



INSTITUTE FOR EMPLOYMENT  
RESEARCH  
The Research Institute of the Federal Employment Agency

# IAB-DISCUSSION PAPER

Articles on labour market issues

---

## 29|2020 TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding

Anna-Carolina Haensch, Jörg Drechsler, Sarah Bernhard



# TippingSens: An R Shiny Application to Facilitate Sensitivity Analysis for Causal Inference Under Confounding

Anna-Carolina Haensch (GESIS Institute for the Social Sciences),  
Jörg Drechsler (Institute for Employment Research),  
Sarah Bernhard (Institute for Employment Research)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
<b>2</b>	<b>Rubin's Causal Model and the assumption of unconfoundedness</b>	<b>10</b>
2.1	The Rubin Causal Model	10
2.2	Quasi-experiments and the assumption of unconfoundedness	11
<b>3</b>	<b>Sensitivity analysis in the context of causal inference</b>	<b>12</b>
<b>4</b>	<b>The Rosenbaum-Rubin Sensitivity Analysis and the TippingSens App</b>	<b>15</b>
4.1	Technical Details	15
4.2	The TippingSens App as a Visualization Tool	16
<b>5</b>	<b>A Practical Example: Sensitivity Analysis for a Quasi-experimental Evaluation of a German Vocational Training Program for the Unemployed</b>	<b>18</b>
5.1	Details about the Study	18
5.2	Sensitivity Analysis with the TippingSens App	20
<b>6</b>	<b>Discussion</b>	<b>26</b>
	References	28
	<b>Appendix</b>	<b>33</b>

## List of Figures

Figure 1:	Sensitivity analysis for the evaluation of vocational training by Bernhard (2016)	22
Figure 2:	Sensitivity analysis for the evaluation of vocational training by Bernhard (2016)	23
Figure 3:	Sensitivity analysis of the evaluation of vocational training by Bernhard (2016)	24
Figure A.1:	Default appearance of the TippingSens app before interacting with any menu/slider	34
Figure A.2:	Example of input data for the app in csv format containing two columns labeled "Outcome" and "Treatment"	35

Figure A.3: Parameter settings used to generate output displayed in Figure A.4 .....36

Figure A.4: Output generated by the TippingSens app based on data from Rosenbaum/  
Rubin (1983) .....39

Figure A.5: Output from the TippingSens app with data taken from Rosenbaum/Rubin  
(1983) .....39

# List of Tables

Table A.1: Upper half of original table containing sensitivity analysis results from Rosen-  
baum/Rubin (1983), Table 2, page 216 .....37

## Abstract

Most strategies for causal inference based on quasi-experimental or observational data critically rely on the assumption of unconfoundedness. If this assumption is suspect, sensitivity analysis can be a viable tool to evaluate the impact of confounding on the analysis of interest. One of the earliest proposals for such a sensitivity analysis was suggested by Rosenbaum/Rubin (1983). However, while it is straightforward to obtain estimates for the causal effect under specific assumptions regarding an unobserved confounder, conducting a full sensitivity analysis based on a range of parameter settings is unwieldy based on the simple forking tables which Rosenbaum and Rubin used. To tackle the multiple parameter problem of the Rosenbaum-Rubin approach, we present an interactive R Shiny application called TippingSens, which visualizes the impact of various parameter settings on the estimated causal effect. Borrowing from the literature on tipping point analysis, the flexible app facilitates manipulating all parameters simultaneously. We demonstrate the usefulness of our app by conducting a sensitivity analysis for a quasi-experiment measuring the effect of vocational training programs on unemployed men. The online supplement accompanying this paper provides a step-by-step introduction to the app using the original illustrative example from Rosenbaum/Rubin (1983).

## Zusammenfassung

Kausalanalysen mit quasi-experimentellem Forschungsdesign bspw. Matching-Verfahren schätzen die Wirkung einer Intervention auf eine Ergebnisvariable. Unverzerrte Schätzungen ergeben sich dabei nur, wenn die Kausalanalyse alle Merkmale mit Einfluss auf die Ergebnisvariable und auf die Selektion in die Intervention berücksichtigt. Wenn nicht sicher ist, ob all diese Merkmale berücksichtigt wurden, helfen Sensitivitätsanalysen, den Einfluss eines nicht beobachteten Merkmals auf die geschätzte Wirkung der Intervention zu bestimmen. Einen der ersten Vorschläge für eine solche Sensitivitätsanalyse lieferten Rosenbaum und Rubin (1983). Danach legen vier Parameter fest, wie das unbeobachtete Merkmal mit der Interventionswahrscheinlichkeit und der Ergebnisvariable (in Interventions- und Kontrollgruppe) zusammenhängt und wie es in der Grundgesamtheit verteilt ist. Einfache tabellarische Darstellungen für den Einfluss des nicht beobachteten Merkmals auf die geschätzte Wirkung für verschiedene Sets dieser vier Parameter stoßen schnell an die Grenze zur Unübersichtlichkeit. Dieses Problem lösen wir mit der interaktiven Web-Anwendung TippingSens. Sie basiert auf Shiny R Studio. TippingSens visualisiert gleichzeitig alle vier Parameter eines unbeobachteten Merkmals und zeigt für unterschiedliche Kombinationen und Ausprägungen dieser Parameter in welchem Ausmaß die Ergebnisse der Kausalanalyse durch das nicht beobachtete Merkmal verändert würden. Wir demonstrieren die Anwendung von TippingSens mit den

Ergebnissen aus einer Wirkungsanalyse beruflicher Weiterbildung von Arbeitslosen. Der Web-Anhang zu diesem Artikel enthält eine Anleitung für TippingSens anhand des ursprünglichen Beispiels von Rosenbaum und Rubin

JEL

C31, C87

Keywords

Endogeneity, Program Evaluation, Unconfoundedness, Vocational Training

# 1 Introduction

Questions of interest in the applied sciences are often questions of causality rather than description or association. The gold standard for estimating causal effects are randomized experiments. With experiments, researchers have full control over the treatment assignment process, substantially facilitating the estimation of causal effects. However, while in the medical sciences, experimental designs are widely used; they are often not feasible or unethical in epidemiological, sociological, or economic research. Instead, empirical research is usually based on pre-existing, observational data. When attempting to make causal claims from observational data, researchers have to make strong assumptions. Arguably the most controversial of these assumptions is the requirement that the assignment to the treatment is unconfounded, i.e., given all the information available in the collected data, the probability of receiving the treatment does not depend on the (potential) outcomes of interest (selection on observables). The unconfoundedness assumption implies that adjusting for differences in observed variables removes biases in causal estimates based on a direct comparison between treated and control groups (Imbens, 2003: 126).

The unconfoundedness assumption is not always realistic, and it can never be tested based on the observed data. A possible way to proceed if the unconfoundedness assumption is suspect is to conduct sensitivity analyses under various assumptions regarding the confounding mechanism.

While intuitive in theory, conducting sensitivity analyses in practice is often difficult. Flexible approaches based on the idea of partial information (Manski, 1990), which try to incorporate the uncertainty about the assignment mechanism directly are usually not helpful in practice since the level of uncertainty about the actual causal effect will often become so large that no meaningful conclusions can be drawn based on the collected data. Thus, specific assumptions need to be postulated about the properties of the confounding variable or the assignment process. Turning these assumptions into parameters can be challenging in practice. In this article, we will focus on a sensitivity analysis approach proposed by Rosenbaum/Rubin (1983), which achieves this difficult task for the particular case of a binary confounding variable.

Despite limiting the range of scenarios in which the sensitivity analysis can be used, binary confounding variables seem plausible in many applications. For example, when analyzing health care records, it might be realistic to assume that the treatment assignment depends on specific health conditions such as obesity or certain blood test results, which are not available in the database for confidentiality reasons. Since treatment guidelines are often based on threshold rules (if the blood pressure is lower than  $x$ , prescribe dose  $y$ , else prescribe dose

$z$ ), the confounding variable can be considered binary. Another example, which we use to illustrate the functionality of the app in Section 5, are confounding variables in the context of assignment to labor market programs. Assignment decisions will be affected by factors such as alcohol abuse or poor personal hygiene, which are observable for the case manager deciding about the treatment assignment but are not recorded in the database used when evaluating the effectiveness of various labor market programs.

Even in the simple case of a single binary confounding variable, Rosenbaum-Rubin sensitivity analyses can be cumbersome since four different parameters need to be specified (see Section 4.1 for details) and the impact on the analysis of interest needs to be evaluated for varying parameter combinations. Rosenbaum/Rubin (1983) work with forking tables to present the results of their sensitivity analysis. However, tables are not suited to explore a broader range of parameter sets since the resulting tables will quickly become very large and complicated to read. For example, evaluating five different values for each of the four parameters would already result in a table with  $5^4 = 625$  table cells.

We postulate that this is one of the main reasons why other strategies, such as the sensitivity analysis proposed by Rosenbaum (1995) or Manski's partial information approach, are often preferred in practice. However, their main advantage – less assumptions about the confounding variables – is also their main problem: without additional assumptions, the plausible range for the causal effect often becomes so wide that no practical recommendations can be given based on the findings. If (partial) information about the confounders is available, the Rosenbaum-Rubin approach offers a flexible tool to integrate this information, helping to narrow the range of plausible values for the causal effect.

Strategies to address the high dimensional table problem of the Rosenbaum-Rubin approach have been suggested in the literature. For example, Imbens (2003) suggests looking at the relationships in the observed data and picking the most extreme values as informed guesses for the unobserved parameters. The underlying assumption is that if the study was carefully designed, it seems unlikely that confounders exist that show higher correlations with the outcome and/or the treatment indicator than any of the observed correlations. Still, researchers are usually not interested in only one set of (extreme) assumptions but try to evaluate how robust their findings are under a whole range of plausible assumptions (if the correct parameters were known, no sensitivity analysis would be required).

To address this problem, we developed an interactive R Shiny app called *TippingSens* that simplifies conducting Rosenbaum-Rubin sensitivity analyses for a large number of parameter sets. Our app enables researchers to gain quick insights regarding the robustness of their findings and to publish comprehensible visualizations of their sensitivity analysis. We tackle the four-dimensional parameter space through the interactive component of our app. The *TippingSens* app is freely available on the R Shiny server



<https://tippingsens.shinyapps.io/TSApp/>.

The remainder of the paper is organized as follows: In Section 2, we give a brief introduction to causal inference with observational data and the problems resulting from self-selection into treatment. Different approaches to sensitivity analysis which are used to examine the robustness of the unconfoundedness assumption are presented in Section 3. Section 4 focuses on the technical details of the sensitivity approach developed by Rosenbaum/Rubin (1983) and introduces our new visualization tool TippingSens. Section 5 offers a practical illustration of the TippingSens app examining the robustness of an analysis by Bernhard (2016), which estimates treatment effects of vocational training programs on unemployed men in Germany. The paper concludes with a discussion of the limitations of the approach and some suggestions for future research. A step-by-step guide how to invoke the TippingSens app with data from Rosenbaum/Rubin (1983) is provided in the online supplementary material accompanying this paper.

## 2 Rubin's Causal Model and the assumption of unconfoundedness

### 2.1 The Rubin Causal Model

Throughout this article, we will discuss causal inference based on the potential outcomes framework developed by Rubin (1974) (see, for example, Pearl/Glymour/Jewell (2016) for a different perspective on this topic). Let  $Y_i$  be the outcome of interest for unit  $i$ ,  $i = 1, \dots, N$ . Let  $W_i$  be an indicator which treatment unit  $i$  received. In many applications  $W_i$  is binary, i.e.,  $W_i = 1$  if unit  $i$  receives the treatment and  $W_i = 0$  otherwise. Under this assumption, the potential outcomes framework defines  $Y_i(1)$  as the potential outcome for unit  $i$ , if unit  $i$  received the treatment, and  $Y_i(0)$  defines the potential outcome if the unit did not receive the treatment. The individual (additive) treatment effect  $\tau_i$  can then be computed as  $\tau_i = Y_i(1) - Y_i(0)$ . However, only one of the two potential outcomes will be observed. The other one will be the ex-post counterfactual, which can never be observed. This is the well-known *fundamental problem of causal inference* (Holland, 1986). Thus, the individual treatment effect can never be observed directly. To be able to draw causal conclusions at least at an aggregate level, further assumptions are required. Besides the stable unit value treatment assumption (SUTVA), which requires that the outcome of unit  $i$  does not depend on whether unit  $j$  receives the treatment and that there is no unobserved variability in the treatment, a common assumption in the context of causal inference is the assumption of a regular assignment mechanism. An assignment mechanism is called regular if the assignment is individualistic, probabilistic, and unconfounded. The first two components require that the probability of assignment to the different treatments for unit  $i$  only depends on its characteristics and not on the characteristics of the other units. Furthermore, the probability of receiving any of the treatments must be strictly positive for all units. The unconfoundedness assumption states that the probability of assignment does not depend on the outcomes of interest given the observed characteristics available in the data. For example, in the case of a quasi-experimental medical drug test, the treatment assignment must not depend on the (expected) survival rates after conditioning on the observed covariates.

Since for randomized experiments, the assignment mechanism is under the control of the researcher, the assumption of a regular assignment mechanism is typically valid for carefully designed experiments. Furthermore, since units generally are randomly assigned to the different treatment groups independent of their characteristics, differences before the treatment are only by chance, and the treatment is called *exogenous*. This further simplifies the analysis, since average treatment effects can be estimated by directly comparing the average outcomes in the different treatment groups.

## 2.2 Quasi-experiments and the assumption of unconfoundedness

In quasi-experiments, assignment to the treatment is no longer under the control of the researcher. Without random assignment, subjects in different groups might not be comparable at the baseline, that is, before treatment. This is because the composition of different treatment groups could result from a selective process. Quasi-experimental evaluations typically impose comparability at the baseline by homogenizing treatment groups on observed characteristics, for example, through propensity score matching.

However, even though various options exist to adjust for differences at the baseline, most adjustment methods assume that the assignment mechanism is regular. While the assignment will often be individualistic and probabilistic, the unconfoundedness assumption is more controversial. This assumption is not verifiable based on the observed data, and differences in unobserved covariates between treated and controls may remain. Bias may arise if the outcome and the unobserved characteristics are correlated.

To address this problem, researchers can evaluate the sensitivity of the results regarding the unconfoundedness assumption. Through simulations, it is possible to explore which properties the unmeasured covariate(s) need to have to substantially change the results and conclusions of the study.

Sensitivity analyses are strongly related to the study of treatment effect robustness after dropping one or more of the observed covariates (Heckman, 1989; Smith/Todd, 2001; Lechner/Wunsch, 2013). Nevertheless, Imbens (2003: 126) stresses one of the main differences between sensitivity analysis and the study of treatment effect robustness:

The attraction of the sensitivity analysis is that it is more directly relevant: one is not interested in what would have happened in the absence of covariates observed, but in biases that are the result of not observing all relevant covariates.

### 3 Sensitivity analysis in the context of causal inference

The general idea of a sensitivity analysis was first proposed by Cornfield et al. (1959) to defend the plausibility of a causal effect of cigarette smoking on lung cancer. The authors demonstrated that the lack of such a relationship was only possible through the existence of an unmeasured confounder with an unrealistically high association with lung cancer and smoking habits. Building on these ideas, several strategies for evaluating the sensitivity of the confoundedness assumption have been proposed in the literature. These strategies can be loosely grouped into three categories based on the assumptions they require.

Perhaps the most radical approach is to drop the assumption of exogeneity/unconfoundedness completely, specifying a range of plausible values for the estimated treatment effect, which accounts for the additional uncertainty regarding the unknown assignment mechanism (Manski, 1990). While this approach is attractive from a theoretical perspective as it requires no untestable assumptions, it also strictly limits the information that can be obtained from observational data. In practice, the uncertainty in the estimated causal effect is typically large, i.e., the uncertainty bounds that define the interval in which the true causal effect might fall are very wide. Thus, the findings often cannot provide useful information, for example, to guide decisions in an evidence-based policy setting.

Since the so-called Manski bounds are often not very informative, the other two general approaches specify limited departures from the unconfoundedness assumption instead of dropping the assumption of exogeneity/unconfoundedness completely. The approaches differ regarding the parameters that need to be specified by the user. The first approach only requires specifying the association between the unobserved confounder(s) and the treatment assignment. The second approach is limited to one confounder and additionally specifies the association between this confounder and the outcome.

Two early proponents of these two approaches are Rosenbaum (1995) (R95) and Rosenbaum/Rubin (1983) (RR83). R95 defines the association between the unobserved confounders and the treatment assignment indirectly by setting a threshold parameter, which specifies the maximum difference between the estimated treatment propensity based on the observable data and the true propensity. The researcher still takes a Manski-style approach regarding the associations between the hidden confounder and the potential outcomes. The RR83 approach additionally requires to explicitly specify the relationship between the confounder and the potential outcomes and assumes that confounding is limited to a single binary variable. Furthermore, both approaches implicitly assume that the relationships between the unobserved confounder(s) and the treatment does not vary as a function of the observed

data (RR83 requires a similar assumption regarding the confounder and the outcome in the two treatment groups).

Other approaches for sensitivity analysis proposed in the literature can mostly be classified as belonging to one of these three types. For example, Manski/Pepper (2000, 2009) extend the approach of Manski (1990) and applications are discussed in Blundell et al. (2007); Kang (2011), and Hof (2014) among others. Gastwirth/Krieger/Rosenbaum (1998); Rosenbaum (2010), and Rosenbaum (2018) discuss several extensions of R95 for specific settings. Practical applications of the strategy are discussed in Rosenbaum (1999, 2003, 2007, 2010); Kitahata et al. (2009); Zubizarreta/Cerdá/Rosenbaum (2013); Rosenbaum (2018) and software implementations are available in Stata Gangl (2004); Becker/Caliendo (2007) and R Keele (2010); Rosenbaum (2014). Extensions of the ideas of RR83 are presented for example in Harding (2003); Greenland (1996), or VanderWeele/Arah (2011) and applications are discussed in Imbens (2003) and Ichino/Mealli/Nannicini (2008).

Comparing the RR83 and R95 approaches, a major advantage of R95 is that less assumptions regarding the confounding variable are required. However, not surprisingly, the uncertainty bounds for the treatment effect will be wider, limiting the conclusions that can be drawn. If additional information is available, which allows limiting the range of plausible values for the correlation between the outcome and the potential confounder, RR83 allows tightening the uncertainty bounds. Since extreme cases, such as perfect positive or negative correlation, can typically be ruled out, RR83 can often be helpful in practical settings.

A downside of RR83 is the required assumption that the confounder is univariate and binary. Still, as discussed in the introduction and as illustrated in the application in Section 5, this assumption can be plausible in many circumstances (see also Liu/Kuramoto/Stuart (2013), who point out that the binary confounder can be seen as a combination of unobserved confounders). As Imbens/Rubin (2015) illustrate, under this assumption, the approaches of Manski (1990) and R95 can be seen as special cases of RR83, fixing some of the parameters at extreme values. Thus, under the assumption of a binary confounder, RR83 offers more flexibility to evaluate the impacts on the estimated causal effect under various assumptions regarding the relationship between the confounder and the outcome.

As pointed out above, we believe that the main reason for the popularity of the approaches akin to R95 is their simplicity. Since only one parameter needs to be specified, visualizing the impact of various assumptions about this parameter is straightforward. RR83-type approaches require monitoring several parameters simultaneously. While calculating the uncertainty bounds is still straightforward when fixing the parameters at specific (extreme) values as done in Imbens (2003) and Liu/Kuramoto/Stuart (2013), evaluating the impacts under various assumptions can be cumbersome. However, this will be a common scenario in practice. For example, in the labor market context it seems prudent to assume that alcohol abuse

has a negative impact on both, the probability of receiving the treatment, i.e., the probability of being assigned to a labor market program, and on the outcome, i.e., on the probability of finding a job. It is exactly in such a situation, where RR83 offers an advantage over R95. Even if the exact relationship between alcohol abuse and the outcome might be unknown, the knowledge that the correlation will most likely be negative can be used in the RR83 approach to narrow the uncertainty regarding the causal effect of the labor market program compared to R95.

However, monitoring the impact over a whole range of parameter settings can be difficult with RR83. Simplifying the sensitivity analysis for such a scenario through useful visualization tools was the main motivation for developing the TippingSens app. Before we describe the app in more detail and illustrate its features in an application, we briefly review the details of the RR83 approach in the next section.

## 4 The Rosenbaum-Rubin Sensitivity Analysis and the TippingSens App

### 4.1 Technical Details

The Rosenbaum-Rubin sensitivity analysis assumes that the unconfoundedness assumption holds given an additional unobserved binary covariate  $U_i$ :

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i, U_i,$$

where  $X_i$  denotes the vector of observed covariates. To evaluate the impact of this unobserved covariate on the analysis of interest, we can postulate parametric models for the marginal distribution of the unobserved covariate  $U$ , the conditional distribution of the treatment  $W$  given  $U$  and  $X$ , and the conditional distributions of the two potential outcomes  $Y(w)$ ,  $w \in \{0, 1\}$ , given  $U$  and  $X$ . In the following, we drop the observed variables  $X$  to avoid clutter. The formulae provided below can easily be extended to include observed variables (see, e.g., Imbens (2003: 127)) or more complex models.

Given that  $U_i$  is binary, we specify

$$q = Pr(U_i = 1) = 1 - Pr(U_i = 0).$$

The probability of receiving the treatment given  $U_i$  is modeled using a logistic regression

$$Pr(W_i = 1 | U_i = u) = \frac{\exp(\gamma_0 + \gamma_1 \cdot u)}{1 + \exp(\gamma_0 + \gamma_1 \cdot u)}.$$

Similarly, a logistic relationship is assumed between the binary outcome and  $U_i$  in both treatment groups:

$$Pr(Y_i(1) = 1 | U_i = u) = \frac{\exp(\alpha_0 + \alpha_1 \cdot u)}{1 + \exp(\alpha_0 + \alpha_1 \cdot u)}$$

and

$$Pr(Y_i(0) = 1 | U_i = u) = \frac{\exp(\beta_0 + \beta_1 \cdot u)}{1 + \exp(\beta_0 + \beta_1 \cdot u)}. \quad (4.1)$$

The parameter  $q$  as well as the parameters  $\gamma_1$ ,  $\beta_1$ , and  $\alpha_1$  are sensitivity parameters. It is not the goal to estimate these parameters from the data set. Instead the researcher can postulate

plausible ranges for them, gained from the literature or previous analyses. Conditional on the specified values for  $(q, \gamma_1, \alpha_1, \beta_1)$ , the remaining parameters  $(\gamma_0, \alpha_0, \beta_0)$  can be estimated through maximum likelihood. Once all parameters are estimated, standard statistics such as the average treatment effect  $\tau$  can be calculated accounting for the unobserved confounder  $U$  (see for example Imbens/Rubin (2015: 502) for details). Imbens (2003: 128) notes that the sensitivity parameters  $\alpha_1, \beta_1$ , and  $\gamma_1$  are difficult to interpret directly as they refer to log odds ratios. Exponentiating simplifies the interpretation, e.g., when  $e^{\alpha_1} = 2$  the odds for  $Y = 1$  in the treatment group double under  $u = 1$  compared to  $u = 0$ .

## 4.2 The TippingSens App as a Visualization Tool

The sensitivity analysis approach developed by Rosenbaum/Rubin (1983) is an excellent way to examine specific violations of the unconfoundedness assumption. However, in most practical settings, the true values for the different parameters are unknown, and researchers are interested in evaluating the robustness of their findings under a whole range of plausible assumptions regarding the associations between the unobserved variable and the outcome or the treatment assignment. The number of simulated possible treatment effects rises quickly with the number of chosen values for the four different parameters and examination of results through tables becomes very cumbersome.

The interactive R shiny app called TippingSens, which we introduce in this paper greatly simplifies conducting sensitivity analysis in this situation. It visualizes the impacts of various assumptions regarding the unknown parameters on the estimated average treatment effect. Thus, the app can be a handy tool to evaluate under which conditions the analytic conclusions would change. Snapshots of the visualization can be downloaded and integrated into any research output to offer easily comprehensible robustness checks of the research findings.

With the TippingSens app, it is possible to specify ranges for all parameters. The app requires a two-column matrix containing the binary treatment indicator value and the binary outcome of interest value for all units. Note that the app assumes that the data are already balanced regarding the observed covariates, that is, any steps for achieving balance, such as matching or trimming, need to be conducted before using the app. Once the data have been loaded, one can freely choose which two of the four parameters should be treated as fixed and specify the values for these parameters. Drop-down menus and sliders allow changing the settings interactively. The drop-down menus specify which parameters should be treated as fixed and which parameters should be displayed on which axis. The sliders are a convenient tool for adjusting the values for the fixed parameters as well as for specifying the range of values considered for the free parameters. We acknowledge that it would also be possible to fix only



one parameter and present the results in a three-dimensional plot. However, we feel that this would sacrifice the intuitiveness of the visualization. The design of the output grid was inspired by the tipping point analysis of Liublinska/Rubin (2014) for missing data sensitivity.

We will demonstrate the use of the TippingSens App in the next section with an example. The app is available at <https://carohaensch.shinyapps.io/TippingSens/>, and the code behind the app can be accessed at <https://github.com/CaroHaensch/TippingSensApp>.

# 5 A Practical Example: Sensitivity Analysis for a Quasi-experimental Evaluation of a German Vocational Training Program for the Unemployed

To illustrate the usefulness of the TippingSens app, we use a quasi-experimental evaluation of vocational training for unemployed in Germany conducted by Bernhard (2016). Another example based on the data used in Rosenbaum/Rubin (1983) is given in the online supplemental materials accompanying this paper. The online supplement also contains a step-by-step illustration how to invoke the app.

Quasi-experimental labor market program evaluations like the one in Bernhard (2016) typically include socio-demographics, information on labor market histories, and regional characteristics to control for the selectivity regarding the assignment to the treatment (see also Lechner (1999) or Heckman et al. (1998)). But they typically cannot incorporate personality traits, skills, preferences, attitudes, or social networks since this information is not available. However, these variables can still influence the job prospects (Bayer/Ross/Topa, 2008; Heckman/Stixrud/Urzua, 2006; Mueller/Plug, 2006; Pannenberg, 2010). They may also be a key driver of selection into training (Heckman/Ichimura/Todd, 1997). Estimates of causal effects using only the observed data can be biased in this situation since the unobserved variables are correlated with both, the outcome of interest as well as the treatment assignment. Thus, sensitivity analyses should be conducted to evaluate how strong these correlations need to be to change the research findings.

## 5.1 Details about the Study

The study of Bernhard (2016) uses the Integrated Employment Biographies (IEB, Dorner et al. (2010)) of the Institute for Employment Research in Germany.

The IEB is a large administrative database integrating five different sources of information collected by the Federal Employment Agency in Germany through different administrative procedures: the Employment History, the Benefit Recipient History, the Participants-in-Measures History, the Unemployment Benefit II Recipient History, and the Jobseeker History. It contains socio-demographic characteristics and individual daily information on employment, unemployment, benefit receipt, and participation in programs of active labor market policy

for the universe of German employees and unemployed.

Based on these data, the treatment group was defined as the total inflow of unemployed welfare recipients into vocational training within a three-months-period in 2005. The control group consisted of a 20 percent random sample of unemployed welfare recipients a day before this three-months-period. The controls did not enter vocational training within this three-months-period, but they could start vocational training afterward to avoid conditioning on future events (Fredriksson/Johansson, 2008).

The following observable information was used to model the assignment into treatment: individual socio-demographic characteristics (age; migration background; disability; qualification), characteristics of the household (single/partner; children), individual labor market history over the last five years (e.g., duration of employment, characteristics of the previous job such as wage, full-/part-time position, time in unemployment since last employment), labor market history of the partner and local labor market characteristics (Rüb/Werner, 2007). The data were stratified by several socio-demographic characteristics such as gender, age group, migration background, etc. and separate propensity score models for the treatment assignment were estimated for each stratum. Within each stratum, the final model was obtained using a stepwise selection procedure. Caliper matching based on the estimated propensity scores was used to get the final data set (see Bernhard (2016) for further details on the matching procedure).

For more than eight years after the (hypothetic) start of the training, labor market outcomes of participants and the matched control group were compared on three dimensions: share of welfare recipients, share of employees, and average monthly real wage. As common with matching approaches for causal inference, the study estimated the average treatment effect on the treated (ATET) and not the average treatment effect (ATE). Focusing on the ATET was useful in this context since it allowed to evaluate the effects on those for whom the program was intended (Heckman/Ichimura/Todd, 1997). However, the switch from ATE to ATET affects the interpretation of the sensitivity parameters as we will discuss below.

The overall results of Bernhard (2016) closely resemble previous findings not only for German data but also in an international context: The beginning of vocational training is an investment phase. During training, the search intensity for new jobs decreases and employment prospects and wages of training participants are lower in comparison to non-participants. This fact is known as Ashenfelter's dip (Ashenfelter, 1978). However, after a few months, positive effects of training on employment prospects, wages, and further welfare receipts can be observed, and these effects persist for up to eight and a half years after the training started. These results are in line with other quasi-experimental evaluations that find positive impacts of training on employment outcomes (Card/Kluve/Weber, 2010).

In our illustrative application of the TippingSens app, we will focus on men in West Germany. Our outcome of interest will be the employment status (employed vs. unemployed) over time. For the subgroup of men in West Germany, this outcome follows the overall trend described above. Assuming no confounding, unemployed West German men have a nine percentage points higher chance to be employed two years after starting a vocational training compared to West German men that did not participate in the training program.

## 5.2 Sensitivity Analysis with the TippingSens App

Specifying the required parameters and information for the TippingSens app is straightforward. Assuming the matching procedures described above resulted in a well-balanced data set regarding the observed characteristics, we only need two vectors from the matched data: the binary treatment indicator and the binary outcome indicator (see the online supplement for details regarding how to set up the app). In our case, the percentage of men in Western Germany that have found a job two years after starting vocational training is 0.46 for the treatment group (those who participated in vocational training) and 0.37 for the control group (those who did not participate in vocational training).

We conduct a Rosenbaum-Rubin sensitivity analysis because we are concerned about bias from unobservables. When evaluating vocational programs, such unobserved variables could be health conditions like alcoholism, obesity, or other factors. These variables are not recorded in the administrative data, but since the placement officer at the labor market agency might observe some of them, they might not only correlate with future job perspectives but might also affect the probability of attending vocational training. We will concentrate on alcoholism as an example here.

We need to think about four sensitivity parameters, i.e., we have to think about the association between alcoholism and participation in the training program ( $\gamma$ ), between alcoholism and employment after two years in the treatment group ( $\alpha$ ) and in the control ( $\beta$ ) group, and we need an estimate for the prevalence of heavy alcohol drinking ( $q$ ).

As mentioned previously, we do not estimate or extract the sensitivity parameters from the data. Instead, we use other sources of information to narrow down the range of plausible values for the parameters. A literature review reveals that a majority of studies suggest that heavy alcohol consumption has negative effects on employment probabilities in Western industrialized states (Popovici/French, 2016). Mullahy/Sindelar (1996) and a reanalysis by Terza (2002) controlling for endogeneity found a strong negative effect on the probability of being employed in the US. MacDonald/Shields (2004) used data from the Health Survey of England and found heavy alcohol drinking negatively associated with the likelihood of employment,

Johansson et al. (2007) obtained similar results for Finland. Devaux/Sassi (2015) estimate that heavy drinking has a strong negative effect on employment in men (white-collar men, OR: 0.54 [0.29; 1.00]). We will take these findings as a reference and use it for the sensitivity parameters  $\alpha_1$  and  $\beta_1$  and we assume that both parameters have an upper bound of zero since based on the literature it seems implausible that alcoholism will have a positive effect on job perspectives irrespective of whether a unit belongs to the treatment or the control group.

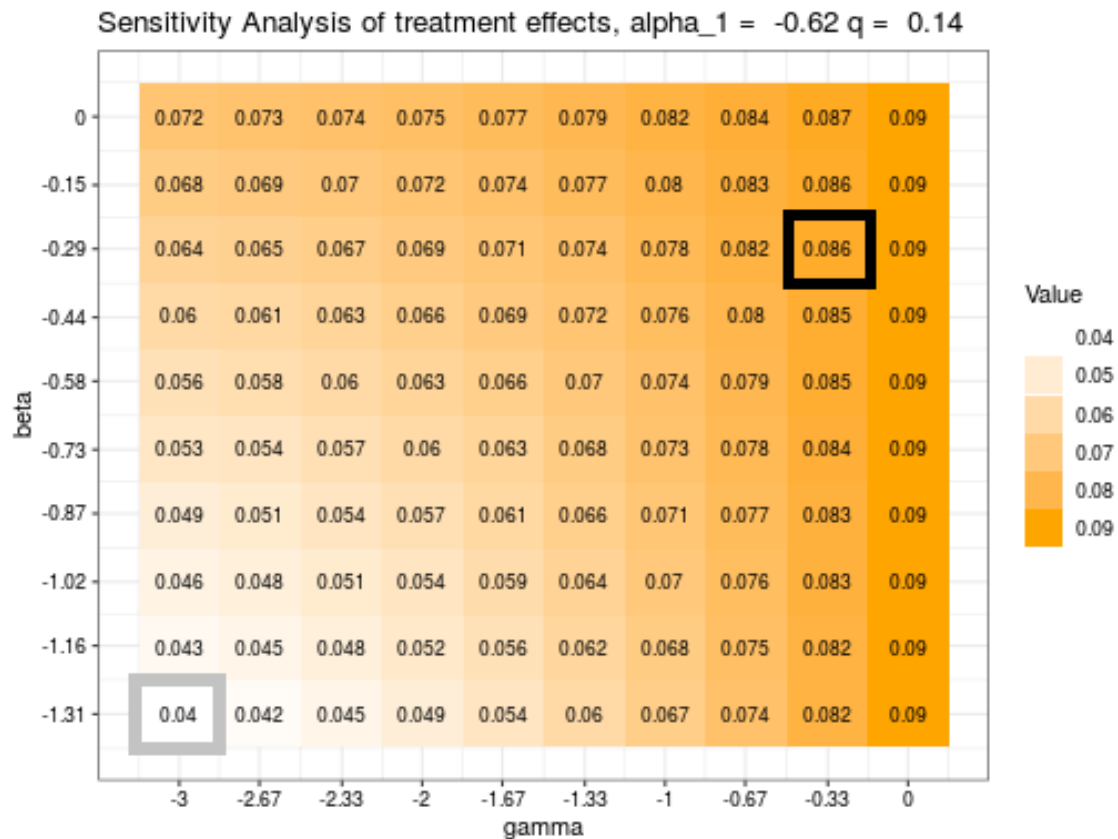
Next, we need to define a plausible range for the parameter  $q$ . If we were interested in the ATE, the parameter would represent the prevalence of alcoholism among men in West Germany. However, given that the study focuses on the ATET, the interpretation of  $q$  changes. It is now the prevalence of alcoholism among individuals with observable characteristics similar to those of the treatment group. Given that the treatment group might be a selective subset of the population (for example, all of them are unemployed at the start of the training program), it is more difficult to narrow down the range of plausible values for  $q$ . To specify an appropriate range for  $q$ , we use a literature review of unemployment and substance use by Henkel (2011). He puts the percentage at 14 percent for unemployed German men, but estimates vary across studies and measurements. We will examine a range of possible values from 2.0 percent to 18 percent.

Defining a suitable range for  $\gamma_1$  is most difficult since there are no studies – at least to our knowledge – that investigate how alcohol abuse influences the probability of being assigned to a labor market program. We will thus evaluate a wide range of values for  $\gamma_1$ .

To visualize the results from the Rosenbaum-Rubin sensitivity analysis, the app requires to fix two of the parameters at specific values, while plausible ranges can be specified for the remaining parameters. It is up to the user to decide which of the four parameters to fix.

We start our evaluation with the setting that appears to be most plausible concerning the literature. We set  $q = 0.14$  and  $\alpha_1 = \log(0.54) = -0.62$ . We note that by fixing  $\alpha_1$  at  $-0.62$  we implicitly assume that the odds ratio of getting a job for heavy drinkers for the treated is the same as for the general population of unemployed German men. We will evaluate the impacts of loosening this assumption below. Given that we expect negative effects of alcoholism on employment perspectives and it seems prudent to assume that alcoholism also has negative effects on the probability of receiving the treatment, we can fix the upper bound for the range of plausible values for the other parameters at zero. Defining meaningful lower bounds is more difficult. In our illustrative application, we set the lower bound for  $\gamma_1$  to  $-3$ , implying that we believe that the odds ratio of being assigned to the training program will not be less than 0.05 for individuals with a drinking problem relative to individuals without drinking problems.

Figure 1: Sensitivity analysis for the evaluation of vocational training by Bernhard (2016)



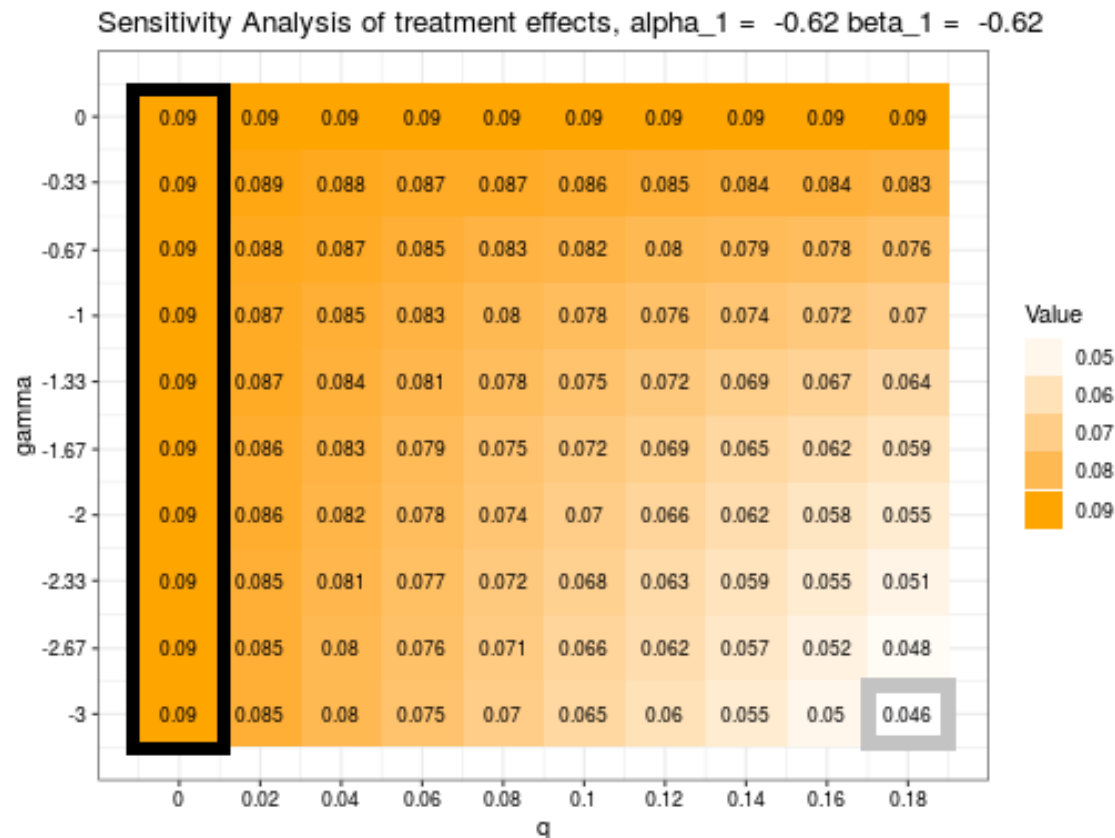
Notes: The gray box contains the treatment effect assuming the following values for the sensitivity parameters:  $\alpha_1 = -0.62$ ,  $\beta_1 = -1.31$ ,  $\gamma_1 = -3$ ,  $q = 0.14$ . The black box contains the treatment effect assuming the following values for the sensitivity parameters:  $\alpha_1 = -0.62$ ,  $\beta_1 = -0.29$ ,  $\gamma_1 = -0.33$ ,  $q = 0.14$ .

Source: Data from Bernhard (2016), Design from TippingSens App.

For the parameter  $\beta_1$ , we set the lower bound to  $-1.31$ . We choose this value, as we assume that the odds ratio of getting a job for individuals with a drinking problem in the control group will never be less than half the odds ratio for those participating in the training program (solving  $\exp^{\beta_1} / \exp^{\alpha_1} \geq 0.5$  for  $\beta_1$  gives  $\beta_1 \geq -1.31$ ). The output of the app based on our data, and these parameter settings are depicted in Figure 1. Lighter colored tiles imply smaller treatment effects. When assuming that the odds ratio of employment for heavy drinkers is about 0.75 ( $\beta_1 = -0.29$ ) in the control group and that the odds ratio for treatment assignment given alcoholism is about 0.72 ( $\gamma_1 = -0.33$ ), while the prevalence rate of alcoholism among individuals with observable characteristics similar to those in the treatment group is 0.14, the treatment effect of vocational training drops from 0.09 in the analysis based on the unconfoundedness assumption to 0.086 (black box in Figure 1). The stronger the negative effect of alcoholism on the outcome for the matched controls and on the treatment assignment, the more important becomes the unobserved covariate in explaining differences be-

tween treatment and control group leading to smaller estimated treatment effects. We note that even under the most extreme setting considered in Figure 1 ( $\gamma = -3, \beta = -1.31$ ) the estimated treatment effect remains positive (gray box in Figure 1). Thus, the results of the analysis based on the unconfoundedness assumption seems to be quite stable.

Figure 2: Sensitivity analysis for the evaluation of vocational training by Bernhard (2016)



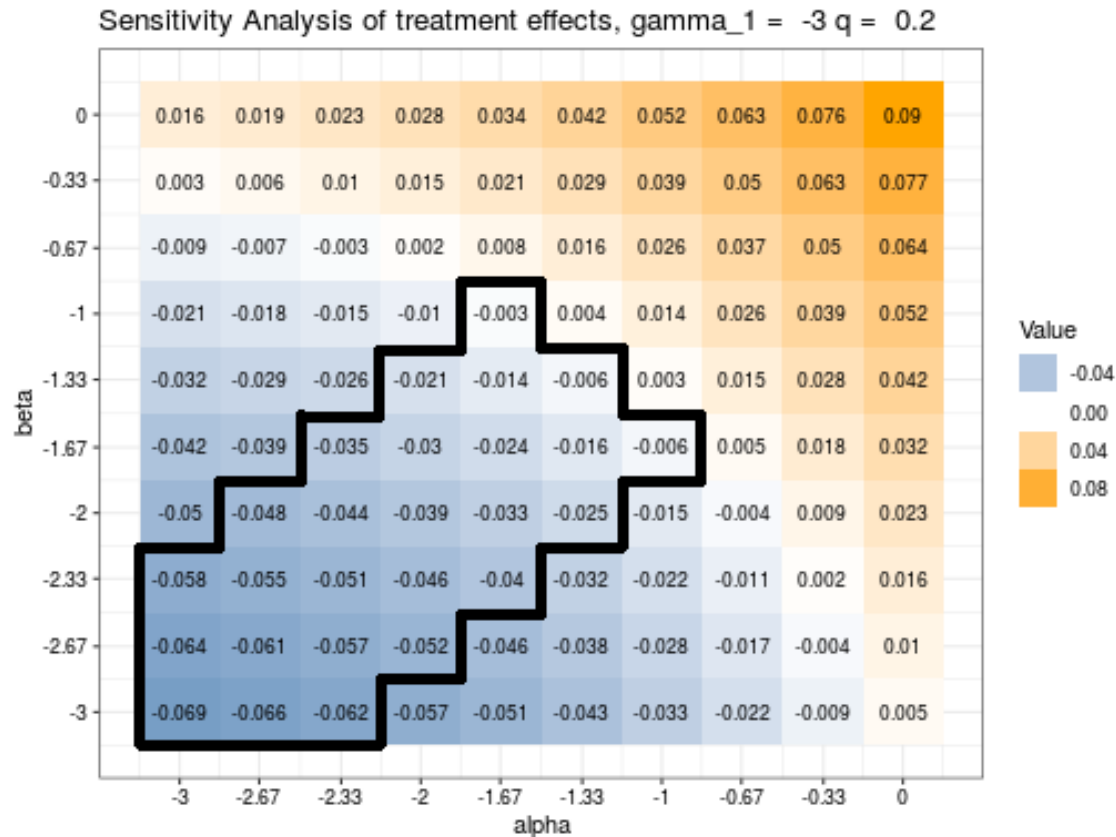
Notes: The black box contains the treatment effect assuming the following values for the sensitivity parameters:  $\alpha_1 = \beta_1 = -0.62, \gamma_1 \in [-3, 0], q = 0$ . The gray box contains the treatment effect assuming the following values for the sensitivity parameters:  $\alpha_1 = \beta_1 = -0.62, \gamma_1 = -3, q = 0.18$ .

Source: Data from Bernhard (2016), Design from TippingSens App.

With the TippingSens plot, we can also examine the estimated treatment effects in other settings. We can, for example, take a closer look at the effects of different percentages of individuals with a drinking problem in our subpopulation of interest. We switch axes in the app and receive a plot with  $\gamma$  on the vertical axis and  $q$  on the horizontal axis. We fix  $\alpha_1$  and  $\beta_1$  at  $-0.62$  implicitly assuming that the odds ratio of employment given alcohol abuse is the same in both treatment groups. The results are depicted in Figure 2. If we assume there are no men with drinking problems, the unobserved covariate should have no effect. This is confirmed in the first column of Figure 2 (where  $q = 0$ ), showing a constant estimated treatment effect of 0.09, which is the estimated treatment effect based on the unconfoundedness assumption

(black box in Figure 2). Moving to higher percentages for  $q$  we can see that the treatment effect diminishes. When  $q$  is close to zero (or one), treatment and control groups are still well balanced regarding the unobserved variable  $q$ , just because there are not enough subjects with  $u = 1$  (or  $u = 0$ ) that can be selected into either group. Thus, the effects of changing  $\gamma$  are small for small values of  $q$ . This changes as  $q$  increases. More variation in the outcome is now attributed to the unobserved covariate. Still, the treatment effect does not drop below 0.046 for the considered settings (gray box in Figure 2).

Figure 3: Sensitivity analysis of the evaluation of vocational training by Bernhard (2016)



Notes: Odds ratios regarding alcoholism do not differ by more than a factor of two between the treatment and the matched control group in the area highlighted by black lines.

Source: Data from Bernhard (2016), Design from TippingSens App.

We can also explore which conditions would be necessary to alter the sign of the estimated treatment effect based on the unconfoundedness assumption. Imbens/Rubin (2015: 506) remark that there is a trade-off between the different parameter: more extreme values for  $\gamma_1$  do not require quite as extreme value for  $\alpha_1$  and  $\beta_1$  to see the same shift in the estimated treatment effect. We thus choose rather extreme values for the fixed parameters. In Figure 3, we fix  $q$  at 20 percent and  $\gamma_1$  at -3, which means that the odds for treatment are about twenty times higher for German men without drinking problems compared to subjects with drinking prob-



lems. But even with such a strong selection mechanism, the estimated treatment effect only changes its sign if  $\alpha_1$  and  $\beta_1$  get very small (blue tiles indicate negative treatment effects). If we also take into account that it seems unrealistic that the odds ratios differ by more than a factor of two between the treatment and the matched control group, the range of admissible values for the combination of the two parameters is limited further. Admissible value combinations that would lead to a negative treatment effect are highlighted by black lines in Figure 3. The figure shows that the values for both parameters would need to be smaller than -1.33 (odds ratio less than 0.26) to change the sign of the estimated treatment effect. This is a much smaller odds ratio than the odds ratio of 0.54 found by Devaux/Sassi (2015). Thus, to obtain a negative treatment effect, we would need to assume that the target population for the training program is a highly selective group, for which the odds ratio of finding a job for heavy drinkers is much smaller and for which the prevalence of alcoholism is much higher than in the reference population. We can conclude that the findings in Bernhard (2016) are robust to the assumption that alcoholism is an unobserved confounder, which influences both the probability of assignment to the training program and the probability of being employed two years after baseline.

## 6 Discussion

The sensitivity analysis proposed by Rosenbaum and Rubin (1983) allows flexible modeling of the violation of the unconfoundedness assumption through four different parameters. In their initial illustrative application, Rosenbaum and Rubin worked with forking tables to convey their idea and their results. The table format forced the authors to evaluate only a limited set of combinations for the parameters. With the R Shiny app introduced in this paper, we simplify the sensitivity analysis by creating interactive visualizations instead. Since it is typically unrealistic to assume that exact values are known for all four sensitivity parameters, the possibility of specifying ranges of plausible values for the parameters is a major advantage of the app compared to previously proposed solutions, such as using the most extreme relationships found in the observed data. With the app, the estimated treatment effects are visualized over the two-dimensional space spanned by the range of plausible values for the free parameters. Color coding helps to identify the relationships between the different parameters and the treatment effect. It is up to the user to decide which parameters should be held fixed and which intervals should be considered for the free parameters. Once the data are loaded, all settings can be changed easily: parameters can be exchanged between the axes or from being fixed to being free, and vice versa, ranges of plausible values can be adjusted independently, and the values of the fixed parameters can be modified using separate sliders for each of the parameters (see also the step-by-step illustration in the online supplement). The interactive flexibility of the app also provides quick insights which (possibly extreme) sets of parameter combinations would be required to substantially change the research findings derived under the unconfoundedness assumption.

The illustrative application based on a quasi-experimental evaluation of vocational training in Germany discussed in Section 5 highlights the benefits of the TippingSens app. Evaluating different plausible scenarios regarding the association between the confounder and the outcome and the treatment, we found that the substantial findings in Bernhard (2016) are robust regarding the effects of alcoholism as a possible confounder. Of course, the evaluation has important limitations. It only focuses on one confounder. Other health problems might also have negative effects on both the outcome and the treatment assignment. Thus, it would not be appropriate to conclude based on this limited sensitivity analysis that the findings in Bernhard (2016) are robust to any form of confounding. However, the application illustrates the basic idea. Similar sensitivity analyses could be conducted for other health parameters in practice.

Still, it must be noted that the applicability of the app is limited by the requirements of the Rosenbaum-Rubin sensitivity analysis: both, the outcome, as well as the unobserved confounder, have to be univariate and binary. While it has been argued that this framework

might still apply in contexts with more than one confounder (Liu/Kuramoto/Stuart, 2013), the assumption of a binary confounder might be too restrictive in other contexts. For example, with continuous confounders, it seems more realistic to assume that the effect of the confounder changes (non)linearly with the value of the confounder instead of assuming only a single change in the effect at a certain threshold value. Whether the app could be extended to allow obtaining useful insights in this more general context beyond the Rosenbaum-Rubin approach would be an interesting area for future research.

# References

- Ashenfelter, O. (1978): Estimating the effect of training programs on earnings. In: *Review of Economics and Statistics*, Vol. 6, No. 1, p. 47–57.
- Bayer, P.; Ross, S.L.; Topa, G. (2008): Place of work and place of residence: informal hiring networks and labor market outcomes. In: *Journal of Political Economy*, Vol. 116, No. 6, p. 1150–1196.
- Becker, Sascha; Caliendo, Marco (2007): Sensitivity analysis for average treatment effects. In: *Stata Journal*, Vol. 7, No. 1, p. 71–83, URL <https://EconPapers.repec.org/RePEc:tsj:stataj:v:7:y:2007:i:1:p:71-83>.
- Bernhard, Sarah (2016): Berufliche Weiterbildung von Arbeitslosengeld-II-Empfängern. Langfristige Wirkungsanalysen. In: *Sozialer Fortschritt*, Vol. 65, No. 7, p. 153–161, URL <http://ejournals.duncker-humblot.de/doi/abs/10.3790/sfo.65.7.153>.
- Blundell, Richard; Gosling, Amanda; Ichimura, Hidehiko; Meghir, Costas (2007): Changes in the distribution of male and female wages accounting for employment composition using bounds. In: *Econometrica*, Vol. 75, No. 2, p. 323–363.
- Card, D.; Kluve, J.; Weber, A. (2010): Active labour market policy evaluations: a meta-analysis. In: *The Economic Journal*, Vol. 120, No. November, p. F452–F477.
- Cornfield, Jerome; Lilienfeld, Abraham M.; Hammond, E. Cuyler; Wynder, Ernst L.; Shimkin, Michael B.; Haenszel, William (1959): Smoking and lung cancer: recent evidence and a discussion of some questions. In: *JNCI: Journal of the National Cancer Institute*, Vol. 22, No. 1, p. 173–203, URL <https://doi.org/10.1093/jnci/22.1.173>.
- Devaux, Marion; Sassi, Franco (2015): The labour market impacts of obesity, smoking, alcohol use and related chronic diseases. In: *OECD Health Working Papers*, Vol. 86, p. 1–50, URL </content/workingpaper/5jrqn5fpv0v-en>.
- Donner, Matthias; Heining, Jörg; Jacobebbinghaus, Peter; Seth, Stefan (2010): The sample of integrated labour market biographies. In: *Schmollers Jahrbuch*, Vol. 130, No. 4, p. 599–608.
- Fredriksson, Peter; Johansson, Per (2008): Dynamic treatment assignment. In: *Journal of Business and Economic Statistics*, Vol. 26, No. 4, p. 435–445.
- Gangl, Markus (2004): RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated. Statistical Software Components, Boston College Department of Economics, URL <https://ideas.repec.org/c/boc/bocode/s438301.html>.

- Gastwirth, Joseph L.; Krieger, Abba M.; Rosenbaum, Paul R. (1998): Dual and simultaneous sensitivity analysis for matched pairs. In: *Biometrika*, Vol. 85, No. 4, p. 907–920, URL <https://dx.doi.org/10.1093/biomet/85.4.907>.
- Greenland, S. (1996): Basic methods for sensitivity analysis of biases. In: *International Journal of Epidemiology*, Vol. 25, No. 6, p. 1107–1116.
- Harding, David J. (2003): Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. In: *American Journal of Sociology*, Vol. 109, No. 3, p. 676–719, URL <http://www.jstor.org/stable/10.1086/379217>.
- Heckman, J.; Ichimura, H.; Smith, J.; Todd, P. (1998): Characterizing selection bias using experimental data. In: *Econometrica*, Vol. 66, No. 5, p. 1017–1098.
- Heckman, James J. (1989): Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. Working Paper 2861, National Bureau of Economic Research, URL <http://www.nber.org/papers/w2861>.
- Heckman, J.J.; Ichimura, H.; Todd, P.E. (1997): Matching as an economic evaluation estimator: Evidence from evaluating a job training programme. In: *The Review of Economic Studies*, Vol. 64, No. 4, p. 605–654.
- Heckman, J.J.; Stixrud, J.; Urzua, S. (2006): The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. In: *Journal of Labor Economics*, Vol. 24, No. 3, p. 411–482.
- Henkel, Dieter (2011): Unemployment and substance use: a review of the literature (1990–2010). In: *Current Drug Abuse Reviews*, Vol. 4, No. 1, p. 4–27.
- Hof, Stefanie (2014): Does private tutoring work? The effectiveness of private tutoring: A non-parametric bounds analysis. In: *Education Economics*, Vol. 22, No. 4, p. 347–366.
- Holland, Paul W. (1986): Statistics and causal inference. In: *Journal of the American Statistical Association*, Vol. 81, No. 396, p. 945–960, URL <http://www.jstor.org/stable/2289064>.
- Ichino, Andrea; Mealli, Fabrizia; Nannicini, Tommaso (2008): From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? In: *Journal of Applied Econometrics*, Vol. 23, No. 3, p. 305–327.
- Imbens, Guido W.; Rubin, Donald B. (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, chap. 22. Cambridge: Cambridge University Press, p. 496–510.
- Imbens, Guido W. (2003): Sensitivity to exogeneity assumptions in program evaluation. In: *American Economic Review*, Vol. 93, No. 2, p. 126–132, URL <http://www.aeaweb.org/articles?id=10.1257/000282803321946921>.

- Johansson, Edvard; Alho, Hannu; Kiiskinen, Urpo; Poikolainen, Kari (2007): The association of alcohol dependency with employment probability: Evidence from the population survey 'Health 2000 in Finland'. In: *Health Economics*, Vol. 16, No. 7, p. 739–754.
- Kang, Changhui (2011): Family Size and Educational Investments in Children: Evidence from Private Tutoring Expenditures in South Korea. In: *Oxford Bulletin of Economics and Statistics*, Vol. 73, No. 1, p. 59–78, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0084.2010.00607.x>.
- Keele, L. (2010): Rbounds: An R package for sensitivity analysis with matched data. R package, URL <http://www.polisci.ohio-state.edu/faculty/lkeele/rbounds.html>.
- Kitahata et al., Mari M. (2009): Effect of early versus deferred antiretroviral therapy for HIV on survival. In: *New England Journal of Medicine*, Vol. 360, No. 18, p. 1815–1826, URL <https://doi.org/10.1056/nejmoa0807252>.
- Lechner, Michael (1999): Earnings and employment effects of continuous off-the-job training in East Germany after unification. In: *Journal of Business and Economic Statistics*, Vol. 17, No. 1, p. 74–90.
- Lechner, Michael; Wunsch, Conny (2013): Sensitivity of matching-based program evaluations to the availability of control variables. In: *Labour Economics*, Vol. 21, No. C, p. 111–121, URL <https://EconPapers.repec.org/RePEc:eee:labeco:v:21:y:2013:i:c:p:111-121>.
- Liu, Weiwei; Kuramoto, S. Janet; Stuart, Elizabeth A. (2013): An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. In: *Prevention Science*, Vol. 14, No. 6, p. 570–580, URL <https://doi.org/10.1007/s11121-012-0339-5>.
- Liubinska, Victoria; Rubin, Donald B. (2014): Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. In: *Statistics in Medicine*, Vol. 33, No. 24, p. 4170–4185, URL <http://onlinelibrary.wiley.com/doi/10.1002/sim.6197/abstract>.
- MacDonald, Ziggy; Shields, Michael A. (2004): Does problem drinking affect employment? Evidence from England. In: *Health Economics*, Vol. 13, No. 2, p. 139–155.
- Manski, Charles F. (1990): Nonparametric bounds on treatment effects. In: *The American Economic Review*, Vol. 80, No. 2, p. 319–323, URL <http://www.jstor.org/stable/2006592>.
- Manski, Charles F; Pepper, John V (2009): More on monotone instrumental variables. In: *The Econometrics Journal*, Vol. 12, No. S1, p. 200–216.
- Manski, Charles F; Pepper, John V (2000): Monotone Instrumental Variables: With an Application to the Returns to Schooling. In: *Econometrica*, Vol. 68, No. 4, p. 997–1010.
- Mueller, G.; Plug, E. (2006): Estimating the effect of personality on male and female earnings. In: *Industrial Labor Relations Review*, Vol. 60, No. 1, p. 3–22.

- Mullahy, J.; Sindelar, J. (1996): Employment, unemployment, and problem drinking. In: *Journal of Health Economics*, Vol. 15, No. 4, p. 409–434.
- Pannenberg, M. (2010): Risk attitudes and reservation wages of unemployed workers: Evidence from panel data. In: *Econometric Letters*, Vol. 106, No. 3, p. 223–226.
- Pearl, Judea; Glymour, Madelyn; Jewell, Nicholas P. (2016): *Causal inference in statistics: A primer*. John Wiley & Sons.
- Popovici, Ioana; French, Michael T. (2016): Substance use and school and occupational performances. In: Sher, Kenneth J. (Ed.) *The Oxford Handbook of Substance Use and Substance Use Disorders: Volume 2*, Oxford: Oxford University Press.
- Rosenbaum, P. R.; Rubin, D. B. (1983): Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 45, No. 2, p. 212–218, URL <http://www.jstor.org/stable/2345524>.
- Rosenbaum, Paul R. (2018): Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. In: *Annals of Applied Statistics*, Vol. 12, No. 4, p. 2312–2334, URL <https://doi.org/10.1214/18-AOAS1153>.
- Rosenbaum, Paul R. (2014): Two R packages for sensitivity analysis in observational studies. In: *Observational Studies*, Vol. 1, No. 1, p. 1–17.
- Rosenbaum, Paul R. (2010): *Design of Observational Studies*, chap. 14. New York: Springer, p. 257–274.
- Rosenbaum, Paul R. (2007): Sensitivity analysis for m-Estimates, tests, and confidence intervals in matched observational studies. In: *Biometrics*, Vol. 63, No. 2, p. 456–464, URL <https://doi.org/10.1111/j.1541-0420.2006.00717.x>.
- Rosenbaum, Paul R. (2003): Does a dose-response relationship reduce sensitivity to hidden bias? In: *Biostatistics*, Vol. 4, No. 1, p. 1–10.
- Rosenbaum, Paul R. (1999): Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects. In: *Biometrics*, Vol. 55, No. 2, p. 560–564, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.1999.00560.x>.
- Rosenbaum, Paul R. (1995): *Observational Studies*. Springer New York, URL <https://doi.org/10.1007/978-1-4757-3692-2>.
- Rubin, Donald B. (1974): Estimating causal effects of treatments in randomized and nonrandomized studies. In: *Journal of Educational Psychology*, Vol. 66, No. 5, p. 688–701.
- Rüb, Felix; Werner, Daniel (2007): Typisierung von SGB II-Trägern. In: *Institute for Employment Research. Research Report*, , No. 1/2007, p. 1–35.

- Smith, Jeffrey A.; Todd, Petra E. (2001): Reconciling conflicting evidence on the performance of propensity-score matching methods. In: American Economic Review, Vol. 91, No. 2, p. 112–118, URL <http://www.aeaweb.org/articles?id=10.1257/aer.91.2.112>.
- Terza, Joseph V. (2002): Alcohol abuse and employment: A second look. In: Journal of Applied Econometrics, Vol. 17, No. 4, p. 393–404, URL <http://onlinelibrary.wiley.com/doi/10.1002/jae.671/abstract>.
- VanderWeele, Tyler J.; Arah, Onyebuchi A. (2011): Bias formulas for sensitivity analysis of unmeasured confounding for general Outcomes, treatments, and confounders. In: Epidemiology, Vol. 22, No. 1, p. 42–52, URL <https://doi.org/10.1097/ede.0b013e3181f74493>.
- Zubizarreta, José R.; Cerdá, Magdalena; Rosenbaum, Paul R. (2013): Effect of the 2010 chilean earthquake on posttraumatic stress. In: Epidemiology, Vol. 24, No. 1, p. 79–87, URL <https://doi.org/10.1097/ede.0b013e318277367e>.



# Appendix

## Supplementary material

### A Step-by-step Guide for the TippingSens R Shiny App

In their illustrative application, Rosenbaum and Rubin worked with forking tables to convey their idea for a sensitivity analysis regarding an unobserved binary confounder. This approach is only useful if a very limited set of plausible values needs to be evaluated per parameter (Rosenbaum and Rubin choose two values for  $\alpha_1$ ,  $\beta_1$ , and  $\gamma_1$  and three values for  $q$ ). More parameter sets would result in complex and hard to read tables. The app simplifies comparisons by creating visualizations instead of tables, which also allow looking into more possible parameter combinations. For the implementation, we borrow ideas from the plot design developed by Liublinska/Rubin (2014) and adapt it for our purpose. Liublinska/Rubin (2014) developed tipping point plots to visualize how different assumptions about missing data affect statistics, such as the average treatment effect. Assuming a binary outcome, the two axes of their plot represent the number of positive outcomes among the missing cases for the treatment and the control group, respectively. The plot is divided into tiles, and the statistic of interest is computed for each tile.

We adopt this design to plot the results of a Rosenbaum-Rubin sensitivity analysis. Since the sensitivity analysis depends on four parameters instead of the two parameters considered in Liublinska/Rubin (2014), we choose an interactive display, an R shiny app. The app allows selecting two of the four parameters for the axes. The values for the other two parameters are fixed at user specified values when generating the plot. Drop down menus allow manipulating which parameters are displayed on the axes and sliders help adjusting the constant values for the fixed parameters and specifying the range for the parameters displayed on the axes. The plot is updated automatically each time any of the specifications is adjusted. The TippingSens app is available at <https://tippingsens.shinyapps.io/TSApp/>.

### The general setup

The default setup when loading the app is displayed in Figure A.1. The panel, which will eventually display the sensitivity plots (Panel 1 in Figure A.1) contains a brief summary regarding the interpretation of the different sensitivity parameters. A first sensitivity plot is created as soon as the user interacts with any menu or slider. Default data for this first plot are based on the illustrative example in Rosenbaum/Rubin (1983). In addition to Panel 1, we have a panel containing two drop-down menus to choose the sensitivity parameters printed on the

axes (Panel 2). We also have two sliders to set the range limits for the parameters on the axes (Panel 3) and two sliders for the fixed sensitivity parameters (Panel 4). The user can upload data in Panel 5. The last drop-down menu in Panel 6 offers two different color fillings for the plot.

**Figure A.1: Default appearance of the TippingSens app before interacting with any menu/slider**

TippingSens App for Rosenbaum-Rubin Sensitivity Analyses

Please move any slider to create your first sensitivity plot. Default data are taken from Rosenbaum and Rubin (1983).

Remember:  
alpha - Log odds ratio of the confounder regarding the outcome in the treatment group  
beta - Log odds ratio of the confounder regarding the outcome in the control group  
gamma - Log odds ratio of the confounder regarding the treatment assignment  
q - Prevalence of the binary confounder

Download Plot

Choose which sensitivity parameters should be displayed on the axes of the output (the other two parameters will be treated as fixed).

alpha

beta

Choose a range for the parameters displayed on the axes.

Range for alpha\_1

Range for beta\_1

Choose a value for the parameters treated as fixed.

Value for gamma\_1

Value for q

Choose CSV File for the data (treatment and outcome).

Browse... No file selected

Take care to specify column names correctly. The column names should be 'Treatment' and 'Outcome'. No rownames or missing values are allowed. Separators are allowed to be comma, semicolon, or tab. Decimal separators are allowed to be comma or point.

An example file can be found here (<https://github.com/CaroHaensch/TippingSensExampleFiles>).

Separator

☒ Comma  
☐ Semicolon  
☐ Tab

Decimal separator

☒ Point  
☐ Comma

The colour filling

range

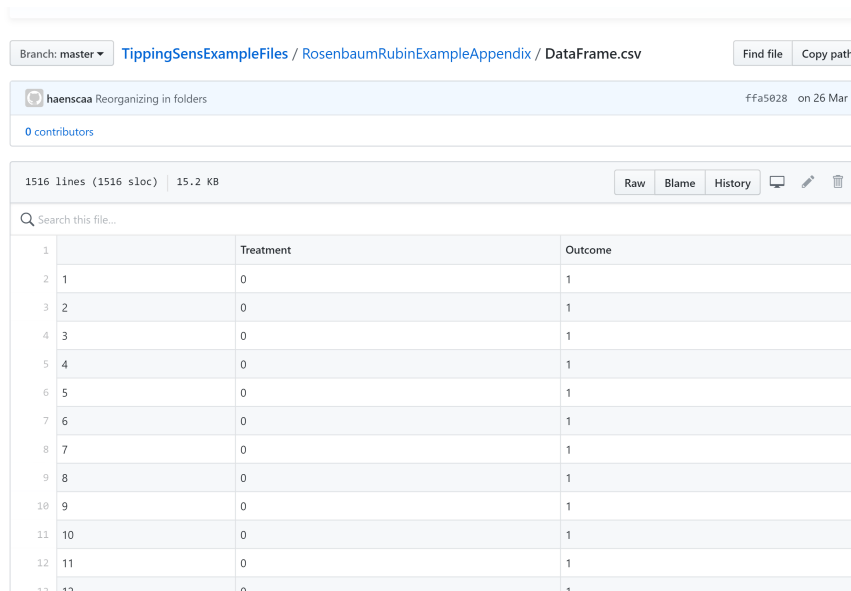
When choosing range the largest value will be orange and the smallest white, when choosing zerotomax negative values will be blue, positive ones orange and values near zero will be white.

Notes: Red numbers added by the authors.  
Source: TippingSens App.

## Using the App

To illustrate how to interact with the app, we use data from the illustrative example in Rosenbaum/Rubin (1983). The authors use data from a clinical study investigating the effects of two different treatments (coronary artery bypass surgery or medical therapy) on symptomatic relief from coronary artery disease. Of the 1515 patients contained in the study, 590 received surgery while 925 underwent medical treatment. The outcome variable in the study is an indicator of improvement six months after cardiac catheterization (1 = improvement, 0 = no improvement). This dataset is also used as a default if no other data are provided.

**Figure A.2: Example of input data for the app in csv format containing two columns labeled “Outcome” and “Treatment”.**



	Treatment	Outcome
1		
2	1	0
3	2	0
4	3	0
5	4	0
6	5	0
7	6	0
8	7	0
9	8	0
10	9	0
11	10	0
12	11	0
13	12	0

Source: TippingSens App.

## Uploading the data

Note that the app assumes that the data are well balanced, i.e., any steps to ensure the balance between the treatment group and control group regarding the observed covariates have been performed before loading the app. Under this assumption, the only information required is the data on the outcome and the treatment indicator. The TippingSens app expects that this information is provided in the form of a csv-file containing two columns: one labeled “Treatment” and the other labeled “Outcome”. Commas, semicolons, or tabs can be used as column separators, and points or commas can be used as decimal separators, but the selected separators need to be specified in the data upload panel (see Panel 5 in Figure A.1). An example, how the csv-file would look like for the data from Rosenbaum/Rubin (1983)

is depicted in Figure A.2 (following the notation in the original article, surgery is coded as 1, and medical treatment is coded as 0). The data are also available at <https://github.com/CaroHaensch/TippingSensExampleFiles/>.

The csv-file containing the data can be uploaded by clicking on the browse button and selecting the file from the appropriate folder.

**Figure A.3: Parameter settings used to generate output displayed in Figure A.4**

Source: TippingSens App

## Adjusting the parameters

Once the data from the (quasi-)experiment are uploaded, researchers have to think about appropriate sensitivity parameters. In their illustration, Rosenbaum and Rubin used three different values for the sensitivity parameter  $q$ : 0.1, 0.5, and 0.9. They also assumed that an unobserved variable could double or half the odds of a recovery in the treatment (surgery) and the control (medical therapy) group. Doubling the odds of recovery in the treatment group is equivalent to  $e^{\alpha_1} = 2$ , therefore the sensitivity parameter needs to be  $\alpha_1 = \ln(2) = 0.693$ . Reducing the odds by half is equivalent to  $e^{\alpha_1} = 1/2$ , therefore  $\alpha_1 = \ln(1/2) = -0.693$ . Finally, they assumed that the unobserved variable could double or triple the odds of getting the treatment.

To compare the results from the app with those of Rosenbaum/Rubin (1983), we keep  $\alpha_1$  and  $\beta_1$  as the parameters to be displayed on the axes and change  $q$  and  $\gamma_1$  from their default values to  $q = 0.5$  and  $\gamma_1 = 0.69$  (the app only allows two decimal places). We also limit the ranges

for  $\alpha_1$  and  $\beta_1$  to  $[-0.69, 0.69]$  to allow for a direct comparison with the results presented Table 2 from Rosenbaum/Rubin (1983). The selected parameter settings are displayed in Figure A.3. The app generates a sensitivity plot, as displayed in Figure A.4. The plot can also be downloaded as a png-file to allow easy integration into technical reports or research papers.

We can compare the TippingSens plot with the table from Rosenbaum and Rubin by subtracting the probability in the row labeled “M” (for medical treatment) from the probability in the row labeled “S” (for surgery) in Table A.1. Corresponding effect sizes are marked by letters a, b, c and d in Table A.1 and Figure A.5.

**Table A.1: Upper half of original table containing sensitivity analysis results from Rosenbaum/Rubin (1983), Table 2, page 216**

Effect of u=1 vs u=0 on treatment assignment z	Effect of u=1 vs u=0 on response under M	Effect of u=1 vs u=0 on response under S	Fraction of patients with u=0: $\pi$					
			0.1		0.5		0.9	
Doubles the odds of surgery $exp(\alpha) = 2$	Halves the odds of improvement $exp(\delta_0) = \frac{1}{2}$	Halves the odds of improvement $exp(\delta_1) = \frac{1}{2}$	S	0.67	S	0.68	S	0.68
			M	0.36	M	0.35	M	0.36
		Doubles the odds of improvement $exp(\delta_1) = 2$	S	0.66	S	0.65	S	0.66
			M	0.36	M	0.35	M	0.36
		Doubles the odds of improvement $exp(\delta_0) = 2$	S	0.67	S	0.68	S	0.68
			M	0.36	M	0.37	M	0.36
	Doubles the odds of improvement $exp(\delta_0) = 2$	Halves the odds of improvement $exp(\delta_1) = \frac{1}{2}$	S	0.67	S	0.68	S	0.68
			M	0.36	M	0.37	M	0.36
		Doubles the odds of improvement $exp(\delta_1) = 2$	S	0.66	S	0.65	S	0.66
			M	0.36	M	0.37	M	0.36
		Doubles the odds of improvement $exp(\delta_0) = 2$	S	0.67	S	0.68	S	0.68
			M	0.36	M	0.37	M	0.36

Notes: Results for  $exp(\alpha) = 3$  omitted. Parameters  $\alpha, \delta_0, \delta_1$  in the notation of Rosenbaum and Rubin are  $\gamma_1, \beta_1, \alpha_1$  in the notation used in this paper. Red letters correspond to the results in Figure A.5 and were added by the authors.

Source: Rosenbaum/Rubin (1983)

A final feature of the app, which we do not illustrate further, is the option to change the color filling (see Panel 6 in Figure A.1). When choosing range, the largest value will be orange, and the smallest value will be white, when choosing zerotomax negative values will be blue, positive values will be orange, and values near zero will be white (see Figure 3 in the main text for an illustration of the latter setting).

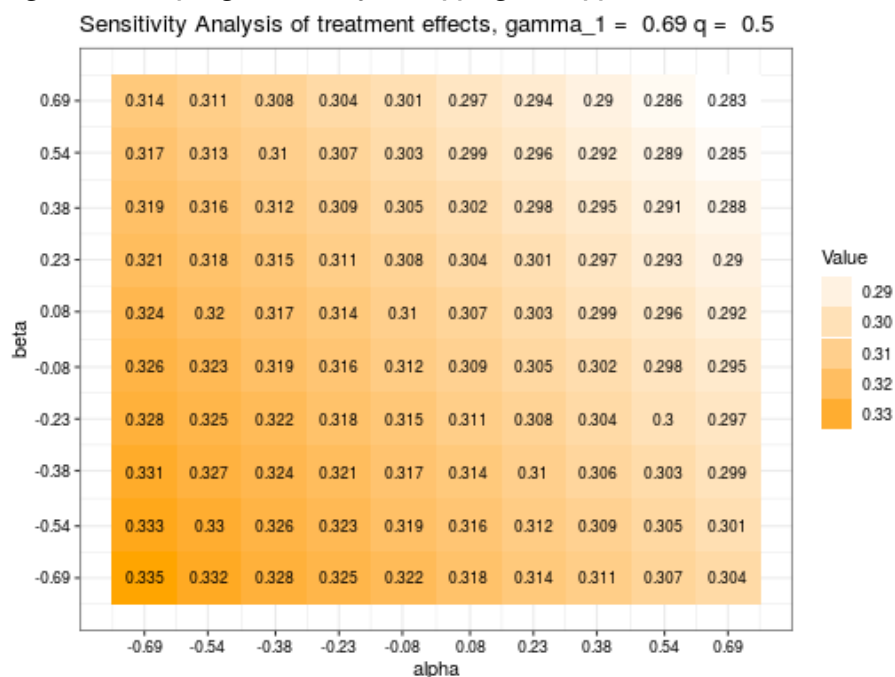
## Limitations of the App

To ensure that the handling of the app is intuitive, we deliberately limited the flexibility of the parameter settings:

- The maximum range for the parameters  $\alpha_1$ ,  $\beta_1$ , and  $\gamma_1$  is  $[-10;10]$  (the range for  $q$  is naturally bounded between 0 and 1).
- Parameter values can only be specified up to the second decimal point.
- Two of the four parameters need to be fixed.

We feel that these limitations are necessary to ensure that useful and easily interpretable output can be generated with just a few clicks once the data have been uploaded. At the same time, we tried to strike a balance between user friendliness and broad applicability. For example, there is an inherent trade-off between the bounds for the parameters and the granularity that can be offered when specifying the values for the fixed parameters. Large bounds on the sliders will allow picking from a very wide range of plausible values. However, it will be more difficult to fix the parameters in specific settings. We feel that fixing the bounds at  $[-10;10]$  offers a good compromise. The bounds imply that the assumed odds-ratios are bounded roughly between  $4.55 \cdot 10^{-5}$  and 22,000. We believe that these bounds are sufficiently extreme for most practical purposes. At the same time, the bounds ensure that the users can conveniently pick any value between the bounds in incremental steps of 0.01 (users can click on the button of the slider and use the up and down arrow on the keyboard for fine-tuning). Of course, it would also have been possible to let the user specify the values of the fixed parameters directly. However, the interactive property of the app would have been lost. We believe that it is one of the attractive features of the app that the researchers can use the sliders to directly evaluate how increasing or decreasing the values of the fixed parameters impacts the estimated treatment effect.

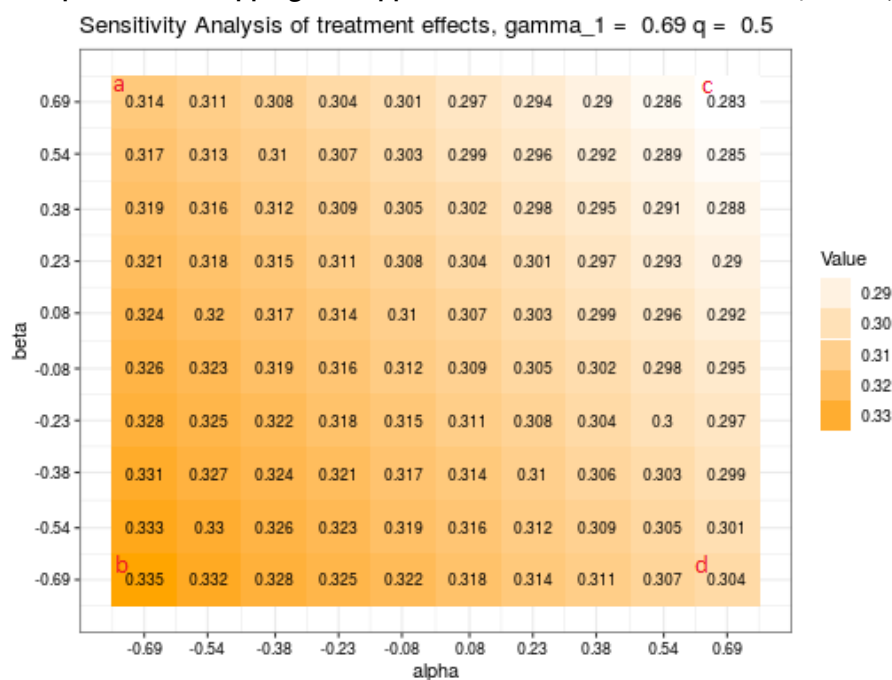
Figure A.4: Output generated by the TippingSens app based on data from Rosenbaum/Rubin (1983)



Notes: These data are also the default data used for the app.

Source: Data from Rosenbaum/Rubin (1983), Design from TippingSens App.

Figure A.5: Output from the TippingSens app with data taken from Rosenbaum/Rubin (1983)



Notes: Red letters correspond to the results in Figure A.1 and were added by the authors.

Source: Data from Rosenbaum/Rubin (1983), Design from TippingSense App

# Imprint

## **IAB-Discussion Paper 29|2020EN**

### **Publication Date**

01. October 2020

### **Publisher**

Institute for Employment Research  
of the Federal Employment Agency  
Regensburger Straße 104  
90478 Nürnberg  
Germany

### **All rights reserved**

Reproduction and distribution in any form – also in parts – requires the permission of the IAB

### **Download**

<http://doku.iab.de/discussionpapers/2020/dp2920.pdf>

### **All publications in the series “IAB-Discussion Paper” can be downloaded from**

<https://www.iab.de/en/publikationen/discussionpaper.aspx>

### **Website**

[www.iab.de/en](http://www.iab.de/en)

---

### **Corresponding author**

Sarah Bernhard  
Phone +49 (911) 179 3079  
E-Mail [Sarah.Bernhard@iab.de](mailto:Sarah.Bernhard@iab.de)

Jörg Drechsler  
Phone +49 (911) 179 4021  
E-Mail [Joerg.Drechsler@iab.de](mailto:Joerg.Drechsler@iab.de)