

Causal Effect of Unpaid Caregiving on Recent Emotional Distress Among Older Adults

Evidence from the Wisconsin Longitudinal Study

Devanshi Shah Jiadong Zhu Zuchen Li

April 22, 2026

1 Introduction

In the United States, older adults frequently assume unpaid caregiving roles for spouses, family members, and friends. According to the U.S. Bureau of Labor Statistics, approximately 17 percent of caregivers are aged 65 or older, and nearly one-quarter of women over 55 provide unpaid care.⁴ A large observational literature links caregiving to increased emotional distress,^{1–3} but whether this relationship is causal remains unclear.

The central challenge is that selection into caregiving is shaped by demographic, socioeconomic, and social-network factors that may independently affect mental health. Exposure to others’ care needs and prior caregiving experience may both increase the likelihood of becoming a caregiver and be associated with emotional distress. Simple comparisons between caregivers and non-caregivers may therefore yield biased estimates. Understanding the causal impact is nevertheless important: unpaid caregiving is a substantial and growing component of the U.S. care system, with implications for public health, labor supply, and the sustainability of informal care networks. Distinguishing the effect of caregiving itself from selection is essential for informing policy and support interventions.

We estimate the effect of being in the state of current unpaid caregiving on recent emotional distress using data from the Wisconsin Longitudinal Study (WLS), focusing on Round 6 (2010–2012) respondents with covariates drawn primarily from Round 5 (2003–2006). We compare caregivers to observationally similar non-caregivers using matching on observed covariates, while recognizing that treatment and outcome are measured in the same survey wave.

2 Dataset

We use data from the Wisconsin Longitudinal Study (WLS),⁶ a long-running panel following a cohort of 1957 Wisconsin high school graduates and selected siblings. Our primary data come from the public Round 6 (R06, 2010–2012) release, supplemented with pre-treatment covariates from Round 5 (R05a, 2003–2006).

The unit of analysis is an individual respondent observed at R06. Each row includes current caregiving status, recent mental health, and demographic, socioeconomic, and baseline health characteristics. Respondents are typically in their early 70s at R06, reflecting the aging population that often assumes unpaid caregiving roles.

The **treatment** is a binary indicator for whether the respondent is currently providing unpaid personal care at R06, constructed from `z_hv600re`. Original survey coding is mapped to $\{0, 1\}$ so that $Z_i = 1$ denotes current caregiving and $Z_i = 0$ denotes no current caregiving. The **outcome** is a binary indicator of recent emotional distress, based on `z_hx333re`, which asks whether the respondent felt fretful, angry, irritable, anxious, or depressed in the past four weeks, recoded as $Y_i \in \{0, 1\}$ where $Y_i = 1$ is a “yes”. With this recoding, sample means correspond to probabilities and estimated treatment effects are interpretable as differences in probabilities (risk differences).

We include covariates drawn primarily from R05a, along with age from R06, to capture factors influencing both caregiving and recent emotional distress. Most covariates are pre-treatment; only

age is measured at R06, and rural–urban residence is proxied using the 2013 Rural–Urban Continuum Codes (RUCC). Age is effectively predetermined, while residential context is likely stable but should be interpreted cautiously as contemporaneous. Race is excluded due to limited variation in the predominantly White WLS sample.

Continuous covariates include age at R06 ($z_ha003re$), and at R05a: years of education ($z_gb103red$), number of children ($z_gd001kd$), SF-12 physical and mental health summary scores ($z_ixsflrec$, $z_ixsf2rec$), and net worth ($z_gr100rpc$), transformed using $\log(1+\max(\text{net worth}, 0))$ to reduce skewness. Categorical covariates include sex (z_sexrsp , time-invariant), marital status from R05a ($z_gc001re$), and the 2013 RUCC ($z_ha930re$). We also include a binary indicator for prior caregiving experience from R05a ($z_gv001res$).

We restrict the sample to respondents with non-missing treatment, outcome, and covariates, recoding binary variables into $\{0, 1\}$ indicators and removing invalid codes for categorical variables; covariates with no variation are dropped, though none arise in our analytic sample. Restricting on non-missingness and a common support region leaves $N = 6,585$ respondents: (1,872 caregivers and 4,713 non-caregivers). Under matching, this results in a final matched sample of $K = 1,870$ treated–control pairs. Each observation corresponds to a respondent with caregiving status, recent emotional distress, and the full set of covariates; summary statistics for this are reported in Table 1.

3 Potential Outcomes Framework

We define the treatment as an indicator for whether respondent i is currently providing unpaid caregiving at the time of the Round 6 (R06) interview. Let $Z_i \in \{0, 1\}$, where $Z_i = 1$ denotes current unpaid caregiving and $Z_i = 0$ denotes not providing such care.

The outcome of interest is whether respondent i reports feeling fretful, angry, irritable, anxious or depressed in the past four weeks, serving as a proxy for recent emotional distress. Specifically, this is a binary indicator equal to one if the respondent reports feeling fretful, angry, irritable, anxious, or depressed, and zero otherwise. Let “recent emotional distress” denote whether respondent i reports feeling fretful, angry, irritable, anxious, or depressed in the past four weeks. We formally define the potential outcomes as:

$$Y_i(1) = \mathbf{1}\{\text{respondent } i \text{ reports recent emotional distress under caregiving}\}, \text{ and}$$

$$Y_i(0) = \mathbf{1}\{\text{respondent } i \text{ reports recent emotional distress under no caregiving}\}.$$

Our estimand of interest is the average treatment effect on the treated (ATT), defined as $\mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = 1]$, that is, the effect of current unpaid caregiving on recent emotional distress among caregivers. We estimate this quantity for the matched sample.

3.1 Stable Unit Treatment Value Assumption (SUTVA)

Identification of this causal effect relies in part on the Stable Unit Treatment Value Assumption (SUTVA), which has two components: (i) no hidden versions of treatment and (ii) no interference between units, which we address with respect to our proposed design below.

No Hidden Versions of Treatment. This assumption requires that the treatment be well-defined and comparable across respondents. In this setting, however, “current unpaid caregiving” is a broad and heterogeneous treatment. Respondents may differ substantially in the number of hours of care provided, the duration of caregiving, whether they care for one or multiple individuals, whether they reside with the care recipient, whether they are the primary caregiver, the severity of

the care recipient’s needs, and the types of tasks performed (e.g., personal care, household help, coordination of services).

Because of this heterogeneity, the treatment cannot be interpreted as a single, uniform intervention. Instead, we interpret $Z_i = 1$ as placing respondents in the general state of currently providing unpaid care, encompassing a range of caregiving intensities and responsibilities. The estimand should therefore be understood as the average effect of being in this broad caregiving state, rather than the effect of a narrowly specified caregiving intervention. Implicit in this interpretation is the assumption that variation in the type and intensity of caregiving is not explicitly modeled and is therefore treated as irrelevant for defining the potential outcomes, in the sense that we do not distinguish between different “versions” of caregiving when defining $Y_i(1)$.

At the same time, this irrelevance assumption is unlikely to hold exactly. Variation in caregiving intensity and type may influence both selection into treatment and the magnitude of the outcome effect. Individuals may sort into caregiving roles based on anticipated burden, while more intensive caregiving situations may have larger effects on recent emotional distress. As a result, the binary treatment indicator aggregates over heterogeneous caregiving experiences that differ systematically across respondents. The estimated effect should therefore be interpreted as an average across these varying caregiving regimes, and may obscure meaningful differences by caregiving intensity or type. This unobserved heterogeneity represents an important limitation of the design.

No Interference Between Units. The no-interference assumption requires that one respondent’s potential outcomes depend only on their own treatment status and not on the treatment status of others. This assumption could be violated in several ways. For instance, two respondents may belong to the same household or family, and caregiving responsibilities may be shared or substituted between them. Furthermore, one individual’s caregiving burden may reduce or increase the responsibilities of another, or emotional spillovers may occur if respondents communicate their stress to friends or relatives in the sample, potentially affecting others’ psychological well-being.

Although the data do not allow for direct verification of these relationships, it is reasonable in this applied setting to assume that most respondents’ recent emotional distress outcomes are driven primarily by their own caregiving status rather than by the treatment status of others. We therefore proceed under the assumption that interference is limited, while acknowledging that spillover effects may introduce bias in the estimated treatment effect.

4 Methods

Estimation Strategy. As noted earlier, our estimand is the ATT, which we estimate for the matched sample. This captures the effect of current unpaid caregiving on recent emotional distress among respondents who are caregivers. To estimate this effect, we use one-to-one nearest-neighbor matching with replacement, matching each treated unit to a control unit based on Mahalanobis distance within a propensity score caliper. The propensity score is estimated using a logistic regression of treatment on observed pre-treatment covariates. We also considered full matching, which can improve efficiency and balance, but it requires weighting within matched subclasses, reducing transparency. Given strong overlap and a large pool of control units, one-to-one matching provides close matches for most treated units while maintaining a simple and interpretable design.

Specifically, we first estimate the propensity score $e(X_i) = \Pr(Z_i = 1 \mid X_i)$ and compute its logit transformation. We then impose a caliper equal to 0.1 times the standard deviation of the logit propensity score, restricting matches to treated–control pairs whose propensity score difference

falls within this range. Among these admissible matches, each treated unit is paired with the control unit that minimizes Mahalanobis distance computed on the covariates.

This approach combines two complementary goals. Matching on Mahalanobis distance promotes similarity across observed covariates directly, while the propensity score caliper ensures that matches occur only between units with comparable probabilities of treatment. Although the propensity score is itself a function of the covariates, imposing a caliper helps prevent poor matches in regions where covariate distributions differ and reduces reliance on extrapolation.

We estimate the ATT as the mean difference in outcomes between each treated unit and its matched control. For inference, we use McNemar’s test, which is appropriate for matched pairs with binary outcomes and relies on discordant pairs to test for systematic differences between treated and control units under the null of no treatment effect.

Why Matching. Alternative approaches are inverse probability weighting (IPW) or augmented IPW, which reweight observations using the estimated propensity score to create a pseudo-population where treatment is independent of observed covariates. In finite samples, however, IPW can be unstable: when propensity scores are close to 0 or 1, inverse weights can be large, increasing variance and yielding imprecise estimates. IPW is also sensitive to misspecification of the propensity score model; incorrect functional form can prevent covariate balance and amplify bias. Matching relies on the same ignorability assumption, but directly compares treated units to similar controls and allows us to assess balance explicitly, often yielding more stable estimates in practice.

Covariate Balance and Diagnostics. We assess balance using standardized differences in means before and after matching, using the common guideline that absolute standardized differences below 0.1 indicate good balance. Prior to matching, most covariates are reasonably balanced, although sex and prior caregiving experience exhibit moderate imbalance.

Assumptions for Identification via 1-1 Matching. We assume conditional ignorability: after controlling for demographic, socioeconomic, family-structure, baseline health, and prior caregiving variables, caregiving status is as-if random with respect to recent emotional distress. This assumption is not directly testable and represents our primary identification challenge in this application, as several unobserved factors (e.g., severity of care recipients’ conditions, family caregiving norms, availability of alternative caregivers, individual traits) may influence both caregiving and recent emotional distress but are not fully captured in the data. In particular, we do not observe detailed, time-relevant health measures for care recipients, and individuals may enter caregiving in response to unobserved shocks.

To mitigate these concerns in part, we include a rich set of pre-treatment covariates, including baseline physical and mental health (SF-12) and prior caregiving experience, which help capture underlying health and caregiving propensity. However, time-varying changes in care needs and social context remain unobserved, and available proxies from earlier waves may be too temporally distant to be informative. As a result, conditional ignorability is likely only approximately satisfied, and our estimates should be interpreted with caution.

We also assume overlap so that treated units can be meaningfully compared to control units with similar covariate values. Overlap would fail if some individuals had no realistic opportunity to become caregivers. Because our estimand is the ATT, overlap is required primarily over the covariate distribution of treated units. We assess this empirically using propensity score distributions and find that the support of the treated and control groups is highly similar (Figure 1). This in-

dicates that comparable non-caregiver units exist across the relevant range of covariates, allowing matching without substantial reliance on extrapolation.

Sensitivity to Unobserved Confounding. Because matching addresses only observed confounding, we assess robustness to hidden bias using Rosenbaum sensitivity analysis, as it evaluates how strongly an unobserved confounder would need to influence treatment assignment within matched pairs to overturn the estimated ATT. Unobserved factors affecting both selection into caregiving and recent emotional distress are a key concern. We implement the analysis using a matched-pair sign test and examine how inference changes as Γ increases, where Γ captures the strength of hidden bias. For each Γ , we focus on the upper bound, representing the worst-case p-value. If statistical significance is lost at values of Γ close to 1, even under this worst-case scenario, the results are sensitive to plausible unobserved differences between caregivers and non-caregivers.

5 Results

Matching. After matching, all absolute standardized differences fell below 0.033, well within the conventional threshold of 0.1 (Figure 2). The largest improvement was for prior caregiving status: the absolute standardized difference decreased from 0.201 pre-matching to 0.000 post-matching.

ATT Estimation. We estimate an ATT of 0.046, corresponding to a 4.6 percentage-point increase in the probability of feeling fretful, angry, irritable, anxious, or depressed during the past four weeks among caregivers, relative to comparable non-caregivers. A McNemar test without continuity correction on the $K = 1,870$ matched pairs yields $\chi^2 = 8.18$ ($p = 0.0042$), with a 95% Wald confidence interval of $[0.015, 0.077]$ for the paired risk difference. Among the 904 discordant pairs, 495 favored the treated and 409 favored the control (odds ratio $a/b = 1.21$).

Sensitivity Analysis. Table 3 reports bounds on the p -value for increasing values of Γ , which captures the degree of hidden bias in treatment assignment within matched pairs. Figure 3 plots the corresponding upper-bound p -values.

Focusing on the upper bound (i.e., the worst-case p -value), the estimated effect remains statistically significant at the 5% level up to $\Gamma = 1.075$ (upper-bound $p = 0.0451$), but loses significance by $\Gamma = 1.100$ (upper-bound $p = 0.0903$). This implies that within a matched pair, two individuals who are identical on observed covariates could differ in their odds of becoming a caregiver by a factor of up to 1.075 due to an unobserved confounder.

Analysis. While the results indicate a positive and statistically significant ATT, the estimated effect is modest, implying that caregiving increases recent emotional distress among the treated but with limited magnitude. This contrast between statistical significance and effect size is important: the estimate is precise, but its substantive impact is small. Moreover, any causal interpretation is conditional on the identifying assumptions of SUTVA, conditional ignorability, and overlap. The sensitivity analysis sharpens this point by showing that significance is maintained only up to $\Gamma = 1.075$ and is lost by $\Gamma = 1.100$, so even modest hidden bias could overturn the result. Because Rosenbaum bounds quantify the strength of unobserved confounding required to invalidate the estimate—not whether such confounding exists—the low threshold implies that the inference is not robust in environments where unobserved selection into caregiving is plausible.

The timing of measurement further complicates causal interpretation. Treatment and outcome are observed in the same wave, so the temporal ordering required for a clear potential outcomes interpretation is not directly observed. In this setting, caregiving may influence recent emotional

distress, but such distress may also influence caregiving, or both may respond to contemporaneous shocks (e.g., sudden deterioration in a family member’s health). This creates ambiguity in defining $Y(1)$ and $Y(0)$ as outcomes that follow treatment assignment, weakening the interpretation of the ATT as a causal effect. At the same time, alternative designs are not clearly superior: using earlier caregiving would estimate the effect of past caregiving rather than current caregiving, while combining caregiving across waves would redefine treatment as a mixture of histories. The chosen specification therefore aligns with the substantive estimand—current caregiving on recent emotional distress—but does so at the cost of stronger identifying assumptions.

Violations of identifying assumptions are substantively plausible. Key drivers of caregiving, such as the severity of the care recipient’s condition, the availability of alternative caregivers, and recent emotional distress just prior to caregiving, are unobserved, particularly given the gap between R05a covariates and the treatment period. If individuals enter caregiving in response to unobserved shocks that also affect recent emotional distress, then conditional ignorability fails in a direction that is difficult to diagnose from observed data. In addition, caregiving is not a single intervention but a collection of treatment versions that differ in duration, intensity, and type. This violates the no-hidden-versions component of SUTVA and implies that the ATT aggregates across heterogeneous treatment regimes. If treatment intensity is correlated with both selection and outcomes, the estimate reflects both averaging over heterogeneous effects and residual confounding.

Taken together, the modest ATT likely reflects both a small effect size and limitations in the available data that restrict identification. Unobserved confounding may attenuate or bias the estimate, while treatment heterogeneity may dilute larger effects present in specific caregiving contexts. The sensitivity of the results indicates that the estimate is better interpreted as evidence consistent with a positive effect, rather than a precise causal quantity. Matching provides a transparent comparison within regions of common support, but it cannot resolve these underlying identification challenges. Accordingly, the findings should be interpreted as suggestive, with their credibility determined less by statistical significance and more by the plausibility of the underlying assumptions.

6 Limitations, Conclusion, and Future Work

This project highlights the strengths and limits of applying causal inference methods to observational data. While matching provides a transparent, well-balanced comparison, our analysis shows that identification depends critically on assumptions, particularly conditional ignorability, that are difficult to verify in this setting. The sensitivity of our estimates to modest hidden bias, together with contemporaneous measurement of treatment and outcome, indicates that statistical significance reflects precision under the model rather than robustness to plausible violations of these assumptions. With additional resources, we could strengthen the design. Obtaining access to more detailed WLS data, particularly measures collected closer to the treatment period, would allow better control of time-varying confounders that likely drive selection into caregiving. Separating the measurement of caregiving and recent emotional distress across waves would impose a clearer temporal ordering, reducing ambiguity in defining potential outcomes and limiting concerns about reverse causality. Richer data on caregiving intensity, duration, and context would allow us to move beyond a binary treatment toward a multi-arm design capturing distinct caregiving regimes, addressing concerns about hidden treatment versions. Finally, leveraging the panel structure, such as exploiting transitions into caregiving, could help isolate variation in treatment that is less confounded by time-invariant differences across individuals, strengthening causal identification.

References

- [1] Chang, S., Zhang, Y., Jeyagurunathan, A., Lau, Y. W., Sagayadevan, V., Chong, S. A., & Subramaniam, M. (2016). Providing care to relatives with mental illness: reactions and distress among primary informal caregivers. *BMC Psychiatry*, 16(1), 80.
- [2] Gaston-Johansson, F., Lachica, E. M., Fall-Dickson, J. M., & Kennedy, J. M. (2004). Psychological distress, fatigue, burden of care, and quality of life in primary caregivers of patients with breast cancer undergoing autologous bone marrow transplantation. In *Oncology nursing forum* (Vol. 31, No. 6, pp. 1161-1169).
- [3] Fujinami, R., Sun, V., Zachariah, F., Uman, G., Grant, M., & Ferrell, B. (2015). Family caregivers' distress levels related to quality of life, burden, and preparedness. *Psycho-Oncology*, 24(1), 54-62.
- [4] U.S. Bureau of Labor Statistics (2025). Unpaid Eldercare in the United States–2023-2024 Summary. *U.S. Bureau of Labor Statistics*, <https://www.bls.gov/news.release/elcare.nr0.htm>.
- [5] Fogarty, C. Course Lecture Notes for STATS 571/671, University of Michigan, 2026.
- [6] Wisconsin Longitudinal Study. Public use data, documentation and codebooks for R05a (grad and sib) (2003–2006) and R06 (grad and sib) (2010–2012).

Figures and tables

Table 1: Analytic sample sizes for propensity-score matching.

Quantity	Value
Pre-matching analytic sample (N)	6,599
After treated common support on logit $\hat{e}(X)$ (N)	6,585
Caregivers / non-caregivers (n_1/n_0)	1,872 / 4,713
Matched pairs used for ATT estimation (K)	1,870

Table 2: Estimated ATT and matched-pair outcome summary for 4-week (recent) emotional distress.

	Value
Matched pairs	1,870
Treated (distressed)	55.4%
Control (distressed)	50.8%
Difference	4.6 percentage points
Discordant pairs: a (T+, C−)	495
Discordant pairs: b (T−, C+)	409
Odds ratio (a/b)	1.21
McNemar’s test statistic	$\chi^2 = 8.18$
p -value (two-sided)	0.0042

Table 3: Rosenbaum sensitivity analysis for the matched-pair effect of caregiving on recent emotional distress using the sign test.

Γ	Discordant Pairs	Treated Better	Treated Worse	Lower Bound p -value	Upper Bound p -value
1.000	958	523	435	0.00246	0.00246
1.025	958	523	435	0.000700	0.00754
1.050	958	523	435	0.000180	0.01980
1.075	958	523	435	0.0000422	0.04511
1.100	958	523	435	9.05×10^{-6}	0.09029
1.125	958	523	435	1.79×10^{-6}	0.16063
1.150	958	523	435	3.27×10^{-7}	0.25702

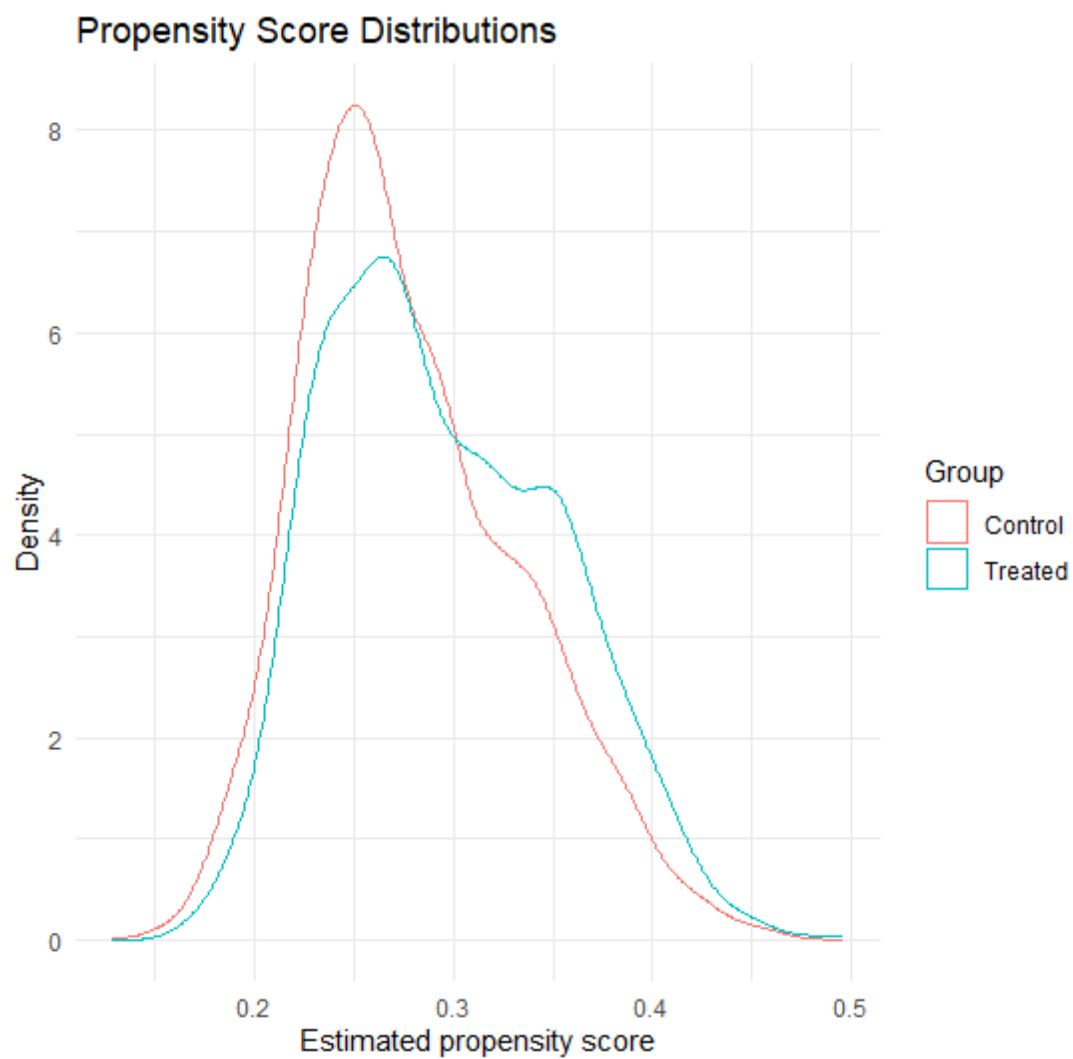


Figure 1: Estimated propensity score distributions by treatment group (pre-matching).

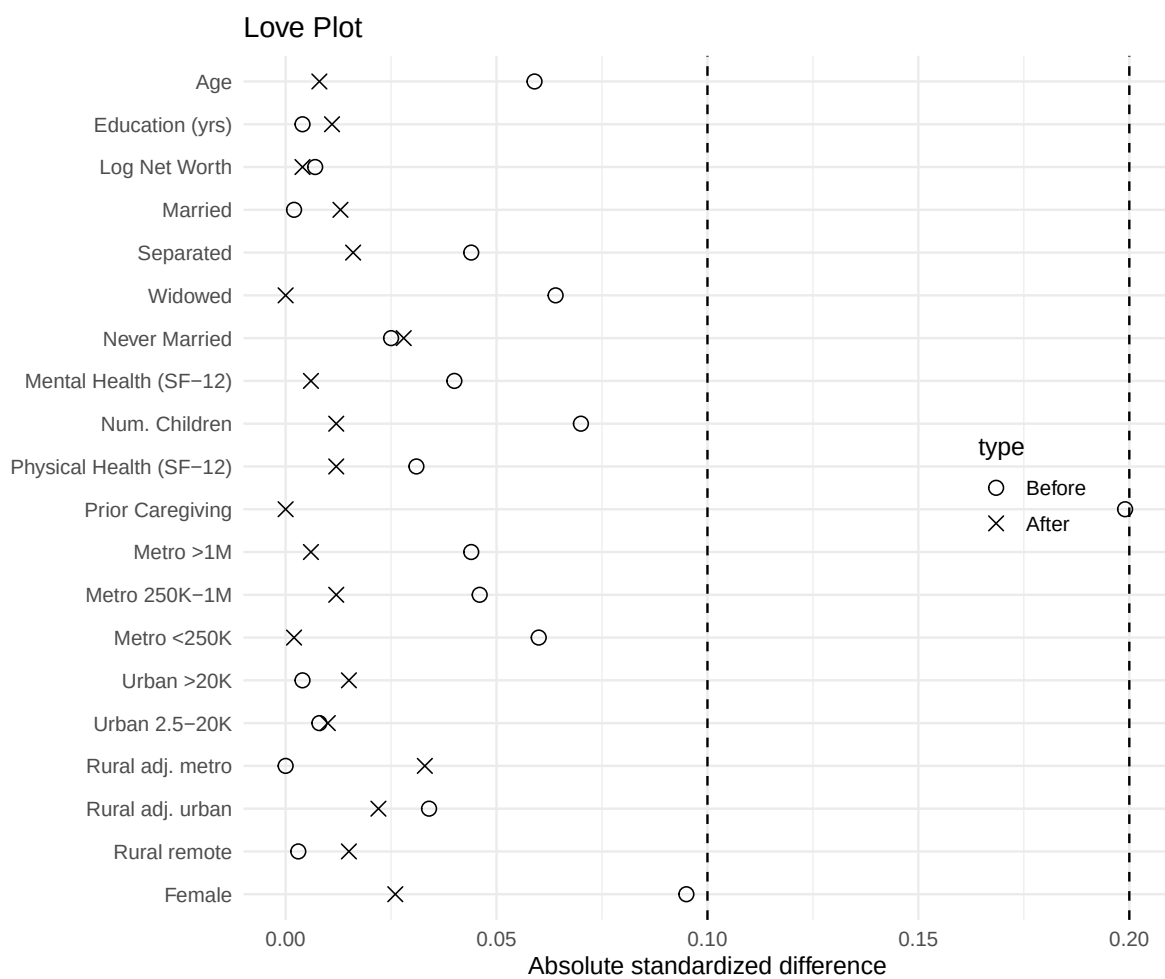


Figure 2: Covariate balance: absolute standardized differences before (circles) and after (crosses) matching. Dashed lines at 0.10 and 0.20.

Figure 2: Covariate balance: absolute standardized differences before (circles) and after (crosses) matching. Dashed lines at 0.10 and 0.20.

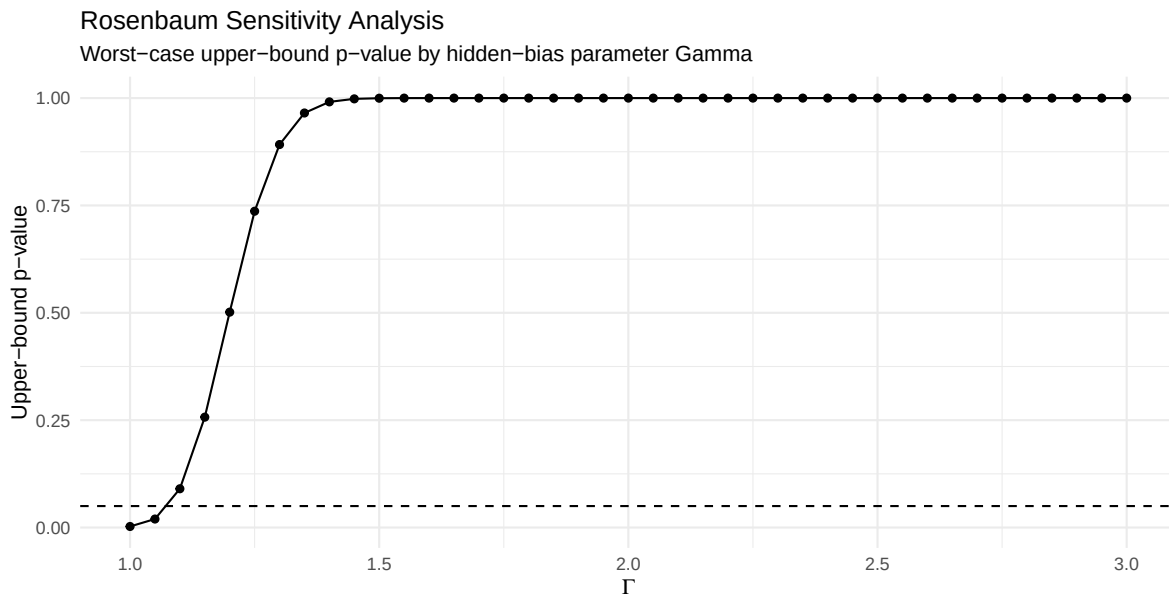


Figure 3: Sensitivity Analysis: upper-bound ("worst case") p-values plotted as level of hidden bias (Gamma) increases. Dashed lines at 0.05, indicating the maximum allowed Type I error for a test at the 5% level.