Reviewer’s report

Title: Two distinct do-not-resuscitate protocols leaving less to the imagination: an observational study using propensity score matching

Version: 2
Date: 8 May 2014

Reviewer: Joseph Massaro

Reviewer’s report:

Major compulsory revisions.

I agree with the authors response to the previous review about the benefit of propensity scores approach over a multivariate logistic regression approach.

Matching, while an appropriate approach, reduces the generalizability of results to the entire non-DNR sample. For example, the mean non-DNR age in the entire sample is 54, but it is 64 in the matched sample. This needs to be discussed as a limitation (in addition to the limitation the authors already discuss regarding the lack of generalizability due to the single-center nature of the study) or the fact that the authors do not consider it a limitation needs to be discussed. On a related matter, the authors state that despite the single-center nature of the trial, there results may be generalizable because of the age and gender distribution of Cuyahoga county is similar to the US. However, the age mean of 40.2 in Cuyahoga county does not match the age mean of the matched sample (mean age of 64-66), so one could argue that the results may not be generalizable even to the entire Cuyahoga county. Please discuss.

The authors should strongly consider an N:1 matching (instead of 1:1) to use as much Non-DNR data as possible. The value of “N” to use is the value that provides a larger number of Non-DNR to be included while still yielding a good match.

The statistical methods that were used to compare matched groups on the outcomes were Student’s t-test and chi-square test. This is not necessarily appropriate as it does not take into account the potential dependency between the groups being compared due to the matching. The group comparisons on the outcomes should be performed using appropriate methodology for matched pairs.

The authors state that, in Model 1, a DNRCC patient was matched to a non-DNR patient with the “nearest propensity score”. Please state further details (e.g., was a "greedy” algorithm used or an “optimal” matching algorithm?) If a greedy algorithm was used, the authors may find even a better match with an optimal matching algorithm, or they may find a better match if they incorporate a greedy algorithm propensity matching with Mahalanobis’ distance.

A more appropriate and traditional way to assess whether matching worked (e.g.,
in Table 2) is not to compare p-values, but calculate standardized differences between groups. A standardized difference <10% generally is considered to indicate good matching. Please add standardized differences to Table 2.

As a supportive analysis, the authors should provide a propensity-adjusted analysis using ALL available patients, to assess if the results are consistent with the matched analysis. Such results could be provided in a supplementary appendix.

The authors use of 40 confounding variables is not a problem; in propensity score creation, as many clinically relevant variables as possible should be used, even if there is a chance they may not be related to the outcome. However, Table 1 does not seem to contain 40 variables but only approximately 15 variables. What were the 40 variables used? All such variables should be included in Tables 1, 2 and 4.

Minor essential revisions are:

Please state why data collection was suspended from March to May 2004.

Please present percentages in Table 3.

Highly skewed items such as daily cost and ICU stay should be presented with median and quartiles (in addition to or instead of mean and standard deviation).

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests.