# Statistical Analysis Plan

## Enhancing the QUality of psychological Interventions delivered by Telephone (EQUITy):

## A cluster randomised trial of an 'enhanced telephone therapy' quality improvement intervention in NHS Talking Therapies

| | |
|---|---|
| Short Title | EQUITy |
| Trial Registration Number | ISRCTN22583714 |
| Protocol Version | v11.0 (16/10/2023) |
| SAP Version | v9.0 (21/07/2025) |
| Chief Investigator | Penny Bee (University of Manchester) |
| Health Economists | Sol Yates (University of Manchester) |
| | Gemma Shields (University of Manchester) |
| Statisticians | Colin Everett (University of Manchester) |
| | Mark Hann (University of Manchester) |
| SAP Contributors | Peter Bower (University of Manchester) |
| | Judith Gellatly (University of Manchester) |

## Trial Design

The EQUITy trial is a two-arm, 1:1 allocation, cluster randomised, parallel-group, superiority trial (https://fundingawards.nihr.ac.uk/award/RP-PG-1016-20010). It compares Step 2 NHS Talking Therapies services (a cluster) receiving an 'enhanced telephone therapy' quality improvement intervention consisting of multiple components (service guidelines, team workshop and follow-up meetings; practitioner telephone training; resources for patients) with NHS Talking Therapies services following usual care.

The intervention has been developed to help NHS Talking Therapies services improve the quality of their telephone treatments. In the intervention clusters, staff will receive a team workshop, telephone training, educational materials to help patients understand treatments delivered by telephone and service guidelines to enhance practice. Patients will receive additional resources such as a leaflet introducing telephone treatment, an appointment card and a poster.

Patients in usual care will continue to receive routine NHS Talking Therapies care (routine telephone therapy).

## Objectives

The primary objective is to assess the clinical effectiveness of the intervention:

> Does the EQUITy intervention improve the mental health outcomes of patients compared to usual NHS Talking Therapies care?

A secondary objective relates to cost effectiveness:

> Does the EQUITy intervention improve the cost-effectiveness of treatment compared to usual NHS Talking Therapies care?

## Sample Size and Power

There are approximately 107 NHS Talking Therapies providers in England, and we aimed to recruit a minimum of 26 providers (clusters). With 13 clusters per arm, assuming an intra-cluster correlation (ICC) of 0.027 and a baseline-endpoint correlation (r) of 0.5, we would have 84% power to detect a standardised mean difference of 0.2 (at the 5% level of significance) assuming a minimum of 100 patients per provider.

However, by May 2024, it became apparent that we were not going to be able to meet the sample size of 26 clusters. There was also concern relating to the larger-than-expected variability in cluster sizes, which would affect study power via the large coefficient of variation. Subsequently, we decided that a fixed number of patients would be randomly selected from each provider. So long as this number is not less than 200 and the coefficient of variation in cluster size did not exceed 0.3, the study would retain at least 80% power to detect a standardised mean difference of 0.2 (with the ICC and r as above) with 22 clusters.

**Randomisation**

As a cluster trial, allocation to either the intervention or usual care took place at the provider level, via minimisation. The minimisation factors will be:

- NHS Talking Therapies clinical performance (Recovery Rate: % of annual NHS Talking Therapy referrals for persons aged 18+ finishing a course of treatment who are 'moving to recovery'; <52 vs. ≥52)

- Site size (count of NHS Talking Therapies patients receiving low-intensity therapy; <815 vs. ≥815)

- Use of telephone delivery (pre-pandemic); ≤33% vs. >33%

- Deprivation score (IMD); <20.8 vs. ≥20.8

Minimisation factor dichotomies were chosen in order that groups would be of roughly equal size, based on data available from NHS Talking Therapies services. Allocation was undertaken sequentially and independently by Manchester Clinical Trials Unit (CTU) to maintain concealment of allocation. A random weighting element of 4:1 allowed for allocation to the group which does not minimise the imbalance between minimisation factors (i.e. a 20% chance).

**Outcome Assessment**

The primary outcome measure will be the Patient Health Questionnaire (PHQ-9) for depression (Kroenke et al., 2001). The PHQ-9 will be collected at baseline and at the end of treatment, via routine NHS Talking Therapies data minimum data set.

The PHQ-9 is a 9-item measure of the severity of depression. Patients are asked how often they have been bothered by the nine symptoms of depression on a 4-point scale from 'not at all' (scoring 0) to 'nearly every day' (scoring 3). A total score, ranging from 0 to 27, indicates the overall severity of depression. Total scores of 5, 10, 15 and 20 represent cut-points for mild, moderate, moderately severe and severe depression, respectively.

The Generalised Anxiety Disorder (GAD-7) for anxiety (Spitzer et al., 2006) will be treated as a secondary outcome. The GAD-7 will be collected at baseline and at the end of treatment, via routine NHS Talking Therapies data minimum data set.

The GAD-7 is a 7-item measure of the severity of anxiety. Patients are asked how often they have been bothered by the seven symptoms of anxiety over the past two weeks. Responses are on a 4-point scale from 'not at all' (scoring 0) to 'nearly every day' (scoring 3). The GAD-7 total score ranges from 0 to 21. Total scores of 5, 10 and 15 are taken as cut-off points for mild, moderate and severe anxiety respectively.

**Baseline Data**

The NHS Talking Therapies minimum data set contains minimal patient covariates - age, sex, ethnicity and employment status. The completeness of the latter two varies between data sets and, as it is not viable to impute this type of information in the volume in which it is sometimes missing, they will not be used. Age and sex will be included in selected regression models (see *Inferential Analyses*).

Patient and/ or general practice postcodes are available either in part (e.g. M13) or in full and can be linked to the Index of Multiple Deprivation (IMD) at Lower Super Output Area level (as

per Stochl et al 2022). Where postcodes are only available 'in part', we will use a weighted average of all IMDs associated with such postcodes to derive the level of deprivation. NHS Talking Therapies service-level IMD has been used in the allocation process and so all models will include some control for this variable (along with the other minimisation factors).

Data on mode of intervention delivery (e.g. face-to-face, via telephone, etc.) are available and will be summarised for each treatment group separately. Data on treatment session attendance and/ or treatment completion will also be summarised by treatment group as appropriate (e.g. median; mode; range; % completing). Talking Therapies services define treatment completion as the attendance of at least two sessions recorded as either '*assessment and treatment'* or '*treatment'*, followed by being discharged (https://www.england.nhs.uk/mental-health/adults/nhs-talking-therapies/service-standards/).

**Statistical Analysis**

*Selection of Patients and Weighting*

As per the amendment to the power calculation described earlier, for each cluster and separately for patients presenting with depression and anxiety, patients who are eligible to receive NHS Talking Therapies services will be ordered by their sex, age and date of first appointment. A simple random sample of 400 will then be drawn without replacement. Only patients whose baseline PHQ/ GAD, age and sex are not missing will be included in the sample (as it is unlikely that this information can be imputed accurately).

The sample with depression will be drawn first and 400 patients who present to each NHS Talking Therapy service (and are eligible for step 2 telephone-based psychological therapy) with mild, moderate or more severe depression (at baseline: i.e. PHQ-9 ≥ 5) will be randomly selected to form an 'analysis sample'.

A second, non-overlapping, random sample of 400 patients with anxiety will be selected to analyse this outcome. Selection will also be based on patients who present to each NHS Talking Therapies service with mild, moderate or more severe anxiety (at baseline: i.e. GAD-7 ≥ 5). If there are fewer than 400 patients identified for either sample, all such patients will be retained.

As NHS Talking Therapies service teams will see different numbers of patients, our primary approach will involve deriving probability weights (the inverse of the probability of being selected) in order that the patients selected at random better represent the wider population of patients seen by that particular team. We will also investigate unweighted analyses.

We have deliberately 'over-sampled' (400 patients) as it is not known what percentage of patients will not have engaged with their NHS Talking Therapies and, therefore, have missing outcome data (post-baseline). Although we will impute any missing 'end-of-treatment' PHQ-9/ GAD-7 scores (defined as the last-recorded score, acknowledging that this may not be the end of treatment per-se), the reliability of such analyses is likely to decrease with increasing amounts of missing data. For example, if 50% of end-of-treatment PHQ-9/ GAD-7 scores are not available, a complete case analysis will still retain 80% power to detect the desired effect.

*Descriptive Analyses*

As per CONSORT recommendations (Hopewell et al 2025), we will report the number of NHS Talking Therapies services approached and the number subsequently recruited and allocated

to each arm of the trial. Reasons for declining participation in the study will be noted. The number of eligible patients (the Talking Therapy service-level mean or median number, the range and the coefficient of variation), the number of eligible patients randomly selected from each Talking Therapy service and the number of randomly selected patients from each service who have outcome data at follow-up/ treatment completion (in each trial arm) will be reported.

Outcome variables (PHQ-9 and GAD-7) and model covariates (age, sex and, if sufficiently complete, ethnicity, employment status and level of deprivation) will be described, by trial arm, for the entire population of patients in our participating Talking Therapy services (at baseline only), all sampled participants (at baseline only) and the sub-sample (of sampled participants) who provide outcome data at follow-up (at baseline and follow-up) using (as appropriate) absolute frequency/ percentage, mean/ standard deviation, median/ interquartile range, minimum/ maximum and skewness/ kurtosis. The percentage of missing data will also be reported for each variable, overall and by treatment group. Data on mode of intervention delivery (frequency and percentage of each type), treatment session attendance and components of the intervention delivered will also be summarised.

*Inferential Analyses*

Null hypothesis: There is no significant difference in the severity of depression (primary)/ severity of anxiety (secondary) between patients receiving the 'enhanced telephone therapy' intervention and patients receiving usual care from NHS Talking Therapies services at the conclusion of their treatment.

Analysis Method: Analyses will be conducted using Stata (Vers.14 or later, StataCorp 2015) or R statistical software. Statistical tests will be performed with a 2-sided type I error rate of 5%. Patients will be analysed according to the trial arm to which their cluster was allocated (intention-to-treat). The analyses will be conducted following completion of treatment. It is possible that a number of patients in the database will not have completed their treatment. However, we will not know if this is the case. Therefore, we will analyse the last available data point as if it was their end of treatment. Although it may not be, this should not introduce any bias between arms.

Data on the outcome measures will be assumed to be continuous; multivariable mixed effects linear regression will be conducted (with the last recorded PHQ-9 score or the last recorded GAD-7 score as the dependent variable and the corresponding baseline values as a fixed effect covariate).

Models to be fitted: The following models will be reported. All models will include the minimisation factors and the specified covariates as fixed effects. For each proposed model, the number of participants with observed data for all variables included in the model will be reported. We will also report the adjusted difference in follow-up PHQ-9/ GAD-7 scores between trial arms (i.e. the regression parameter for trial arm), along with its 95% confidence interval and corresponding p-value.

| Outcome | Model | Covariate(s) |
|---|---|---|
| PHQ-9/ GAD-7 at the end of treatment | 0 | Trial Arm Allocation only |
| | A | Trial Arm Allocation<br>PHQ-9/ GAD-7 at baseline |
| | B1 (all participants: primary analysis model) | Trial Arm Allocation<br>PHQ-9/ GAD-7 at baseline<br>Age and Sex |

| | | |
|---|---|---|
| | B2 (subset of participants, equivalent to C) | Trial Arm Allocation<br>PHQ-9/ GAD-7 at baseline<br>Age and Sex |
| | C | Trial Arm Allocation<br>PHQ-9/ GAD-7 at baseline<br>Age and Sex<br>Any Other Usable Variables * |

* We will investigate the data completeness of Employment Status and Ethnicity. We will only include them in Model C above if they are at least 80% complete, as the imputation of such variables is virtually impossible. If Model C is fitted, then we will also fit Model B2 on the same subset of participants (allowing comparison). If Model C is not fitted, Model B2 will not be either.

The data structure is strictly hierarchical (patients within NHS Talking Therapies services) and, therefore, the NHS Talking Therapies service will be treated as a random effect. We will estimate regression parameters using the method of maximum likelihood (MLE). Whilst restricted maximum likelihood estimation may result in less biased parameter estimates, it is uncertain whether this approach can be used in conjunction with probability weighting (see *Selection of Patients and Weighting*). However, given the large sample size (potentially 22 x 400 = 8,800), MLE will be approximately unbiased. We will also use an unstructured covariance matrix.

Missing data at follow-up: Dependent on the extent of missing data, the method of multiple imputation (MI) by chained equations will be employed, with the extent of missing data determining how many MI datasets will be generated (as a guide, one 'imputation' per 1% of missing data, but a minimum of 10). Imputation of missing 'follow-up' values will be done separately in each arm using, as a minimum, age, sex and the baseline values of both the PHQ-9 and the GAD-7. This will be our primary modelling approach (Model B1). As a sensitivity analysis, we will also conduct an analysis of complete cases (by which we mean where outcome data are missing for a particular patient at follow-up, we will assume that it is missing (completely) at random, and we will not replace it). Such patients will, therefore, be excluded from this analysis. No subgroup analyses are planned.

Alternative tests if distributional assumptions are violated: The distribution of the responses to the PHQ-9 and GAD-7 will be examined, prior to conducting the regression analyses, via a histogram and summary statistics. If the absolute value of the skewness exceeds 1 or the kurtosis falls outside the range 2 to 4, we will consider it to be non-normal and p-values will be validated using a non-parametric bootstrapped estimate of the standard error. A randomly generated seed will be used as a starting point for the bootstrapping process and 1,000 replications will be generated. One consideration here is that some statistical software does not permit bootstrapping in conjunction with probability weighting. If this is the case, it may be necessary to rely on the properties of large samples and assume that the sampling distribution of the mean is, in fact, normal (and that bootstrapping is not required). If the skewness is in a positive direction, we will also consider modelling a logarithmic transformation of the outcome. If a satisfactory transformation cannot be found, categorisation of the PHQ-9 and/ or GAD-7 will be considered.

The focus of our intervention will be on improving patient outcomes and engagement among those offered telephone treatments. However, there is a risk that our intervention will affect the types of patients who are offered telephone treatment in intervention sites, compared to controls. We will conduct a range of robustness checks to assess any potential impact, including exploring interview data with professionals to understand changes that might have been implemented and a comparison of the types of patient referred for telephone-treatment

before and after the intervention (via comparison of the baseline characteristics, such as depression/ anxiety severity, of patients in the intervention and usual care groups).

It is also possible that the intervention could influence the duration of treatment (people may be more likely to stay engaged with telephone therapy if it is of higher quality). We will initially compare the groups descriptively (e.g. median; inter-quartile range; range) in terms of treatment length (and hence timing of outcome) to assess this and report if there is any evidence of bias in the variation of time of follow up. As treatment duration cannot be known at baseline, but may be related to treatment, it could be a potential mediator variable (that is, the intervention increases treatment duration which, in turn, leads to better clinical outcomes). We will investigate this, but not formally as part of this Statistical Analysis Plan.

**References**

Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RCW, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JPA, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegried NL, Staniszewska S, Taylor RS, Thabane L, Torgerson D, Vohra S, White IR, & Boutron I. CONSORT 2025 statement: updated guideline for reporting randomised trials.. BMJ. 2025; 388:e081123. https://dx.doi.org/10.1136/bmj-2024-081123

Kroenke K, Spitzer RL, & Williams JB. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine. 2001;16(9), 606– 613. doi: 10.1046/j.1525-1497.2001.016009606.x.

Spitzer RL, Kroenke K, Williams JB, & Löwe BA. A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine. 2006;166(10), 1092–1097. doi: 10.1001/archinte.166.10.1092.

StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LLC.

Stochl J, Soneson E, Stuart F, Fritz J, Walsh AEL, Croudace T, Hodgekins J, Patel U, Russo DA, Knight C, Jones PB, & Perez J. Determinants of patient-reported outcome trajectories and symptomatic recovery in Improving Access to Psychological Therapies (IAPT) services. *Psychological Medicine*. 2022;52(14):3231-3240. doi: 10.1017/S0033291720005395.