

APPENDIX

Supplementary Materials to “Variable selection in high dimensions for discrete-outcome
individualized treatment rules: A case study in reducing severity of depression symptoms”

by

EEM MOODIE, Z BIAN, J COULOMBE, Y LIAN, AY YANG, and SM SHORTREED

These Supplementary Materials contain details of the missing data procedures and results of secondary analyses including additional tailored treatment analyses that compare results of initiation with the antidepressant medication classes SSRIs and SNRIs, MOIs, mirtazapine, TCAs, and bupropion. These materials also report results on secondary analyses defining severe depression symptoms as a PHQ score of 20 or greater.

A. ADDITIONAL COMPUTATIONAL DETAILS

As an alternative to the use of the Clarke subdifferential of the SCAD penalty as in Equation (2.5), we can make a local quadratic approximation (LQA, Fan and Li, 2001; Johnson *and others*, 2008) or a local linear approximation (LLA, Zou and Li, 2008) to the penalty. As LLA has been shown to be computationally superior to LQA (Zou and Li, 2008), we only show how the LLA approximation can be incorporated into our method. The LLA approximates the penalty using

$$\rho_\lambda(\psi_j) \approx \rho_\lambda^{\text{LLA}}(\psi_j) \rho_\lambda(\tilde{\psi}_j) + \rho'_\lambda(\tilde{\psi}_j)(|\psi_j| - |\tilde{\psi}_j|),$$

where $\tilde{\psi}_j$ is an initial estimate satisfying certain conditions (Zou and Li, 2008); this leads to the following fixed-point problem with a weighted LASSO soft-thresholding operator

$$\hat{\boldsymbol{\psi}}^{\text{RALF}} = f(\hat{\boldsymbol{\psi}}^{\text{RALF}}), \text{ where } f(\boldsymbol{\psi}) = S_{\tau \rho'_\lambda(\tilde{\boldsymbol{\psi}})_\lambda}(\boldsymbol{\psi} - \tau \mathbf{U}_1(\boldsymbol{\psi})). \quad (\text{A.1})$$

Accordingly, the formulations of the REE in Equations (2.6) and (A.1) can both be very efficiently solved by a fixed-point algorithm (Yang *and others*, 2021). Since these two approaches (exact and approximation) are both computationally efficient and statistically similar, we focus on the exact approach throughout this paper. Following Zhang *and others* (2020) and Lian (2022), we incorporate a Type-I AA technique during the iterative update to accelerate the computation. Specifically, the update at the k -th iteration has the form

$$\boldsymbol{\psi}^{(k+1)} = \sum_{r=0}^{m_k} \alpha_r^{(k)} f\left(\boldsymbol{\psi}^{(k-m_k+r)}\right),$$

where $m_k = \min(m, k)$ for some maximum memory size, $m > 0$, and coefficients $\boldsymbol{\alpha}^{(k)} = (\alpha_0^{(k)}, \alpha_1^{(k)}, \dots, \alpha_{m_k}^{(k)})^\top$ sum to one. A memory size, m , ranging from 2 to 50 is found to be a reasonable choice. In this paper, we choose $m = 10$ as suggested by Zhang *and others* (2020), which shows that the influence of m is mostly on the convergence rate rather than solutions. The coefficients are designed to minimize the residuals $g(\boldsymbol{\psi}) \equiv \boldsymbol{\psi} - f(\boldsymbol{\psi})$ of the previous m_k iterations,

$$\begin{aligned} & \arg \min_{\boldsymbol{\alpha}^{(k)}, \mathbf{1}^\top \boldsymbol{\alpha}^{(k)} = 1} \left\| \sum_{r=0}^{m_k} \alpha_r^{(k)} \left(\boldsymbol{\psi}^{(k-m_k+r)} - f(\boldsymbol{\psi}^{(k-m_k+r)}) \right) \right\|_2, \\ &= \arg \min_{\boldsymbol{\alpha}^{(k)}, \mathbf{1}^\top \boldsymbol{\alpha}^{(k)} = 1} \left\| \sum_{r=0}^{m_k} \alpha_r^{(k)} \boldsymbol{g}^{(k-m_k+r)} \right\|_2. \end{aligned}$$

Zhang *and others* (2020) showed that the optimal coefficients can be acquired in closed form so the AA update is

$$\boldsymbol{\psi}^{(k+1)} = \boldsymbol{\psi}^{(k)} - \left(I + (\boldsymbol{S}_k - \boldsymbol{V}_k)(\boldsymbol{S}_k^\top \boldsymbol{V}_k)^{-1} \boldsymbol{S}_k^\top \right) \boldsymbol{g}^{(k)}, \quad (\text{A.2})$$

where

$$\begin{aligned} \boldsymbol{V}_k &= [(\boldsymbol{g}^{(k-m_k+1)} - \boldsymbol{g}^{(k-m_k)}), \dots, (\boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)})], \\ \boldsymbol{S}_k &= [(\boldsymbol{\psi}^{(k-m_k+1)} - \boldsymbol{\psi}^{(k-m_k)}), \dots, (\boldsymbol{\psi}^{(k)} - \boldsymbol{\psi}^{(k-1)})]. \end{aligned} \quad (\text{A.3})$$

We see that update Equation (A.2) avoids expensive computation of the estimating function gradient and its inverse and requires only computation of the estimating function itself, which significantly reduces the computation cost per iteration, especially in the high-dimensional case. Overall, RALF is built on a fixed-point based algorithm for solving general REEs (Yang *and*

others, 2021) that was shown to provide higher estimation accuracy, computational efficiency, and scalability (Yang *and others*, 2021; Lian, 2022) over existing REE algorithms (Fu, 2003; Johnson *and others*, 2008; Wang *and others*, 2012). As discussed in Section 2.1, when the sample size is small, it may be difficult to obtain a well-behaved initial estimator for PDR. If a well-behaved estimator cannot be identified, performance may be poor. However, with the help of RALF, a reasonable initial estimator can be obtained. Specifically, after $\hat{\psi}^{RALF}$ is estimated as in Expression (2.4), it can be plugged into the PDR minimization objective function in Expression (2.3) to further obtain the blip estimator.

B. ADDITIONAL INFORMATION ON THE KPWA DATA

Cohort construction and setting

KPWA provides care and coverage to patients in Washington state, U.S.A. Kaiser Permanente Washington Health Research Institute uses a virtual data warehouse created for research that brings together electronic health records and insurance billing information, including demographics, prescription fills, patient-reported outcomes, and health care utilization including for serious outcomes such as hospitalizations and deaths. The KPWA Institutional Review Board approved waivers of consent for use of records data in this research.

In addition to the inclusion criteria listed in the main paper, cohort construction excluded some individuals. Individuals were excluded if they had a diagnosis of personality, bipolar, or psychotic disorder in the year prior to treatment initiation. Also excluded from analyses were individuals who initiated treatment with more than one antidepressant medication, determined by extracting data on all antidepressant medications that were filled at the time of medication initiation.

Covariates, potential tailoring variables, outcome definitions

Demographic information was extracted from health records data on all patients. Age in years at treatment initiation was calculated from date of birth. At the time of the data pull, information on patient sex (male or female) most likely represented sex assigned at birth. Health records data on race and ethnicity information are usually self-reported at the time of the first outpatient medical appointment within the health system. Race and ethnicity information was combined to categorize all individuals who self-reported Hispanic ethnicity, while all other individuals were classified using the following race categories: American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/Pacific Islander, or White. Additional demographic information included insurance type (commercial, Medicaid, Medicare, or private) and information obtained from patient addresses and the 2010 Census, including neighborhood educational attainment (less than 25% college degrees), income (median lower than 40,000 USD), and level of poverty (20% of households below federal poverty level). We also scored if patients lived in an urban or rural area (1 to 6, with 1 the most urban and 6 the most rural). General medication and mental health information at the time of treatment initiation was extracted from health records, including the Charlson score, a general measure of comorbidity (Charlson *and others*, 1987), and tobacco use in the year prior. Height in inches and weight in pounds were collected from the visit closest in time to treatment initiation, looking back up to 5 years for height information and up to 2 years for weight information. We gathered information on the following mental health diagnosis in the past year: anxiety, alcohol use disorder, autism spectrum disorder, obsessive compulsive disorder, opioid use disorder, personality disorder, post-traumatic stress disorder, and sedative use disorder. In this population, mental health conditions other than anxiety were very rare, so all diagnoses other than anxiety were combined into a single indicator of a mental health or substance use disorder. We collected the number of suicide attempts and the number of psychiatric hospitalizations in the 6 months prior to treatment initiation. Additionally, we collected the number

of different antidepressants taken in the 5 years prior, if the patient had received psychotherapy in the 5 years prior, and the number of PHQ measurements recorded in the medical record in the year prior to treatment initiation. Baseline depression symptom severity was measured using the PHQ recorded closest to treatment initiation looking back up to 15 days and forward up to 3 days, to allow for data lags. All these covariates were considered potential tailoring variables and were used in the propensity score to account for potential confounding. We also added the calendar year of treatment initiation as a potential confounder in the treatment model.

Covariates considered in secondary analyses

In secondary analyses, we extended the study to other pairwise comparisons of treatments across SSRI, SNRI, monoamine oxidase inhibitors (MOIs), mirtazapine, tricyclic antidepressants (TCA), and bupropion, always with the aim of reducing the risk of severe depression symptoms. Of the 73,103 episodes, 56,876 (78%) corresponded to initiation of an SSRI, 4,056 (5.5%) to an SNRI; 22 (<1%) to an MOI; 1,747 (2%) to mirtazapine; 2,011 (3%) to a TCA; 8,330 (11%) to bupropion; and 61 (<1%) corresponding to initiation of an antidepressant in a medication class not included in any analyses

In secondary analyses, we also considered more severe symptoms using the threshold of PHQ greater than or equal to 20, along with other outcomes associated with severe depression, including self-harm, hospitalization for depression, and treatment failure, as well as remission of depression symptoms and the potential side effect of weight gain.

REFERENCES

- CHARLSON, ME, POMPEI, P, ALES, KL AND MACKENZIE, R. (1987). Charlson comorbidity index. *Journal of Crohn's and Colitis* **40**(5), 373–383.
- FAN, JIANQING AND LI, RUNZE. (2001). Variable selection via nonconcave penalized likelihood

- and its oracle properties. *Journal of the American statistical Association* **96**(456), 1348–1360.
- FU, WENJIANG J. (2003). Penalized estimating equations. *Biometrics* **59**(1), 126–132.
- JOHNSON, BRENT A, LIN, DY AND ZENG, DONGLIN. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**(482), 672–680.
- LIAN, YI. (2022). Some new computational methods in high-dimensional statistical learning in biostatistics [Ph.D. Thesis]. McGill University.
- WANG, LAN, ZHOU, JIANHUI AND QU, ANNIE. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**(2), 353–360.
- YANG, ARCHER YI, ZHAO, YU, GU, YUWEN, LIAN, YI AND FAN, JUN. (2021). Regularized estimating equations: Some new perspectives. *arXiv* **doi:10.48550/ARXIV.2110.11074**.
- ZHANG, JUNZI, O’DONOGHUE, BRENDAN AND BOYD, STEPHEN. (2020). Globally convergent type-I anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization* **30**(4), 3170–3197.
- ZOU, HUI AND LI, RUNZE. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**(4), 1509 – 1533.

Supplementary Material

C. ADDITIONAL RESULTS FROM THE KPWA ANALYSIS

Table S1: Crude risk of severe depression symptoms, average time to the first severe depression symptoms observed (in months), and average time to the last severe depression symptoms observed (in months) by drug class in the first year of follow-up, computed using Rubin's rule, KPWA health data, 2008-2018

Drug class	Crude risk PHQ		Time to PHQ		Time to the last PHQ	
	≥ 15	≥ 20	≥ 15 (SD)	≥ 20 (SD)	≥ 15 (SD)	≥ 20 (SD)
SSRI	82.3 %	43.4 %	4.4 (2.8)	5.7 (3.0)	8.4 (2.6)	7.6 (2.8)
SNRI	84.8 %	49.3 %	4.1 (2.8)	5.4 (3.0)	8.5 (2.6)	7.6 (2.9)
Mirtazapine	77.0 %	39.6 %	4.3 (2.8)	5.5 (2.9)	8.1 (2.7)	7.3 (2.8)
Bupropion	82.3 %	41.8 %	4.5 (2.9)	5.9 (3.0)	8.5 (2.6)	7.7 (2.8)
TCA	84.8 %	45.7 %	4.7 (2.9)	6.2 (2.9)	8.8 (2.4)	8.1 (2.6)
MOI	66.9 %	26.1 %	5.2 (3.1)	6.4 (2.8)	7.9 (2.9)	7.7 (2.8)

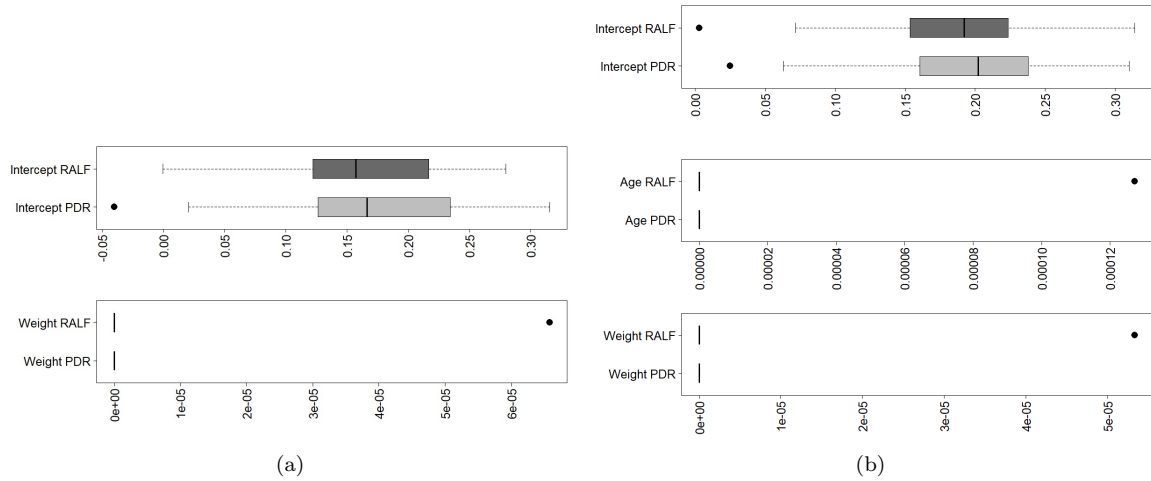


Fig. S1: Distribution across the 25 imputed datasets of the coefficients on age and weight, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SSRI and SNRI** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found).

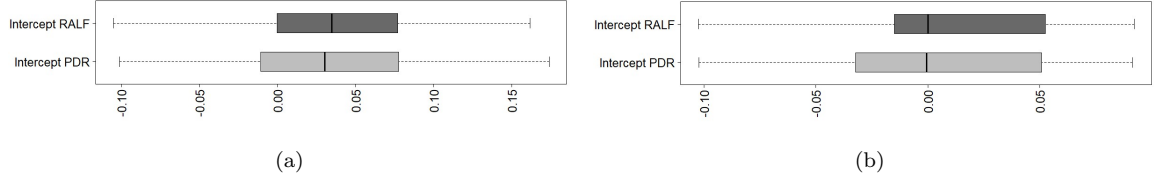


Fig. S2: Distribution across the 25 imputed datasets of the intercept coefficients, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SSRI and bupropion** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). Here, the intercept is positive, meaning that the average treatment effect of SSRI is preferable to that of bupropion and that SSRI would always be the recommended treatment.

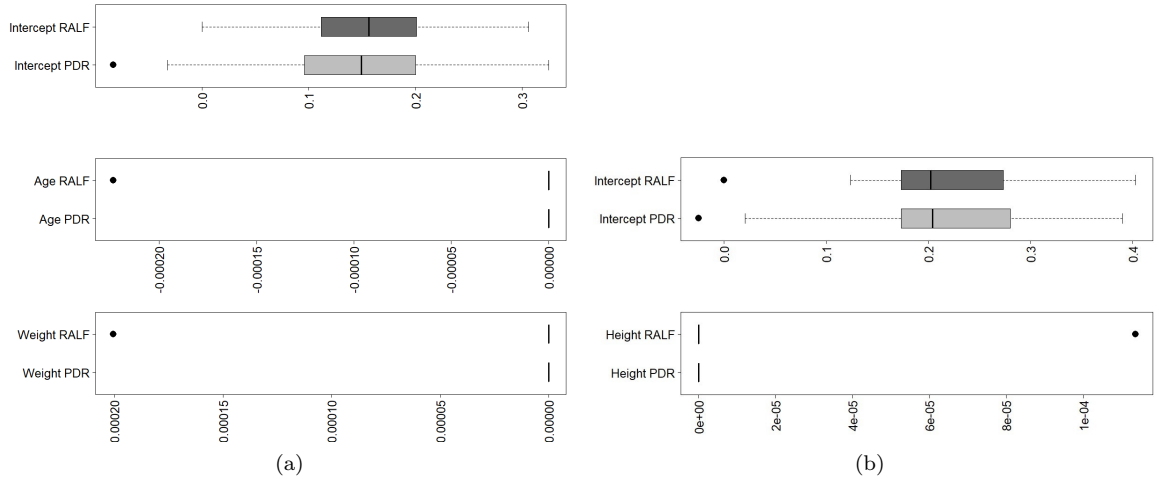


Fig. S3: Distribution across the 25 imputed datasets of the coefficients on age, weight and height, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SSRI and mirtazapine** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found).

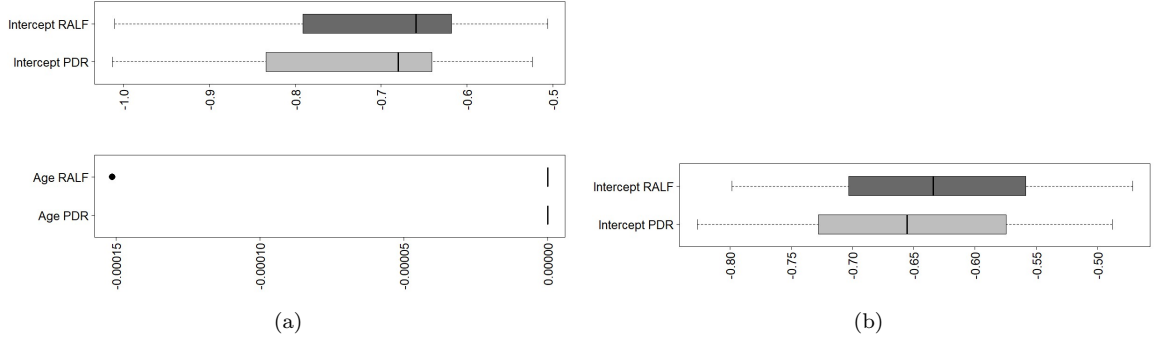


Fig. S4: Distribution across the 25 imputed datasets of the coefficients on age, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **bupropion and TCA** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). The distribution in b) contains only 19 coefficients due to lack of convergence in 6 of the imputed datasets.

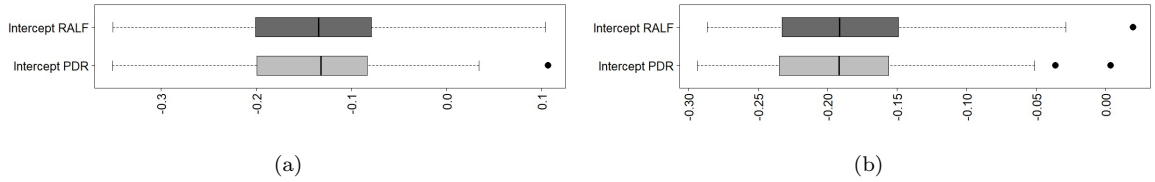


Fig. S5: Distribution across the 25 imputed datasets of the intercept coefficients, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SNRI and bupropion** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). Here, the intercept is negative, meaning that the average treatment effect of bupropion is preferable to that of SNRI and that bupropion would always be the recommended treatment.

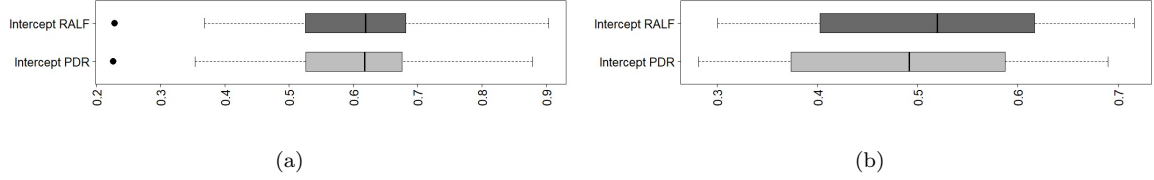


Fig. S6: Distribution across the 25 imputed datasets of the intercept coefficients, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SNRI and TCA** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). The distribution in b) contains only 20 coefficients due to lack of convergence in 5 of the imputed datasets. Here, the intercept is positive, meaning that the average treatment effect of SNRI is preferable to that of TCA and that SNRI would always be the recommended treatment.

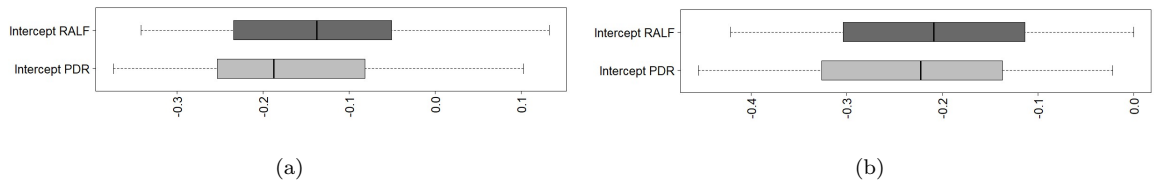


Fig. S7: Distribution across the 25 imputed datasets of the intercept coefficients, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **mirtazapine and bupropion** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). Here, the intercept is negative, meaning that the average treatment effect of bupropion is preferable to that of mirtazapine and that bupropion would always be the recommended treatment.

Supplementary Material

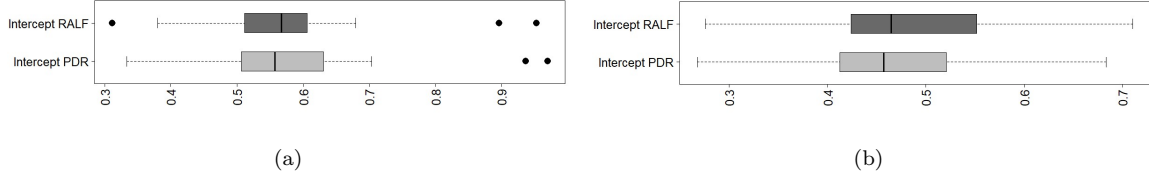


Fig. S8: Distribution across the 25 imputed datasets of the intercept coefficients, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **mirtazapine and TCA** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). The distribution in b) contains only 19 coefficients due to lack of convergence in 6 of the imputed datasets. Here, the intercept is positive, meaning that the average treatment effect of mirtazapine is preferable to that of TCA and that mirtazapine would always be the recommended treatment.

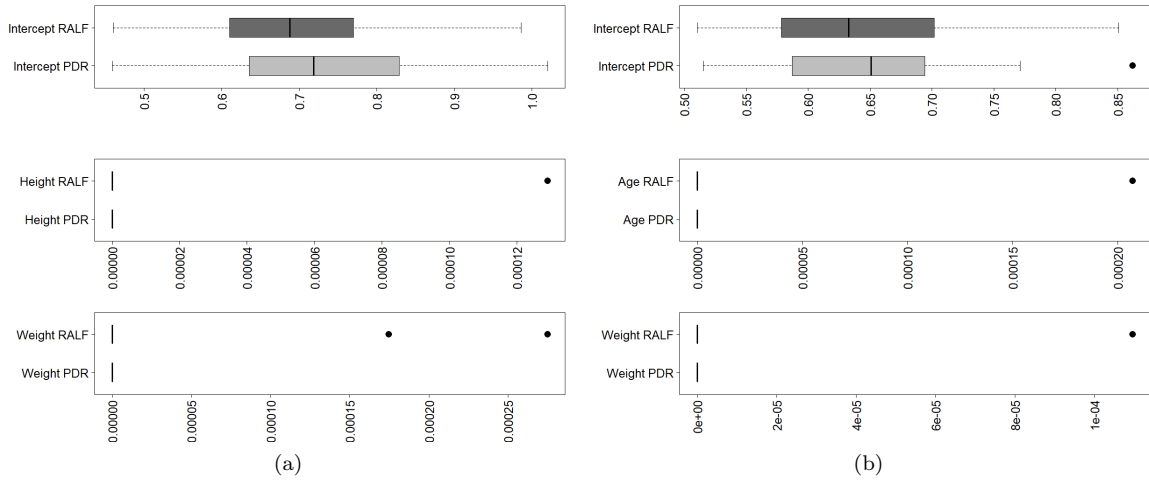


Fig. S9: Distribution across the 25 imputed datasets of the coefficients on age, height and weight, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SSRI and TCA** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). The distribution in b) contains only 20 coefficients due to lack of convergence in 5 of the imputed datasets.

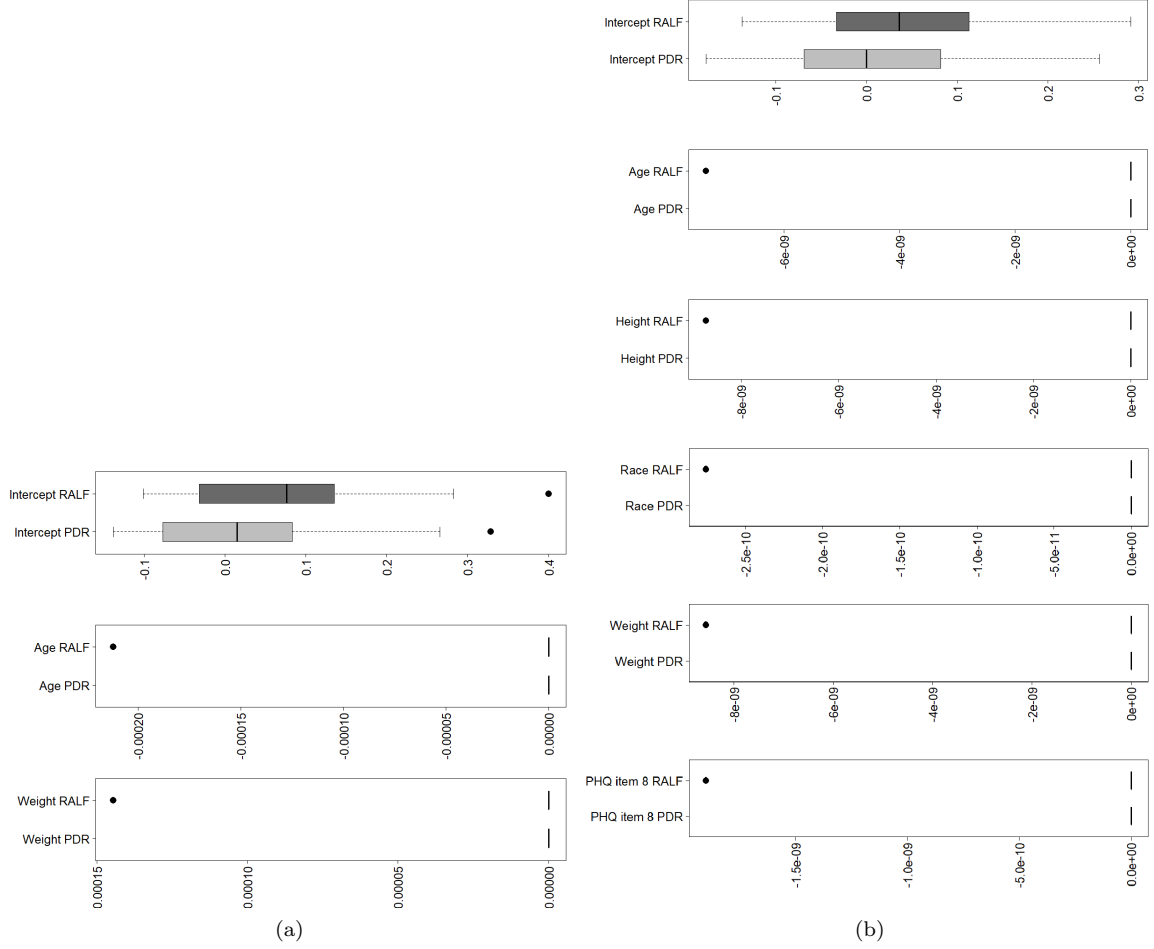


Fig. S10: Distribution across the 25 imputed datasets of the coefficients on age, race, height, weight and PHQ8, the only nonzero blip coefficients among the 23 potential effect modifiers in the comparison of **SNRI and mirtazapine** to minimize the risk of **a) a PHQ greater than 15**; and **b) a PHQ greater than 20** (RALF, fixed-point regularized A-learning; PDR, penalized doubly robust; the first term in the labels on the Y-axis corresponds to the effect modifier for which non-null coefficient(s) were found). The distribution in b) contains only 24 coefficients due to lack of convergence in 1 of the imputed datasets.