S1 - Supplementary material to: Assessment of patient-specific surgery effect based on weighted estimation and propensity scoring in the re-analysis of the Sciatica Trial.

Mertens, B. J. A.\textsuperscript{1}, Jacobs\textsuperscript{2}, W.C.H., Brand\textsuperscript{1}, R., and Peul\textsuperscript{2}, W.

October 3, 2014

Departments of Medical Statistics\textsuperscript{1} and Neurosurgery\textsuperscript{2}, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, Netherlands

1 Inverse probability treatment weighted analysis

Marginal structural models are estimated in two stages. At the first stage, a treatment prediction model is fit, which is used to calculate the patient-specific probabilities to receive treatment (surgery in this case) at any point in time while the patient has not yet received the treatment. At the second stage, the inverses of these treatment probabilities are used as weights in the calibration of the substantive model, such as a regression model or other on

*Correspondence to: Bart J. A. Mertens, b.mertens@lumc.nl, http://www.lumc.nl. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
the outcome of interest, which is used for the assessment of the treatment effect.

For the fitting of the treatment prediction model we may only use the observed data up to and including the time of surgery, or end of follow up, whichever comes sooner. All data subsequent to the surgery is discarded. Let \( k \in \{1, \ldots, 9\} \) denote the visit numbers in follow-up and \( k = 0 \) denotes the baseline. Then we write \( L(i, t_{ik}) \) for the set of all observed scores (VAS1, VAS2, Roland and Likert) for any \( i^{th} \) patient at the \( k^{th} \) visit, with corresponding visit time \( t_{ik} \) after baseline. A patient who remains free of surgery till end follow-up will contribute the full set of measurements \( \{L(i, t_{i1}), \ldots, L(i, t_{i9})\} \) across all visits to the fitting of the treatment prediction model. If an \( i^{th} \) patient receives surgery at time \( t_{is} \) and there are \( k - 1 \) scheduled visits prior to surgery for this patient, then the patient contributes the set \( \{L(i, t_{i1}), \ldots, L(i, t_{i(k-1)}), L(i, t_{is})\} \) to the data used for fitting the weights model. Should surgery take place prior to the next scheduled visit, then we define the visit number corresponding to the final measurement to be \( k(t_{is}) = k \). In addition, we have for each \( i^{th} \) patient the baseline set of covariates \( V(i) \), which consists of the above mentioned four baseline scores \( L(i, 0) \), besides age, gender, height and weight. In analogy to the above, we define an additional set \( M(i, t_{ik}) \) of time-dependent measures on visits \( k > 0 \) defined as the changes in the VAS1, VAS2, Roland and Likert scores from baseline at each time point as well as the changes in these scores from each previous time point (lag values). Finally, we may define the treatment (surgery) indicators \( A(i, t_{i}) \) for any patient such that \( A(i, t_{i}) = 1 \) for any time point from surgery onwards \( (t_{i} \geq t_{is}) \) and zero otherwise or when surgery never took place for that patient. Note \( A(i, 0) = 0 \) by definition and
that treatment can only be given once.

1.1 Treatment probability prediction model for IPTW weighting

Calibration of a treatment probability model is formally a survival analysis problem on time till treatment and thus Cox regression could be used. Hernan, Brumback and Robins [1][2][3][4] propose pooled logistic regression models for the estimation of the treatment probabilities. While the main arguments seem to be simplicities of the approach and ease of implementation, we refer readers to the above cited papers for full discussion of the pros and cons involved. To remain consistent with these publications, we decided to apply the same methodology on the sciatica data in this paper. Let \( P_{i,t} \) be the treatment probability for the \( i^{th} \) patient at time \( t \) and conditional on the patient not yet having received treatment at that point. We then formulate the model as

\[
\log \left( \frac{P_{i,t}}{1 - P_{i,t}} \right) = \alpha + h(t) + [L(i,t)\beta_L] + [M(i,t)\beta_M] + [V(i)\beta_V]. \tag{1}
\]

The term \([L(i,t)\beta_L]\) is a shorthand notation for the linear combination of the scores within the predictor set \( L(i,t) \) with regression coefficients in the vector \( \beta_L \) and similarly for the baseline component \([V(i)\beta_V]\) and for \([M(i,t)\beta_M]\). The function \( h(t) \) is a function of time since randomization only and defined as

\[
h(t) = \gamma t + \alpha I(k(t) = 5),
\]
where $I()$ denotes the indicator function having a value of 1 when its argument is true and 0 otherwise and $k(t)$ denotes the visit interval number which corresponds to time $t$. The last term represents a probability spike at the fifth visit. The latter is to account for a sudden increase in surgeries carried out as a consequence of the trial actively offering the surgery to patients who had not yet received it at that stage and did not show recovery of function or pain. We carried out a confirmatory analysis on our choice of the time component by starting with the more general model

$$h(t) = \gamma t + \sum_{k=1}^{9} \alpha_k I(k(t) = k)$$

and then incrementally merging time points and their coefficients $\alpha_k$ from the last visit backwards or from the first visit onwards. By checking deviances we found that all time points could indeed be merged within the model, except for the fifth visit. While this approach is simple, it is however essential to enforce parsimony on the calibration of the weights as much as possible, because inverse probability weighted effect estimation could be highly sensitive to excess variation of the weights (see subsequent results and discussion).

1.2 Inverse probability treatment weights

Once the treatment probabilities $P_{i,t}$ have been calibrated at the observed visit and treatment times, the estimated weight $W(i)$ for any $i^{th}$ person are defined as

$$W(i) = \frac{W_{num}(i)}{W_{denom}(i)} \quad (2)$$
with the denominators defined as

\[ W_{\text{denom}}(i) = P_{i,t_{is}} \prod_{l=1}^{k-1} (1 - P_{i,t_{il}}) \]  

(3)

if the individual did have treatment at time \( t_{is} \). Otherwise we have

\[ W_{\text{denom}}(i) = \prod_{l=1}^{9} (1 - P_{i,t_{il}}) \]

(4)

for those people never taking treatment up to end of the trial. The numerators \( W_{\text{num}}(i) \) are stabilizing weights calculated in the same manner from complementary logistic models to the above described. They have identical definition as the above model, except that all the time-dependent predictors are removed from the model (keeping only the above defined 8 baseline predictors and the time component). An inverse probability of treatment weighting approach then applies these weights to the likelihood as in

\[ \prod_{i=1}^{n} [lik_{i}]^{W(i)}, \]

where the \( lik_{i} \) are the likelihood components for each \( i^{th} \) patient (see also [1], page 442).

Tables 1 and 2 show median of bootstrap effect estimates as well as a 95% confidence interval from both an inverse treatment probability weighted analysis (IPTW, first 3 columns) as well as a classical unweighted analysis (next 3 columns, adjusted for baseline variables only) of the above covariance regression model for Roland and Likert scores respectively, based on 10000 bootstrapping repetitions, to account for the variability of the weight esti-
mation. The top row in the table (‘all preds’) gives results when including all predictors.

The weights model may leave unknown confounders unaccounted for. Nevertheless, the model may include too many predictors, particularly as these are highly correlated (such as the baseline and subsequent change from baseline measures) and thus may inflate the variance of weight prediction, which is undesirable. Some balance must be struck between the need to enforce parsimony of weight estimation as much as possible with capturing the required confounding information in the weight prediction. We address the problem by successively reducing the weight prediction model by removing predictors and then re-calculating the weighted estimates, removing the change scores last. The results from this procedure are shown in subsequent rows to the first where predictors are removed from the model as shown in the last column. For consistency, we always remove the same predictors from the covariance regression models (both IPTW and Classic). The last column in both tables shows deviances from the logistic regression weight model 1 on treatment probability, which give an indication of the impact of removing predictors on the change in calibrated weights relative to the previous model.
Table 1: Median of bootstrap estimates and 95% confidence intervals based on 10000 bootstrap replications for the treatment effect on Roland scores at end of study, based on application of IPTW weighted analysis (columns 1-3) as well as unweighted analysis (classic, columns 4-6) of a covariance adjusted regression of Roland score at end of study on treatment indicator, adjusted for baseline covariables. A marginal weighted estimate is also presented (columns 7-9). The analysis is repeated in subsequent rows to the first by removing predictors from the treatment probability prediction model (column 10). Deviances of each treatment probability prediction model are shown in the last column.

<table>
<thead>
<tr>
<th>IPTW</th>
<th>2.5%</th>
<th>med</th>
<th>97.5%</th>
<th>2.5%</th>
<th>med</th>
<th>97.5%</th>
<th>2.5%</th>
<th>med</th>
<th>97.5%</th>
<th>Model</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.21</td>
<td>-2.10</td>
<td>0.07</td>
<td>-2.72</td>
<td>-1.47</td>
<td>-0.14</td>
<td>-2.94</td>
<td>-1.39</td>
<td>0.42</td>
<td>all preds</td>
<td>315.50</td>
</tr>
<tr>
<td></td>
<td>-4.74</td>
<td>-2.40</td>
<td>0.01</td>
<td>-2.42</td>
<td>-1.25</td>
<td>-0.02</td>
<td>-3.44</td>
<td>-1.66</td>
<td>0.20</td>
<td>-gender, weight, height</td>
<td>318.14</td>
</tr>
<tr>
<td></td>
<td>-5.07</td>
<td>-2.49</td>
<td>0.06</td>
<td>-2.51</td>
<td>-1.32</td>
<td>-0.07</td>
<td>-3.59</td>
<td>-1.69</td>
<td>0.23</td>
<td>-age</td>
<td>318.35</td>
</tr>
<tr>
<td></td>
<td>-5.07</td>
<td>-2.52</td>
<td>-0.03</td>
<td>-2.46</td>
<td>-1.32</td>
<td>-0.09</td>
<td>-3.61</td>
<td>-1.74</td>
<td>0.21</td>
<td>-VAS1/2 lags</td>
<td>323.35</td>
</tr>
<tr>
<td></td>
<td>-4.77</td>
<td>-2.51</td>
<td>-0.16</td>
<td>-2.48</td>
<td>-1.31</td>
<td>-0.09</td>
<td>-3.42</td>
<td>-1.72</td>
<td>-0.08</td>
<td>-Roland (chg/lags)</td>
<td>325.32</td>
</tr>
<tr>
<td></td>
<td>-4.23</td>
<td>-2.17</td>
<td>-0.19</td>
<td>-2.50</td>
<td>-1.31</td>
<td>-0.11</td>
<td>-3.06</td>
<td>-1.61</td>
<td>-0.12</td>
<td>-Likert (chg/lags)</td>
<td>340.40</td>
</tr>
<tr>
<td></td>
<td>-2.47</td>
<td>-1.32</td>
<td>-0.09</td>
<td>-2.47</td>
<td>-1.32</td>
<td>-0.09</td>
<td>-1.83</td>
<td>-0.84</td>
<td>0.14</td>
<td>-VAS1/2 chg</td>
<td>374.98</td>
</tr>
</tbody>
</table>
Table 2: Median of bootstrap estimates and 95% confidence intervals based on 10000 bootstrap replications for the treatment effect on Likert scores at end of study, based on application of IPTW weighted analysis (columns 1-3) as well as unweighted analysis (classic, columns 4-6) of a covariance adjusted regression of Likert score at end of study on treatment indicator, adjusted for baseline covariates. A marginal weighted estimate is also presented (columns 7-9). The analysis is repeated in subsequent rows to the first by removing predictors from the treatment probability prediction model (column 10). Deviances of each treatment probability prediction model are shown in the last column.

<table>
<thead>
<tr>
<th>IPTW</th>
<th>Classic</th>
<th>Marginal weighted</th>
<th>Model</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>med</td>
<td>97.5%</td>
<td>2.5%</td>
<td>med</td>
</tr>
<tr>
<td>-0.72</td>
<td>-0.21</td>
<td>0.24</td>
<td>-0.62</td>
<td>-0.28</td>
</tr>
<tr>
<td>-0.69</td>
<td>-0.20</td>
<td>0.25</td>
<td>-0.55</td>
<td>-0.24</td>
</tr>
<tr>
<td>-0.68</td>
<td>-0.19</td>
<td>0.28</td>
<td>-0.57</td>
<td>-0.26</td>
</tr>
<tr>
<td>-0.70</td>
<td>-0.23</td>
<td>0.23</td>
<td>-0.57</td>
<td>-0.26</td>
</tr>
<tr>
<td>-0.75</td>
<td>-0.29</td>
<td>0.16</td>
<td>-0.58</td>
<td>-0.26</td>
</tr>
<tr>
<td>-0.73</td>
<td>-0.29</td>
<td>0.15</td>
<td>-0.57</td>
<td>-0.26</td>
</tr>
<tr>
<td>-0.58</td>
<td>-0.26</td>
<td>0.05</td>
<td>-0.58</td>
<td>-0.26</td>
</tr>
</tbody>
</table>
1.3 Marginal weighted analysis

For comparison, we may consider a marginal analysis, using the new weights

\[ W_m(i) = \frac{1}{\sum_{j=1}^{n} \frac{1}{W_{denom}(j)}}, \]  

(5)

where the \( W_{denom}(i) \) are defined as before and use the same baseline and
time-dependent covariates and the sum is over all individuals. We can now
define a weighted mean difference in an outcome measure \( X \) (such as \( L(t_9) \)
in this paper) as

\[ \Delta = \sum_{i=1}^{n} W_m(i)X(i)I(T(i) = 0) - \sum_{i=1}^{n} W_m(i)X(i)I(T(i) = 1). \]  

(6)

A (bootstrapped) standardized version of \( \Delta \) can be considered an analogue
of a t-statistic. See the tables for these results.

1.4 Propensity scores for stratification

We define the propensity scores \( P(i) \) as the probability of treatment for each
\( i^{th} \) patient, based on model 1 with all predictors added, but now only retain-
ing either the last calculated probability of treatment \( P_{i,t_{is}} \) from the model if
the person has treatment at time \( t_{is} \) and otherwise \( P_{i,t_{is}} \) for untreated indi-
viduals at the end of the trial (last visit = 9). The corresponding estimates
are simply obtained from the fit of the full pooled logistic model discussed
before for the MSM approach, hence no new calculations are required.
1.5 Regression model with propensity scores

Adjusted estimates of treatment effect are obtained by adding the propensity scores and baseline variables as covariables to a simple linear regression model on the final observed outcome scores at visit 9, besides the treatment indicator. So the final regression model for the propensity-based regression analysis is

\[ L(i, t_{9}) = \alpha + \beta T(i) + \gamma P(i) + \delta T(i)P(i) + [V(i)\delta] + \varepsilon \] (7)

where \( P(i) \) is the propensity score treated as a continuous covariate and \( T(i) \) the treatment indicator for treatment prior to last visit and \( T(i)P(i) \) the interaction term.

References


