May 13, 2015

Richard Irwin, MD

*CHEST* Global Headquarters

2595 Patriot Blvd  
Glenview, IL 60026

Re: The Effect of Indwelling Arterial Catheters in Hemodynamically Stable Patients With Respiratory Failure: A Propensity Score Analysis (CHEST-15-0516)

Dear Dr. Irwin,

We would like to thank you and the reviewers for the insightful comments regarding our manuscript, “The Effect of Indwelling Arterial Catheters in Hemodynamically Stable Patients With Respiratory Failure: A Propensity Score Analysis,” submitted to *CHEST*. In addition, we have added Dr. Kenneth Chen, MD as a co-author. Dr. Chen has played a critical role in helping us refine our propensity score model and aid in several sensitivity analyses. Below, please find our comments and responses to the Reviewers.

REVIEWER 1

MINOR COMMENTS:

1.  Propensity score variables, Appendix.  Classically, parsimony is not particularly desirable for a balancing score [Brookhart, American Journal of Epidemiology 163(12):1149-1156,2006; Caliendo, IZA Discussion Papers 2005(August 1), 2005].  A related issue is that one is not seeking the best fit of group assignment, but rather good balancing of covariates.  Clearly these are highly technical considerations and your choice of generating a matching score using a genetic matching scheme is perfectly OK, but please supply a reference in the main text (in addition to the Appendix).  I particularly find Diamond & Sekhon’s paper to be great (The Review of Economics and Statistics 95(3):932-945,2013), either in addition to the textbook you quote, or instead of it.

We would like to thank the reviewer for providing the Diamond and Sekhon reference{Diamond:2013tt}, which we have added as a reference in the main text.

But I do have a concern about the choice of the AUC as the “loss function” to maximize.  That choice implies that you sought to maximize predictive power of the final propensity score, rather than to explicitly obtain the best possible covariate matching.

The reviewer provides an insightful point regarding the appropriate diagnostics to perform for propensity score matching. The AUC or c-statistic, as we have reported, assesses the accuracy of a logistic regression model for predicting the exposure (i.e. IAC placement). As the reviewer has pointed out, there is debate within the literature as to the utility of the AUC in this context.{Rubin:2004cn}{Brookhart:2006gx} In building our propensity score model, we consulted several biostatisticians and epidemiologists with regards with regards to optimization that would best balance the confounders and performance of the propensity score model to predict insertion of an indwelling arterial catheters. Due to the large number of candidate variables considered, we had concerns about over-fitting the propensity score model using a genetic algorithm, but we are reassured based on our cross-validation, as well as the fact that the clinicians involved in this project were satisfied with the biologic plausibility of the included covariates. To further clarify our model building and associated diagnostics, we have included an additional diagnostic assessment in the Supplement (eTable 2), where we report the distributional balance of all 57 covariates used to develop our propensity score model, as well as univariate measures of association. As shown, there were no differences in propensity matched groups for all 57 original covariates, indicating adequacy of balancing of covariates. We have also performed an additional sensitivity analysis utilizing propensity score weights (PSW) to perform a weighted regression for outcome estimation. The PSW were generated by an algorithm that aimed at optimizing post-weighting balance of covariates between the treatment and the control group. We have included these results, which also show no difference in the primary outcome (28-day mortality) in the supplement and main text.

2.  Table 1.  In the Appendix you describe a set of 60 original variables.  Presumably, this set met the usual expectations that they would be expected to influence the outcome +/- the group assignment (Brookhart reference as above).  But in this table you only show matching for 28 of those variables.  Despite the fact that the Genetic Matching algorithm doesn’t include ALL those variables, the concept seems to imply that matching should be good for all of them.

We agree with the reviewer that the propensity score is sufficient to match for all the covariates in the original 57 we used to create our PS model. We have included in the Supplement eTable 2, which shows pre and post matching univariate measures of association for all 57 covariates. As shown, there were no differences in PS-matched groups for all 57 original covariates.

3.  P.8, line 29.  It is not clear why you refer to Figure 1 here.

We thank the reviewer for pointing out this incorrect reference to Figure 1, which we have removed from the main text.

4.  Table 2, last data row.  These days there are many patients who get venous blood gases (in addition to noninvasive pulse oximetry), instead of arterial gases.  Is this outcome just arterial?  Did you look at the number of venous gases?

The MIMIC-II dataset contains multiple blood gas data elements that include arterial blood gas and venous blood gas groups. After further examining the MIMIC-II dataset, we felt that the most reliable blood gas data element was a combined arterial and venous blood gas grouping. We apologize for the change, and have clarified this in the main text, tables, and appendix. The reviewer makes a very astute point that there might be an inverse relationship between the rate of venous blood gas measurements and arterial blood gas measurements, although we feel that the single combined blood gas outcome is a valid marker of blood gas sampling utilization.

Reviewer: 2

**Major Concerns:**

1. The Authors used 28-day mortality as the primary outcome of their study, however in many respects the secondary outcomes examined are of greater interest. Most clinicians do not expect the placement of an arterial line to be a life-saving procedure, but rather that it will improve intermediate outcomes, such as patient comfort (sparing them from repeated venous puncture), more rapid titration of vasopressors, avoidance of even brief periods of hypotension, and better management of mechanical ventilation. In an important sense, the primary question asked may not be as relevant as the other questions asked.

We would like to thank the reviewer for pointing out the relevance of the primary outcome (28-day mortality) as compared to secondary outcomes, which is central to this paper. We believe that assessment of both 28-day mortality, used in our and in the Gershengorn et al study, and secondary outcomes that seek to assess utilization, optimization, and burden are both central to patient-centered outcomes research. Our findings would suggest that not only is there no association between IAC use and an improvement in mortality, but also that there is no association between IAC use and utilization and optimization secondary outcomes relevant to our study population (length-of-stay, duration of mechanical ventilation, blood gas measurements). The reviewer is correct in pointing out that titration of vasopressors and avoidance of hypotensive episodes are interesting and valid outcomes to assess, although these outcomes do not apply to our study population of patients with respiratory failure without hemodynamic compromise. With regards to ventilator management, we feel that duration of mechanical ventilation is a good indicator of the quality of ventilator management. Also, patient comfort is a vitally important patient-centered outcome, although balancing increased phlebotomy rates (either arterial or venous) with longer mechanical ventilation time, ICU LOS and hospital LOS is a larger issue that is not addressed in our study.

1. There may be significant unmeasured confounders that were not addressed by the genetic algorithm used to derive the propensity score. These include such factors as previous ICU admissions, a history of prolonged mechanical ventilation, the reason for intubation, initial ventilator settings, and treating physician (since the practice patterns vary in this area). These and other considerations could play a central role in the decision to place an IAC, and could also be correlated with the outcomes that were measured. The Authors note the possibility of residual confounding in the Discussion, but do not elaborate on what these might be or how important an effect they may have.

We have added to the Discussion section specific covariates that may be unmeasured confounders, including prior history of ventilation, prolonged mechanical ventilation, and treating physician(s). With regards to indication for mechanical ventilation, our propensity score model includes PO2 and PCO2 measurements, as well as co-incident diseases including COPD, heart failure, and pneumonia. We believe that all these covariates reflect the initial indication for mechanical ventilation. We have further added language regarding the potential effects of negative confounding on our results into the main text.

1. There as on to focus only on patients who were intubated in the ICU, those who were hemodynamically stable, and those without a diagnosis of sepsis, is not adequately explained. Though there is mention in the Discussion that hemodynamic compromise and sepsis were excluded as alternate reasons to place an IAC, the rationale for focusing only on pure respiratory failure is not provided. Is there existing literature examining this question in patients with sepsis and hemodynamic compromise?

There is an absence of literature on the use of IAC in all subgroups of critical care patients. Due to the heterogeneity of the ICU population, our analysis of only patients with respiratory failure was intended to create as homogenous a group of ICU patients as possible in the context of this heterogeneity. By eliminating other potential indications for IAC placement, such as sepsis and/or hemodynamic compromise, we are attempting to minimize effect modification or interaction, as the association between IAC placement and patient outcomes may be different in these patient groups.

Further, the heterogeneity of in the ICU is also relevant to optimal propensity score modeling. The relationship between covariates in the propensity score model and outcome (28-day mortality) may vary based on indications for IAC placement (e.g. acute respiratory failure vs septic shock vs acute respiratory failure and septic shock). As Reviewer 1 has pointed out, inclusion of variables related to the outcome but not to the exposure should be included as this will decrease the variance of the exposure effect. In contrast, inclusion of covariates in the propensity score model that are related to the exposure but not the outcome will increase the variance of the estimated exposure effect without a decrease in bias.{Brookhart:2006gx} By limiting our study sample to a single indication for IAC placement, we are attempting to optimize the PS model both for exposure prediction standpoint and for assessment of the exposure to the outcome. Of note, we plan on performing subsequent analyses in MIMIC-II and larger EHR-derived datasets for other ICU sub-groups with different indications for IAC placement. We have added text to clarify this issue in the Discussion section.

1. The methodology used does not address the fact that the decision to place an IACs at the time of admission is not the only relevant decision in this regard. Ideally, the decision is revisited daily on rounds, to see if the IAC should be removed. Is there a way to use the MIMIC II data to estimate the length of time that the IAC was in place? This might add an additional “dose effect” that could be useful in strengthening the conclusions.

We agree with the reviewer that frequent re-assessment of the need for continued IAC use is a crucial part of daily multidisciplinary rounding in the ICU setting. The question of whether the duration of IAC use has a “dose effect” on mortality will be difficult. Even with controlling for severity of illness (such as with daily calculated SOFA scores), we feel that it will be very difficult to separate out the IAC effect from severity of critical illness. Thus, while the MIMIC-II database does have time-stamped IAC information that allows for calculation of the duration of IAC placement, we feel that duration of IAC placement and the association with outcomes is best examined as a separate project.

**Minor Concerns:**

INTRODUCTION 1. There is no explanation provided for why this particular subset of patients was examined. Why exclude hemodynamically unstable patients, or patients not being ventilated? Is there some *a priori* hypothesis about why this group of patients might have a mortality benefit from IAC use?

MATERIALS AND METHODS

1. Why include only those patients who had IACs placed after intubation? How did the  Authors determine when the IAC was placed (it might be useful to describe this as  part of the supplements)? Were there data on why the IAC was placed?

The MIMIC-II database includes time-stamped data indicating time of IAC placement by critical care nursing. Placement of an IAC after endotracheal intubation was an inclusion criteria in order to optimize the propensity score model as explained above.

1. Why were patients with sepsis excluded?

As detailed above, patients with sepsis were excluded in order to minimize interaction and in order to optimize our propensity score model.

1. On page 6, it says that 30 features were used to estimate the propensity score but in  the Appendix where the model is shown, it says that the model consisted of 29  features.

We would like to thank the Reviewer for pointing out this discrepancy. We have revised the main text on page 6 to indicate that 20 covariates were used in our final propensity score model.

1. The statistical analysis is complex and would benefit from a dedicate review. Why  was the Fisher’s exact test chosen instead of Chi-square (given the large sample size) or McNemar’s test (given that the patients are paired)?

We would like to thank the reviewer for pointing out the paired relationship of the matched cohort analyses. As the matched cohorts are not independent, we have updated our univariate comparison of baseline variables in the matched cohort by displaying results of a McNemar’s test for categorical variables and Wilcoxon Signed Rank Test for continuous variables (Table 1). For assessment of outcomes, we have updated our results by performing McNemar’s test for categorical outcomes and paired t-tests for continuous outcomes (with assumption of the Central Limit Theorem). These findings have been updated in the main text, as well as Table 1, Table 2, and eTable 2.

RESULTS

1. Page 8, line 30 - reference to Figure 1 should be Table 1?

We thank the reviewer for pointing this confusion out. The reference to Figure 1 has been removed from page 8, line 30.

2. While Table 1 is useful in showing that the matched cohorts are similar, the reliance on the non-significant p-values may be misleading since the newer cohorts are smaller in size than the original cohorts. A comparison of boxplots and/or histograms of the baseline characteristics may be useful in evaluating the success of matching (perhaps as a part of the supplement)

We have added an additional figure to the appendix (eFigure 1) displaying histograms of all covariates considered in developing the propensity score model.

DISCUSSION

1. The discussion of technology adoption in the ICU is useful and well presented

We thank the reviewer for his or her comment.

1. The Authors state that RCTs of the use of new devices are unlikely to take place because of the cost and logistical challenges of performing RCTs in the ICU. However a great many RCTs have been conducted in ICU settings, including complex multi-center studies. Moreover, the cost barriers should be less in these cases as for the most part, studies of new devices are likely to be industry sponsored.

The argument that databases such as MIMIC II should be explored to identify subpopulations is more compelling. These databases might also be useful in identifying appropriate outcomes for RCTs, since measuring 28-day mortality might not be the most useful approach. Many of these devices will be used for convenience, comfort, cost, and efficiency, rather than because they save lives.

We would like to thank the reviewer for pointing out the multitude of ICU-specific RCTs. We believe that the use of large clinical databases, such as MIMIC-II, are useful in providing pilot data to drive the development of RCTs. For example, we feel that our analysis, taken in conjunction with the Gershengorn et al analysis, provide justification for a randomized clinical trial of indwelling arterial catheters. We further believe that the use of databases such as MIMIC-II can help identify sub-populations of patients that may benefit most from these interventions, thus further aiding in the development of RCTs, especially pragmatic RCTs. We agree with the reviewer that databases such as MIMIC-II can further informing RCT design by helping to establish appropriate primary and secondary outcomes.

1. Page10, second to lastline“...contains comprehensive electronic health record electronic health record data...” (repeated phrase)

APPENDIX 1. Demographic covariates include “ethical group”, which perhaps should be “ethnic group”?

We thank the reviewer for pointing out these typographical errors. We have corrected these errors in the main text and appendix.