Staged Bilateral Total Joint Arthroplasty Patients in Registries. Immortal Time Bias and Methodological Options.
Stéphanie L. van der Pas, MSc, MA, Rob G.H.H. Nelissen, MD, PhD and Marta Fiocco, PhD.

Investigation performed at the Departments of Orthopaedics and Medical Statistics of the Leiden University Medical Center, and at the Mathematical Institute, Leiden University, Leiden, The Netherlands.

Background: Patients with staged bilateral total joint arthroplasty (TJA) in arthroplasty data pose a problem in the statistical analysis. Subgroup analysis, in which patients with unilateral and bilateral TJA are studied separately, is sometimes considered an appropriate solution to the problem; we aim to show that it is not, due to the immortal time bias.

Methods: We consider staged (at any time) bilateral patients. The logical fallacy leading to the immortal time bias is explained through a simple artificial data example. The cumulative incidences of revision and death are computed by subgroup analysis and by landmark analysis, both on hip replacement data from the Dutch Arthroplasty Register and on simulated data sets.

Results: For unilateral patients, subgroup analysis can lead to an overestimate of the cumulative incidence of death, and an underestimate of the cumulative incidence of revision. The reverse conclusion holds for staged bilateral patients. The immortal time bias can be prevented by using landmark analysis.

Conclusions: Staged bilateral patients should be analysed with caution during analysis of arthroplasty registry data. An appropriate statistical method to address the research question should be selected.

Worldwide more than 3 million total hip and knee arthroplasties are performed annually, and this number is predicted to increase substantially within the next decades. In general outcome is measured in terms of mortality, complications or as revision rate differences between implants. Data on total joint arthroplasties (TJAs) are collected in an increasing number of arthroplasty registries around the world, and the resulting data has proven to be valuable in improving the outcome of TJA. However, since during the follow-up period patients tend to have additional TJAs, these "new" primary TJAs will interfere with analysis on the first primary joint. The number of patients with bilateral TJAs is not negligible; in The Netherlands 20% of total hip arthroplasty surgeries in 2014 concerned the placement of a second prosthesis, in Sweden 20.5% of patients became staged bilateral between 1992-2014, and in Norway, 23.6% of patients became bilateral within 10 years.

A basic principle in survival analysis is that subgroups defined by patient characteristics not known at the start of follow-up (such as receiving a second TJA), can only be compared with the greatest caution. The reason is the immortal time bias, a well-known phenomenon in observational studies, resulting from flawed statistical analysis. Immortal time refers to a period of follow-up during which the study outcome, which may be death or another event, cannot occur. It was first described in the context of heart transplant data, when it was noted that the observed improved survival of heart transplant patients was due to selection bias: only patients who survived long enough to receive a heart transplant were included in the transplant group. Analyses of arthroplasty data risk being affected by the immortal time bias as well. The immortal time bias arises when patients with staged bilateral TJA are studied as a separate subgroup, because only those patients who survive long enough to be able to receive the second implant are observed. The bias occurs when either revision of one of the implants or death are taken as an endpoint. This is a different methodological problem than that of competing risks, although that should be accounted for as well.

Problems with the statistical analysis of arthroplasty data which include patients with staged bilateral TJA, who are fundamentally different from unilateral TJA patients, have been
noted. The focus so far has been on the dependence between two TJAs in one patient, and not on the time-dependent bilateral status. No consensus has been achieved on how to analyse these data in a methodological sound way. A review of papers in seven high-impact orthopaedic journals showed that 42% of clinical studies with multiple observations per patient used these multiple observations inappropriately. One recommendation by Bryant et al. 2006 was to analyse patients with bilateral TJA as a separate subgroup, which is done in practice, or they are excluded from analysis. However, both methods may lead to misinterpretation due to the immortal time bias.

The aim of this article is to explain why subgroup analysis is inappropriate for estimating revision rates in TJA, and to suggest alternative statistical techniques. Which of the methods is appropriate depends on the research question. Landmark analysis is explained in detail, because it is a relatively new method that is not commonly applied to arthroplasty data, even though it is designed to prevent immortal time bias. It is especially well suited for research questions in which unilateral and staged bilateral patients are compared, and also for comparisons of subgroups within the staged bilateral patient population. A second type of research question concerns analyses on the entire arthroplasty patient population, with no particular interest in bilateral patients. This type of research question is more common than the first. Leaving out the bilateral patients is not appropriate in this case either. We review some statistical methods for this second type of research question in the Discussion.

Methods

We show the mechanism behind the immortal time bias and potential consequences on three datasets: an artificial data example, simulated datasets, and data from the Dutch Arthroplasty Register. The analyses are for total hip arthroplasty (THA), but the same reasoning holds for any joint of which there are at least two, such as knees, shoulders and ankles. We only consider staged (at any time during follow-up of the index TJA) bilateral TJA patients, as one-stage (i.e. within one surgical session) bilaterals do not contribute to the immortal time bias. One-stage bilateral TJA will be discussed in the Discussion.

Competing risks

A patient may die before experiencing revision of the implant. If this competing risk of death is not accounted for, the risk of revision surgery will be overestimated. This is especially important for arthroplasty analyses given the age of most patients (average age at index surgery in The Netherlands is 69 years for THA; 68 for total knee arthroplasty). The need for competing risks is illustrated in Figure 1, which shows a strong competing risk of death for patients older than 80. Competing risks methods for unadjusted and adjusted analyses include the Aalen-Johansen estimator, Fine-Gray regression, or cause-specific Cox regression.

Artificial data example

The underlying mechanism of the immortal time bias is illustrated through an artificial example, in which 50% of patients will become bilateral exactly 2 years after index surgery. The first-placed implants of unilateral and bilateral patients are compared. The implants of all patients behave the same: they have a 30% probability of revision after exactly 3 years. Furthermore, each patient has a 20% probability of dying after 1 year. The process is visualized in Figure 2. All percentages are chosen for illustrative purposes and are not meant to be realistic. We consider a second scenario with 5% mortality, 20% bilateral and 5% revision. We assume independence for all events.

Landmark analysis
Landmark analysis allows for comparison of unilateral and staged bilateral patients without the risk of immortal time bias.\textsuperscript{18,19} The first step is to choose a landmark time of for example 2 years. The choice of landmark time should be guided by the research question, and should be long enough that both subgroups have sufficient members for analysis. Only patients who are still alive and have not experienced revision yet at the landmark time are included in the analysis. This ensures a fair comparison between the two groups, as both need to survive for the same minimum amount of time in order to be included in the analysis.

The next step is to create the subgroups: patients who have become bilateral by the landmark time, and patients who were unilateral at the landmark time. As only each patient’s status at the landmark time is considered, the latter group includes patients who may receive a second implant after the landmark time. The procedure is illustrated in Figure 3.

When the landmark subgroups have been made, the appropriate statistical method can be applied.\textsuperscript{15,17} The interpretation of the resulting models is conditional on the landmark time. Thus, conclusions can be drawn for comparisons between unilateral and staged bilateral patients, conditional on the fact that these patients were still alive and did not undergo revision until the landmark time. This is a limitation to this method, which will be discussed further in the Discussion.

\textit{Simulation study}

Two thousand datasets were generated consisting of 10,000 patients each, in three settings: 10, 20 or 30 percent potentially bilateral patients. The simulation parameters are in Table 1.

In all scenarios, there is no difference in the risk of revision for the first prostheses of unilateral and staged bilateral patients. All patients had the same risk of death. These mechanisms are unlikely to be realistic, but serve to clearly illustrate the potential bias induced by performing a subgroup analysis.

The two groups under comparison are unilateral patients and staged bilateral patients. For the landmark analysis, a landmark time of 3 years was used. The cumulative incidences were estimated by the Aalen-Johansen estimator and a two-sided hypothesis test was performed.\textsuperscript{17,20} By construction of the simulated data, every significant test result is a false positive.

\textit{Dutch Arthroplasty Registry data}

So far, the discussion has focused on comparing unilateral and staged bilateral patients. However, the bias is not limited to this situation alone, but may occur in any comparison based on patient characteristics that develop over time, such as becoming bilateral within a specified number of years. Despite this bias, such comparisons are carried out in practice.\textsuperscript{6,12,21} We illustrate this phenomenon with data from the Dutch Arthroplasty Registry.\textsuperscript{22} All patients who underwent their first primary THA between 2007 and 2014, and became staged bilateral within at most 3 years after the first THA were included, except after diagnoses of trauma or a tumor. In total, 14,067 staged bilateral patients were analyzed.

Two groups were compared: patients who had staged bilateral TJA within 1.5 years, or after 1.5 to 3 years, respectively. The landmark analysis was performed at a 3 year landmark point. The numbers of patients and events are in Table 2.

In both analyses, the Aalen-Johansen estimator of the unadjusted cumulative incidences was computed.\textsuperscript{17} A two-sided hypothesis test was performed at the 5% level.\textsuperscript{20}

\textit{Source of funding}

The Netherlands Organization for Scientific Research provided funding for this study, but did not play any role in this investigation.
Results

Artificial data example
Results are reported in Table 3. With subgroup analysis, the cumulative incidence of revision is underestimated and the cumulative incidence of death overestimated for unilateral patients. The reverse conclusion holds for staged bilateral patients.

Simulation study
The percentages of false positives resulting from all analyses are in Figure 4. Figure 5 illustrates the results, and is representative for all simulated data sets.

Dutch Arthroplasty Registry data
The subgroup analysis finds a significant difference between the groups of patients (revision: \( p = 0.001 \), death: \( p < 0.0001 \)); the landmark analysis does not (revision: \( p = 0.89 \), death: \( p = 0.50 \)). The cumulative incidences at 6 years are in Table 4, and the results are illustrated in Figure 6.

Discussion

General
When the outcome of interest is revision of a joint prosthesis, the statistical analysis is complicated by two issues: the presence of staged bilateral patients and the competing risk of death. The immortal time bias lies at their intersection. In general, immortal time bias can occur when a ‘responder’ and ‘non-responder’ group are studied, but the responder needs to survive long enough to be able to respond. With arthroplasty data, when the outcome of interest is revision, the bias is subtle. Revision of the first hip does not prevent a patient from joining the staged bilateral group, and thus there is no obvious immortal time bias. However, there is the competing risk of death. The resulting bias is explained via the artificial data example.

Artificial data example and simulation study
The artificial data example shows the basic mechanism behind the immortal time bias. As illustrated in Figure 2, when the unilateral and staged bilateral subgroups are created at the end of follow-up, patients that would have become bilateral at the 2-year mark but died before realizing that potential, are observed to be unilateral. This leads to an estimate of a zero probability of death for staged bilateral patients, while the cumulative incidence of death is overestimated for unilateral patients. The reverse happens for revision: the cumulative incidence of revision is overestimated for staged bilateral patients, as their deaths are not observed, while it is underestimated for unilateral patients, because the risk set is made artificially large by the inclusion of patients who would have become staged bilateral if they had not died before the second surgery could take place.

The main advantage of a simulation study is that we can decide on an underlying truth and then study the effects of the subgroup and landmark approaches. The simulation study shows that a subgroup analysis can be misleading, leading to more false detections of a difference than landmark analysis. The estimate of the cumulative incidence of death is more affected than the estimate of the cumulative incidence of revision. The immortal time bias increases as the percentage of bilateral TJA patients increases, as is to be expected.

The severity of the effect of the immortal time bias depends on the revision, mortality and bilaterality rates, and also on the research question. With 5% revision, 5% mortality and 20% bilaterality, the bias in the artificial data example is inconsequential for the cumulative incidence of revision, but still relatively large for death. Moreover, statistically significant differences in implant survival between two groups can be very small, even less than 1%, when the follow up is short. In such a case, even a small bias may be large enough to give the false impression of a difference between subgroups where there is actually none. In addition, the
Swedish Hip Arthroplasty Register reports 23-year revision rates of up to 38.5% for men who are 50-59 years at index surgery. After such a long follow up, immortal time bias may significantly affect analyses, and thus clinical results without proper statistical analysis should be interpreted with caution.

Dutch Arthroplasty Register data
The analysis of the THA data from the Dutch Arthroplasty Registry shows the potential effects on outcome estimates of performing a subgroup analysis. The detection of a difference in survival and implant survival between patients who become staged bilateral within 1.5 or after 1.5-3 years may be caused by the immortal time bias. The patients whose second implant is placed after at least 1.5 years need to survive longer to be able to be included in the analysis than patients whose second implant is placed within 1.5 years. The lower left panel of Figure 6 should alert the researcher to the potential bias in the subgroup analysis, as no deaths are observed within the first two years in the 1.5-3 years group.

The landmark analysis finds no difference in the cumulative incidences of death and revision for both groups of staged bilateral patients. The result should be carefully interpreted however, as the research question is a conditional one: the analysis estimates the cumulative incidence for those patients who survive for at least 3 years and do not undergo revision within that time. Figure 6 shows that the cumulative incidence of revision resulting from the landmark analysis is much lower than that in the subgroup analysis, because revisions before the landmark time are not included. Thus, we cannot tell whether there is truly no difference between the two subgroups, or if there is a difference that disappears before the 3 year landmark time.

Limitations of landmark analysis and alternative methods
The main advantages of landmark analysis are that it avoids immortal time bias, and it is simple to implement. On the other hand, the main disadvantage is that the conclusions only hold for patients who are still alive and unrevised until the landmark time point. However, this does not need to be a disadvantage, as the resulting answer will be of interest to a patient who has survived some time unrevised since the index TJA. However, excluding the first few postoperative months or even years from analysis may not be satisfactory in a situation where mortality or revision risk are high immediately following surgery. We thus review some alternative suggestions, and distinguish two types of research questions: questions related to all patients within the cohort, and questions in which bilateral patients are studied in particular.

We first consider methods for research questions about all patients. Two reviews of arthroplasty studies found that is commonly believed that the bilateral patients do not affect the results of the analyses too much, and thus the dependence of their observations is often ignored. Recommendations from these reviews are (i) using statistical methods that model the correlation within patients, (ii) excluding the second limb, (iii) randomly choosing one limb per bilateral patient, or (iv) analyzing bilateral patients as a subgroup. Options (i) and (iv) do not account for the time between the two surgeries and are only suitable for one-stage bilaterals, and (iii) may introduce bias through the dependence on the particular sample. Option (ii), only using the first TJA per patient, seems most appropriate, as this removes all dependence issues and ensures that all patients are on the same time scale. A drawback is that not all data is used, but this is not necessarily problematic in registry studies, where the number of observations can be in the hundred thousands.

There is a fifth option, which is to include every observation and do the analysis as if all joints are independent, and to perform some sensitivity analysis. This is the recommendation of Ranstam et al., and indeed, when the question is about all TJA patients, when revision and mortality are low and sample sizes are large, incorrectly assuming independence may not be of
any consequence. For example, the mortality rate for patients younger than 50 is very low (Figure 1), in which case the immortal time bias is likely negligible.

However, when the research question is concerned with bilateral TJA patients, caution is warranted. The main potential pitfalls lie in the time dependence of a patient's bilateral status. Extra care needs to be taken when the proportion of bilateral patients is high, and when unilateral TJA patients become staged bilateral relatively long after the first TJA surgery. If landmark analysis is not suitable for answering the research question, an appropriate method for modelling the path a patient may take from unilateral to bilateral TJA to or via revision and/or death, is a multistate model.\textsuperscript{17,18} Multistate models have been applied to the Australian National Joint Replacement Registry and are able to incorporate the complicated dependence structures which are often present in arthroplasty data.\textsuperscript{23,24}

Concluding remarks
Analyses of arthroplasty data in which staged bilateral TJA patients are excluded show overestimates of surgical risks due to a difference in these risks between unilateral and staged bilateral TJA patients.\textsuperscript{25} This paper shows that the immortal time bias can be added to the list of arguments against this practice. In any analysis of arthroplasty registry data, researchers should carefully consider the impact of the immortal bias on their results, and select the statistical method accordingly.
References

Figure legend

Figure 1. Cumulative incidence of revision and death for patients younger than 50 and older than 80, estimated by the Aalen-Johansen estimator, using data from the Dutch Arthroplasty Register (n = 6,361 patients under 50 years old, n = 21,891 patients over 80 years old). Each patient's first placed hip was used. The competing risk is negligible for the youngest patients, but very strong for the older patients.

Figure 2. Visualization of the artificial data example. Circles denote patients who will only have one implant, while squares indicate patients who will become bilateral at the 2-year mark. Green indicates event free patients, black patients who die before experiencing revision, and orange patients whose prosthesis has been revised. The subgroup analysis ignores the fact that some patients will have died before realizing their potential of becoming bilateral, and thus some potentially bilateral patients will be considered unilateral.

Figure 3. A landmark time is chosen, in this case after the patients become bilateral. All patients who died or were revised before the landmark time are excluded from the analysis.

Figure 4. Simulation results. Percentage of false positives for a difference in cumulative incidence of death or revision, at significance level 5%. Based on 2000 simulation replicates.

Figure 5. Representative results from one simulated data set, in the setting where 30% of patients has the potential to become bilateral. The plots show the cumulative incidence of revision (top panels) and death (bottom panels) for staged bilateral (black line) and unilateral (red line) patients. The subgroup analysis finds a significant difference between the groups of patients (revision: p = 0.0004, death: p < 0.0001) while the landmark analysis does not (revision: p = 0.84, death: p = 0.41).

Figure 6. Cumulative incidence of revision and death for staged bilateral patients. Data from Dutch Arthroplasty Registry, n = 14,067 patients included in the subgroup analysis, n = 8,898 in the landmark analysis. Cumulative incidence of revision of the first implant (top panels) and death (bottom panels) for patients who had bilateral TJA within 1.5 years (red line) and between 1.5 and 3 years (black line). The panels on the left and on the right show the results from a subgroup analysis and from landmark analysis with a 3 year landmark time respectively.
Figure 1
Figure 3

Figure 4
Figure 5

Figure 6
<table>
<thead>
<tr>
<th>Percentage potential bilateral (Bernoulli)</th>
<th>Average number of observed bilateral patients</th>
<th>Common simulation parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>2462 (25%)</td>
<td>n = 10,000 patients</td>
</tr>
<tr>
<td>20%</td>
<td>1641 (16%)</td>
<td>N = 2,000 datasets</td>
</tr>
<tr>
<td>10%</td>
<td>820 (8%)</td>
<td>Time to revision: Weibull, shape = 1.8, scale = 20 (average of 2704 revisions)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time to death: Weibull, shape = 1.2, scale = 20 (average of 3601 deaths)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time to bilateral: Exponential, rate = 1/3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time of follow up: Uniform, between 0 and 30</td>
</tr>
</tbody>
</table>

Table 1. Simulation parameters, and observed average number of bilateral patients in each setting.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Patients</th>
<th>Revisions</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup, 0-1.5 yr</td>
<td>10,393</td>
<td>169</td>
<td>355</td>
</tr>
<tr>
<td>1.5-3 yr</td>
<td>3,674</td>
<td>104</td>
<td>100</td>
</tr>
<tr>
<td>Landmark, 0-1.5 yr</td>
<td>5,990</td>
<td>46</td>
<td>171</td>
</tr>
<tr>
<td>1.5-3yr</td>
<td>2,908</td>
<td>23</td>
<td>75</td>
</tr>
</tbody>
</table>

Table 2. Numbers of patients and events included in the analysis of the Dutch Arthroplasty Registry data.

<table>
<thead>
<tr>
<th>Cumulative incidence of revision at 3 years.</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>24%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Estimated (unilateral)</td>
<td>20%</td>
<td>4.7%</td>
</tr>
<tr>
<td>Estimated (staged bilateral)</td>
<td>30%</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cumulative incidence of death at 3 years.</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>20%</td>
<td>5%</td>
</tr>
<tr>
<td>Estimated (unilateral)</td>
<td>33%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Estimated (staged bilateral)</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3. Results of the subgroup analysis of the artificial data example.

<table>
<thead>
<tr>
<th>Subgroup, 0-1.5 yr</th>
<th>Revision (95% CI)</th>
<th>Death (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5-3 yr</td>
<td>2.4% (2.0-2.9)</td>
<td>6.4% (5.6-7.2)</td>
</tr>
<tr>
<td>Landmark, 0-1.5 yr</td>
<td>3.4% (2.8-4.3)</td>
<td>4.2% (3.3-5.2)</td>
</tr>
<tr>
<td>1.5-3yr</td>
<td>1.2% (0.8-1.6)</td>
<td>4.3% (3.6-5.0)</td>
</tr>
</tbody>
</table>

Table 4. Estimated cumulative incidences of revision and death at 6 years of follow up, with 95% confidence intervals, based on the Dutch Arthroplasty Registry data.