Formal sensitivity analysis in observational studies

Dan Hedlin¹, Mahmood Ul Hassan¹ Hong Xu²,³, Sara Hägg³, Juraj Secnik²,⁴, and Dorota Religa²,⁵

¹Department of Statistics, Stockholm University, Stockholm, Sweden;
²Division of Clinical Geriatrics, Center for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Huddinge, Sweden;
³Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;
⁴University Hospital Motol, Prague, Czechia;
⁵Theme Aging, Karolinska University Hospital, Huddinge, Sweden

Corresponding author: Dan Hedlin, Department of Statistics, Stockholm University, email: dan.hedlin@stat.su.se.

Abstract

Objective: The positive effect of antidiabetic medication on cognitive decline has been given some support by, among others, Secnik et al (2021) and Secnik et al (2022). However, as they are observational studies, it is not clear whether the effect is causal.

Research design and methods: Using the Swedish Dementia Registry and supplementary Swedish registers/databases, we identified 1,873 patients (4,732 observations) with diagnosis of diabetes and Alzheimer’s disease or mixed-pathology dementia who were followed-up at least once after dementia diagnosis. The association of use of metformin with Mini-Mental State Examination scores in patients with diabetes and dementia was studied in two ways. 1) The difference between the last and the first score for each patient was compared with treatment (use of metformin) and subjected to a new sensitivity analysis. 2) The difference between scores for each patient at the points in time when there was a change in use of metformin (either start of use, or discontinuation of use) was studied.

Results: There is an association between cognitive decline and use of metformin. However, any conclusion of a causal relationship is tenuous.

Conclusion: The present study offers no basis for causal conclusions, but given the association, further examination of cognitive effects of metformin is warranted.

Keywords: sensitivity, observational study, dementia, metformin

Introduction

The real interest in many medical studies centres on cause and effect. Does a specific treatment cause the effect that we are interested in? Randomised studies, in which treatment is randomised, are widely considered gold-standard. Despite the debate on reproducibility randomised controlled studies are often thought of as being capable of providing evidence on cause and effect. This is debatable, as the power of a statistical test does not take features of data collection into account. To create a “clean” dataset, onto which a statistical test is employed, a large number of statistical and practical activities need to be performed. They
include finding and selecting subjects, devising measurement procedures, and taking those measurements. Also, the researcher often needs to impute missing values. After the test has been done there are further issues, such as choice of method to address misleading p-values if the study contains several tests. Whether a treatment actually causes the observed effect can rarely, if ever, be established in a single study.

The same is true for observational studies. Contrary to randomised studies, it widely thought that observational studies are not capable of saying anything on cause and effect. This is exaggerated, and we show why in this article.

In an observational study, individuals are matched on a set of variables. Denote these variables by $x_k$ for individual $k$. One individual, who has been given a specific treatment, is matched with another individual, who has not been given the treatment. A response variable is measured for all individuals at time $t$. Assume for the moment that the study variable is measured at the same time $t$ for all individuals and the matching is pairwise. The research hypothesis is that the treatment has an observable effect on the study variable.

Consider the motivating example. The research hypothesis is that metformin prescribed for diabetes may also have a beneficial effect on dementia. An observational study is conducted to examine the hypothesis. There may be bias ensuing from unknown factors.

The analysis in an observational study relies heavily on the assumptions that

1. the probability that an individual has been given the treatment does not depend on anything else than $x_k$ and that
2. the probabilities that $k$ and $l$ have been given the treatment are not associated (“no interference between units”).

These are strong assumptions that require support. Even if we have collected a rich dataset and managed to match people on more variables than the two mentioned in the example, there is always a risk that there is a further variable $u_k$ that the probability depends on ($u$ for unobserved or unmeasured).

Consider as an example an observational study that analyses a treatment’s effect on dementia. The following conceivable scenario challenges the assumptions. If the same doctor treats both $k$ and $l$, and this doctor tends to favour the treatment under study to other treatments, then the probability that $k$ and $l$ have been given the treatment may larger than the probability that individual $m$ has been given the treatment, if $m$ has been treated by another doctor. Unless doctor is included in $x_k$, there is a variable $u_k$ (doctor) that affects the probability. Moreover, there is a dependence introduced through the doctor.

It may not be possible to test the assumptions involving unobserved variables in an observational study. One thing you can do, is to explore the sensitivity of the conclusions to inevitable departures from the assumptions. Weak sensitivity is no proof of a causal effect but it does mitigate the concern of the lack of randomisation.

Motivated by the studies in Secnik et al (2021, 2022) we ask if these observational studies may allow for tentative causal conclusions. We approach the issue with Rosenbaum’s sensitivity test described below. Other approaches would be variants of the study of differential effects (Rosenbaum, 2006, and Rosenbaum, 2020, p. 146). For example, we could study effects of change in medication for each patient who has been subjected to such a
change. If there is an effect of metformin on cognitive decline (less sharp decline than without metformin), then we should see a difference when a prevalent user of metformin discontinues the use as opposed to the other way round, when a patient starts using metformin. Another approach would be to pair patients with similar values of all covariates but taking different diabetes medication (Rosenbaum, 2020, p. 147, 290).

**Rosenbaum’s sensitivity test**

First, as a preparation, how would you test a treatment effect if treatment was randomised within each pair? Let \( r_{i1} \) and \( r_{i2} \) be a continuous study variable for a matched pair \( i \) of two individuals 1 and 2. Having been matched they have the same value of covariates used for matching, that is, \( x_k = x_l \). One of them has been treated and the other one has not. In a random experiment the probability that \( k \) has been given the treatment equals the probability that \( l \) has been given the treatment.

Define a variable that captures the response from pair \( i \):

\[
y_i = (z_{i1} - z_{i2})(r_{i1} - r_{i2})
\]

where \( z_{i1} = 1 \) if individual 1 has been treated, \( z_{i1} = 0 \) if not, and similarly for individual 2. Note that \( z_{i1} - z_{i2} \) is either 1 or \(-1\). Equation (1) can be rewritten as \( y_i = \pm(r_{i1} - r_{i2}) \) with plus sign if individual 1 is treated, minus sign if 2 is treated.

If the number of pairs is large, the nonparametric Wilcoxon signed rank statistic is approximately normal. With this statistic the absolute value of \( y_i \) does not matter, what matters is only the rank of the absolute value of \( y_i \). If the null hypothesis of no treatment effect is true, there should be no association between the sign of \( y_i \) and the rank of \( |y_i| \). If the data indicates such an association, then that is taken as evidence against the null hypothesis.

Once you have computed the value of the test statistic, deduct the expected value of the statistic and divide by the standard deviation to obtain a standardised statistic (Rosenbaum 2020, Ch. 2), and compare with a standard normal distribution, just as the usual procedure goes.

Let’s say you have conducted an observational study with a continuous study variable and found that the treatment effect is statistically significant with a p-value of \(3.1 \cdot 10^{-7}\). But can we trust this, given the strong assumptions 1 and 2?

Denote the probability that individual \( k \) has been given the treatment by \( \pi_k \), the probability that individual \( l \) has been given the treatment by \( \pi_l \), and the odds ratio for individuals \( k \) and \( l \) in pair \( i \) by \( OR = \frac{\pi_k/(1-\pi_k)}{\pi_l/(1-\pi_l)} \). There will be an odds ratio of the same form for each pair.

The minimum value of an odds ratio is zero, and it can be arbitrary large. \( OR = 1 \) is attained for \( \pi_k = \pi_l = 1/2 \), which is what you would have if treatments were administered randomly within pairs. This is the assumption under which the p-value of \( 3.1 \cdot 10^{-7} \) was computed.

The distance of the odds ratio from 1 can thus be interpreted as a measure of departure from the most favourable scenario for an observational study. It is convenient to introduce a number \( \Gamma \) that can take any positive number greater than or equal to 1, and the model \( 1/\Gamma \leq OR \leq \Gamma \):
\[ \frac{1}{\Gamma} \leq \frac{\pi_k/(1 - \pi_k)}{\pi_l/(1 - \pi_l)} \leq \Gamma \]  

For example, \( \Gamma = 3 \) is a large number, because it means that the odds for individual \( k \) to be given the treatment may be a third or three times of the odds of individual \( l \). For example, if \( \pi_k = 1/2 \) and \( \pi_l = 1/4 \), then \( \Gamma = 3 \), which is a large departure from a random experiment. In many practical situations it should be possible to determine whether \( \Gamma \geq 3 \) for a large proportion of the pairs is unlikely or plausible.

For a given \( \Gamma \), for example \( \Gamma = 2 \), you can identify the \( \pi_k \) and \( \pi_l \) that turn the left and right hand sides of (2), respectively, to equalities. And from these \( \pi_k \) and \( \pi_l \) it is possible to compute the minimum and the maximum value of the Wilcoxon statistic that can occur within the limits of the chosen \( \Gamma \). From the minimum and the maximum value of the Wilcoxon statistic you can go on to identify the minimum and the maximum \( p \)-value.

Wilcoxon’s signed rank test is simple and robust, even too robust. There are (slightly) better tests, for example an \( m \)-test (Rosenbaum 2007). If you have matched each treated subject with more than one control you can use an \( m \)-test (Rosenbaum 2007).

The temporal dimension has not been considered so far. Let’s say we have formulated a model where a function of the response variable captures the response, \( y_k \), of the treatment under study. The explanatory variables in the model are \( x_k \) and \( t_k z_k \), with \( z_k = 1 \) if \( k \) has been treated, \( z_k = 0 \) otherwise, and \( t_k \) time after baseline. When the model is fitted to data, the coefficient \( \beta \) in front of \( t_k z_k \) is estimated and a \( p \)-value under the null hypothesis that \( \beta = 0 \) is computed. Now there at least two strategies. If we have a very large dataset, it may be feasible to select a subset of individuals with approximately equal time after baseline. We can make a Wilcoxon test (or some other appropriate test) on the study variable as above. Another strategy is to take the heterogeneity induced by \( t_k z_k \) into account. If \( \beta < 0 \), then \( \beta t_k z_k \) will give a linear decline in the study variable over time. There are methods for this scenario. One benefit of the first approach is that it also tests the model that postulates that the decline is linear.

**Matching**

We performed propensity score matching where scores were estimated with cognition (clock test), medication for vascular risk factors, consultation with a counsellor, gender, age, BMI and presence of home health care. Some other covariates were involved in the modelling (see Table 1) but they were less important. Home health care and BMI were imputed due to a large proportion of missing values. We did no adjustment for attrition. Figure 1 shows the covariate balance between treatment and non-treatment groups. The ‘unadjusted’ curve depicts the mean difference for each covariate between patients receiving metformin and those who did not use metformin. The ‘adjusted’ curve is the mean of pairwise differences after matching.
Table 1. Unadjusted curve: the mean difference for each covariate between patients receiving metformin and those who did not use metformin. Adjusted curve: the mean of pairwise differences after matching

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unadjusted mean difference</th>
<th>Adjusted mean difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coresident present</td>
<td>0.0376</td>
<td>0</td>
</tr>
<tr>
<td>Cognition</td>
<td>-0.0034</td>
<td>-0.004</td>
</tr>
<tr>
<td>Vascular</td>
<td>0.2416</td>
<td>0</td>
</tr>
<tr>
<td>Antidepressive</td>
<td>0.0061</td>
<td>0</td>
</tr>
<tr>
<td>Neuroleptica</td>
<td>-0.0001</td>
<td>0</td>
</tr>
<tr>
<td>Anxiety suppressor</td>
<td>-0.0064</td>
<td>0.0009</td>
</tr>
<tr>
<td>Barbituates</td>
<td>0.0001</td>
<td>0.0059</td>
</tr>
<tr>
<td>Counsellor</td>
<td>-0.003</td>
<td>0</td>
</tr>
<tr>
<td>Codependent support</td>
<td>-0.0001</td>
<td>0</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.0867</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.2128</td>
<td>-0.011</td>
</tr>
<tr>
<td>BMI</td>
<td>0.359</td>
<td>0.0085</td>
</tr>
<tr>
<td>Home health care</td>
<td>0.0129</td>
<td>0.0101</td>
</tr>
</tbody>
</table>

Results

Although the decline in Mini-Mental State Examination scores is significantly less for the treatment group (use of metformin), it can only be established for $\Gamma = 1.2$, that is if treatment/no treatment was administered randomly. For $\Gamma = 1.3$ the effect is no longer significant. Table 2 lists the maximum $p$-values that the Wilcoxon test statistic can result in for several values of $\Gamma$, which are measures of departure from the most favourable scenario for an observational study ($\Gamma = 1$ is no departure). The number $\Gamma = 1.3$ means that the odds for treatment for one patient may be 30% higher than for another patient with the same covariates. It is a small odds, which is entirely conceivable. That is, the study Secnik et al (2021) offers no basis for causal conclusions. We also used the m-test, as opposed to the Wilcoxon test, but as that test gave similar results we only report the results of the Wilcoxon test. However, there is an association, albeit not necessarily causal, between metformin and less sharp decline in cognitive ability, which seems to warrant further studies.
Table 2. The maximum p-value of that can occur within limits of the odds-ratio (\(\Gamma\))

<table>
<thead>
<tr>
<th>(\Gamma)</th>
<th>Maximum p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.1</td>
<td>0.0000</td>
</tr>
<tr>
<td>1.2</td>
<td>0.0003</td>
</tr>
<tr>
<td>1.3</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

References


Rosenbaum, P.R. (2020). Design of observational studies. 2nd ed. Cham: Springer.
