Exploration, Confirmation, and Replication in the Same Observational Study: A Two Team Cross-Screening Approach to Studying the Effect of Unwanted Pregnancy on Mothers' Later Life Outcomes

Samrat Roy
Operations and Decision Sciences, Indian Institute of Management Ahmedabad
Marina Bogomolov
Data and Decision Sciences, Technion - Israel Institute of Technology
Ruth Heller
Department of Statistics and Operations Research, Tel-Aviv University
Amy M. Claridge
Child Development and Family Science, Central Washington University
Tishra Beeson
Department of Health Sciences, Central Washington University
and
Dylan S. Small[*]
Department of Statistics and Data Science, University of Pennsylvania

January 14, 2025

Abstract

The long term consequences of unwanted pregnancies carried to term on mothers have not been much explored. We use data from the Wisconsin Longitudinal Study (WLS) and propose a novel approach, namely two team cross-screening, to study the possible effects of unwanted pregnancies carried to term on various aspects of mothers' later-life mental health, physical health, economic well-being and life satisfaction. Our method, unlike existing approaches to observational studies, enables the investigators to perform exploratory data analysis, confirmatory data analysis and replication in the same study – this is a valuable property when there is only a single data set available with unique strengths to perform exploratory, confirmatory and replication

[*] The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

analysis. In two team cross-screening, the investigators split themselves into two teams and the data is split as well according to a meaningful covariate. Each team then performs exploratory data analysis on its part of the data to design an analysis plan for the other part of the data. The complete freedom of the teams in designing the analysis has the potential to generate new unanticipated hypotheses in addition to a prefixed set of hypotheses. Moreover, only the hypotheses that looked promising in the data each team explored are forwarded for analysis (thus alleviating the multiple testing problem). These advantages are demonstrated in our study of the effects of unwanted pregnancies on mothers' later life outcomes.

Keywords: Causal Inference, Multiple Testing, Sample Splitting, Replicability.

1 Introduction

1.1 An Observational Study Assessing the Effect of Unwanted Pregnancies on Mothers' Later Life Outcomes

Unwanted pregnancies are prevalent worldwide (Postlethwaite et al., 2010; Finer and Zolna, 2016; Bearak et al., 2018). For unwanted pregnancies that are carried to term, it is crucial to assess their effect on the child that is born and on the mother in order to design policies that best support the child and the mother. For the mother, there is considerable evidence of short term effects such as less emotional attachment to her baby in pregnancy (Pakseresht et al., 2018), more parenting stress at 6 months and one year postpartum, and less effective parenting strategies (Mark and Cowan, 2022). However, there is little evidence about the long term effects. The only study we are aware of in this context is Herd et al. (2016) which used linear regression to examine one aspect of mental health, depression. In this work, we aim to study the possible effects of an unwanted pregnancy carried to term on various aspects of a mother's later-life mental health, physical health, economic well-being and life satisfaction. A data set that has unique strengths to carry out such an observational study is the Wisconsin Longitudinal Study (WLS) (Herd et al., 2014). The WLS took a random sample of one third of people graduating from high school in Wisconsin in 1957 and had longitudinally followed them approximately every decade since. Strengths of the WLS include:

– The WLS asked about whether the pregnancy was wanted somewhat contemporaneously at age approximately 34. Specifically, women who had experienced pregnancies by age 34 were asked for each pregnancy, "Before you became pregnant did you want to become pregnant at that time?", and if the woman answered no, she was asked, "Did you want to have another baby sometime?". If the woman answered no to both questions, it was coded as an unwanted pregnancy. This way of identifying an unwanted pregnancy when the woman is of age 34 is better than asking a woman when a child is completely grown up, but not as good as asking a woman about her pregnancy intention prior to her pregnancy as her recall could be colored by the events of the intervening years between the pregnancy and the time the woman turned 34. However, we proceed with this way of identification due to lack of any other reliable data sources.

- The WLS measured a wide range of variables prospectively before the births of participants' children that could confound the relationship between unwanted pregnancies and later life outcomes. These variables include family background, adolescent characteristics, educational and occupational achievement, and aspirations.

- The WLS has long term longitudinal follow-up of its respondents when they were around age 53, 65 and 72, enabling the study of later life outcomes.

Note that women in the WLS experienced most of their pregnancies before the 1973 *Roe v. Wade* decision that made abortion legal throughout the United States. Prior to 1973, abortion was illegal in Wisconsin. Thus, this data is most relevant to the question of what is the effect of unwanted pregnancy in places where abortion is illegal. This is an important current question because today abortion is completely illegal in 24 countries, is legal only when the women's health is at risk in 37 other countries, and is illegal in some states in the U.S. following the 2022 Dobbs vs. Jackson Supreme Court decision.

1.2 Exploratory Data Analysis, Confirmatory Data Analysis and Replication in Observational Studies

Exploratory data analysis, confirmatory data analysis and replication are three important aspects of building strong evidence from observational studies. Exploratory data analysis (EDA) helps in forming hypotheses that one might not have anticipated and in removing outliers that do not represent natural variation in the data (Tukey, 1977; Diaconis, 2006). It can also help in identifying a variable that does not measure what one thought it did, and it may suggest a different variable which better measures the construct one is actually interested in. EDA is like detective work about a crime, following hunches and leads to establish pretrial evidence (Tukey, 1977). Confirmatory data analysis, on the other hand, is a scientific trial about the hypotheses suggested by EDA. John Tukey, the "father of exploratory data analysis" (Donoho, 2017), emphasized that confirmatory data analysis is critical for science (Tukey, 1980):

> *"Important questions can demand the most careful planning for confirmatory analysis."*

Replication of results is also important for science (National Academies of Sciences et al., 2019):

> *"Repeated findings of comparable results tend to confirm the veracity of an original scientific conclusion, and, by the same token, repeated failures to confirm throw the original conclusions into doubt."*

For establishing replicability in observational studies, it is important to avoid just repeating the same hidden bias (Rosenbaum, 2001). The most valuable replication in observational studies is consistency among studies that may suffer from different hidden biases. For example, if a study in New York City found that fish eaters live longer than non-fish eaters, then repeating the study in Philadelphia might not be particularly helpful because both studies could be biased by fish being a comparatively expensive food that is favored by those seeking to maintain a healthy diet (Rosenbaum, 2015). A more helpful replication would be the one that Lund and Bønaa (1993) discussed, where they compared wives of fishermen in Norway to wives of other workers with similar income. This study might also be biased, e.g., wives of fishermen might be more likely to live in rural areas and get more exercise than wives of other workers of similar income. However, the reason why some individuals are getting treatment (eating fish) and others are not, is somewhat different in the two studies – in the Norway study, it may be because the wives of fishermen sometimes eat fish that they are unable to sell, whereas in the case of New York City, it may be because the fish eaters have higher income. Thus the potential sources of bias in the two studies are different. If both studies found evidence of a treatment effect, then in order to just explain this away as bias without any treatment effect, one would need to posit that two different biases were present. On the other hand, if one had studied New York City and Philadelphia, an ostensible treatment effect in both studies could be explained away by one bias. In our study of the effect of unwanted pregnancy on mothers, Catholic and non-Catholic women could have unwanted pregnancies for somewhat different reasons. Some Catholic women could have unwanted pregnancies because they followed the Catholic Church's opposition to contraceptive methods (Miller and Gur, 2002; Jones and Dreweke, 2011) [note that it would be good if we could verify that the Catholic Church's opposition actually had an effect on contraception among Catholic women, but unfortunately there are no data in the WLS to verify this]. On the other hand, non-Catholic women could be relatively more likely to have unwanted pregnancies because a woman (or her partner) either willingly did not use contraception, or improperly used contraception. If we were to find that unwanted pregnancy was associated with increased depression among both Catholic and non-Catholic women, it could not be explained away as non-causal just by Catholic women who followed the Catholic Church's opposition to contraceptive methods having a more depressive personality than Catholic women who ignored the Church's opposition or just by non-Catholic women who improperly used contraception having a more depressive personality than non-Catholic women who properly used contraception. Both biases (or some other biases) would have to be present to explain away the associations as non-causal. In general, an observational study can effectively replicate results by splitting the data based on a factor such that the individuals with one level of that factor receive treatment for one reason and the other individuals with the other level of that factor receive treatment for a different reason (Rosenbaum, 2001, 2015).

1.3 Exploration, Confirmation and Replication in the Same Observational Study: A Two Team Cross-Screening Approach

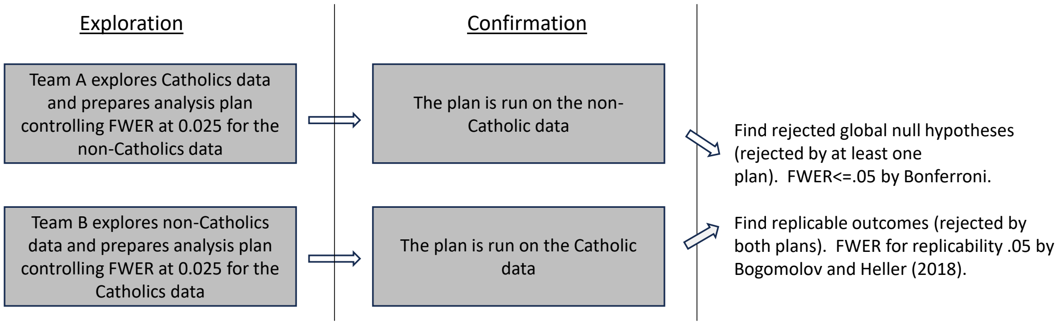

Exploration
Confirmation
Team A explores Catholics data and prepares analysis plan controlling FWER at 0.025 for the non-Catholics data
The plan is run on the non-Catholic data
Team B explores non-Catholics data and prepares analysis plan controlling FWER at 0.025 for the Catholics data
The plan is run on the Catholic data
Find rejected global null hypotheses (rejected by at least one plan). FWER<=.05 by Bonferroni.
Find replicable outcomes (rejected by both plans). FWER for replicability .05 by Bogomolov and Heller (2018).

Figure 1.1: Our proposed two team cross-screening approach: each team explores their part and prepares an analysis plan that will be run on the other part of the data.

Typically, exploratory data analysis, confirmatory data analysis and replication are performed on separate studies. However, for settings where randomized experiments are impossible to conduct for ethical or other reasons, and observational studies must be relied on, there is sometimes a single data set available with unique strengths to perform the study. In such cases, it is valuable to be able to perform exploratory analysis, confirmatory analysis and replication on that single data set.

In this paper, we use the WLS which has unique strengths for studying the long term effects of unwanted pregnancy on mothers as discussed in Section 1.1, and develop an approach called two team cross-screening to perform exploratory data analysis, confirmatory data analysis and replication using this one data set. As displayed in Figure 1.1, we consider two subgroups of women in the WLS data – Catholics and non-Catholics – and then divide ourselves into two independent teams, each having access to data on only one subgroup, either Catholics or non-Catholics. The first team, say team A, performs exploratory analyses on the Catholic data to plan the analyses which will be performed on the non-Catholic data. On the other hand, team B explores the non-Catholic data to design the analyses that will be performed on the Catholic data. Each team constructs a testing plan that has familywise error rate (FWER) control at level .025 so that by Bonferroni, if we reject any hypothesis that was tested by either team, the overall FWER is controlled at .05 (Zhao et al., 2018). Such rejected hypotheses are called the rejected global null hypotheses. If the same hypothesis is tested and rejected on both Catholic and non-Catholic data, then it is a replicable finding, and the probability to make at least one false replicability claim is at most .05 (Bogomolov and Heller, 2018). As discussed in Section 1.2, the Catholic and non-Catholic women may have had unwanted pregnancies for somewhat different reasons, and thus replicable evidence for both Catholics and non-Catholics experiencing an effect strengthens the case for it being a causal effect rather than just a spurious association (Rosenbaum, 2001, 2015).

Why do we need two teams – couldn't we achieve exploration, confirmation and replication with just one team? One team automated cross-screening (Zhao et al., 2018) is a method involving one team in which the team divides the data into two parts, I and II, and prespecifies, without exploring the data, an automatic way in which part I will be examined to choose a .025 FWER controlling plan for testing hypotheses in part II of the data and an automatic way in which part II will be examined to choose a .025 FWER controlling plan for testing hypotheses in part I of the data. For example, the automatic way to choose the plan for part II of the data based on part I of the data might be to examine all prespecified outcomes and test the null hypothesis of no treatment effect for each outcome and if M of the outcomes give a p-value less than .025 in part I of the data, test those M outcomes at level .025/M on part II of the data. For valid FWER control, this one team cross-screening approach requires that the way in which part II of the data will be analyzed to choose an analysis for part I be automatic, i.e., specified before looking at the data. One team cross-screening cannot accommodate exploring both parts of the data. The reason is that if we used EDA on part I of the data to choose hypotheses on part II of the data and then went to do EDA on part II of the data to chooses hypotheses on part I of the data, our EDA on part II would be biased by already knowing what the EDA on part I of the data yielded. This bias may result in an inflated FWER for the global null hypotheses as well as for replicability findings. Having two independent teams prevents this bias, enabling unbiased exploration on both parts of the data and FWER control.

The remainder of the paper is organized as follows. Section 2 summarizes relevant work in the existing literature. In Sections 3 and 4, we provide a brief description of the exposure, covariates, matching procedure, and the pre-analysis discussion that we had before splitting into two teams. Section 5 provides detailed description of the EDA performed and analysis plans prepared by team A and team B and Section 6 summarizes the findings of applying our method to data. In Section 7, we compare the results from our method to the ones obtained by some other existing approaches. Section 8 discusses how to use the idea of two team cross-screening to obtain confidence intervals for effect sizes. We conclude with a discussion in Section 9.

2 Literature review

There has been work that combines some but not all of exploration, confirmation and replication in the same observational study.

EDA but no replicability: In a single split sample design (Cox, 1975; Heller et al., 2009), the data is randomly split into a smaller planning sample and a larger analysis sample, where the planning sample is used to plan the analysis. EDA can be used on the planning sample and the analysis sample allows for confirmatory data analysis, but this approach does not offer the opportunity for replication.

Replicability but no EDA: In one team automated cross-screening (Zhao et al., 2018), as mentioned in Section 1.3, the data is split into two parts and each part is used to propose an analysis for the other part like in two team cross-screening. However, as discussed in Section 1.3, when there is only one team of investigators, in order to guarantee FWER control, the investigators need to decide before seeing the data how they will use one part of the data to design the analysis for the other part, and thus the analysis needs to be automated. Another approach that allows for replicability but not EDA is evidence factors (Rosenbaum, 2015, 2017; Karmakar et al., 2019) which are statistically independent tests of a treatment effect in the same observational study where the potential biases of the tests are different. The tests can be based on two separate pieces of the data, like Catholics and non-Catholics, or different aspects of the whole data. If the tests concur in rejecting a null hypothesis of no treatment effect after accounting for multiplicity, then the finding is replicated. The tests are specified in advance of examining the data, so evidence factors allow for replicability but not EDA.

Unlike both one team automated cross-screening and evidence factors, our two team cross-screening approach allows the investigators to perform EDA prior to making the analysis plan, which can help in producing efficient designs by spending α in a sensible way. Also the EDA in two team cross-screening can generate new unanticipated hypotheses in addition to the prefixed set of hypotheses that the investigators initially planned to test.

In observational studies, besides uncertainty due to potential hidden bias, another source of uncertainty is due to how different analytical teams approach an analysis – a finding could be more the result of how a particular team approaches the analysis than something strongly indicated by the data. For example, Breznau et al. (2022) found that when different teams examined the same social science question using the same data set, results varied greatly. Our two team cross-screening reduces such analyst uncertainty when a finding is replicated since the replication means two different analytical teams found the same thing. A related method for addressing analyst uncertainty is Yu et al. (2011) in which independent teams analyze portions of a data set. Unlike Yu et al. (2011), in the two team cross-screening approach presented in this paper, the two teams have to agree on a covariate to use to split the data and thus the analyses are not completely independent. One could however adapt the idea of Yu et al. (2011) and have multiple pairs of teams do two team cross-screening and then combine the results as in Yu et al. (2011).

3 Controlling for Confounders: Risk-Set Matching

In an observational study, a direct comparison of treated to control subjects may be biased because of confounders, i.e., pre-treatment differences between the treatment and control groups that affect the outcome. One way to reduce potential bias from potential confounders is to match treated subjects to control subjects on observed potential confounders, compare treated to control subjects within matched sets and then aggregate these comparisons (Rosenbaum, 2020b). If all potential confounders have been matched on, then the matched observational study can be analyzed like a stratified randomized experiment (Rosenbaum, 2020b). In usual matching, treatment occurs at one time and treated subjects are matched to control subjects on all observed potential confounders at the time of treatment. However, in our study, unwanted pregnancy (which we designate as the treatment) can occur at different times in a woman's life. Usual matching can be biased when treatment occurs at different times and treatment can affect future values of confounding variables. Suppose we want to match on education and use usual matching to match a woman who had an unwanted pregnancy at age 18 to a woman who did not have an unwanted pregnancy. The woman who had an unwanted pregnancy at age 18 might have not gone to college because of the unwanted pregnancy and so her education is not comparable to a woman who did not have an unwanted pregnancy. Here education is partially a post-treatment variable and matching on it can cause bias (Rosenbaum, 1984). To avoid such bias, we employ risk-set matching which matches sequentially over time only on the potential confounders that have been observed up to the time of treatment (Lu and Greevy, 2023). In risk-set matching, we would match a woman who became pregnant at age 18 to a woman who had not yet become pregnant by age 18 on education up to age 18, the time of the pregnancy. In addition to education up the time of the pregnancy, we

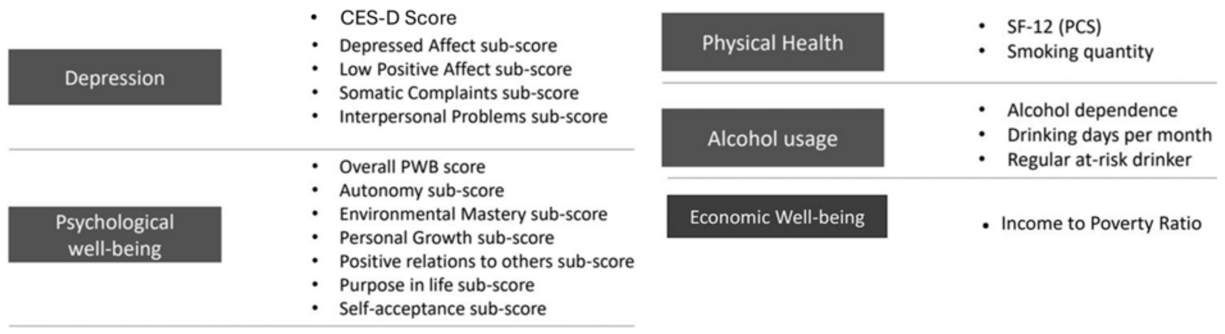

Depression
CES-D Score
Depressed Affect sub-score
Low Positive Affect sub-score
Somatic Complaints sub-score
Interpersonal Problems sub-score
Psychological well-being
Overall PWB score
Autonomy sub-score
Environmental Mastery sub-score
Personal Growth sub-score
Positive relations to others sub-score
Purpose in life sub-score
Self-acceptance sub-score
Physical Health
SF-12 (PCS)
Smoking quantity
Alcohol usage
Alcohol dependence
Drinking days per month
Regular at-risk drinker
Economic Well-being
Income to Poverty Ratio

Figure 4.1: List of outcomes decided during pre-analysis discussion.

match on a woman's sociopsychological characteristics up to the time of the pregnancy and her childhood socioeconomic status. Section A of Supplementary Material-1 provides a detailed description of the variables we match on and the steps of the risk-set matching that produced 325 matched pairs for the Catholics and 383 matched pairs for the non-Catholics. Figure D.4 in Supplementary Material-1 depicts the Love plots of the absolute values of the pre-matching and post-matching standardized differences (that is, average difference between the treated units and matched controls in estimated standard deviation units) for both the Catholics and Non-Catholics subgroups. The absolute values of the post-matching standardized differences were all less than 0.2, which is considered to indicate acceptable balance of the covariates across the treated and matched control groups (Silber et al., 2001).

4 Pre-Analysis Discussion Before Splitting into Teams

Before splitting into two teams, we met to discuss our thinking on outcomes that might be of our interest. This enhances the chances of replicability by increasing the chance that both teams would consider at least some of the same outcomes for inclusion in their analysis plans. We decided to focus on five aspects of later-life outcomes – depression (a measure based on CES-D score and its sub-scale scores, Orme et al. (1986)), psychological well-being, physical health, alcohol usage, and economic well-being. For each of these aspects, we considered several outcomes as summarized in Figure 4.1 (see Section A of Supplementary Material-1 for detailed descriptions of these outcomes). However, as discussed in Sections 2 and 5, using our method, the investigating teams will be allowed to perform EDA prior to preparing the analysis plan, which could generate some new outcomes of interest besides the ones shown in Figure 4.1 that were identified in this pre-analysis discussion (the WLS contains thousands of outcomes). During this pre-analysis discussion, we also decided on a plan for how to use different existing approaches (see Section C of Supplementary Material-1) to which we compare our method later in Section 7.

5 Two Team Cross-Screening: EDA and Proposed Data Analyses Plans

We divided ourselves into two independent teams. Each team contained two statisticians and one maternal health researcher. Team A analyzed the Catholic data to plan the analysis for the non-Catholic data. Team B analyzed the non-Catholic data to plan the analysis for the Catholic data. To maintain the overall FWER control at $\alpha = 0.05$, we split the whole α equally between the two teams, each having an allowance of spending 0.025 in total. Below we provide a detailed description of the analysis plans prepared by both teams.

5.1 Analysis planned by team A

In this section, we, team A, explored the data on the Catholic subgroup to plan the analysis for the non-Catholic data. We first examined the variables in Figure 4.1 when the women were approximately 53 years old. As depicted by the boxplot in Figure D.2 and summarized in Table D.1 in Supplementary Material-1, the depression score turned out to be significantly higher among the treated group (the women with an unwanted pregnancy) with p-value 0.017 (one-sided Wilcoxon signed rank test). We considered the sub-scales of the depression score as well in addition to the overall score, but decided based on the p-values and interpretability, that it would be best to just focus on the overall score in our analysis plan. For psychological well being, the self-acceptance score was significantly lower among the treated group (see Table D.2 and Figure D.2 in the Supplementary Material-1), and we decided to include self-acceptance score in our analysis plan. We investigated depression and self-acceptance at ages 65 and 72 and found the same direction of effects as at age 53, but decided based on the p-values and for clarity of focus, just to test depression and self-acceptance at age 53 in our analysis plan. For physical health, we found that the mean SF-12 PCS score for the treated and matched control groups were quite similar, 48.01 and 48.37 respectively, and the corresponding p-value was 0.53 (one-sided Wilcoxon signed rank test). Hence we decided not to include the testing of SF-12 score in our plan. For smoking quantity, defined by the mean number of packs of cigarette the woman usually smokes (or, used to smoke) per day, the corresponding p-value was 0.22, and hence we did not find any evidence of differential behaviour in smoking, and the corresponding test was not included in our plan. For alcohol usage too, none of the three variables, namely, drinking days, regular ‘at risk’ drinker and possible alcohol dependence (see Section A of Supplementary Material-1 for the definitions), showed any significant difference across the treated and matched control groups.

For economic well-being, we initially found significantly higher income to poverty level ratio (see Section A of Supplementary Material-1 for the definition) among the treated group. We also found significantly higher total personal income and total household income (in last 12 months prior to the moment when they answered the questionnaire) among the treated group. To have a better understanding, we considered the following potential mediators: a) Social participation, that is the level of involvement with social organizations such as church connected groups, lodges, sports teams, neighborhood improvement organizations and so on (number of organizations in which the woman had at least some involvement ranged from 0 to 14), b) Number of job spells, this is a measure of the job stability of the woman, c) Number of divorces or separations, d) Additional years of education after the index time (time at which treated woman had first unwanted pregnancy), and e) Additional number of children from further unwanted pregnancies after the index time. Among these five variables, we found that, unwanted pregnancy significantly decreased job stability (that is, more job spells), increased additional children from further unwanted pregnancies, and increased number of divorces.

The examination of potential mediators of the unwanted pregnancy group having higher income than matched controls produced interesting findings, but we remained surprised that income could be higher for the unwanted pregnancy group. We could see how decreased job stability, increased additional children from further unwanted pregnancies and increased divorces could lower income but not how it could increase income. We decided to further investigate the different components of income which revealed that the two key components, namely (i) wages and (ii) earnings from own business, had no significant difference between the two groups. However, the group with unwanted pregnancies had significantly higher pensions, annuities and survivor's benefits as compared to the other group. This difference in income only due to pensions, annuities and survivor's benefits was not interpretable to us and we decided not to focus on it. Thus, though we initially found the overall income difference was statistically significant, later, a more careful data exploration revealed that the overall income difference was not actually meaningful to us. Hence, as highlighted in Section 2, this was an instance where EDA prevented us from spending α on a hypothesis that is not worthwhile. However, our investigations of possible mediators of income difference led us to additional hypotheses of interest about unwanted pregnancy's effects, that we did not plan initially. This is an advantage of looking deeper into the data that lets one's mind wander creatively (MacEachern and Van Zandt, 2019).

In addition to exploring different hypotheses, cross-screening gives us a chance to select different test statistics for the chosen hypotheses to achieve better power. To that end, as employed in Zhao et al. (2018), we explored the family of U-statistics introduced in Rosenbaum (2011) and formally defined in Section B of the Supplementary Material-1. Specifically, we explored U statistic with three sets of values for m, $\underline{m}$, and $\overline{m}$ as $(8, 5, 8)$, $(8, 6, 7)$ and $(8, 7, 8)$, along with the Wilcoxon signed rank test. For each of the hypotheses, we selected that test statistic for which the corresponding p-value was minimum (see Table D.6 in Supplementary Material-1).

To control for the multiple testing of the five hypotheses listed in Table D.6 and put priority on the most important hypotheses, we followed the idea of serial gatekeeping procedures in which the hypotheses were divided into families and the priority families served as 'gatekeepers' for subsequent families in the sense that all hypotheses in the current family must be rejected to start testing the hypotheses in the next family (Dmitrienko and Tamhane, 2007). After considering the importance of different hypotheses and their chances of leading to significant findings, we decided on the following analysis plan which is shown as a flowchart in Figure 5.1.

(i) Start with testing whether depression score is higher among the women giving births from unwanted pregnancies, using U-statistic with $(m, \underline{m}, \overline{m})$ as $(8, 6, 7)$ with $\alpha =$

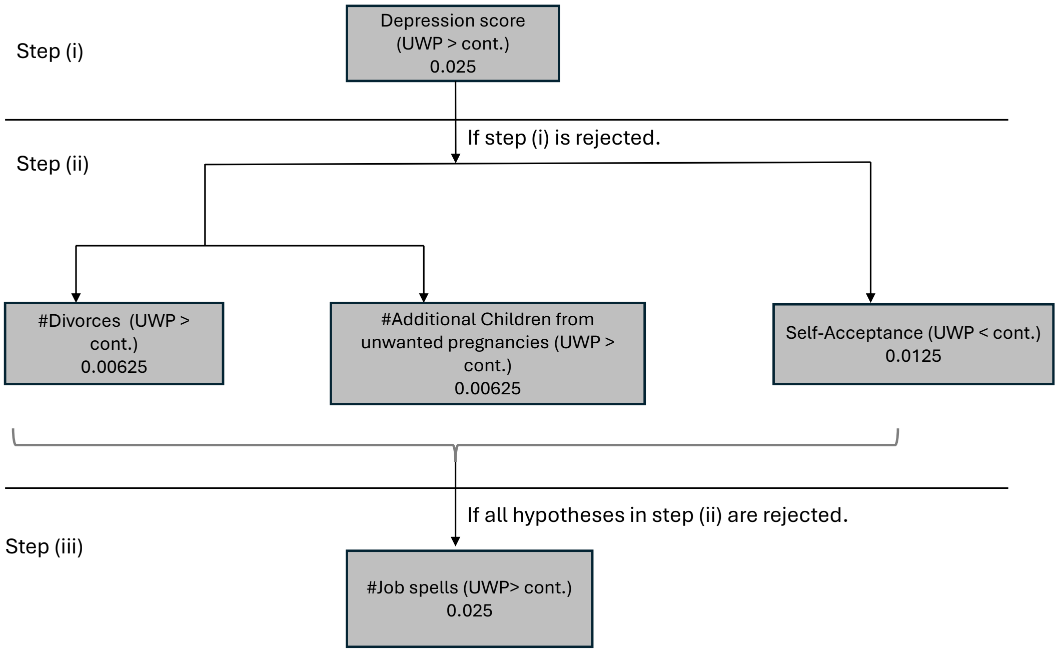


Figure 5.1: Analysis plan prepared by Team A based on the Catholic data, which will be performed on the non-Catholic data: at each step, the hypothesis (or the family of hypotheses) is depicted along with the direction and the corresponding amount of α that will be spent for the testing of that hypothesis.

0.025 (see step (i) of Figure 5.1). We chose to put the highest priority on depression because of its clinical importance.

(ii) If the above test at step (i) is rejected, carry the whole $\alpha = 0.025$ forward to the next step, and perform the following multiple testing procedure. In the first path, use the first half of α , that is 0.0125, to test whether the self-acceptance sub-scale score is significantly lower in the treated group using U-statistics with $(m, \underline{m}, \overline{m})$ as $(8, 5, 8)$.

In the second path, further split the corresponding α, that is 0.0125, into two equal parts and test in parallel the following hypotheses with the Wilcoxon signed rank test: (1) whether the women with unwanted pregnancy have significantly higher number of additional children from further unwanted pregnancies after the index time, and

(2) whether the women with unwanted pregnancies have significantly higher number of divorces or separations after the index time.

(iii) If all three hypotheses in step (ii) are rejected, then proceed further to test whether the number of job spells is significantly higher among the mothers who gave births to unwanted pregnancies with $\alpha = 0.025$ (see Step (iii) in Figure 5.1). For this testing, use U-statistic with $(m, \underline{m}, \overline{m})$ as $(8, 7, 8)$.

5.2 Analysis planned by team B

In this section, we, team B, analyze the non-Catholic data, in order to design the analysis for the Catholic data. Our first step was to examine the scores and sub-scale scores depicted in Figure 4.1. We used one-sided tests for each outcome assuming no hidden bias (that is, sensitivity parameter $\Gamma = 1$ in Rosenbaum (2007)), as well as with the arguably more reasonable assumption that there was bias due to nonrandom treatment assignment up to $\Gamma = 1.2$. The parameter Γ quantifies the maximum ratio of the odds of having an unwanted pregnancy between two women with similar observed covariates at the time when one of them experienced an unwanted pregnancy. Formally, for women i and j with the same observed covariates X, the relationship between probabilities of receiving treatment, π_i and π_j for women i and j respectively, are bounded by

$$\frac{1}{\Gamma} \leq \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} \leq \Gamma \tag{1}$$

$\Gamma = 1$ means that there is no unmeasured confounding and $\Gamma = 1.2$ means that because of an unmeasured confounder, the odds of having an unwanted pregnancy could increase 20% among women in the same matched pair. We decided to incorporate $\Gamma = 1.2$ into our proposed analysis in addition to $\Gamma = 1$ in order to allow for the possibility of moderate bias from unmeasured confounders. This particular choice of $\Gamma = 1.2$ was inspired from relevant literature (Rosenbaum, 2007; Imbens and Rubin, 2015; Ding and VanderWeele, 2016; Zhao and Small, 2018).

The tests were run using the function "*senmv*" from R package "*sensitivitymv*" available in CRAN (Rosenbaum (2007)). Specifically, we used the permutation t-test statistic "t", Wilcoxon's signed rank test statistic "W", as well as a test that ignores absolute pair differences smaller than half the median, "i". The p-values for one-sided hypotheses in the expected direction of effect for the questionnaires are reported in Table D.3 of Supplementary Material-1. The smoking quantity, alcohol usage, and economic well-being were also examined, and turned out to be all non-significant with large p-values (omitted for brevity). The only outcomes with a p-value at most 0.025 at $\Gamma = 1.2$ were depression score, the low-positive affect sub-scale score, and the interpersonal problems sub-scale score. The low-positive affect sub-scale score stood out since it had the strongest evidence, by a great margin, over all other outcomes, with all the considered tests. The low-positive affect sub-scale score was the only one with positive phrasing of questions in the depression score: "I felt as good as other people; I was happy; I enjoyed life"; see Table D.4 in Supplementary Material-1. So from our analysis, it appeared that there was strong evidence that non-Catholic women who had an unwanted pregnancy agreed significantly less with these statements than their paired controls. Since difference in a sub-scale score implies difference in the aggregate depression score, and since the evidence of this sub-scale score was by far greater than all other sub-scale scores as well as the aggregate depression score, this was the primary outcome we considered testing. Concluding that the treatment had an effect on the sub-scale score implies that there is an effect in the depression assessment, and moreover it provides a greater resolution of in which dimension of depression the effect lies.

The exploratory stage allowed us to examine more thoroughly our decision to concentrate on the low-positive affect sub-scale score. Our next step was to repeat the p-value calculations for all the outcomes on important subgroups, as well as using different methods to impute missing data. Specifically, we considered only those matched pairs that, at the index date, were at most 2 years apart in age, or had the exact same number of years of education, or had the exact same number of prior children. Moreover, we considered the answers from survey year 2003 in the WLS database (when women were age 65) whenever it was missing in the survey year 1992 (when women were age 53). The qualitative conclusions remain unchanged. The quantitative conclusions were similar to the original analysis, so we finalized our decision to test the null hypothesis of no low-positive sub-scale score effect using the permutation t-test on all pairs, without imputing missing data. Table D.3 in Supplementary Material-1 shows that the p-value using this test on the non-Catholics is 0.0003 with $\Gamma = 1$ and 0.0082 with $\Gamma = 1.2$.

Next, we searched for possible effect modification with the covariates considered for the risk set matching in Figure D.4 of the Supplementary Material-1. Our aim was to identify the covariates that are associated with the difference in low-positive affect sub-scale score between the women who had an unwanted pregnancy and their paired controls. We discovered moderate evidence that the aforementioned difference varied across the age when unwanted pregnancy occurred. As anticipated, the greatest effect appeared when the unwanted pregnancy occurred early in life, see Figure D.3 in Supplementary Material-1. We tested the null hypothesis that the effect of unwanted pregnancy on the low-positive affect sub-scale for women aged below their mid-twenties was higher than the effect for women aged above their mid-twenties using a one-sided Wilcoxon rank sum test and p-value was

0.011. We decided to include this test in our analysis plan for the Catholics.

Based on the above exploratory data analysis on non-Catholic data, our suggested analysis plan for the Catholics data is to test the following hypotheses in order (Rosenbaum, 2008), each at leavel 0.025:

(1) Test the null hypothesis that the low-positive affect sub-scale score distribution is the same for women that had and did not have an unwanted pregnancy, using a one-sided permutation t-test.

(2) If (1) is rejected, test the same hypothesis at $\Gamma = 1.2$, using the one-sided permutation t-test.

(3) If (2) is rejected, test the null hypothesis of no effect modification by age, using the one-sided Wilcoxon rank sum test.

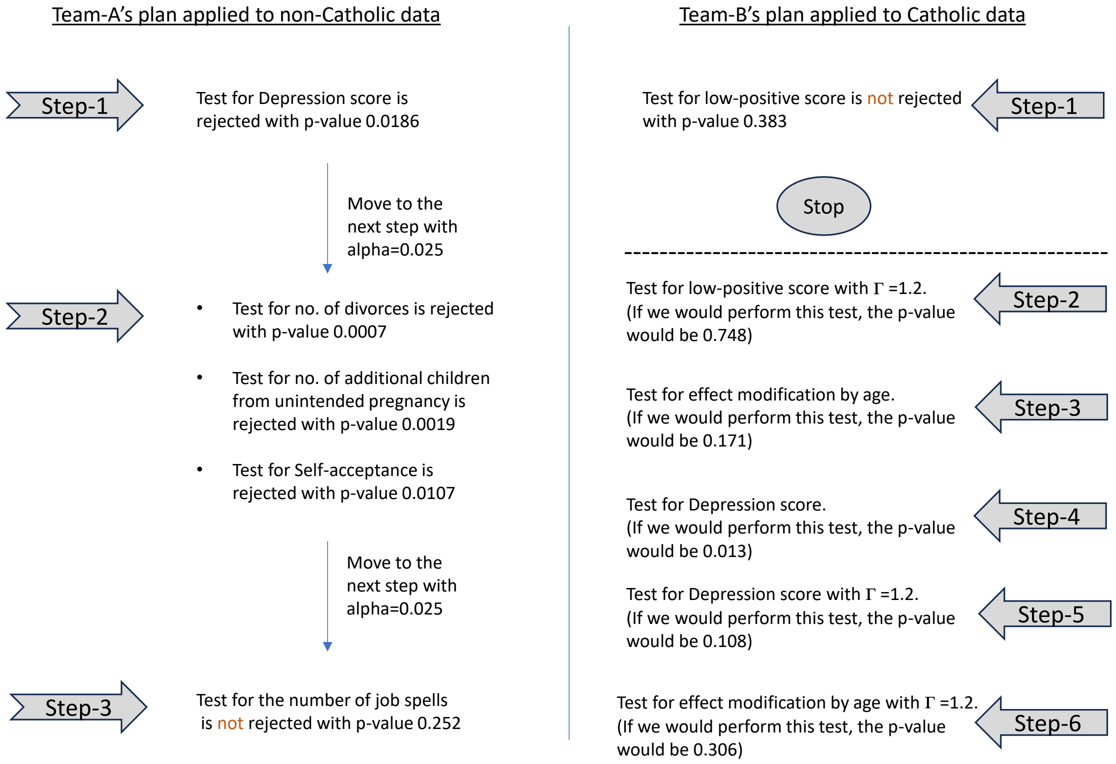
(4) If (3) is rejected, test the null hypothesis that the depression score distribution is the same for women that had and did not have an unwanted pregnancy, using a one-sided permutation t-test.

(5) If (4) is rejected, test the same hypothesis at $\Gamma = 1.2$, using the one-sided permutation t-test.

(6) If (5) is rejected, test the hypothesis in (3) at $\Gamma = 1.2$, using the one-sided Wilcoxon rank sum test.

We remark that even though the tests for low-positive affect sub-scale score in (1) and (2) can also serve to test for the aggregate depression score, we still suggested performing the latter tests in items (4) and (5), because their result may be of interest to investigators that are used to reporting of the effect of depression using the aggregate score. Moreover, we put (6) last, since at $\Gamma = 1.2$ it was clear that we could not establish the effect modification by age among the non-Catholics that we examined in our EDA, and so we did not want to waste the α of previous hypotheses on this hypothesis. On the other hand, if the testing of this hypothesis is reached for the Catholics, and it is rejected, then it will be possible to state that there is strong evidence towards effect modification by age in the Catholic subgroup, which is interesting.

6 Results of Two Team Cross-Screening



Team-A's plan applied to non-Catholic data

Test for Depression score is rejected with p-value 0.0186

Move to the next step with alpha=0.025

- Test for no. of divorces is rejected with p-value 0.0007
- Test for no. of additional children from unintended pregnancy is rejected with p-value 0.0019
- Test for Self-acceptance is rejected with p-value 0.0107

Move to the next step with alpha=0.025

Test for the number of job spells is not rejected with p-value 0.252

Team-B's plan applied to Catholic data

Test for low-positive score is not rejected with p-value 0.383

Test for low-positive score with Γ =1.2. (If we would perform this test, the p-value would be 0.748)

Test for effect modification by age. (If we would perform this test, the p-value would be 0.171)

Test for Depression score. (If we would perform this test, the p-value would be 0.013)

Test for Depression score with Γ =1.2. (If we would perform this test, the p-value would be 0.108)

Test for effect modification by age with Γ =1.2. (If we would perform this test, the p-value would be 0.306)

Figure 6.1: Results of the two team cross-screening: depression score, self-acceptance, divorces and additional number of children from unwanted pregnancy are rejected global nulls. No replicable outcomes are detected.

We applied each team's plan on the other team's data. As depicted in Figure 6.1, while applying team A's plan (see Figure 5.1) on the non-Catholics data, the test for depression score was rejected in the first step, and the testing procedure moved forward to the second step. In the second step, the test for the number of divorces, the test for number of additional children from unwanted pregnancies, and the test for self-acceptance were rejected. Thus the process moved further to the third step. However, in this step, the test for the number of job spells was not rejected. On the other hand, while applying team B's plan on the Catholics data, the test for low-positive affect score was not rejected at the first step, and the testing procedure ended there.

The above results provided some significant new discoveries on the lasting effects of unwanted pregnancies on mothers. So far, the only previous study that considered long term effects of unwanted pregnancies we are aware of was Herd et al. (2016) that found an adverse effect of unwanted pregnancies on mental depression, using the depression score. However, as indicated in Figure 6.1, our study, besides depression score, also delved into self-acceptance and found that giving births to unwanted pregnancies results in significantly less life satisfaction among mothers. This finding strengthens a similar discovery based on a much more short-term study (Biggs et al., 2014) that showed that women who were denied an abortion reported lower self-esteem and life satisfaction than women who sought and obtained an abortion. We also found that the number of divorces was significantly higher among women with unwanted pregnancies, indicating an adverse effect on the marital lives of the mothers. As argued in Logan et al. (2007), due to increased divorces, children born after unwanted pregnancies are likely to live apart from one or both of their parents, usually their father, sometime during childhood. Finally, the women with an unwanted pregnancy were found to be more likely to have further unwanted pregnancies in the future – a useful finding for policy makers in designing suitable plans and creating awareness among the mothers.

The above effects were found only using team A's plan on non-Catholic data, and the other team's plan was terminated in the first step (Figure 6.1). Hence we did not produce any replicable findings. It is not clear why the effect of unwanted pregnancy on the low-positive affect subscale was only significant among the non-Catholics and not among the Catholics. The subscale assesses general feelings of happiness and life enjoyment (e.g., "I was happy" and "I enjoyed life"). Some prior literature suggests that women with religious beliefs who experience unwanted pregnancy may have viewed it as a "blessing" or "unplanned gift" and this perspective of divine agency over family size and reproductive outcomes could moderate the impact on one's positive affect (Seeman et al., 2016). Perhaps although there was an overall effect of unwanted pregnancy on depression scores, depressive symptoms manifested differently across the Catholics and non-Catholics. This could be explored in future research.

7 Comparing the Results to Existing Approaches

We compared our findings with the ones obtained by some other relevant methods which we had chosen during our pre-analysis discussion (see Section 4). As described in Section C of Supplementary Material-1, in simulations, automated cross-screening turned out to be the most powerful competing method for replicability analysis, and 'Holm - full data' was the most powerful competing method for testing the global nulls. Hence we chose to compare our findings with the ones obtained by automated cross-screening for replicability and Holm - full data for global nulls.

Figure 7.1 depicts the findings of applying automated cross-screening for replicability to our data. At the first stage of automated analyses in each part of the data to choose hypotheses to test in the other part of the data, the tests for depression score, self-acceptance and income to poverty level ratio were rejected on the Catholics data and thus, these three

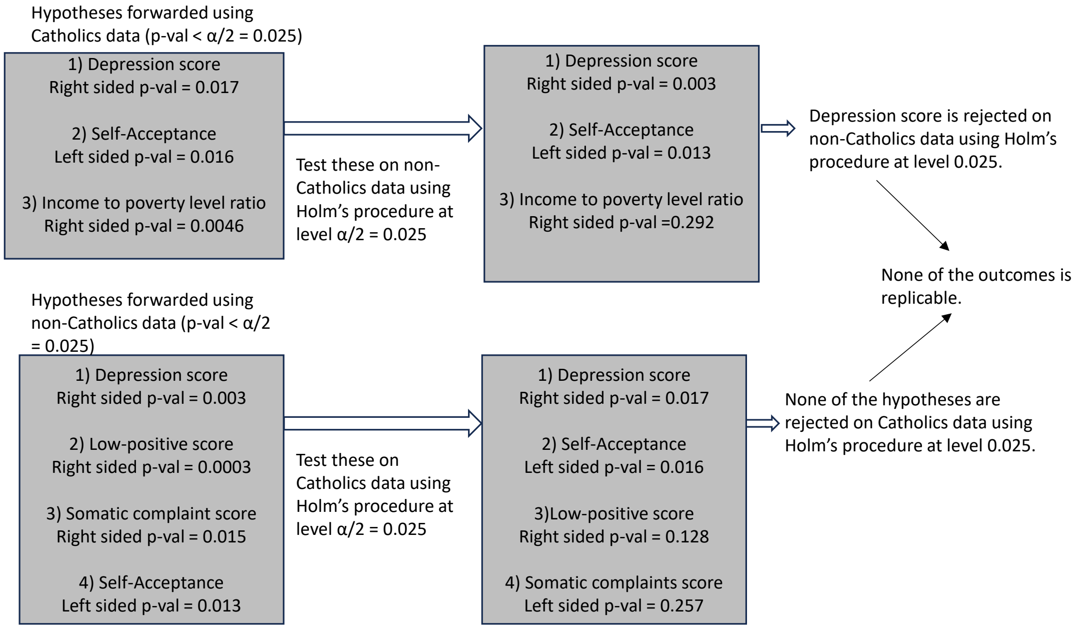

Hypotheses forwarded using
Catholics data (p-val < α/2 = 0.025)
1) Depression score
Right sided p-val = 0.017
2) Self-Acceptance
Left sided p-val = 0.016
3) Income to poverty level ratio
Right sided p-val = 0.0046
Test these on non-Catholics data using Holm's procedure at level α/2 = 0.025
1) Depression score
Right sided p-val = 0.003
2) Self-Acceptance
Left sided p-val = 0.013
3) Income to poverty level ratio
Right sided p-val =0.292
Depression score is rejected on non-Catholics data using Holm's procedure at level 0.025.
None of the outcomes is replicable.
Hypotheses forwarded using
non-Catholics data (p-val < α/2
= 0.025)
1) Depression score
Right sided p-val = 0.003
2) Low-positive score
Right sided p-val = 0.0003
3) Somatic complaint score
Right sided p-val = 0.015
4) Self-Acceptance
Left sided p-val = 0.013
Test these on Catholics data using Holm's procedure at level α/2 = 0.025
1) Depression score
Right sided p-val = 0.017
2) Self-Acceptance
Left sided p-val = 0.016
3)Low-positive score
Right sided p-val = 0.128
4) Somatic complaints score
Left sided p-val = 0.257
None of the hypotheses are rejected on Catholics data using Holm's procedure at level 0.025.

Figure 7.1: Results of automated cross-screening designs. Right-sided (or, left-sided) p-value corresponds to the alternative that the outcome is higher (or, lower) for the women with unwanted pregnancies compared to their controls.

hypotheses were tested on the non-Catholics data using Holm's procedure at level 0.025. In the second stage of testing the hypotheses chosen in the first stage from the other part of the data, depression score had a significant effect in the non-Catholic data and there were no significant effects in the Catholic data. Thus, there were no replicable outcomes.

For testing of global nulls, Holm on the full data found that depression score and low-positive subscale score were significantly higher among the women with unwanted pregnancies.

As mentioned in Section 2, one of the key advantages of our method compared to automated cross-screening and Holm-full data, which both must be prespecified before looking at the data, is that our approach allows the teams to explore their parts of the data to prepare an analysis plan for the other part of the data. This exploration created the opportunity to delve into new unanticipated hypotheses in addition to a prefixed set of outcomes. Hence, while the prespecified methods discovered an effect of unwanted pregnancy only on depression score and low-positive sub-scale score, two team cross-screening made several other discoveries, namely effects on self-acceptance, divorces and number of additional children from unwanted pregnancies (see Section 6).

Another advantage of two team cross-screening over the prespecified methods is evident from Figure 7.1. As depicted in left panel of the figure, income to poverty level ratio turned out to have a strong effect among the Catholics, and the automated method forwarded this hypothesis to be tested on the non-Catholics data. However, in our two team approach, team A, while exploring the Catholics data, found that this seemingly significant difference in the income to poverty level ratio was actually only due to pensions, annuities and survivor's benefits, which was not interpretable to us and we decided not to focus on it (see Section 5.1). Thus, our method, unlike the other methods, explored the data carefully prior to making the analysis plan and prevented the investigators from spending α on a hypothesis that was not worthwhile.

8 Two Team Cross-Screening for Effect Sizes

We have focused on using two team cross-screening for testing hypotheses. However it can also be used for confidence intervals for effect sizes. Each team can analyze its data to determine the outcomes for which to report a confidence interval on the other part of the data. If 95% confidence intervals are reported, then out of the list of all confidence intervals that are reported, we expect only 5% not to include the true value. The selection criterion the teams use might be different than for testing hypotheses. Any outcome that seems to have signal and is of scientific interest might be worth including when forming confidence intervals even if its one-sided p-value is greater than 0.025. Although we did not choose outcomes with this in mind, for illustrative purposes, Table D.5 in the Supplementary Material-1 shows 95% confidence intervals for the effect sizes on the outcomes of the non-Catholics, where the outcomes were those chosen by team A from analyzing the Catholic data, and the 95% confidence intervals for the effect sizes on the outcomes of Catholics, where the outcomes were those chosen by team B from analyzing the non-Catholic data. The confidence intervals and point estimates for the effect of unwanted pregnancy on the depression score are similar for Catholics (3.2, 95% CI: 1.0, 5.6) and non-Catholics (2.9, 95% CI: 0.1, 5.5). To provide a sense of magnitude, these effects on depression are similar to the estimated effects on depression of having a high school degree vs. a college degree or being married vs. unmarried (using the WLS data and assuming no unmeasured confounders) (Herd et al., 2016).

The expected value for the number of the 11 confidence intervals in Table D.5 that do not contain the true value is $.05 \times 11 = 0.55$. If we would like stricter control of the error rate, we could control the overall simultaneous coverage or the false coverage rate (Benjamini and Yekutieli, 2005). For example, the following would guarantee 95% overall simultaneous coverage (and more strongly at most 2.5% chance of not covering at least one outcome in each subgroup): for the n_1 outcomes specified by Team A for examining on Team B's data, form $(1 - .025/n_1)$ confidence intervals (e.g., confidence intervals based on the $(1-.0125/n_1)$ quantiles of a pivotal statistic) on Team B's data, and for the n_2 outcomes specified by Team B for examining on Team A's data, form $(1-.025/n_2)$ confidence intervals on Team A's data.

9 Discussion

We proposed a novel two team cross-screening approach that enabled us to perform exploratory data analysis, confirmatory data analysis and replication in the same observational study. Our method made some significant new discoveries on lasting effects of unwanted pregnancies that are carried to term on mothers. However, these effects were found only using team A's plan on non-Catholic data; the other team's plan was terminated in the first step and hence, our method did not find any replicable outcomes. There was an informal hint of replicability though as a significant effect on depression score was found both while team A explored the Catholic data and also while team A's plan was applied on the non-Catholic data. However, this is only an informal hint as it does not control for the multiple testing in team A's exploration of the Catholic data. Our method, unlike the existing approaches, allowed both investigating teams to perform exploratory data analysis using their parts of the data which helped in generating some new unanticipated hypotheses and prevented the investigators from spending α on an outcome that was not worthwhile.

The chance of finding replicable outcomes could be improved in future studies by having more pre-analysis discussion and using more adaptive tests. Team A was more familiar with the WLS database than team B and considered more additional outcomes beyond those decided during the pre-analysis discussion; a lengthier pre-analysis discussion might have led to more uniform choices. Both teams chose their tests for the other subgroup based on minimizing p-values on the subgroup they analyzed. It might have been better to use more adaptive tests, recognizing the role of chance in making one test look best on a subgroup as well as the fact that effect sizes might differ on the other subgroup. For example, in testing for an effect of depression, Team B put all its eggs in one basket by testing the low-positive affect subscore before the overall depression score, and the non-significant p-value of 0.383 for low-positive affect subscore meant overall depression score was never tested on the Catholic data. If team B had considered using an adaptive test with the test statistic being the minimum of the Wilcoxon test p-values for the low-positive affect subscore and overall depression score, they would have obtained a p-value for the adaptive test of 0.029, not quite but close to achieving replicability for unwanted pregnancy increasing later-life depression. A way to use information from the EDA on the other subgroup while also being adaptive is Rosenbaum (2020a)'s test that combines a planned linear combination of outcomes based on the EDA and an adaptively chosen linear combination of outcomes, correcting for multiple testing.

There can be more than one feature to split the data on. In our study, in addition to Catholics and non-Catholics having unwanted pregnancies for relatively different reasons, women who graduated from college and women who did not might have unwanted pregnancies for relatively different reasons. One option for making use of both ways to split the data is the following. Team A can explore Catholics with college education and non-Catholics without college education and specify two different analysis plans, one for Catholics without college education, and another for non-Catholics with college education, each controlling for 0.05/4 FWER. On the other hand, team B can explore Catholics without college education, and non-Catholics with college education and specify two different analyses plans – one for Catholics with college education and another for non-Catholics without college education – each controlling for 0.05/4 FWER. The FWER is .05 for a finding in the confirmatory analysis in any of the four subgroups. Another option for making use of two ways to split the data is 4-team cross-screening, where each of four teams can design a 0.05/4 analysis plan for one subgroup using the data from the remaining three subgroups. Like the first option, the FWER is .05 for a finding in the confirmatory analysis in any of the four subgroups. These methods of considering multiple splits allow one to consider whether there is replicability across multiple splits. However, the multiple splits could dilute power and we recommend sticking with just one split unless there is enough power for multiple splits. Even finding replicability across one split in a given observational study raises the standard of evidence compared to the common practice of not seeking any replicability.

While having many advantages, two-team cross-screening has some limitations. Obviously, two team cross-screening is not feasible when there is only one person working on the project. Moreover, if there is effect modification by the covariate used to split the data, then using one subgroup to design the analysis of the other subgroup may not be effective since learning the effect pattern in one subgroup may tell us little or nothing about the effect pattern in the other subgroup. If the researchers worry about effect modification, the following variant of two team cross-screening may be considered instead: split the data randomly into three parts – say 20%, 20% and 60% – and then team A will use the first 20% to design the analysis for the remaining 80% and team B will use the second 20% to design the analysis for the other 80% (first 20% and last 60%). Team A could propose tests for both subgroups and thus have a chance to achieve replicability; likewise, team B could propose tests for both subgroups. Like in the two team cross-screening approach discussed in this paper, this new variant also uses all the data for inference, but here each team uses less data for planning but more for inference. A useful topic for future research would be to compare these two different two team cross-screening approaches.

Supplementary Material

Supplementary Material-1 provides detailed description on the outcomes and simulation studies. Supplementary Material-2 contains the protocol document.

References

Bearak, J., A. Popinchalk, L. Alkema, and G. Sedgh (2018). Global, regional, and subregional trends in unintended pregnancy and its outcomes from 1990 to 2014: estimates from a bayesian hierarchical model. *The Lancet Global Health 6*(4), e380–e389.

Benjamini, Y. and D. Yekutieli (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association 100*(469), 71–81.

Biggs, M., U. D. Upadhyay, J. R. Steinberg, and D. G. Foster (2014). Does abortion reduce self-esteem and life satisfaction? *Quality of Life Research 23*, 2505–2513.

Bogomolov, M. and R. Heller (2018). Assessing replicability of findings across two studies of multiple features. *Biometrika 105*(3), 505–516.

Breznau, N., E. M. Rinke, A. Wuttke, H. H. Nguyen, M. Adem, J. Adriaans, A. Alvarez-Benjumea, H. K. Andersen, D. Auer, F. Azevedo, et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences 119*(44), e2203150119.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika 62*(2), 441–444.

Diaconis, P. (2006). Theories of data analysis: From magical thinking through classical statistics. *Exploring data tables, trends, and shapes*, 1–36.

Ding, P. and T. J. VanderWeele (2016). Sensitivity analysis without assumptions. *Epidemiology 27*(3), 368–377.

Dmitrienko, A. and A. C. Tamhane (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 6*(3), 171–180.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics 26*(4), 745–766.

Finer, L. B. and M. R. Zolna (2016). Declines in unintended pregnancy in the united states, 2008–2011. *New England journal of medicine 374*(9), 843–852.

Heller, R., P. R. Rosenbaum, and D. S. Small (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association 104*(487), 1090–1101.

Herd, P., D. Carr, and C. Roan (2014). Cohort profile: Wisconsin longitudinal study (wls). *International journal of epidemiology 43*(1), 34–41.

Herd, P., J. Higgins, K. Sicinski, and I. Merkurieva (2016). The implications of unintended pregnancies for mental health in later life. *American journal of public health 106*(3), 421–429.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Jones, R. K. and J. Dreweke (2011). *Countering conventional wisdom: New evidence on religion and contraceptive use.* Alan Guttmacher Institute New York, NY.

Karmakar, B., B. French, and D. S. Small (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika 106*(2), 353–367.

Logan, C., E. Holcombe, J. Manlove, and S. Ryan (2007). The consequences of unintended childbearing. *Washington, DC: Child Trends and National Campaign to Prevent Teen Pregnancy 28*, 142–151.

Lu, B. and R. A. Greevy (2023). Risk set matching. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pp. 169–184. Chapman and Hall/CRC.

Lund, E. and K. H. Bønaa (1993). Reduced breast cancer mortality among fishermen's wives in norway. *Cancer Causes & Control 4*, 283–287.

MacEachern, S. N. and T. Van Zandt (2019). Preregistration of modeling exercises may not be useful. *Computational Brain & Behavior 2*, 179–182.

Mark, N. D. and S. K. Cowan (2022). Do pregnancy intentions matter? a research note revisiting relationships among pregnancy, birth, and maternal outcomes. *Demography 59*(1), 37–49.

Miller, L. and M. Gur (2002). Religiousness and sexual responsibility in adolescent girls. *Journal of Adolescent Health 31*(5), 401–406.

National Academies of Sciences, E., Medicine, et al. (2019). Reproducibility and replicability in science.

Orme, J. G., J. Reis, and E. J. Herz (1986). Factorial and discriminant validity of the center for epidemiological studies depression (ces-d) scale. *Journal of clinical psychology 42*(1), 28–33.

Pakseresht, S., P. Rasekh, and E. K. Leili (2018). Physical health and maternal-fetal attachment among women: Planned versus unplanned pregnancy. *INternational Journal of Womens Health and Reproduction Sciences 6*(3), 335–41.

Postlethwaite, D., M. A. Armstrong, Y.-Y. Hung, and R. Shaber (2010). Pregnancy outcomes by pregnancy intention in a managed care setting. *Maternal and child health journal 14*(2), 227–234.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A: Statistics in Society 147*(5), 656–666.

Rosenbaum, P. R. (2001). Replicating effects and biases. *The american statistician 55*(3), 223–227.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics 63*(2), 456–464.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika 95*(1), 248–252.

Rosenbaum, P. R. (2011). A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics 67*(3), 1017–1027.

Rosenbaum, P. R. (2015). How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application 2*, 21–48.

Rosenbaum, P. R. (2017). The general structure of evidence factors in observational studies.

Rosenbaum, P. R. (2020a). Combining planned and discovered comparisons in observational studies. *Biostatistics 21*(3), 384–399.

Rosenbaum, P. R. (2020b). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application 7*(1), 143–176.

Seeman, D., I. Roushdy-Hammady, A. Hardison-Moody, W. W. Thompson, L. M. Gaydos, and C. J. R. Hogue (2016). Blessing unintended pregnancy. *Medicine Anthropology Theory 3*(1).

Silber, J. H., P. R. Rosenbaum, M. E. Trudeau, O. Even-Shoshan, W. Chen, X. Zhang, and R. E. Mosher (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Medical care*, 1048–1064.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The american statistician 34*(1), 23–25.

Yu, Q., S. N. MacEachern, and M. Peruggia (2011). Bayesian synthesis: Combining subjective analyses, with an application to ozone data.

Zhao, Q. and D. S. Small (2018). Sensitivity analysis for unmeasured confounding under a quasi-experimental design. *Journal of the American Statistical Association 113*(522), 527–540.

Zhao, Q., D. S. Small, and P. R. Rosenbaum (2018). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association 113*(523), 1070–1084.