Dear Editor,

"How does correlation structure differ between real and fabricated data-sets?"

Thank you very much for processing the above manuscript for publication in the BMC Medical Research Methodology. The manuscript is now revised based on the comments made by the reviewer 2. A point-by-point response is prepared for your consideration.

We would like to thank the reviewer for his time and comments.

Sincerely,

Noori Akhtar-Danesh Mahshid Dehghan-Kooshkghazi

--------------------------------------------------------------------

Response to Reviewer 2:

1- Some more changes have been made to improve the writing style.

2- Specifying a cutoff point would be very difficult or even impossible because the correlation coefficient itself has a wide range (theoretically from -1 to +1). However, as mentioned in the manuscript, it is highly unlikely to be greater than 0.80 in absolute value (except for longitudinal data or repeated measures).

As has been shown, for calculating the false positive and false negative rates a reasonable number of fabricated data-sets (more than the current 34) and real data-sets are needed. Nonetheless, for Example 1 and Example 2 the following statements are included in the manuscript (amended in the manuscript):

Example 1: "In addition, in these made up data-sets 19 (56%) of them had correlation coefficients greater than 0.43. The correlation coefficient in 18 (53%) data-sets was greater than 0.70, and in ten (29%) was 0.90 or higher. In comparison, only 5.5% of the simulated data-sets had correlation coefficients statistically different from 0.43. Thus, the made up (fabricated) data-sets yielded considerably higher correlation coefficients than the corresponding real or randomly generated data-sets."

Example 2: "Furthermore, in the 22 made up data-sets with correlation coefficients statistically different from zero, 20 (59%) of them had a positive correlation coefficient and only in two the correlation coefficient was negative. Indeed, in 13 (38%) of them the correlation coefficient was greater than 0.70 and in 5 (15%) more than 0.90. Only 4.4% of the random samples from the real data-set had correlation coefficient different from zero."

Therefore, in the first example 53% of the fabricated data-sets have correlation coefficient greater than 0.70 which can be considered as the true positive rate for the cutoff point of 0.70, hence, the false negative rate is 47% (as the reviewer suggested). In the same time, only about 5.5% of
simulated data-sets have correlation coefficient different from the real data-sets resulting that the maximum rate of the false positive rate is 5.5%. Similarly, false positive and false negative rates for Example 2 are 4.4% and 62%, respectively.

Of course for the cutoff point 0.70 the false negative rate is still too big. However, it is not suggested as a diagnostic test but only as one method which could suggest further scrutiny for finding fabricated data. In any case, other methods such as checking the scatter plot should be taken into account.