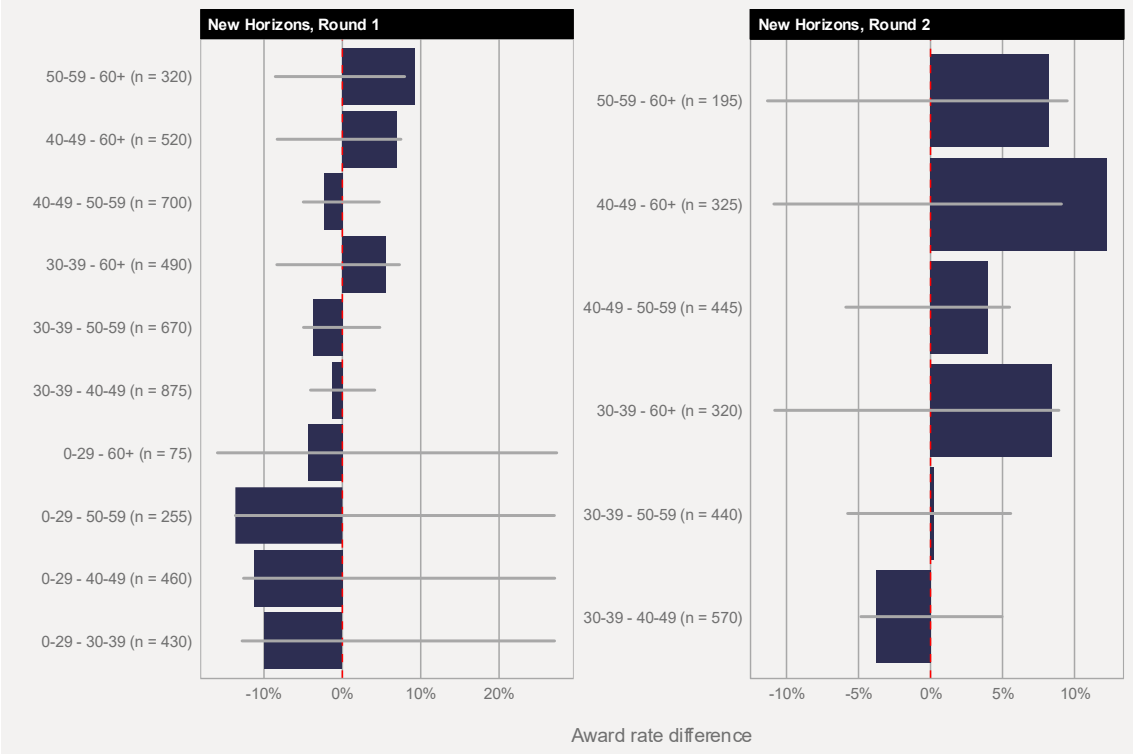

Figure 5 Differences between award rates across age groups, including 95% expected ranges. Total number of individuals involved in each comparison is shown in parentheses on the y axis, and is rounded to the nearest five.

In most cases the observed inter-group rate difference is within expected ranges. There are two exceptions, both involving applicants in the 60+ range, who had the lower of the two rates. As there are 16 individual comparisons, there is quite a good chance (~37%) of seeing one significant result by chance alone, and about a ~15% chance of seeing two. Given this, and the inconsistency of the effect, it would not be entirely safe to conclude that there is a systematic age bias against the oldest applicants influencing these decisions, but the data is suggestive.

In Round 1, the very youngest applicants experienced award rates which were lower than those of all the other age groups. This could indicate some bias against younger applicants, but as the differences fall within expected ranges it is not possible to draw a sound conclusion for this age group.

Annex – explanation of statistical tests

The data we have is not a sample of all possible data. It includes the whole 'population' of funding decisions, and there is nothing more that can be known. The most familiar statistical tests are used, broadly speaking, to infer something about a population from a sample, given that the process of sampling leads to uncertainty about the population. Here there is no uncertainty about the population.

Instead we are faced with uncertainty about how these outcomes might have played out if things had been different. If we know the range of possible outcomes, we can place the actual outcome within that range and see whether the observed outcome was unusual. This is what the tests applied in this analysis do. They have been used in two ways.

For the analysis of success rates within and across groups sharing a characteristic, and their differences, we are interested in the question of whether outcomes are independent of that characteristic. For example, does an applicant's ethnicity matter, in the sense of being associated with the outcome they experienced? We can test for this by randomly assigning ethnicities to applicants in the data (in the proportion found in the data) and then calculating the award rate for each group, or the award rate difference.

That is one instance of the sort of award rate (or difference in award rates) we might expect to see for that group if ethnicity had no influence on outcomes. If we repeat this process many times we create a distribution of the outcomes possible, conditional on the assumption that the characteristic is not associated with the outcome.

We can use this distribution to identify a range of outcomes that is compatible with a belief that outcome and characteristic are not associated. By comparing the observed result with this range of plausible outcomes, we can decide whether we believe that the observed outcome is too unusual for us to discount the possibility that outcome and characteristic are associated. To ensure that the range is a reliable reflection of what might have been, the sampling process has been repeated 10,000 times.

This is known as randomisation or permutation testing. The horizontal lines on the charts are a visual representation of the range of outcomes plausible under the null hypothesis of no association between outcome and characteristic. Sometimes these outcomes are award rates, sometimes they are differences in award rates and sometimes they are differences in differences. The logic of the test is the same in each case.

An interval covering 95% of the possible outcomes is used in the charts as it has some familiarity in relation to the traditional $p < .05$ statistical testing threshold. But it is an arbitrary choice, and others could be used which might be more or less conservative.

For tests associated with a partial randomisation process, the intervals show the range of outcomes that would arise if the randomisation had been repeated many times rather than just once. These intervals then show the extent to which (partial) randomisation of funding decisions might have had the ability to influence outcome, and where the actual randomisation used to make funding decisions sits in that spectrum.

Permutation tests make few assumptions, the main one being that of 'exchangeability'. This is simply the requirement that the labels associated with a person can be swapped freely with the labels associated with another person. The data we have does not strictly meet this requirement because if the same person appears in the data more than once, presumably with the same age etc, each instance of their appearance ought to have the same label. The same person cannot have two

different ages, and their characteristics may be associated with each other. In reality though, the instance of duplicated applicants will be rare, and the calculated ranges will be affected only imperceptibly by this violation (which will tend to make the ranges larger than they ought to be, making the error a conservative one.)