November 18, 2021

Integrating a Stepped Care Model of Screening and Treatment for Depression into Malawi's National HIV Care Delivery Platform

(ID: R01MH117760**)**

November 18, 2021

Statistical Analysis Plan

Analysis of Primary Outcomes

This analysis plan concerns the primary outcomes of the IC3D intervention trial. However, analyses of the secondary outcomes will follow the same procedure with appropriate changes for the nature of the outcome.

We will have primary outcomes at two levels. At the aggregate cluster-time level we have

1. Prevalence of depression

As we do not capture data on those who do not have depression that would enable us to link individual screening outcomes between time periods, we will only be able to collect total number screened and total number with depression for each cluster at each time period (i.e. "step").

At the individual level, among those who are diagnosed with depression we will collect:

2. Depression severity among those with depression (PHQ-9)
3. WHODAS score among those with depression
4. EQ-5D-3L utility score and those with depression

If the intervention is successful then it should have the effect of reducing the overall prevalence of depression among those with chronic conditions, and also reduce the severity of depression and improve the quality of life among those with depression in this group of patients.

The data derived from the trial will be analysed using a generalised linear mixed model framework standard for stepped-wedge cluster designs.[1] Several trial guidelines, including the influential CONSORT statement,[2,3] recommend the specification of a single primary outcome to avoid a multiple comparisons problem. However, given the complex, multifaceted nature of this health service delivery intervention, no single outcome can adequately describe the effect of the intervention and be used to judge its effects. As a result, we have specified several primary outcomes and will "correct" or modify our analyses to account for the multiple outcomes problem.

There are two main types of approach to dealing with multiple outcomes: multivariate methods and univariate methods. The former method works by jointly modelling all the outcomes at once. However, the joint model of the four outcomes is highly complex, potentially with non-linear sub-models, two-stage models, and shared random effects to allow for correlation (see Appendix), which makes inference difficult or even infeasible without development (and evaluation) of custom software (e.g. a Monte Carlo Maximum Likelihood sampler), which is out of the scope of this project. Instead, we will model the outcomes separately and use a method to correct the family-wise error rate (a univariate methodology).

Statistical model

Aggregate cluster-time outcome

For cluster $j = 1, \dots, J$ at time period $t = 1, \dots T$ we will screen n_{jt} patients of whom $y_{jt}^{(1)}$ will be diagnosed with depression such that

$$y_{jt}^{(1)} \sim Binomial(n_{jt}, p_{jt})$$

where we will specify a logistic regression model:

$$\log\left(\frac{p_{jt}}{1-p_{jt}}\right) = z_{jt}'\gamma_1 + \delta_1 D_{jt} + \alpha_{j,1} + \phi_{jt,1}$$

where z_{jt} is a vector of cluster-level covariates, and $\alpha_j \sim N(0, \sigma_\alpha^2)$ and $\phi_{jt} \sim N\left(0, \sigma_\phi^2\right)$ are cluster and cluster-time random effects, respectively.

Individual-level outcomes

For patient $i = 1, \dots, N$ who has depression the outcome is $y_{ijt}^{(k)}$ for outcomes $k = 2,3,4$. For each outcome we specify the linear model:

$$y_{ijt}^{(k)} = x_{ijt}'\beta_k + z_{jt}'\gamma_k + \delta_k D_{(i)jt} + \theta_{i,k} + \alpha_{j,k} + \phi_{jt,k} + e_{ijt,k}$$

where x_{ijt} is a vector of patient-level covariates $\theta_i \sim N\left(0, \sigma_\theta^2\right)$ is an individual-level random effect and e_{ijt} is an independent, identically distributed error term.

The parameters of interest are δ_k, which are a (log) odds-ratio for the prevalence outcome and a mean absolute difference in score for the three individual-level outcomes. We will take both intention to treat (ITT) and per protocol approaches for the individual level outcomes. For the ITT analysis, the variable D_{jt} is an indicator equal to one if cluster j is in the intervention state at time t and zero otherwise. Thus, δ represents our estimated treatment effect. For the per protocol analysis, D_{ijt} will be a patient-level variable equal to one if the clinic was in the intervention state and the patient completed the treatment as defined above, and zero otherwise. As we describe below, we will use a randomisation test method to derive exact inferential statistics (p-values and confidence intervals) for the parameters of interest.

Outcomes 2-4 described above are only recorded for those who screen positive for depression at their scheduled visit. Thus, the data do not all cover the full patient population: data for outcomes 2 to 4 are a subset of those for outcome 1. Based on this, we anticipate the data to be normally distributed. However, we will test this assumption by examining QQ plots and conducting KS (Kolmogorov Smirnov) and SW (Shapiro Wilk) tests. In the event the normality assumption is violated, we will consider an alternative specification, such as a two-part semi-continuous model.

Statistical Inference

All models will be estimated separately using restricted maximum likelihood, using the R package lme4. We will report point estimates, confidence intervals, and p-values but not make any claims of "statistical significance", given recent strong arguments against doing so [2].

P-values will be based on the null hypotheses $H_0: \delta_k = 0$ versus the two-sided alternatives $H_1: \delta_k \neq 0$ in each of the models defined above. Given that there are multiple primary outcomes, we will adjust reported p-values for multiple testing using a cluster-based stepdown method, which provides an efficient means of controlling the family-wise error rate [18]. The full methodology is described elsewhere, but we describe it here briefly.

Let there be P hypotheses to be tested $H_1, \dots, H_P$ with associated test statistics T_p. We let the ordered test statistics be:

$$T_{[1]} \geq T_{[2]} \geq \cdots \geq T_{[P]}$$

corresponding to hypotheses $H_{[1]}, \dots, H_{[P]}$. The family of hypotheses being tested is $K \subset \{1, \dots, P\}$. We first describe a stepdown procedure for the decision to accept/reject given a value for the type I error rate α, as this provides a clear way of deriving the associated p-value that we will report.[4] The stepdown method works by firstly testing if the joint null hypothesis is true by comparing the largest test statistic to the critical value $c_K\left(1 - \frac{\alpha}{2}\right)$. If it is smaller than this critical value, we accept all null hypotheses; otherwise, we reject $H_{[1]}$ and test the remaining hypotheses as a new family in the same way. More specifically, the algorithm is:

1. Let $K_1 = \{1, \dots, P\}$. If $T_{[1]} \leq c_{K_1}\left(1 - \frac{\alpha}{2}\right)$, then accept all hypotheses and stop, otherwise reject $H_{[1]}$ and continue.
2. Let K_2 be the indices of hypotheses not previously rejected. If $T_{[2]} \leq c_{K_2}\left(1 - \frac{\alpha}{2}\right)$ the accept all remaining hypotheses, otherwise reject $H_{[2]}$ and continue.

The critical values here are the $1 - \frac{\alpha}{2}$ quantiles of the distribution of the largest test statistic for the relevant family of hypotheses. Exact distributions may not exist; however, we can derive them using a pseudo-permutation testing approach by enumerating a large number of permutations (e.g. 100,000 permutations) following a randomisation test approach based on the cluster randomisation.

At each stage of the stepdown algorithm, we are testing the largest test statistic, that is: $T_K = \max_{q \in K} T_q$. If the associated hypothesis is rejected, we create a new family of hypotheses that excludes the rejected hypothesis. If we re-randomise the clusters, then for each permutation $m = 1, \dots, M$, we can re-calculate the desired test statistic, $T_K^{(m)}$, to create our reference distribution. The p-value is then:

$$p_{[k]} = \frac{1}{M} \sum_{m=1}^{M} I\left(T_K^{(m)} \geq T_K\right)$$

That is to say that the reference distribution for the largest test statistic is the largest test statistic from the permutations of the remaining hypotheses, and so on (see also [4–6]). The test statistics we will use are the studentised, generalised residuals from each of the models, which are optimal for randomisation tests for cluster-based analyses and exponential family models,[7] and also ensure each of the individual tests in the family of tests has the same power since the test statistics will be of equivalent scale.[5]

The procedure above can be used to derive a set of confidence intervals for the parameters of interest that has nominal “family-wise coverage”, as described by Watson et al [TBA]. We will estimate 95% confidence intervals for the set of treatment effect parameters following this method.

Missing Data

We foresee the main risk in terms of missing data to be patients whose depression status may cause them to miss appointments and hence have fewer observations during the trial. This mechanism would generate data missing *not at random* (MNAR). Alternative mechanisms include data entry error and (conditionally) random patient no-shows to appointments (e.g. older patients are less likely to attend), both of which would be missing at random (MAR) or missing completely at random (MCAR). It is not possible to distinguish between MNAR and

MAR empirically.[8] Thus, to provide information about the possible mechanism of missingness in the data, CHWs tasked with following-up with patients will be asked to identify a main reason for the missed appointments in an anonymous reporting form. For example, reasons may include “mental health/emotional issues” or alternative such as “was not able to attend due to other commitments”. A qualitative judgement will be made on the basis of these data by the trial team as to what the most likely/common cause of missingness is likely to be. We also note that the CHW follow-up process should also limit the extent of missing data in this study.

If the data are considered to be (mostly) MAR or MCAR then they do not present a risk of bias for our analyses; however, the uncertainty may be under-estimated. If this is the case, and there exists a substantial proportion (>10% missing) of missing data, then a secondary analysis will be conducted using multiple imputation methods. If the data are determined to be (mostly) MNAR then a secondary analysis that attempts to jointly model the outcome, covariates, and missingness mechanism will be conducted: in particularly, a random coefficient pattern mixture model approach will be used, which has previously been shown to provide reliable inferences for longitudinal data with MNAR data, particularly in the case where drop out may be due to a decline in functioning over time.[9]

Secondary Outcomes

The analysis plan will also include several secondary analyses, in which we focus on the relationships between the intervention and physical health outcomes—such as HIV, diabetes, and hypertension—as well as theoretical moderators, including fidelity to protocols among providers, levels of social support received by participants, and perceived internalized stigma among patients. For physical health outcomes, the same model specification will be used as described above. For evaluation of the effect of the intervention on potential moderators, the relationships will predominately be examined through interaction models, in which we will include an interaction term of the treatment effect with variables measuring fidelity, social support and stigma in the models. While examining the relationships among these variables, the research team may also consider mediator pathways for social support and stigma, whereby the intervention leads to changes in social support and stigma, which (in turn) lead to reductions in depression symptoms.

Addendum: Added 18 November, 2021

Implementation outcomes

We propose to add a more detailed and explicit assessment of the implementation of the intervention as a secondary analysis. The primary aim of the trial remains to estimate the effectiveness of the intervention from an intention-to-treat perspective. However, given the novelty of our approach, understanding the context for implementation and the barriers to successful adoption and integration of mental health interventions is an important secondary aim. The trial is thus a "hybrid effectiveness-implementation trial" according to the schema outlined by Wolfenden et al. (2021) [10]

The implementation outcomes are divided into three categories:

1. Screening:
    a. Proportion of patients screened with the PHQ-2, out of those attending IC3 clinics.
    b. Proportion screening positive with PHQ-2, out of those who were administered the PHQ-2.
    c. Proportion screened on the PHQ-9, out of those who screened positive on the PHQ-2.
    d. Proportion screening positive on PHQ-9, out of those who were administered the PHQ-2.
2. Diagnosis:
    a. Proportion of patients who are administered the BDIS, out of those who screened positive on the PHQ-9.
    b. Proportion of patients who receive a diagnosis using BDIS, out of those administered the BDIS.
3. Treatment:
    a. Proportion of patients attending treatment sessions (PM+, ADT, both), out of those who have been enrolled in the IC3D clinical trial and elected to receive treatment (PM+, ADT, both)
4. Protocol Fidelity:
    a. Proportion of reviewed mastercards appropriately completed by clinical officers
    b. Proportion of patient registration forms appropriately completed by counsellors
    c. Proportion of PM+ components appropriately covered within each PM+ therapy session

Each step is a potential point of failure that could limit the acceptability or effectiveness of the intervention. Analysis of the implementation outcomes constitutes a descriptive analysis intended to support interpretation of trial results.

We will report the overall trial mean proportions of each implementation outcome along with 95% confidence intervals. Note that the screening and diagnosis outcomes are available for both treatment and control conditions, while the treatment outcomes are only available in treatment units. We will also examine the temporal changes in the implementation outcomes by reporting within each three-month block mean proportions and associated uncertainty intervals.

Secondary analyses

In addition to the usual issues of adherence, such as patients refusing treatment or not attending scheduled treatment, we have identified that patients diagnosed in one month or trial "block" may not be treated until the subsequent month or trial block. At an individual level, non-adherence would result in individuals being assigned to treatment but not receiving it at a particular time. While our primary interest is in the “real-world” intention to treat effects, we will conduct a secondary analysis to estimate the complier average causal effects (CACE) (also known as local average treatment effects (LATE)). The CACE provides an estimate of the effect of the treatment on those who receive the treatment as intended. This will provide insight into the potential benefits of treatment, should implementation be improved. The CACE is estimated using an instrumental variable approach where the instrument is the randomised allocation at each time period and the treatment status is the actually received (or not) treatment at the individual level.

References

1. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 2017;18:101.

2. The Lancet. Consort 2010. Lancet. 2010;375:1136.

3. Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, Thompson JA, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ. 2018;:k1614.

4. Romano JP, Wolf M. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. Stat Probab Lett. 2016;113:38–40. doi:10.1016/j.spl.2016.02.012.

5. Romano JP, Wolf M. Stepwise multiple testing as formalized data snooping. Econometrica. 2005;73:1237–82.

6. Romano JP, Wolf M. Exact and approximate stepdown methods for multiple hypothesis testing. J Am Stat Assoc. 2005;100:94–108.

7. Braun TM, Feng Z. Optimal permutation tests for the analysis of group randomized trials. J Am Stat Assoc. 2001;96:1424–32.

8. Molenberghs G, Beunckens C, Sotto C, Kenward MG. Every missingness not at random model has a missingness at random counterpart with equal fit. J R Stat Soc Ser B (Statistical Methodol. 2008;70:371–88. doi:10.1111/j.1467-9868.2007.00640.x.

9. Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Stat Med. 2003;22:2553–75. doi:10.1002/sim.1475.

10. Wolfenden *et al*. Designing and undertaking randomised implementation trials: guide for researchers. BMJ 2021;372:m3721.

APPENDIX

Multivariate model

Given that the outcomes are mixed discrete and continuous, we will use a shared random effects approach to jointly model the outcomes. For patient $i = 1, \dots, N$ in cluster $j = 1, \dots, J$ at time period $t = 1, \dots T$ the outcome is $y_{ijt}^{(k)}$ for outcomes $k = 1,2,3,4$. For each outcome we specify the linear predictor:

$$\eta_{ijt}^{(k)} = x_{ijt}'\beta + z_{jt}'\gamma + \delta_k D_{(i)jt} + \psi_{k,\theta}\theta_i + \psi_{k,\alpha}\alpha_j + \psi_{k,\phi}\phi_{jt} \tag{3}$$

where x_{ijt} is a vector of patient-level covariates, z_{jt} is a vector of cluster-level covariates, $\alpha_j \sim N(0, \sigma_\alpha^2)$, $\phi_{jt} \sim N(0, \sigma_\phi^2)$ and $\theta_i \sim N(0, \sigma_\theta^2)$ are cluster, cluster-time, and individual-level random effects, respectively, and $\psi_{k,.}$ are factor loadings with $\psi_{1,.} = 1$ for identifiability. The parameter of interest is δ_k. We will take both intention to treat (ITT) and per protocol approaches. For the ITT analysis, the variable D_{jt} is an indicator equal to one if cluster j is in the intervention state at time t and zero otherwise. Thus, δ represents our estimated treatment effect. For the per protocol analysis, D_{ijt} will be a patient-level variable equal to one if the clinic was in the intervention state and the patient completed the treatment as defined above, and zero otherwise.

For each outcome, $y_{ijt}^{(k)}$, we specify its distribution with probability density function

$$f_k\left(y_{ijt}^{(k)}; \Theta, \Phi\right)$$

where $f_k(.)$ is a probability density function, $g_k(.)$ a link function and Θ is set of all model parameters and Φ are the random effect terms. For outcome 1 we specify a Binomial-logisitic model, and for the other outcomes a linear model with normally-distributed errors.

Outcomes 2-4 described above are only recorded for those who screen positive for depression at their scheduled visit as these instruments are not used for other patients in the trial. For each time period (quarterly), any patient eligible for screening (i.e. in the "denominator population") will have a value for recorded for Outcome 1 equal to one if they screen positive and meet the pre-defined criteria for entry into the depression care programme, or zero otherwise. The model is thus a "two-stage" type model with only those with defined depression contributing data from Outcomes 2-4 to the likelihood. Letting Y represent all the data, the log-Likelihood is therefore:

$$\begin{aligned}\log L(\Theta|Y,\Phi) &= \log f(Y;\Theta,\Phi) \\ &= \sum_{i=1}^{n}\left(1 - y_{ijk}^{(1)}\right)\log\left(1 - g_1\left(\eta_{ijt}^{(1)}\right)\right) \\ &+ y_{ijk}^{(1)}\left(\log\left(g_1\left(\eta_{ijt}^{(1)}\right)\right) + \log f_2\left(y_{ijt}^{(2)};\Theta,\Phi\right) + \log f_3\left(y_{ijt}^{(3)};\Theta,\Phi\right)\right. \\ &\left.+ \log f_2\left(y_{ijt}^{(3)};\Theta,\Phi\right)\right) + \log f_\Phi(\Phi;\omega)\end{aligned}$$

where ω are the parameters of the joint distribution of random effects. The likelihood function for the parameters alone is determined by integrating out the random effects Φ and then maximising the resulting likelihood. Given the complexity of our model including non-linear effects, two-stage models, and multiple shared random effect terms, there exists no "out of the box" software that could reliably be used. Instead an alternative approach, such as Monte Carlo Maximum Likelihood would be required. Given the technical complexity required, and any simulation-based testing that would be required to ensure any custom-built sampler was functioning correctly, we instead opted for a univariate methodology.