SUPPLEMENT

Supplement 1. Study Protocol

(Approved by the Kyoto University Graduate School of Medicine Ethics Committee (R3574-1), on 9 September 2022)

Estimating the Smallest Worthwhile Difference (SWD) of Antidepressants for Major Depressive Disorder: A Protocol

INTRODUCTION

Depression is the second leading cause of global disability (GBD 2019 Mental Disorders Collaborators, 2022) with point prevalence estimates upwards of 4.4% of the global population (WHO, 2017). Millions seek treatment for depression through antidepressant medications within the United States alone (Luo et al., 2020). Antidepressants are proven efficacious, backed by hundreds of randomized controlled trials and rigorous meta-analyses (Cipriani et al., 2018; Cuijpers et al., 2020). For instance, a network meta-analysis including 21 commonly prescribed antidepressants demonstrated an average standardized mean difference (SMD) of 0.30 (95% CrI = 0.26-0.34) favoring antidepressants over placebo (Cipriani et al., 2018). However, effect sizes such as SMDs in behavioral health research can miss the patient importance of an intervention (Furukawa et al., 2014). For instance, whether antidepressant treatments are a worthwhile option for patients, given the benefits in symptom improvement on the one hand and costs including harms, expenses, and other burdens on the other, remains controverted (Hengartner & Plöder, 2018; Kirsch & Sapirstein, 1998; Moncrieff, 2018).

The minimum important change (MIC), also known as the minimal important difference or minimal important clinical difference, is the smallest change in a health outcome that patients perceive as important (Jaeschke et al., 1989). Defining the MIC is a useful way to interpret patient-reported outcome measures (PROMs) (Carrasco-Labra et al., 2021). The MIC is often calculated through the anchor-based approach, which uses a readily interpretable external criterion to determine patients' importance of changes seen in a health outcome of interest. By definition, the MIC is specific to a particular assessment scale (Devji et al., 2020; Ferreira et al., 2012), generally lacks association with an intervention (McNamara et al., 2015), and does not explicitly account for costs and benefits relative to an alternative (Ferreira et al., 2012; McNamara et al., 2015). The MICs for depression scales have been estimated between a 7 to 9 point reduction for the Montgomery-Asberg Depression Rating Scale (MADRS) (Leucht et al., 2017), a 6 point reduction for the Beck Depression Inventory-II (BDI-II) (Hiroe et al., 2005), and a 7 to 8 point reduction for the Hamilton Depression Rating Scale 17-item version (HAMD-17) (Furukawa et al., 2007; Leucht et al., 2013).

The MIC can thus help determine whether changes in health status are trivial or small but important from the viewpoint of the patients but does not relate to whether differences in changes in one treatment over another are worth the costs of the treatment including its harms, expenses, and other burdens. A conceptually different approach to facilitate interpretation of patient importance in PROM scores in the context of an intervention is to estimate the smallest worthwhile difference (SWD). The SWD is "the smallest beneficial effect of an intervention that justifies the costs, risks, and inconveniences of that intervention" over an alternative intervention including no treatment or placebo (Ferreira et al., 2012). The SWD represents a between-treatment assessment reflecting a trade-off of the benefits and costs of two treatment options (Furukawa, 2020). It is patient-derived, intervention-specific, and expressed as an absolute difference in outcomes between alternative treatments (Barrett et al., 2005, 2008; Ferreira et al., 2012; McNamara et al., 2015). Two methods have been proposed to estimate the SWD: the discrete choice experiment (DCE) and the benefit-harm trade-off method (BHTM). The DCE asks individuals to state their preferences for a series of hypothetical scenarios where attributes and their levels may vary (i.e., differing benefits and costs), which one then analyzes in multivariable regression models to determine the threshold for preference for one treatment over another (Franco et al., 2016). The BHTO asks individuals directly how many benefits they are willing to trade-off for the expected costs of an intervention over another (Barrett et al., 2005). The BHTM has been suggested as an easy-to-apply and useful method and has been used to estimate the SWD in treatments for respiratory disease (McNamara et al., 2015), fall prevention (Franco et al., 2016), and pain reduction therapies (Christiansen et al., 2018; Ferreira et al., 2013). Computer applications of the BHTM have also been successful at estimating SWDs when a treatment decision is to be made versus no treatment (McNamara et al., 2015).

Patient-centered medicine places greater value on the patient perspective (Devji et al., 2020; Laine & Davidoff, 1996). However, to our great surprise and dismay, the SWD of antidepressants for depression has never been estimated in over half a century of research into the psychopharmacology of depression. Identifying the SWD of antidepressants for depression would help understand patient expectations of antidepressant therapies, evaluate

the worthiness of the current antidepressants, and establish evidence-based benchmarks to be aimed at in clinical trials for new antidepressant therapies. The present study will therefore determine the SWD of commonly prescribed antidepressants for depressive disorders using the BHTO method.

METHODS

Participants

We will include respondents aged 18 and older, who reside in the USA, and are fluent in English. To improve data quality, it is suggested to use minimal inclusion criteria variables with panel survey samples, and to instead examine subgroup effects of the variables of interest (Chandler & Shapiro, 2016).

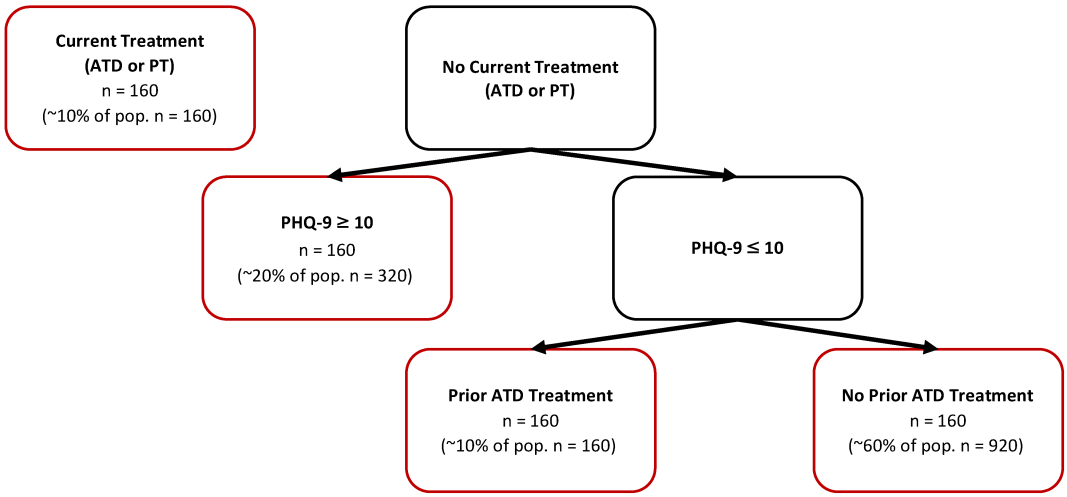


Figure 1. Subgroup combinations with primary interest of no current treatment (antidepressant or psychotherapy), positive screening for depressive symptoms, and no prior antidepressant treatment. Red boxes represent subgroups of interest with necessary minimum of n = 160 to calculate subgroup smallest worthwhile difference, PHQ-9 = Personal Health Questionnaire 9-item, ATD = antidepressants, PT = psychotherapy, pop. = population.

Thus, salient factors that may influence respondent evaluations of the SWD will be accounted for in subgroup combinations based on 1) current treatment status (antidepressant or psychotherapy), current depression symptom status (PHQ-9), and antidepressant history. Figure 1 shows the subgroup combinations of interest.

Respondents will be recruited using Amazon's Mechanical Turk (MTurk). MTurk is an internet-based research crowdsourcing recruitment tool providing monetary compensation for participation. Potential respondents are called “MTurk Workers” and compose an established respondent pool. Researchers upload surveys through MTurk and request participation from workers fitting the inclusion criteria. High-quality data from demographically diverse populations can be drawn from the MTurk crowd with greater generalizability than internet or undergraduate convenience samples (Chandler & Shapiro, 2016). Psychometrics calculated with MTurk samples demonstrate effect sizes equal to standard sampling, high test-retest reliability, and high convergent and concurrent validity with state and trait psychological constructs (Chandler & Shapiro, 2016). For depression symptomology specifically, scores measured with the PHQ-9 were highly correlated (r = 0.78) among MTurk respondents one week apart, suggesting adequate test-retest reliability (Carr, 2014).

In MTurk, workers are rated by researchers and given an approval rating for their trustworthiness. To increase data quality, we will restrict MTurk respondent enrollment to those with approval ratings above 95% (Paolacci & Chandler, 2014). Because of this restriction, surveys will not be modified with attention check validity questions. Attention is generally equivalent to other convenience samples and attention checks do not improve data quality for MTurk respondents with high approval ratings (Chandler & Shapiro, 2016). Respondents will provide e-consent and will be debriefed and compensated after completion. Compensation will be US$1, which is commensurate with similar MTurk study lengths (Nikčević et al., 2020). The initial survey will be piloted with a small number of workers and feedback will be solicited before full survey deployment.

Procedure

Participation is possible only after e-consent is signed (Supplement 2). We will collect the following information from respondents who give consent:

- *Demographic data:* Age, gender, race/ethnicity, education, employment, and insurance status
- *Clinical data:* lifetime diagnosed depression, current depressive symptoms (PHQ-9), family history of diagnosed depression, treatment preference (pharmacotherapy or psychotherapy), previous treatment (pharmacotherapy or psychotherapy) and current treatment (pharmacotherapy or psychotherapy)

Measurements

Lifetime diagnosed depression will be determined with one yes/no question, "Have you ever been diagnosed with a depressive disorder by a healthcare professional?" Current depressive symptoms will be assessed with the Patient Health Questionnaire 9-Item (PHQ-9) (Kroenke et al., 2001). The PHQ-9 provides a good description of current depression severity: none-minimal (1–4 points), mild (5–9), moderate (10–14), moderately severe (15–19), and severe (20–27). The PHQ- 9 is a reliable and valid instrument with good sensitivity (88%) and specificity (88%) for identifying major depression at a recommended cutoff of 9/10 (Kroenke & Spitzer, 2002). Family history of diagnosed depression will be determined with one yes/no question, "Has anyone in your family ever been diagnosed with a depressive disorder by a healthcare professional?" Treatment preference (pharmacotherapy or psychotherapy) will be determined with a two-option question, "If you were to receive treatment for depression, would you prefer drug or talk therapy?" Previous treatment (pharmacotherapy or psychotherapy) will be determined with a multiple-choice question, "Have you ever received treatment for depression (yes: drug therapy/pharmacotherapy, yes: talk therapy/psychotherapy, or no: I have never had treatment)?" Current treatment (pharmacotherapy or psychotherapy) will be determined with a multiple-choice question, "Are you currently receiving treatment for depression (yes: drug therapy/pharmacotherapy, yes: talk therapy/psychotherapy, or no: I am not currently receiving treatment)?"

The questionnaires will be implemented electronically using a dedicated electronic data capturing (EDC) system (Qualtrics) via MTurk. Only the system administrator has direct access to the server and back-ups. Researchers will have access to the MTurk worker ID which could be linked to personal information on Amazon public profile pages. This is dependent on respondents' individual Amazon profile settings. Additionally, Amazon maintains access to individual MTurk IDs and personal information (i.e., social security number, IP address, bank account information). MTurk worker IDs are not shared outside the study team and will only be used for distributing remuneration. MTurk IDs will not be stored with survey responses/data. Worker IDs will be removed from the dataset after collection is completed and the stored dataset will be de-identified. Data will be retained on the Amazon servers per Amazon policy. Both Amazon and Qualtrics data will be deleted upon data collection completion and deidentification. The final deidentified dataset will be maintained on a local password-protected server in a locked room.

Obtaining the smallest worthwhile effect (SWD)

The SWD will be estimated using the Benefit-Harm Trade-off Method (BHTM) (Barrett et al., 2005), presenting the survey respondents with variable, hypothetical magnitudes of antidepressant outcomes to find the smallest acceptable effect over natural recovery. The overall average smallest acceptable effect represents the SWD. Respondents will be asked a series of questions regarding the percentage of improvement they consider to be worthwhile, given the costs of antidepressant treatment. We will use the selective serotonin reuptake inhibitors (SSRIs) and serotonin-norepinephrine reuptake inhibitors (SNRIs) as the exemplar cases in the descriptions of antidepressants, as they are the most frequently prescribed (Luo et al., 2020), have similar efficacy and harm profiles (Cipriani et al., 2018), and are now mostly off-patent and hence similarly inexpensive.

The BHTM is a two-step method. In Step 1, respondents are presented with summaries of clinical depression, benefits, and costs (side effects, expenses, and burdens) associated with antidepressant treatment and with no treatment/natural recovery (see Supplement 2). The summaries will be piloted using two members of the Patient and Public Involvement (PPI) group at Oxford University. The patient members will review the scripts and provide valuable feedback as to the clarity, inclusivity, and accuracy of patient experiences captured by the descriptions. Respondents will be informed about how much improvement can be expected without antidepressant treatment, i.e., natural recovery. After the no-treatment comparison is established, respondents are presented with the potential costs of antidepressant treatment. They will be provided with a list of common (Trindade et al., 1998; U.S. Food and Drug Administration, 2019) and rare side effects (U.S. Food and Drug Administration, 2019).

In Step 2, respondents will determine if antidepressants are worthwhile given a hypothetical symptom reduction percentage compared to the known symptom reduction of natural recovery. They are asked to weigh the cost and benefits and asked if they would take the drug alternating between variable magnitudes of hypothetical

antidepressant outcomes. A detailed BHTM narrative and figures illustrating the variable antidepressant effects from the survey are provided in Supplement 2. There are two separate methods within the BHTM for determining the respondents' SWD; the *high-to-low* method and the *back-and-forth* method. We will randomly employ both methods because neither is shown to be more accurate or result in differential variance. The survey algorithm for both the high-to-low and back-and-forth methods are represented in Supplements 4.

Sample Size

Sample size was calculated to achieve expected precision in the estimates of the SWD. Because there has been no study to estimate SWD for depression and its SD is unknown, we substituted it with the SD of SWD for pain also measured on a scale of 0-100: this study suggested an SD of 22 for SWDs of 20% (Ferreira et al., 2013). Assuming a similar SD, and to obtain a 95% confidence interval within 10 points, approximately 80 respondents per group are needed. We will also randomly assign participants to either the *high-to-low* or *back-and-forth* BHTM methods, which will require twice the respondents per method to determine differences between the methods with 160 respondents per group (80 respondents x 2 methods). Since we are estimating SWD and its 95% CI for 4 subgroups (see respondent recruitment above), 640 respondents are needed (160 respondents x 4 groups). This should be enough to examine all the pre-specified potential predictors (see Data analysis below) (Austin & Steyerberg, 2015).

The survey will be piloted with 300 respondents to determine if the population distribution estimate is appropriate to reach n = 160 in all four subgroups. We have estimated that approximately 20% of the respondents will screen positive for moderate to severe depressive symptoms with the PHQ-9. This is greater than the general population, but still a conservative estimate for MTurk subgroups, which have been shown to demonstrate up to 3.6 times greater depression point prevalence than the general population (Ophir et al., 2020). Based on this estimated MTurk-specific subgroup populations, approximately 1,600 respondents may be necessary to reach 80 in the subgroups with the smallest populations (see Figure 1). If more respondents are needed, recruitment will continue until 80 people per subgroup are reached.

Data analysis

We will first present the distribution of the SWD for all patients and estimate the mean and standard deviation (SD) of the SWD, or the median SWD with its interquartile range (IQR) depending on distribution skewness. The 95% CI will also be presented. Then we will present the distribution of SWD in subgroups for the various patient, treatment, and condition characteristics (predictors). These characteristics will be included in a regression model to evaluate whether they are associated with variability in the SWD. As the SWD is estimated as a continuous variable we will examine whether the distribution is compatible with the normal distribution. If not, transformation using a link function f will be considered. Then a standard linear regression model, using the predictors will be estimated. Model selection will be performed using the least absolute shrinkage and selection operator (LASSO) method (Tibshirani, 1996). When important subgroup differences are found, we will present the SWD for such subgroups. To determine if the two survey methods equally estimate the SWD, we will compare the total average SWDs of the *high-to-low* and *back-and-forth* BHTM methods with a t test. All statistical analyses will be performed using R software (Statistical Computing, Vienna, Austria) or SAS (Cary, NC, SAS Institute Inc).

Expected Results

We expect to find an estimate of the SWD for antidepressants in the treatment of depression for the first time in the long history of human psychopharmacology. Further, we expect that the SWD will vary between clinical and demographic subgroups. In particular, we expect there to be variation between those who demonstrate depression symptoms consistent with a depressive disorder, based on PHQ-9 scores or self-reported diagnosis history. We hope that the findings from this project will inform the controversies about the worthiness of antidepressants for depression.

REFERENCES

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Barrett, B., Brown, D., Mundt, M., & Brown, R. (2005). Sufficiently important difference: Expanding the framework of clinical significance. *Medical Decision Making*, *25*(3), 250–261. https://doi.org/10.1177/0272989X05276863

Barrett, B., Brown, R., & Mundt, M. (2008). Comparison of anchor-based and distributional approaches in estimating important difference in common cold. *Quality of Life Research*, *17*(1), 75–85.

https://doi.org/10.1007/s11136-007-9277-2

Carr, A. (2014). An exploration of Mechanical Turk as a feasible recruitment platform for cancer survivors. In *Unpublished undergraduate honors theses, Paper* (Vol. 59).

Carrasco-Labra, A., Devji, T., Qasim, A., Phillips, M. R., Wang, Y., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Hao, Q., Wong, V., Ye, Z., Yao, L., … Guyatt, G. H. (2021). Minimal important difference estimates for patient-reported outcomes: A systematic survey. *Journal of Clinical Epidemiology*, *133*, 61–71. https://doi.org/10.1016/j.jclinepi.2020.11.024

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology*, *12*, 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Christiansen, D. H., de Vos Andersen, N. B., Poulsen, P. H., & Ostelo, R. W. (2018). The smallest worthwhile effect of primary care physiotherapy did not differ across musculoskeletal pain sites. *Journal of Clinical Epidemiology*, *101*, 44–52. https://doi.org/10.1016/j.jclinepi.2018.05.019

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., Ruhe, H. G., Turner, E. H., Higgins, J. P. T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J. P. A., & Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, *391*(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

Cuijpers, P., Noma, H., Karyotaki, E., Vinkers, C. H., Cipriani, Andrea, & Furukawa, T. A. (2020). A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. *World Psychiatry*, *19*(1), 92–107. https://doi.org/10.1002/wps.20701

Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Pardo-Hernandez, H., Vernooij, R. W. M., Huang, H., Rizwan, Y., … Guyatt, G. H. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: Instrument development and reliability study. *The BMJ*, *369*, 1–11. https://doi.org/10.1136/bmj.m1714

Ferreira, M. L., Herbert, R. D., Ferreira, P. H., Latimer, J., Ostelo, R. W., Grotle, M., & Barrett, B. (2013). The smallest worthwhile effect of nonsteroidal anti-inflammatory drugs and physiotherapy for chronic low back pain: A benefit-harm trade-off study. *Journal of Clinical Epidemiology*, *66*(12), 1397–1404. https://doi.org/10.1016/j.jclinepi.2013.02.018

Ferreira, M. L., Herbert, R. D., Ferreira, P. H., Latimer, J., Ostelo, R. W., Nascimento, D. P., & Smeets, R. J. (2012). A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *Journal of Clinical Epidemiology*, *65*(3), 253–261. https://doi.org/10.1016/j.jclinepi.2011.06.018

Franco, M. R., Howard, K., Sherrington, C., Rose, J., Ferreira, P. H., & Ferreira, M. L. (2016). Smallest worthwhile effect of exercise programs to prevent falls among older people: estimates from benefit-harm trade-off and discrete choice methods. *Age and Ageing*, *45*(6), 806–812. https://doi.org/10.1093/ageing/afw110

Furukawa, T. A. (2020). Responder analyses and the assessment of a clinically relevant treatment effect. *JAMA*, *323*(5), 479. https://doi.org/10.1186/1745-6215-8-31

Furukawa, T. A., Akechi, T., Azuma, H., Okuyama, T., & Higuchi, T. (2007). Letters to the Editors High-Dose Ziprasidone in Treatment-Resistant Schizophrenia and Affective Spectrum Disorders Clozapine and Amisulpride in Refractory Schizophrenia and. *Journal of Clinical Psychopharmacology*, *27*(5), 531–534.

Furukawa, T. A., Scott, I., & Guyatt, G. (2014). Measuring patients' experience. In G. Guyatt, D. Rennie, M. Meade, & D. Cook (Eds.), *Users' guides to the medical literature: a manual for evidence-based clinical practice (3rd ed.)* (Vol. 706, pp. 219–234). New York: The McGraw-Hill Companies, Inc.

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/s2215-0366(21)00395-3

Hengartner, M. P., & Plöder, M. (2018). Statistically significant antidepressant-placebo differences on subjective symptom-rating scales do not prove that the drugs work: Effect size and method bias matter! *Frontiers in Psychiatry*, *9*, 517. https://doi.org/10.3389/fpsyt.2018.00517

Hiroe, T., Kojima, M., Yamamoto, I., Nojima, S., Kinoshita, Y., Hashimoto, N., Watanabe, N., Maeda, T., & Furukawa, T. A. (2005). Gradations of clinical severity and sensitivity to change assessed with the Beck Depression Inventory-II in Japanese patients with depression. *Psychiatry Research*, *135*(3), 229–235. https://doi.org/10.1016/j.psychres.2004.03.014

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status ascertaining the minimal

clinically important difference. *Controlled Clinical Trials*, *10*, 407–415.
Kirsch, I., & Sapirstein, G. (1998). Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medication. *Prevention and Treatment*, *1*(1:0002A). https://doi.org/10.2493/jjspe.63.807
Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06
Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613.
Laine, C., & Davidoff, F. (1996). Patient-centered medicine. A professional evolution. *JAMA*, *275*(2), 152–156. https://doi.org/10.1001/jama.275.2.152
Leucht, S., Fennema, H., Engel, R., Kaspers-Janssen, M., Lepping, P., & Szegedi, A. (2013). What does the HAMD mean? *Journal of Affective Disorders*, *148*, 243–248. https://doi.org/10.1016/j.jad.2012.12.001
Leucht, S., Fennema, H., Engel, R. R., Kaspers-Janssen, M., Lepping, P., & Szegedi, A. (2017). What does the MADRS mean? Equipercentile linking with the CGI using a company database of mirtazapine studies. *Journal of Affective Disorders*, *210*, 287–293. https://doi.org/10.1016/j.jad.2016.12.041
Luo, Y., Kataoka, Y., Ostinelli, E. G., Cipriani, A., & Furukawa, T. A. (2020). National Prescription Patterns of Antidepressants in the Treatment of Adults With Major Depression in the US Between 1996 and 2015: A Population Representative Survey Based Analysis. *Frontiers in Psychiatry*, *11*(February), 1–11. https://doi.org/10.3389/fpsyt.2020.00035
McNamara, R. J., Elkins, M. R., Ferreira, M. L., Spencer, L. M., & Herbert, R. D. (2015). Smallest worthwhile effect of land-based and water-based pulmonary rehabilitation for COPD. *ERJ Open Research*, *1*(1), 1–8. https://doi.org/10.1183/23120541.00007-2015
Moncrieff, J. (2018). What does the latest meta-Analysis really tell us about antidepressants? *Epidemiology and Psychiatric Sciences*, *27*(5), 430–432. https://doi.org/10.1017/S2045796018000240
Nikčević, A. V, Marino, C., Kolubinski, D. C., Leach, D., & Spada, M. M. (2020). *Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information . January*.
Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The Turker Blues: Hidden Factors Behind Increased Depression Rates Among Amazon's Mechanical Turkers. *Clinical Psychological Science*, *8*(1), 65–83. https://doi.org/10.1177/2167702619865973
Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598
Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Trindade, E., Menon, D., Topfer, L. A., & Coloma, C. (1998). Adverse effects associated with selective serotonin reuptake inhibitors and tricyclic antidepressants: A meta-analysis. *Cmaj*, *159*(10), 1245–1252.
U.S. Food and Drug Administration. (2019). Depression Medicines. In *Depression Medicines*. https://www.fda.gov/consumers/free-publications-women/depression-medicines
WHO. (2017). *Depression and other common mental disorders: global health estimates*. World Health Organization. https://apps.who.int/iris/handle/10665/254610

Supplement 2. Changes from the Protocol

Due to our stringent quality screening process, we were unable to recruit enough responses from Mechanical Turk (MTurk) alone. This was first approved by the IRB August 9th, 2022 (R3574). We added two other research participant crowdsourcing services (Prolific, MQ) and included UK residents, which was approved September 9th, 2022 (R3574-1). Second, at the time of protocol writing, we had not anticipated that there would be people who would never consider antidepressants worthwhile, even if they brought about 100% response. We therefore made a post hoc decision to prioritize practical interpretation and exclude such people from our primary analysis but run sensitivity analyses including them. Finally, because the two types of the BHTM scripts originally included in the questionnaire produced similar estimates, we considered all responses to equally represent the SWD in the analyses.

Supplement 3. Smallest Worthwhile Difference Survey Script

PART 1

The next five slides include information on depression and its treatment. Please carefully read them and keep this in mind when answering the questions after the Information.

1. CLINICAL DEPRESSION

Everyone feels "depressed" from time to time. However, clinical depression is a more serious condition lasting much longer. When clinical depression is moderate to severe, symptoms such as the following are present nearly every day for a minimum of two weeks, and often for several months:

- Feelings of sadness, tearfulness, emptiness, or hopelessness
- Loss of interest or pleasure in most or all normal activities, such as hobbies, sports or sex
- Reduced appetite and weight loss or conversely, increased cravings for food and weight gain
- Anxiety, agitation, or restlessness
- Sleep disturbances, including insomnia or sleeping too much
- Slowed thinking, speaking, or body movements
- Tiredness and lack of energy, so even small tasks may require extra effort
- Feelings of worthlessness or guilt, fixating on past failures or self-blame
- Trouble thinking, concentrating, making decisions, and remembering things
- Frequent or recurring thoughts of death, suicidal thoughts, or suicide attempts
- Unexplained physical problems, such as back pain or headaches

Not everyone who is depressed experiences every symptom. Some people experience only a few symptoms while others may experience many. Symptoms are usually severe enough to cause noticeable problems with relationships, work, school, at social activities.

2. TREATMENT OF CLINICAL DEPRESSION

There are many different approaches to treating depression, but in this study, we will only focus on (1) no treatment (i.e., natural recovery) and (2) antidepressant drugs. First, I will describe each treatment option's expected benefits and drawbacks. Then I will ask if you think antidepressant treatment is worthwhile at different levels of patient benefit.

3. No Treatment (Natural Recovery)

By declining all treatment, about 30/100 people can expect to feel much better after 8 weeks, on average.

4. Antidepressant Drugs

Antidepressants are medicines that treat clinical depression. They help regulate the way our brain manages mood and stress. Antidepressants usually take 2 to 4 weeks to work. Depending on insurance plans, monthly expenses can range from $0-$350 USD equivalent (100 USD ≈ 84 GBP/96 EUR/149 AUD), covering office/clinic visits, prescription expenses, and transportation. As with all prescriptions, antidepressants may cause side effects. If side effects occur, they are mostly temporary and mild. The most common side effects listed by the US government (FDA) include:

- nausea
- nervousness
- insomnia
- sexual problems
- tremor (shaking)
- sweating
- agitation
- feeling tired
- dry mouth
- constipation

Serious side effects are very rare, happening in less than 1% of patients. Patients are directed to call their doctor if they experience any of the following rare symptoms:

- seizures
- abnormal bleeding or bruising
- increase in blood pressure

Antidepressants may cause other side effects that were not included in this list. Ending antidepressant prescriptions should be done slowly and under doctor supervision because abruptly stopping can cause withdrawal symptoms like anxiety, insomnia, headaches, tiredness, irritability, or flu-like symptoms.

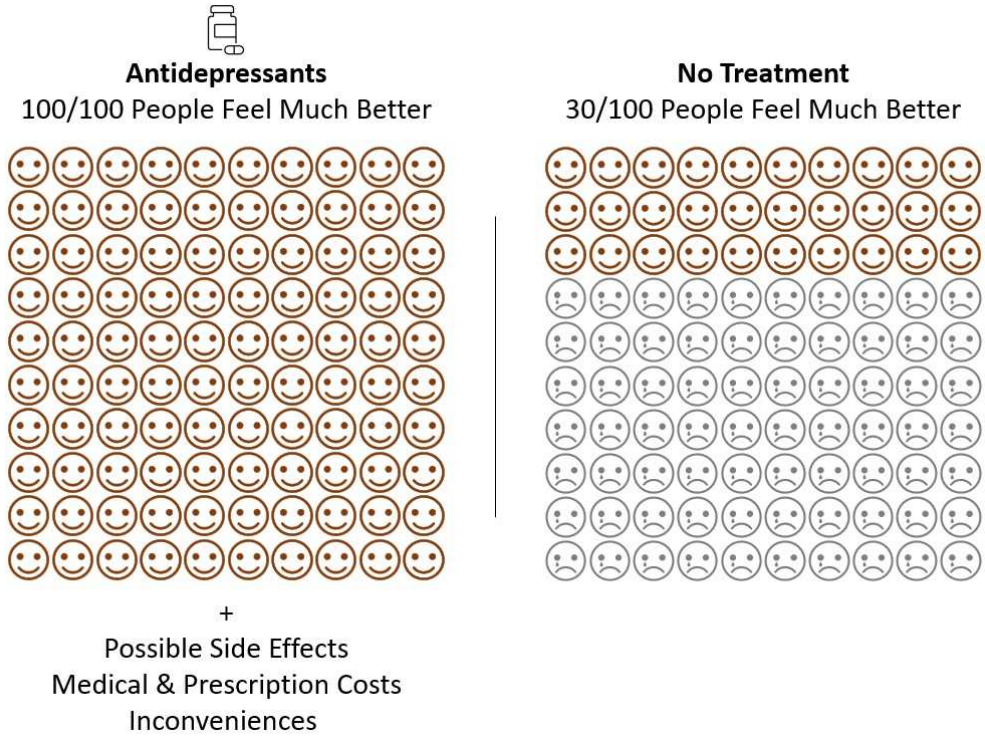
5. Reminder

There are many different approaches to treating depression, but please assume you only have two choices when deciding to relieve depressive symptoms: antidepressant medications versus no treatment. Do not consider any other alternatives in your decision.

First, consider the possible drawbacks that have been mentioned (side effects, medical and prescription expenses, and other inconveniences). Then, weigh the drawbacks relevant to you with the presented hypothetical benefits when answering the following questions.

PART 2

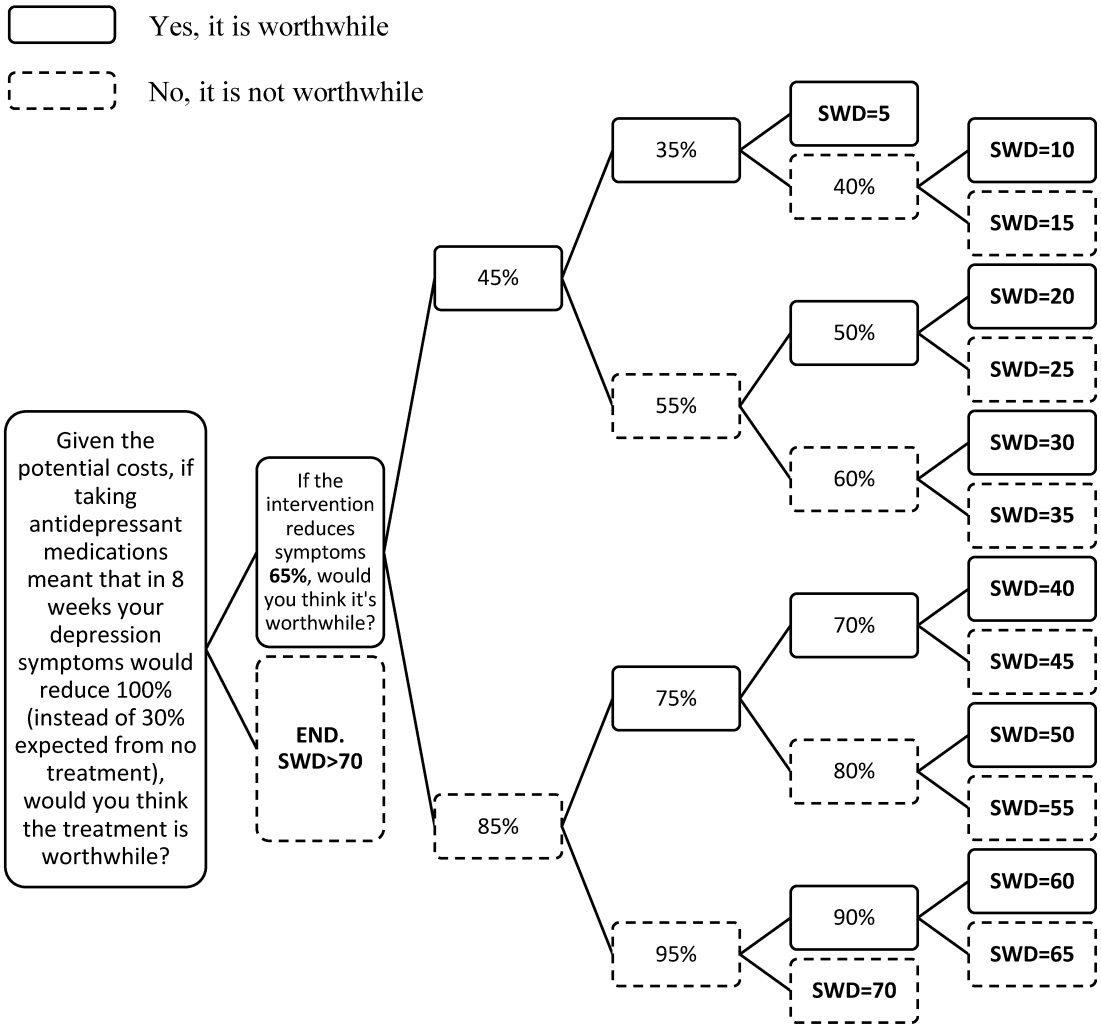
Given the potential drawbacks after 8 weeks, if <u>100/100</u> people taking antidepressant medications felt much better (instead of 30/100 from no treatment), would you think the treatment is worthwhile?



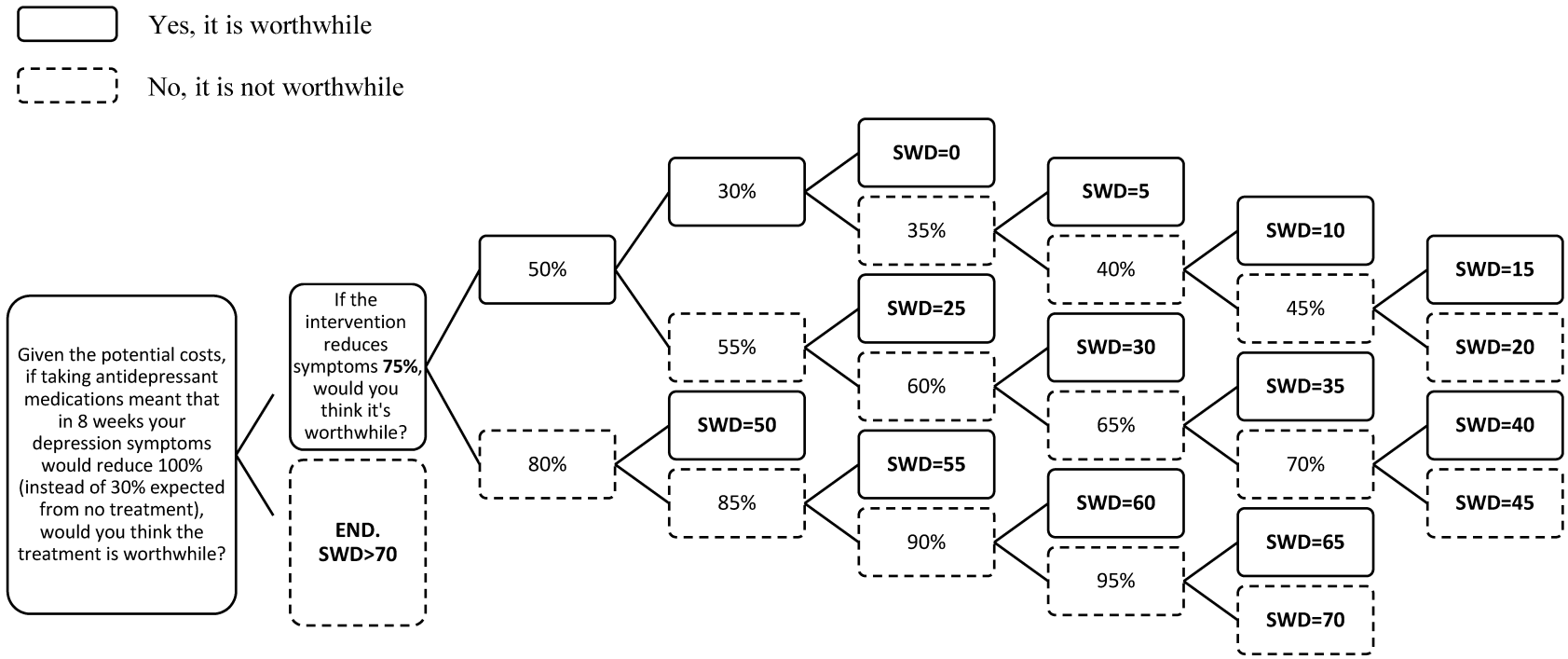
IF NO = Survey complete, respondent does not believe antidepressants are worthwhile at any effect
IF YES = Follow algorithm to next hypothetical effect

Responses follow one of two different randomly assigned algorithms presented in Supplement 4.

Supplement 4. Question Sequence Algorithms
4a. Back-and-Forth Benefit Harm Tradeoff Method

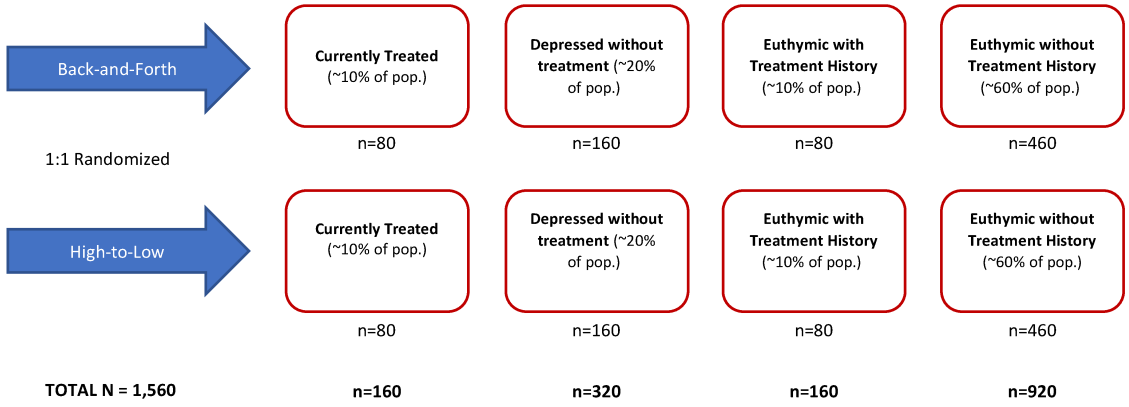


4b. High-to-Low Benefit Harm Tradeoff Method



Supplement 5. Benefit-Harm Tradeoff Method Algorithm Testing and Sample Size Calculation

We used two separate methods within the BHTM for determining the participants' SWD: the *high-to-low* method and the *back-and-forth* method (see Supplements 3 and 4). Both methods have been employed in the SWD literature (See Ferreira et al., 2012 and McNamara et al., 2015), and the most accurate SWD estimation had not been determined. Therefore, we opted to randomly allocated participants 1:1 to each method to see if they differed in their estimation. Our original sample size estimate for the SWD and its 95% CI for 4 clinical subgroups, required 780 participants based on population estimates. However, testing the difference between two separate methods required double the estimate. Based on subgroup populations, approximately N=1,600 participants were thought to be necessary to reach a minimum of n=80 in all 8 subgroups.



After we reached a minimum number of participants per group, we compared the SWD between methods. We found the difference between methods was 3·6 percentage points and we used 5 percentage point increments for the SWD. Thus, we determined both methods to be equally appropriate and all participants were analyzed together. Recruitment was stopped after all subgroups included n=80.

Supplement 6. Associations between demographic and clinical variables and SWD

	Univariable regression			**Multivariable regression**	
	Coefficient	**95%CI**		**Coefficient**	**95% CI**
Demographic Variables					
Age	0·050	-0·021 to 0·12		0·069	-0·03 to 0·16
Sex					
Male	ref			ref	
Female	-0·61	-2·75 to 1·53		-1·66	-4·01 to 0·67
Non-binary	1·66	-7·23 to 10·56		1·43	-7·75 to 10·61
Race					
White/Caucasian	ref			ref	
Black/African American	2·22	-2·53 to 6·97		3·57	-1·29 to 8·43
Asian	2·74	-2·40 to 7·89		3·33	-2·06 to 8·71
Latino/Hispanic	3·69	-1·98 to 9·35		6·55	0·62to 12·47
Multiracial	2·27	-4·45 to 8·90		3·71	-3·00 to 10·42
Other	1·42	-8·86 to 11·70		0·90	-9·40 to 11·19
Education					
Less than high school	3·78	-2·36 to 9·92		2·61	-3·74 to 8·94
High school graduate/Equivalent	0·36	-2·96 to 3·69		0·04	-3·35 to 3·44
Some college	-0·61	-3·18 to 1·96		-1·16	-3·87 to 1·54
2-year degree	-3·23	-6·56 to 0·10		-2·76	-6·15 to 0·63
4-year degree	ref			ref	
Master's degree	-0·86	-4·45 to 2·73		-1·74	-5·34 to 1·86
Doctorate	-4·04	-11·25 to 3·17		-4·09	-11·26 to 3·08
Employment					
Disabled	3·62	-0·36 to 7·60		4·08	-0·23 to 8·40
Homemaker	1·51	-3·50 to 6·51		1·89	-3·21 to 6·98
Retired	2·93	-0·57 to 6·42		2·10	-2·26 to 6·46
Student	1·52	-3·55 to 6·59		1·83	-3·65 to 7·32
Unemployed and looking	4·88	0·21 to 9·54		5·61	0·84 to 10·38
Working full-time	ref			ref	
Working part-time	1·39	-1·29 to 4·06		2·20	-0·65 to 5·04
Country					
UK	ref			ref	
US	-2·39	-4·43 to -0·34		-0·46	-4·32 to 3·41
Insurance					
Affordable Care Act/Obamacare	-3·55	-10·33 to 3·23		-3·92	-11·65 to 3·82
Medicare/Medicaid	-4·21	-7·99 to -0·43		-5·19	-10·52 to 0·14
National Healthcare Insurance	ref			ref	
Private Health Insurance	-2·42	-4·95 to 0·11		-3·03	-7·08 to 1·01
Uninsured	-1·07	-3·66 to 1·53		-1·56	-4·25 to 1·13
Clinical Variables					

Group					
Currently Depressed but Not Treated	ref			ref	
Currently in Treatment	1·96	-1·25 to 5·17		2·70	-0·83 to 6·22
Euthymic with Treatment Experiences	1·62	-2·53 to 5·77		2·30	-2·00 to 6·61
Euthymic without Treatment Experiences	2·80	-0·52 to 6·13		2·23	-1·56 to 6·02
Lifetime Depression	-0·84	-2·85 to 1·17		-0·91	-4·47 to 2·66
Family History of Depression	-0·01	-2·03 to 2·02		1·14	-1·02 to 3·30
Treatment Preference					
Antidepressants	-4·32	-6·35 to -2·30		-4·95	-7·19 to -2·73
Psychotherapy	ref			ref	

Supplement 7. Conversion of the SWD into OR and SMD

An RR can be converted into a corresponding OR, given the control event rate (CER), by the following formula:

OR = RR * (1- CER) / (1 – RR * CER)

When CER=0·30 and SWD=0·20, RR=0·50/0·30 =1·67 and OR=2·33.
Or when CER=0·30 and SWD=0·25, RR=0·55/0·30=1·83 and OR=2·85.

OR can be converted into SMD using Chinn's formula [1]:

SMD = sqrt(3) * Ln(OR) / π = 1·81 Ln(OR)

Then OR of 2·33 corresponds with SMD of 0·47, and OR of 2·85 corresponds with SMD of 0·58.

1. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000;19(22):3127-31.