Short Data Analysis Plan (DAP) for ITAMITED project

ClinicalTrials.gov Register: NCT04127214

The ITA Model of Integrated Treatment of Eating Disorders (ITAMITED)

Date: July 19, 2022

A DAP created 19.ix.2020 but not uploaded then through error.

The information included in this document is also available in: Evans, Chris. (2022). Short Data Analysis Plan (DAP) for ITAMITED project (Version 3). Zenodo. https://doi.org/10.5281/zenodo.6854720

Short data analysis plan for ITAMITED study

Table of Contents

Epistemological position

This work is pragmatic and contextual: the claims to produce evidence turn on the plausible utility within the contexts of evidence of clinical psychology, psychiatry and psychotherapy.

General statistical approach

This is a quantitative programme of Practice Based Evidence (PBE, e.g. Margison et al., 2000) work largely in the spirit of “realist evaluation” (e.g. Van Belle et al., 2016)and not an Evidence Based Practice (EBP) Randomised Controlled Trial (RCT). The design is a cohort follow-up study and the naturalistic data collection means that the statistical methods are exploratory and descriptive even where null hypothesis testing will be used. There is no randomisation or experimental control meaning that there will always be multiple ways of explaining associations and the design is not focused as in an RCT on a single primary contrast of interest and a small set of secondary tests. The terminology of sample and population will be used and estimation methods (implicitly of population parameters) will be central, and population hypothesis tests used where useful and meaningful. As participation was voluntary the data are inevitably, ethically, incomplete and not a true, complete census,

1. Multiple tests and p-value hunting/hacking

We recognise that there are no perfect ways of exploring such data. There are many areas where exploration of associations, e.g. between plausible predictor and response variables can usefully be informed by null hypothesis testing. As there are many such associations, this raises the classical "multiple tests" issue: that if all the null hypotheses about the population were true (and independent) then the expectation of a single test being "significant" at the alpha level (.05) rises rapidly, and markedly above .05, as the number of tests rises. This issue has of course been know from the invention of hypothesis testing, for a readable early discussion in relation to mental health research see (e.g. Grove & Andreason, 1982). This leads to two related issues: the simple one of the elevated "studywise" false positive rate, and the issue of "p-value hunting/hacking" (touched on as part of the recent resurgence of interest in trying to reduce the dangers of naïve use of hypothesis testing, e.g. (Amrhein et al., 2019; APA, 2016; Wasserstein & Lazar, 2016).

There is an extensive literature with arguments for and against application of methods to reduce the overall "studywise" false positive rate to .05, and various methods by which to do this (Primo de Carvalho Alves & Sica da Rocha, 2019 e.g.) . Given the diversity of areas in which we will be exploring associations and where null hypothesis significance testing (NHST) has some value as a decision rule, we do not propose to use any of the methods of controlling for the inevitably elevated false positive rate. However, we do propose to be consistent and clear in recognising the risk. All hypothesis tests will be reported with the actual p value, not as "p < .05" *versus* "NS" (or some other ordinal categorisation of p values by asterixes). Where possible, 95% confidence intervals (CIs) will be reported for all statistics. To be precise, we will report the Bias Corrected Adjusted, (bca) 95% CI from 1000 bootstrap replications unless it is not available in which case the percential CI will be used, see (Carpenter & Bithell, 2000; Puth et al., 2015).

That issue of the inflated type I error rate that is bound to arise from there being multiple questions of interest leads to the option for deliberate or uninformed "p-value

hunting/hacking" with the considerable risk of severely misleading findings. We recognise the multiple tests issue and wish to conduct analyses as likely to result in replicable findings as possible. We also see the only protection against the accusation of p-value hunting/hacking is to be as clear as possible about the distinction between planned, i.e. *a priori* analyses, and "emergent" ones, i.e. those conducted in exploration of, and intended to throw more light on unexpected findings. With a very complex clinical service, multiple important aims and an extensive dataset, there are an almost infinite number of possible analyses and a very large number of important analyses apparent *a priori*. This makes it challenging even to summarise the planned analyses; hence the length of this, Data Analysis Plan (DAP) which is itself a shortened "freeze" of a much larger document. This document is being submitted with the study protocol paper and designed to complement that and enable a moderately thorough coverage of the *a priori* analyses and the methodological/statistical issues that would otherwise overwhelm the paper. In line with our epistemological position, this document has been written trying to be readable to research interested clinicians as well as clinical researchers, the target context for findings, not for statisticians and many technical statistical issues are presented as simply as possible.

2. Planned and emergent analyses

Many analyses will be conducted to address the study objectives and these are summarised below. As noted, the complexity of the data will certainly lead to the performance of emergent analyses. These will always be defined as such, against this DAP, in any presentations, reports or papers. All publications will be made with the (Altman & Moher, 2013) transparency statement which will apply to the distinction between *a priori* and emergent analyses as much as to all important issues to be acknowledged transparently.

The analysis timing has been designed in three tranches.

- The first tranche will focus on the baseline data for participants recruited during the first three years. The most important focus will be to establish the baseline distributions of variables that will be used later to measure change, the psychometric properties of

those measures, the differences between clinical centers, and any strong associations between baseline variables likely to raise issues about confounding in the subsequent analyses of change.

- The second tranche will contain all data collected by the end of October 2020: the endof the intensive data collection phase. When analysing these data, the focus will be on change: changes in sociodemographic variables (assessed annually) and clinical interventions but primarily changes in the key clinical outcome variables of the questionaire scores and BMI.
- The third tranche will add the follow-up information for all the participants recruited to October 2020 as and followed-up, where in continued clinical contact with the services, to November 2025. The main analyses for this tranche will largely follow the explorations of change, its diversity and its predictors as from the second tranche.

Here we present the a priori analyses planned for each objective and broken down by tranche.

3. Primary objective: Effectiveness of ITAMITED as it is applied routinely in its natural context.

We will analyse effectiveness with three complementary analyses performed for data/analysis tranches 2 and 3.

(a) **Baseline/termination change**. This is has been the traditional focus of many existing explorations of effectiveness of treatment. We will report both mean change and also effect sizes reporting Cohen's d2 and dz (Cohen, 1988) (the ratio of mean change to baseline SD and the ratio of mean change to SD of change respectively) and report these with bootstrap CIs. As well as reporting the global aggregate mean shifts with bootstrap 95% CI for each score. We will report the regression of termination scores on baseline scores and the distributions of change scores.

We will explore the impact of baseline variables: gender, age, diagnosis, demographic

variables, medication use and service on change. These predictor effects will be explored entering them as predictors of termination scores/values in ANOVA models with attention paid to the almost certain violation of assumptions of the ANOVA and exploration of mitigation of those violations. As gender and age are such crucial variables in Eating Disorders (Eds) the effects of entering these as predictors (or analysing separately be gender) on effects of other variables will be reported.

To address individual change we will report the Jacobson plot of termination against baseline scores for all scores and report the RCSC breakdown of change on each measure where it is possible to define the CSC and RCI.

(b) **Multilevel models (MLM) of rates of change**. MLM allows analysis of change slopes across diverse durations of therapy/ascertainment and the method handles the non-independence of observations arising from the repeated measures within participant appropriately. These analyses can give a much more nuanced picture of change than baseline/termination analyses while allowing exploration of the effects of the same predictors and covariates analysed in (a). As for (a), durations of therapy are not directly reflected in the analyses though in contrast to (a), all data across all durations of data collection are used rather than just the first and last values.

We will analyse fit to linear, square root and log-linear transforms of time as these have been the most widely reported linearised change/time patterns in the literature. The analyses against the baseline predictors (as fixed predictors) noted above allow some exploration of heterogeneity of change trajectories against those measured variables. Diagnostics after overall MLM analyses will be considered with the earlier analyses (i.e. tranche 1 baseline analyses, and baseline/termination analyses from tranche 2 above) as possible indicators of subgroups differing fundamentally in pertinent statistical characteristics. These subgroup indications will inform analyses exploring these subgroups as possible indicators of mixture model population structures of change.

However, heterogeneity may also relate to variables not measured in this study. (There are a very large number of such potential predictors, no realistic study can measure them all.) In order to explore possible heterogeneity other than arising from measured predictors we will conduct two other analyses of change trajectories relating to MLM estimation: latent profile analysis (LPA) of change and nearest neighbour prediction of change.

There are two complicating issues that will have to be considered here. The first is that not all centres were active at the start of data collection so analysis will be checked for any impacts of this, i.e. confounding by centre. The other challenge arose from the coronavirus pandemic. Social distancing regulations forced all but inpatient therapy to online therapy for (currently one) period in 2020. Time profiles will be inspected carefully against dates as well as against time in therapy to check on the impacts of this and in order to decide how best to prevent this creating artefactual impacts on trajectory analyses.

(c) **Survival analysis** (e.g. Allgulander & Fisher, 1986; Singer & Willett, 2003; Willett & Singer, 1993)for time to terminations by category of termination. These analyses focus on the time spans between entry into treatment and termination which are not so focal in (a) and (b). The particular strength of survival analysis for a naturalistic cohort follow-up study is that it address the inevitable right censoring of follow-up. That is that at termination of intensive data collection some participants will be still in treatment so their "outcome" is only know to this point and how long each has been in treatment to that point will be very varied with some of the very first participants joining the study potentially still in treatment near three years on though many of those early participants will have terminated treatment so not be "right censored". Of course far more of the recently recruited participants will be still in treatment and their change only known to that point.

The uncertainties arising from this censoring of outcomes is considered in, and visible in, the findings of survival analyses. We will analyse time to completion of therapy by category of termination and also analyse how baseline predictor variables impact on the slope of those curves. Different survival analytic methods make different assumptions about the relationships between predictors and the time curves, notably whether there is "parallelism" of curves across groups, or not, and these issues will be explored, and the appropriate analyses reported.

Secondary objectives:

1. Describe the study participants.

To establish the baseline distributions of variables (i.e., BMI, CORE, EAT-26/ChEAT-26, and BITE) that will be used later to measure change (tranche 1). The aims are:

- To describe the distributions of scores at baseline on the change measures and how these relate to other variables:
  - demographic variables, primarily age and gender;
  - clinical variables primarily diagnosis, care type and medication and,
  - both because of the interest in this within ITA, and for the MLM analyses of change (above) and the differences between clinical centres.

2. Study the psychometric properties of the self-report measures.

To explore the baseline, cross-sectional psychometric properties of the self-report measures (tranche 1). The analyses will be as follows.

- Acceptability in terms of omitted items. Where measures have recommended prorating for some proportion of missed items the proportions proratable will be reported. If omission rates are high the possible effects of prorating, and the possible biases if omission is associated with any particular demographic or clinical variables, or tends to occur at higher or lower scores across the completed items will be note.

- Floor and/or ceiling effects for items.
- Internal reliability (Cronbach alpha (Cronbach, 1951) and McDonald's Omega (Revelle & Zinbarg, 2008) of scores both for the total participant group completing all items, but also by gender and diagnostic groups and by age for the young participants (as there is evidence (e.g. Twigg et al., 2016) that reliability, as well as score distributions, can vary importantly and statistically significantly with gender and age through adolescence). There has been much argument against the use of Cronbach's alpha, much of it accusing it of doing things Cronbach was careful in 1951 to say it didn't do, fundamentally these criticisms are that it can both underestimate and overestimate classical test theory (CTT) reliability, the latter particularly when a measure is not unidimensional. McDonald's omega corrects the latter issue particularly but alpha has been, and may remain, the dominant statistic reflecting internal reliability hence we will report both.
- Construct validity in terms of factor analyses within and across measures. We believe that factor analytic methods are often overvalued when applied to short, broad coverage measures of diverse problems, measures designed as much to be good within-individual reflections of change as to be good between-individual comparators. However, both in clinical practice and research all the self-report measures used in the study are used both for comparisons across individuals and to measure change over time hence it is important, as with reporting the internal reliability, to explore factor structure. The within measure exploration will use confirmatory factor analysis (CFA) where either there are widely used scale scores for the measure or where there has been debate in the literature about whether a measure (or scale/score) is unidimesional or has more factors of interpersonal difference. As the indicators (items) are short ordinal scales and as the latent variables are almost certainly not Gaussian in distribution, we will use both maximum likelihood (ML) and diagonally weighted least squares (DWLS) (Li, 2016; Rhemtulla et al., 2012) estimation and appropriate scaled/robust fit

statistics (Satorra & Bentler, 2010) for the estimation method. Assuming the DWLS and robust fit statistics give better fit these will be reported with a note about the ML fit. If numbers allow (n > 250) separate analyses will be conducted by gender, age and diagnosis.

We will explore factorial measurement invariance across first and second completions (exempting items clearly impacted by the restrictions of hospitalisation). We will also explore invariance between groups by gender, sensible age groups (determined partly by cell sizes, here again n > 250) and diagnoses (again, restricted by sufficient cell sizes). For both temporal and between group analyses we expect only quite low levels of invariance as we believe there are very real changes and differences in structure, hence the focus will be more on the magnitude of the impacts on scores of the invariance rather than whether statistically significant invariance is or is not present.

If fit to the scale design is particularly bad, exploratory factor analyses will be conducted and reported.

As well as within-measure analyses, we will also conduct across-measure factor analyses as these allow construct validity to check on where one might expect shared factor structure (i.e. for (Ch)EAT and BITE) or might expect largely unrelated factor structure (i.e. between (YP-)CORE and both (Ch)EAT and BITE). These analyses will be CFA by design structure and EFA if fit to design structure is clearly very poor.

- If some measures/scores show clear fit to unidimensionality, we will conduct Item Response Theory (IRT) analyses to identify cutting points and potential evidence that the conventional scoring of the measures may be suboptimal for cross-sectional comparisons.
- Convergent validity between related measures (EAT-26/CHEAT-26 and BITE) and convergent and divergent validity between them and diagnoses and BMI. These are more conventional explorations than the item level across-measure factor analyses noted above. We will explore scatterplots for marked non-linearity or

clearly non-zero intercepts and, assuming there are not major issues of those kinds, will report Spearman rho and Pearson R correlations with bootstrap confidence intervals. Reporting both correlations maximises comparability with existing literature as Spearman correlations have often been reported in the past, before the widespread availability of bootstrapping, as score distributions are generally not Gaussian. If numbers allow (n > 80) separate analyses will be conducted by gender, age and diagnosis.

- Divergent validity/confounding against gender, age, first language, education. Distributions will be compared by groups (with sensible age groups) and the simple scattergrams against age inspected. Unless clearly inappropriate Pearson correlations with age, and mean differences by groups, will be reported with bootstrapped CIs.

To explore the psychometrics of change (tranche 2 and 3):

- Where participants have sufficient completed measures (n > 20) we will to apply P-technique estimation of Cronbach's alpha within individuals, within and across measures. For the small number of participants with the larger number of sufficient completions (n > 40) we will also explore P-technique "factor" structure (actually Principal Component Analyses (PCA)) to show the structures that emerge where the sequences of completed measures within individuals allow this.
- As well as individual P-technique we will use "chained P-technique" to look for shared intraindividual factor structure of change. These methods are essentially those of the exploration of measurement invariance by CFA with repeated completions over time within individuals being compared for structure.

3. Identify change patterns by predictors.

Analysis of temporal/cohort changes in the variables across all the data collection (tranches 2 and 3). The methodology for this exploration is all within the MLM described above, and the extension of the MLM with LPA and NNA.

4. Evaluation of the care types (inpatient, day hospital, and outpatient).

As with diversity of change, this aim will be explored largely through baseline/termination change broken down by types of care and, if numbers are large enough (to be explored by simulation) also by MLM and survival analysis with type of care as predictor (tranches 1, 2 and 3).
However, this is clearly not a simple predictor/response relationship and there are two main complicating issues: that participants can have multiple trajectories through types of care, and that the types of care offered will reflect clinical need/severity and, probably to a lesser extent, demographic issues including age, occupation and carer responsibilities.

We expect that only a minority of the participants will receive only a single type of care as almost all who start in inpatient or day hospital care will progress to a less intense type of care and inpatient care unless they are very late recruits (and even then, transitions will be seen in the second, follow-up, phase of data collection). Inpatient care will often be succeeded by day hospital care and that in turn by outpatient care. A further complication is that a subset of participants will have "upward" as well as "downward" transitions and the diversity of trajectories across types of care will only permit robust MLM if only the major and simplest trajectories are analysed treating levels of care as time dependent covariates. For this reason, analyses will be only of the "simple transitions" but presented with clear caveats about the dangers of extrapolating overconfidently to all future clients starting in any of the three care types without taking the real complexities of transitions into account.

The second challenge, that types of care offered will undoubtedly be affected by severity (hence probably strongly correlated with scores on measures) and by some aspects of demography, can only really be explored in terms of associations with scores and with

demographics at baseline by type of care, and for the categories of later care type transitions.

5. Describe the main challenges implementing PBE.

We believe this DAP underlines many of the challenges of *analysing* PBE data. One particular challenge, affecting all the above analyses, is that of non-participation. The relatively small issue of missed items has been noted above but other missing data is likely to be far more important. For all the analyses described above we will describe initial refusals and withdrawals and will also report data loss by exclusions and administrative errors and explore how these may have affected data.

Particular attention will be paid to cohort/time changes in missingness, and to differences between centres. It is recognised that recruitment since the coronavirus (CV-19) pandemic will have been affected and this will be studied specifically against the timing of the public health measures imposed and how ITA reacted made operational changes. These analyses are expected to be conducted in all the three tranches.

Given that none of the missingness is likely to be truly "Missing at Random" (MAR) let alone "Missing Completely at Random" (MCAR) we believe that the sophisticated imputation methods, such as Multiple Imputation by Chained Equations (MICE), which are powerful and unbiased for MCAR, and strong for some MAR models, may have limited or no advantages over crude substitutions (means across all participants for the missing data, or means for that participant) or simple deletions (listwise/casewise) all of which can introduce strong biases into analyses. We will use methods, in the light of the observed patterns of missingness, and report their impacts on analyses as robustness checks rather than as solutions for missingness.

In addition to these data analytic challenges, many of which are far more severe than for analysis of trial data, and are consequently less explored, there are simple challenges of implementing PBE data collection without substantial research funding. We will describe qualitatively, the challenges and how we sought to manage and mitigate them, where possible.

Where, as with missingness, quantitative data throw light on success or failure of mitigation, these data will be reported.

References

Allgulander, C., & Fisher, L. D. (1986). Survival analysis (or time to an event analysis), and the Cox regression model—Methods for longitudinal psychiatric research. *Acta Psychiatrica Scandinavica*, *74*, 529–535.

Altman, D. G., & Moher, D. (2013). Declaration of transparency for each research article. *BMJ*, *347*(aug07 3), f4796–f4796. https://doi.org/10.1136/bmj.f4796

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

APA. (2016, March). *P-values under question*. Https://Www.Apa.Org. https://www.apa.org/science/about/psa/2016/03/p-values

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*, 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Grove, W. M., & Andreason, N. C. (1982). Simultaneous tests of many hypotheses in exploratory research. *Journal of Mental and Nervous Diseases*, *170*, 3–8.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Mellor-Clark, J., Audin, K., & Connell, J. (2000). Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *The British Journal of Psychiatry*, *177*(2), 123–130. https://doi.org/10.1192/bjp.177.2.123

Primo de Carvalho Alves, L., & Sica da Rocha, N. (2019). The harm of adjusting for multiple statistical testing in psychiatric research. *Acta Psychiatrica Scandinavica*, *140*(6), 586–588. https://doi.org/10.1111/acps.13103

Puth, M.-T., Neuhäuser, M., & Ruxton, G. D. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology*, *84*(4), 892–897. https://doi.org/10.1111/1365-2656.12382

Revelle, W., & Zinbarg, R. E. (2008). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Satorra, A., & Bentler, P. M. (2010). Ensuring Positiveness of the Scaled Difference Chi-square Test Statistic. *Psychometrika*, *75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Twigg, E., Cooper, M., Evans, C., Freire, E. S., Mellor-Clark, J., McInnes, B., & Barkham, M. (2016). Acceptability, reliability, referential distributions, and sensitivity to change of the YP-CORE outcome measure: Replication and refinement. *Child and Adolescent Mental Health*, *21*(2), 115–123. https://doi.org/10.1111/camh.12128

Van Belle, S., Wong, G., Westhorp, G., Pearson, M., Emmel, N., Manzano, A., & Marchal, B. (2016). Can "realist" randomised controlled trials be genuinely realist? *Trials*, *17*(1), 313, s13063-016-1407–0. https://doi.org/10.1186/s13063-016-1407-0

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should and how you can use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology*, *61*(6), 952–965.