Combining causal inference and within-trial economic evaluation methods to assess comparative cost-effectiveness using real-world data: a tutorial with recommendations based on the quasi-experimental ADAPT study of a redesigned mental health service

Matthew Franklin (matt.franklin@sheffield.ac.uk)
University of Sheffield https://orcid.org/0000-0002-2774-9439
Alice Porter
University of Bristol https://orcid.org/0000-0001-5281-7694
Frank De Vocht
University of Bristol https://orcid.org/0000-0003-3631-627X
Benjamin Kearns
Lumanity https://orcid.org/0000-0001-7730-668X
Nicholas Latimer
University of Sheffield https://orcid.org/0000-0001-5304-5585
Monica Hernández Alava
University of Sheffield https://orcid.org/0000-0003-4474-5883
Tracey Young
University of Sheffield https://orcid.org/0000-0001-8467-0471
Judi Kidger
University of Bristol https://orcid.org/0000-0002-1054-6758

Method Article

Keywords: mental health, real-world evidence, real-world data, causal inference, economic evaluation, quality-adjusted life-years, health technology assessment

Posted Date: September 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3317541/v1

Abstract

OBJECTIVES. Real-world evidence is playing an increasingly important role in health technology assessment, but is prone to selection and confounding bias. We demonstrate how to conduct a real-world within-study cost per quality-adjusted life-year (QALY) analysis. We combined traditional within-trial bootstrapped regression-baseline-adjustment with causal inference methods, using a Target Trial framework, inverse probability weights (IPWs), marginal structural models (MSMs), and g-computation, applied to England's Talking Therapies for anxiety and depression services (TTad) mental-health e-records.

METHODS. The 'Assessing a Distinct IAPT service' (ADAPT) quasi-experimental-study evaluated an Enhanced-TTad-service Vs. TTad-services' treatment-as-usual. TTad-services collect patient-reported PHQ-9-depression and GAD-7-anxiety scores at index-assessment and each treatment session, from which we predicted EQ-5D utilities using a mapping function. Our primary estimands were incremental costs and QALYs for Enhanced-TTad Vs. treatment-as-usual at 16-weeks post-TTad-service-index-assessment.

We prespecified our target trial including eligibility, treatment strategies, assignment procedure, follow-up, outcomes, estimands, and analysis plan. We used stabilised treatment-related and censoring-related IPWs within MSMs to reduce selection and confounding bias due to non-randomised treatment allocation and informative censoring, respectively. Our doubly-robust approach involved MSM-adjusted baseline confounders and g-computation to estimate incremental utilities, costs, and QALYs, with bootstrapped bias-corrected 95% confidence-intervals (95%bCIs) and cost-effectiveness acceptability curves.

RESULTS. Primary analysis sample: Enhanced, N=5,441; treatment-as-usual, N=2,149. Naïve regression-baseline-adjustment and doubly-robust approaches suggested Enhanced-TTad-service dominated treatment-as-usual, with average per-person (95%bCIs) cost-savings of £30.64 (£22.26 to £38.90) or £29.64 (£20.69 to £37.99) and QALYs-gained of 0.00035 (-0.00075 to 0.00152) or 0.00052 (-0.00105 to 0.00277), respectively; probability of cost-effectiveness at £30,000 per QALY was 99% or 95%, respectively. The doubly-robust and naïve results concurred; albeit, the doubly-robust results suggested average QALY gains were higher but less certain. The cost-effectiveness results were driven by potential cost-savings.

CONCLUSION. When treatment allocation is non-randomised, the Target Trial framework alongside doubly-robust analyses should be used to reduce selection and confounding bias.

Article Summary

We describe how to conduct a causal economic evaluation using mental health real-world data based on the 'Assessing a Distinct IAPT service' (ADAPT) study

Highlights

What is already known about the topic? [TBC]

What does the paper add to existing knowledge? [TBC]

What insights does the paper provide for informing health care-related decision making? [TBC]

1. Introduction

Randomised controlled trials (RCTs) are considered the gold standard for causally estimating comparative average treatment effects (ATEs) of alternative care interventions to inform evidence-based decision-making (1, 2). RCTs are key vehicles for cost-effectiveness analyses (CEA), for example based on within-trial economic evaluation methods to account for the costs and consequences of alternative care strategies to assess value-for-money (3–7). However, issues with RCTs include poor external validity, randomisation not always being possible (e.g. unethical), and long-time and high-cost requirements (1, 8–10). As such, there is a need to identify ways to overcome the shortcomings of RCTs, but any alternative must produce an appropriate and trusted evidence-base like RCTs; for example, by emulating RCTs within non-randomised studies based on real-world data (RWD) including electronic health records (EHRs) and registries (9–12).

RCT evidence informs health technology assessment (HTA) processes used to guide evidence-based recommendations about treatments and procedures; for example, by the National Institute for Health and Care Excellence (NICE) in England and Wales (13). NICE's real-world evidence (RWE) framework (published June 2022) states: "Non-randomised studies can be used to provide evidence on comparative effects in the absence of RCTs or to complement trial evidence to answer a broader range of questions about the effects of interventions in routine settings" (8). NICE's RWE framework defines RWE as "evidence generated from the analysis of [RWD] relating to patient health or experience or care delivery collected outside the context of a highly controlled clinical trial[, e.g. non-randomised studies]"(8). Timely and cost-efficient RWD studies are needed, for example related to mental health as outlined in the UK's NHS Long Term plan (14); however, these studies need to be conducted and reported appropriately, for example by following NICE's RWE framework (8).

NICE's RWE framework details structural (e.g. study design) and analytical causal inference methods for reducing/removing or quantifying bias within any within-study derived comparative effect estimates (8). Appendix S1 defines confounding, selection, and information bias based on Hernán and Robins (15) and NICE's (8) definitions; alternatively, refer to the Catalogue of Bias (www.catalogofbias.org). For non-randomised studies based on RWD, it is possible to emulate trial designs by using the Target Trial (TT) framework which reduces bias by clearly articulating seven study dimensions: eligibility criteria, treatment strategies, assignment procedure, follow-up period, outcomes, estimand(s), and analysis plan (12, 15). The TT framework is intended to mimic the ideal 'pragmatic' RCT, which underpins the Cochrane ROBINS-I risk-of-bias tool for non-randomised studies (8, 16). A particular key issue for non-randomised studies compared to RCTs though is confounding, whereby for RCTs randomised treatment allocation avoids confounding at study-entry. As analytical examples when adjusting/conditioning on measured confounders, baseline covariates can be used within traditional adjustment methods including restriction, stratification, exact matching, multivariable regression and propensity-score methods (8, 17). There are also 'doubly-robust' approaches, for example by combining the propensity-score and regression methods means only one need be correctly specified to obtain an unbiased effect estimator (18, 19). However, it is not yet wholly common to combine a range of causal inference methods with economic evaluation methods to conduct CEAs alongside non-randomised studies (6, 7, 20–23); e.g., for cost per quality-adjusted life-year (QALY) analyses as recommended by NICE and HTA agencies internationally (4, 5, 24). For example, Krief's (23) review of CEA statistical practice to address selection and confounding bias when using observational data identified studies mainly used regression alone (51%), with 22% using propensity-score methods. Similarly, a systematic review of non-randomised cardiology CEAs by Guertin, Conombo (22) found only 45% of identified studies adjusted for confounding at all. Both Kreif, Grieve (23) and Guertin, Conombo (22) recommend doubly-robust methods which weren't generally used in their identified CEA studies; additionally, only methods such as inverse probability weights (IPWs) can address time-varying confounding and there are other biases which require consideration/controlling (see Appendix S1).

Our aim is to describe and demonstrate how to conduct a real-world within-study cost-per-QALY analysis. We combine traditional within-trial bootstrapped regression-baseline-adjustment with causal inference methods recommended by the NICE RWE framework, including using a Target Trial (TT) framework, directed acyclic graphs (DAGs), inverse probability weights (IPWs), marginal structural models (MSMs), and g-computation – we describe these methods in reasonable detail to aid the naïve reader understand the basic need and principles that underpin these methods, with appropriate references for further reading and statistical software code. We demonstrate the application of these methods using England's Talking Therapies for anxiety and depression services (TTad) mental-health EHRs for the 'Assessing a Distinct IAPT service' (ADAPT) study – IAPT (Improving Access to Psychological Therapies) is the previous name of TTad-services.

2. NHS Talking Therapies for anxiety and depression services and the ADAPT case-study

NHS England's TTad-service 2021/22 annual report stated 1.8-million people had been referred to TTad-services that reporting year, up 24.5% since 2020/21 (25). Common mental health disorders, like anxiety and depression, affect ≈ 15% of the population, associated with high mortality levels and years lost with disability (26, 27). TTad-services offer a variety of NICE recommended psychological therapies and routinely collects data on its service-users to inform its key performance metric reports, for which national TTad dataset metadata and annual reports are available online (25, 28, 29). TTad health-outcome metrics are based on the service-user-reported Patient Health Questionaire-9 (PHQ-9) depression-scale and Generalised Anxiety Disorder (GAD-7) anxiety-scale from 0 (best state) to 27 or 21 (worst states), respectively, collected at index assessment (i.e. baseline) and each treatment session until service discharge (29). One such health-outcome metric is 'recovery' since baseline, quantified as moving from 'caseness' (GAD-7 ≥ 8; PHQ-9 ≥ 10) on either measure to 'no caseness' (GAD-7 < 8; PHQ-9 < 10) on both measures up to service discharge (29).

In 2021, a new Health and Wellbeing pathway was introduced into the TTad-service in one geographical area of the UK to address the wider determinants of mental health problems (30). This 'enhanced' TTad-service was developed and delivered in three local authorities in South-West England. In the enhanced TTad-service after index assessment for caseness before waiting-list allocation, service-users could be referred to TTad treatment-as-usual (TAU) only (talking therapies, e.g. psychotherapy), TTad plus the enhanced service, or the enhanced service only. The enhanced service consisted of two elements. First, the 'Healthy Living Healthy Minds' programme as a six-session group webinar series offering guided exercise and advice on healthy lifestyles. Secondly as an 'and/or', one-to-one sessions with a 'Wellbeing Navigator', who facilitated access to community organisations to address the wider psychosocial problems individuals were experiencing (e.g., poverty, unemployment, social isolation). Appendix S2 provides a visual representation of the enhanced pathway.

For the ADAPT study, the enhanced TTad-service (intervention) in South-West England was compared to the same TTad-service before the enhanced service was introduced in March 2021 (historical-control), and a standard TTad-service with TAU in South-East England (geographical-control). The intention was to estimate the effectiveness (reported elsewhere) and cost-effectiveness (reported here) of the enhanced service. A published qualitative study provides additional detail about the enhanced service (30). Appendix S3 provides an overview of the TTad-service data requested for the ADAPT study, including data cleaning and structuring. The Health Research Authority and Health and Care Research Wales granted ethical approval (21/PR/0230).

3. Structural and analytical methods to account for bias in the economic evaluation

This within-study CEA focuses on the NICE reference case of cost-per-QALY. The primary comparator is the geographical-control due to TTad-service data coding being the same as for the intervention-group (version-2), whereas the historical-control used an older coding system (version 1.5) which may have influenced the comparative cost analyses (see Appendix S3). The costing perspective is from a TTad-service perspective in 2020/21 Great British Pounds (GBP £), rather than NICE recommended health and social care perspective due to using TTad-service EHR data only. Appendix S4 describes resource-use parameters, unit costs, and associated cost calculations/assumptions.

We followed NICE's methods guide and RWE framework, CHEERS and ROBINS-I tools, and recommended methods for handling utilities, costs, missing data and informative censoring using Stata version 17 (6, 7, 31–39). For drawing DAGs, we used DAGitty: https://dagitty.net/. Application of our analytical methods were particularly informed by Stata code provided by Smith, Mansournia (40), Fewell, Hernán (41), Gabrio, Plumpton (42), among short-courses described in our acknowledgements.

3.1. Target Trial (TT) specification including causal contrasts and estimands

Table 1 provides an overview of our *a-priori* TT specification based on the recommendations of Hernán and Robins (12), with an extended rationale for our TT specification provided in Appendix S5. Hernán and Robins (12) TT framework refers to 'causal contrasts' such as specified intention-to-treat (ITT) or per-protocol (PP) analyses; however, the related estimand framework has grown in importance and popularity to more clearly state what is intended to be estimated from any specific analysis (43–45). The estimand framework's usefulness and importance for patient-reported outcomes (e.g. PHQ-9 and GAD-7), partly to provide a common and transparent language, is described by Lawrance, Degtyarev (46). As the economic evaluation outcome of interest are QALYs, we used mapping functions by Franklin and Hernández Alava (47) to predict EQ-5D-5L cross-walked utilities for the EQ-5D-3L UK value set from PHQ-9 and GAD-7 scores, age, and sex recorded in TTad-service data. The ADAPT economic evaluation primary estimand was defined as:

In new referrals to TTad-services, what is the between-group difference in mean TTad-service costs and QALYs accumulated since index assessment (i.e. baseline), with QALYs based on EQ-5D-5L crosswalk utilities predicted from PHQ-9 and GAD-7 scores, age, and sex, for those referred to the enhanced TTad-service compared to treatment-as-usual (TAU) for those within a geographical-control-site up to 16-weeks after baseline, regardless of TAU received and service discharge?

For shorter descriptive purposes, we define other analyses on traditional principles of ITT and PP. Our primary and secondary ITT analyses are based on an ITT principle that the enhanced service is available for anyone referred to the intervention site; therefore, the ITT sample is all new referrals to the interventions-site if they then received TAU and/or enhanced elements, among other eligibility criteria (see Table 1). In contrast, our PP analyses for the intervention-group is focussed on only people who engaged with the enhanced elements, i.e. people who were in the intervention-group who received TAU only were omitted from these PP analyses.

3.2. Directed Acyclic Graphs (DAGs)

DAGs are an increasingly popular approach for specifying causal pathways and identifying confounding variables that require conditioning when estimating causal effects, for which there is published guidance (48). NICE's RWE framework recommends DAGs when rationalising the causal analyses conducted (8).

DAGs can be developed through structured and repeated discussions between researchers, clinicians, and other relevant experts to understand the causal pathways between intervention/exposure, outcome, and other covariates which may be a confounder, mediator, or collider (see Appendix S1) (49). It is important to identify confounders for conditioning which are variables (L) that causally affects both exposure (A) and outcome (Y), such that L affects A and Y, compared to variables that are mediators (A affects L and L affects Y) or colliders (A and Y affects L). Whereas conditioning on confounders is required to estimate the causal relationship from exposure to outcome; conditioning on a collider or mediator should be avoided, which can be avoided by not conditioning on any covariate recorded after baseline (i.e. post-exposure such as treatment allocation). Although this has complications when accounting for time-updated confounders, it is simpler when controlling for baseline (i.e. time-fixed) confounders. For transparent reporting purposes, our DAG was developed after some initial analyses but before the final analysis had been conducted, presented in Appendix S6.

3.3. Stabilised Inverse Probability of Treatment Weights (IPTWs)

Chesnaye, Stel (50) provides an introduction to adjusting for confounding using IPTWs, with 'best practice' recommendations by Austin and Stuart (51). The essence of IPTW is to balance the differences in prognostic and/or confounding characteristics between comparison groups by reweighting the outcome variable of these individuals by the inverse probability of the treatment actually assigned, i.e. the propensity-score (40). Propensity-scores can be estimated using logistic regression with treatment assignment as the dependent variable alongside other covariates that are prognostically important (related to outcomes) or confounders (related to treatment and outcomes) (51). So-called 'unstabilised' and 'stabilised' weights use the propensity-score as the denominator; however, stabilised weights use the marginal probability of being in the treatment actually assigned as the numerator (i.e. logistic regression with treatment allocation as the only covariate) compared to a numerator of '1' for unstabilised weights – see Smith, Mansournia (40) for example software code. Unstabilised weights tend to have larger values than stabilised weights, forcing the variance to increase and exacerbate the uncertainty of the ATE estimation; as such, stabilised weights are used more in practice (40)

In cross-sectional data, regression adjustment, propensity-score-matching, or IPTW may yield very similar results – for longitudinal data though the advantage lies with IPTW: regression adjustment is not possible if time-varying treatments and time-varying confounders are present (see Appendix S1) (52). Our IPTWs included baseline predicted utility score (e.g. predicted EQ-5D-5L cross-walked utility) and demographic variables (age, sex, ethnicity, Index of Multiple Deprivation [IMD] decile) as independent variables – no time-updated covariates were included due to data and data structure limitations. Balance between groups due to using IPTWs was checked using weighted standardised differences to compare means of baseline covariates and visually using graphical methods (e.g. to compare the propensity-score distribution), presented in Appendix S7 (51).

3.4. Stabilised Inverse Probability of Censoring Weights (IPCWs)

Censoring can be a missing data problem as the outcome isn't fully observed, and if censoring is informative (i.e. related to outcomes) it causes selection bias (8, 15). Informative censoring can be addressed using g-methods, such as inverse probability of censoring weights (IPCW) which are similar to IPTWs with similar assumptions, although for IPCWs we weight by the inverse probability of being observed (rather than censored) – both IPTWs and IPCWs are described (the latter as 'censoring weights') by Fewell, Hernán (41).

TTad-services only collect PHQ-9 and GAD-7 data up to the point of discharge because this data is collected to mainly inform treatment, therefore data post-discharge is not considered relevant nor necessary. However, for QALY estimation this leads to informative censoring as it is hypothesised that discharge is directly related to PHQ-9 and GAD-7 scores which are used to predict health utilities which informs the quality-adjustment part of our QALYs. IPCWs were only considered relevant for utilities/QALYs as costs post-discharge were considered zero costs (not censored) as part of the TTad-service-only costing perspective. Calculating stabilised IPCWs uses similar methods and the same covariates as our IPTWs, but for the IPCWs we use a 'observed/censored' rather than 'treatment (allocation)' binary variable used for the IPTWs. Additionally, for the IPCWs we used a restricted cubic spline to account for time (i.e. weeks since baseline) with 4 degrees of freedom for the time-based knots over the analysis period (e.g. 16-weeks). Appendix S7 describes and presents our IPCW checks.

3.5. Inverse Probability Weights (IPWs) within doubly-robust Marginal Structural Models (MSMs)

Marginal structural models (MSMs), introduced by Robins (53), are a class of causal (regression) models whose parameters are estimated using IPWs estimators (e.g. IPTWs) (54). A key benefit of IPWs within MSMs is that they can deal with time-updated covariates (specified for the IPW), specifically to address time-varying confounders and for analysing time-varying treatments (52). In the causal inference literature, this model is known as a MSM because causal models are often referred to as *structural* in the econometric and social science literature; at the same time, it is a model for the *marginal* distribution of the counterfactual variable (e.g. the counterfactual utility values associated with each treatment allocation) conditional on baseline variables, rather than the joint/conditional distribution of the counterfactual variables (15, 41). Although MSMs rely on the use of IPWs, the regression form we use can be described as either a Generalised Estimating Equation (GEE) or Generalised Linear Model (GLM). Specifically, we use GLMs with a Normal-distribution and identity-link for utilities, and Gamma-distribution with log-link for costs. When using longitudinal data, the use of GLMs can be referred to as GEEs, whereby GEE and GLM are equivalent when using/assuming an independent correlation structure. Hereafter to align with the causal inference literature for descriptive purposes, we refer to MSMs.

Both IPTWs and IPCWs are used within our MSMs by multiplying the IPWs together, for which we use the notation IPTCW (i.e. IPTCW = IPTW * IPCW). When IPWs are combined with baseline covariate adjustment in the MSM, this approach is defined as 'doubly-robust' as only the weights or the regression need be correctly specified to obtain an unbiased effect estimator (18). We use the same baseline covariates for direct-MSM-adjustment as we used for our IPWs. G-computation was subsequently also used as a parametric regression implementation of the g-formula, presented and described by Smith, Mansournia (40). The g-formula allows us to obtain an unconfounded marginal estimation of the ATE, whereas g-computation (simply put) is an algorithm to practically implement the g-formula (40). In essence for g-computation, we follow a two-step process: first fit the outcome MSM (i.e. for utility or costs) for the intervention sample only then predict the outcome for the whole analytical sample, then do the same again but using the outcome MSM for the control-group only. The ATE is the subsequent differences in MSM-predicted mean outcomes between intervention and control (40).

3.6. Calculating QALYs and total costs post-MSM-estimation

QALYs are typically calculated assuming a linear change between utilities overtime by using the trapezoid-based total area-under-the-curve (AUC) method as described by Hunter, Baio (7); note, we also utilise this linear change assumption as part of a linear interpolation to address interval censored utility values to inform the QALY calculation, as described in Appendix S3. Typically, observed utility values are used to calculate QALYs based on the trapezoid-AUC method, with the QALY (not individual utility scores) then being the outcome of interest for the regression model; it is then common (even in RCTs) to condition on baseline utility scores (7, 55). However, here utility scores are the outcome of interest with QALYs calculated post-estimation (i.e. post-regression-adjustment), for which the key reason is that the IPWs can be based on time-updated covariates and therefore the IPWs differ at each time-point such that we need to focus on the time-dependent utilities rather than cross-time QALYs, so QALYs can't be calculated until post-estimation of the adjusted utility scores.

To calculate QALYs post-estimation of the adjusted utility values, we used a week-based-time dummy variable in our MSMs as an interaction term with treatment allocation (other than when using g-computation when time was an independent covariate) such that utilities could be estimated on a per-week basis. Relatedly, we used individually-clustered standard errors to account for intra-individual correlation due to the repeated-measure nature of our longitudinal data. We subsequently used the rectangular-AUC method described by Gabrio, Plumpton (42) to calculate the QALYs from the week-based, adjusted utility values. The main difference in the AUC calculation is that whereas Hunter, Baio (7) trapezoids-AUC calculation is based on the (traditionally defined) sum of multiple trapezoid areas, the calculation can be re-written such that the rectangular-AUC is the sum of rectangular areas – see Gabrio, Plumpton (42) (their Appendix C). An advantage of the rectangular-AUC approach is that is can be used to calculate QALYs post-estimation as weighted linear combinations of the coefficient estimates from the utility-focussed MSM, which can't be done using the trapezoid-AUC as it requires the baseline utility score which becomes null post-estimation. Similarly, total costs are calculated based on weighted linear combinations of the coefficient estimates from the weekly-cost focussed MSMs.

3.7. Confidence intervals, bootstrapping, and cost-effectiveness acceptability curves

Bootstrapping with 1,000 iterations were used to calculate bias-corrected 95% confidence intervals (95% bCIs) for weekly-costs, weekly-utilities, total costs, and QALYs. When using IPWs with MSMs, both generation of the IPWs and the MSMs must be included in the bootstrap procedure to ensure that the appropriate redrawn analytical samples are accounted for in the whole process, e.g. you need to generate IPTWs for each redrawn analytical sample independently as a single IPTW-set will not achieve the necessary balance between comparison groups across all redrawn samples. For CEA, bootstrapping is common for estimating bCIs as it also enables generation of cost-effectiveness planes (CE-planes) and cost-effectiveness acceptability curves (CEACs). However, computationally time-consuming bootstrapped CIs are only used for specific analyses when needed (e.g. to generate CEACs and when using g-estimation as it accounts for the two-step process), otherwise the delta-method is a less computationally time-consuming and suitable method for estimating CIs.

3.8. Result robustness: sensitivity analyses

Our sensitivity analyses included comparing the naïve regression (i.e. GLM with no IPWs or g-computation) and doubly-robust approach (i.e. MSMs with IPWs and g-computation), and stepped-inclusion of different weights (i.e. IPTW, IPCW, then IPTCW) in the MSMs with/out g-computation. Alternative follow-ups were also considered over 12-, 20, and 24-weeks post-baseline, and alternative predicted utility scores for the EQ-5D-5L Value Set for England (VSE) and Recovering Quality of Life Utility Index (ReQoL-UI) UK value set (47).

Weighting-based sensitivity analyses can include restricting analysis to the 'common support' (i.e. restricted to the range of propensity-scores at which we observe both counterfactual group subjects) and further trimming based on the propensity-score which can gain accuracy and precision in mean estimates by excluding subjects with particularly large or small propensity-scores, but at the expense of efficiency as we are removing eligible subjects from the analysis (51): trimming is considered for the IPTWs at both the 5th and 10th centile.

4. Results

Table 2 provides baseline descriptive statistics for the intervention ITT-group and PP-group, and geographical- and historical-controls. Overall, all four groups were quite similar in terms of their means and/or distributional aspects related to age, sex, ethnicity, IMD decile, mental health scores, proportion with anxiety and/or depression caseness, and predicted health-related quality-of-life (aka. utility) scores. The PP-group who received an element of the enhanced service was 10.1% (N = 549) of the ITT-group (N = 5,441) and had slightly worse mean baseline mental health scores and slightly larger proportions of people with anxiety and/or depression caseness compared to the ITT group.

4.1. ITT- or PP-intervention vs geographical-control

Table 3 shows that in the primary analysis for the ITT-intervention-group, the naïve regression-baseline-adjustment and doubly-robust approaches suggested the enhanced service dominated treatment-as-usual, with average per-person (95%bCIs) cost-savings of £30.64 (£22.26 to £38.90) or £29.64 (£20.69 to £37.99) and QALYs-gained of 0.00035 (-0.00075 to 0.00152) or 0.00052 (-0.00105 to 0.00277), respectively; probability of cost-effectiveness at £30,000 per QALY was 99% or 95%, respectively. Although the estimated costs-savings were statistically significantly different, the QALY gains were not.

Table 3 also shows that compared to the ITT-intervention-group, the naïve regression-baseline-adjustment and doubly-robust approaches suggests that for the PP-sample who specifically received the enhanced elements (i.e. PP-intervention-group) that although similar cost-savings were estimated of £34.09 (£22.70 to £45.95) or £33.33 (£21.82 to £45.24), there was an estimated mean QALY loss of 0.00478 (0.00233 to 0.00729) or 0.00157 (-0.00322 to 0.00736), respectively; probability of cost-effectiveness at £30,000 per QALY was 31% or 46%, respectively. Similar to the ITT analyses, the cost-savings were statistically significant. The key difference with the ITT analyses was the PP analyses suggested a mean QALY loss; however, although the naïve regression suggested this was statistically significant, this was not confirmed by the doubly-robust analysis. Figures 1 and 2 presents the relevant CE-planes and CEACs, respectively.

4.2. ITT- or PP-intervention vs historical-control

The TTad-service data for the historical-control is based on an older coding system than the intervention-group: version 1.5 vs version-2, respectively. Although we hypothesised it shouldn't have an effect of QALY estimates/comparisons, there were some concerns with the appointments data/details which may impact the comparative cost analyses (see Appendix S3 for further details). Taking this concern into consideration, Table 3 suggests that the cost-savings against the historical-control were almost 2.5-times the size of that against the geographical-control; although the suggestion of cost-savings is confirmatory of the geographical-control comparison results, we opt not to focus too much on the cost-saving magnitude given our known concern with this historical-control comparison.

In terms of QALY differences, comparisons against the historical-control generally suggested higher mean QALYs than estimated against geographical-control. ITT-intervention comparisons suggest higher QALY gains which are now statistically significant compared to what was estimated against the geographical-control. PP-intervention comparisons suggest lower QALY losses which are non-statistically significant, noting against the geographical-control only the naïve regression PP analysis suggested statistically significant QALY losses.

4.3. Other secondary and sensitivity analyses

Given the methodological focus of the article, Table 4 presents how the results from the TT-intervention (and PP-intervention for comparison) versus geographical-control changes when introducing the different aspects which enhances our naïve regression-based analysis into the doubly-robust approach.

For example, compared to naïve regression alone, when IPTWs were introduced the mean QALY gain increased due to higher estimated QALYs in the ITT-intervention and lower estimated QALYs in the geographical-control (Table 4, analysis-1 vs analysis-2). Before IPTWs were used, we had unacceptable imbalance in our baseline utility score between ITT-intervention and geographical-control (0.643 vs 0.609, respectively; standardised difference: 0.251) whereas after weighting we had acceptable balance (0.627 vs 0.627, respectively; standardised difference: -0.002). As higher baseline utility limits ability for utility gain post-exposure, we hypothesise re-balancing baseline utility in particular may have led to the higher estimated QALY gain compared to naïve regression alone which may not have as sufficiently addressed this imbalance (56).

Alternatively, when IPCWs were introduced the mean QALY gain dropped although there was higher estimated QALYs in both ITT-intervention and geographical-control (Table 4, analysis-1 vs analysis-3). IPCWs address informative censoring, thus it makes logical sense that the QALYs would be higher for both groups if there had been no censoring and overall suggests this censoring had a bigger impact on the estimated QALYs in the geographical-control compared to ITT-intervention.

Although overall the change in QALYs (and costs) are minimal when stepping through the analyses, Table 4 provides an indication that the methods are adjusting the estimates in a way that post-hoc make logical sense when considering what each method is attempting to achieve. Additionally, an appropriately designed TT should limit the adjustments required through analytical methods, which may also explain why a minimal difference was observed when moving from naïve regression to our doubly-robust approach; although, both structural and analytical methods are required to adjust for potential bias, but also improve trust in the analysis conducted. A range of other secondary and sensitivity analyses are described and presented in Appendix S8.

5. Discussion

In the primary economic evaluation focussed on the ITT-intervention vs geographical-control, the doubly-robust and naïve results concurred that the cost-effectiveness results were driven by the potential for the enhanced service to provide cost-savings, with statistically insignificant QALY gains. Our PP-intervention analyses also suggested potential cost-savings, but suggested a mean QALY loss which was statistically significant for the naïve regression but insignificant for the doubly-robust approach. Although the historical-control has limitations particularly related to the comparative cost-analyses, the results generally suggested the same thing as the geographical-control comparison: potential for the enhanced service to provide cost-savings, but uncertainty around the QALY gains/losses dependent on the analysis and comparison. Additionally, our DAG suggests potential unmeasured confounding in our analyses which may have affected our results, partly due to the data available for analysis. Sensitivity analyses related to these results are described and presented in Appendix S8.

5.1. Structural aspects to account for bias including data structuring and cleaning

The TT framework's elements are perhaps commonly known for those used to working alongside RCTs, including those who conduct within-trial economic evaluation; it is the emulated observational study based on RWD and developing DAGs which are perhaps not as familiar. DAGs are a useful tool for identifying confounders for conditioning and being transparent around hypothesised causal pathways, which included a recognition that unmeasured confounding is a potential issue for our analyses. Although the usefulness of DAGs to choose/specify confounders for conditioning has been questioned, the transparency they provide around the covariates included (or not) in the causal analyses conducted is a valuable asset (48). The implications of unobserved confounding could be explored using analytical methods such as e-values and quantitative bias analyses, which were not used for this study mainly due to time-restrictions to learn and use another set of methods despite their usefulness (57, 58).

The time to clean and structure RWD is a considerable task which takes knowledge and expertise to accomplish (e.g. see Appendix S3), even more so when using linked datasets (e.g. TTad-service linked with hospital data). The ADAPT study data and associated structure was not perfect, meaning that some desirable analytical aspects could not be achieved, e.g. including time-updated covariates in the IPWs. As such there is potentially time-varying confounding that has not been accounted for in the analysis, mainly due to the data not being available/structured in a way to enable these analyses. There was no funded time to request further data and allow additional data cleaning and structuring which could have resolved/reduced this issue; however, our results are still informative for decision-makers, while recognising such limitations (Section 5.4).

5.2. Combining within-trial economic evaluation and causal inference analytical methods

For those familiar with conducting within-trial CEAs, it is perhaps g-methods (e.g. IPWs and g-computation) which are unfamiliar, as regression-based analyses and bootstrapping are common practice (6, 7, 20, 38). Baseline adjustment for covariates is still common for RCTs, because if a covariate is prognostic (i.e. related to outcomes) accounting for its effect will improve power to detect a treatment effect; in fact, IPTWs still have a place in RCTs to complement or replace direct regression-based adjustment (59). Although methods like using IPWs may represent a learning curve, their use are recommended/required when accounting for baseline confounding in non-randomised studies, and necessary if accounting for time-varying confounding. Although there are other g-methods (noting g-computation was also used in ADAPT), becoming familiar with IPWs and doubly-robust methods is a good-step forward to better conducting causal economic evaluation in non-randomised studies (40).

5.3. Limitations with real-world data for causal inference and economic evaluation

Reliance on single or even linked RWD sources (e.g. linked EHRs with surveys) have limitations, as there is only a finite amount of data which can be collected/reported at any given point and then used for public benefit (e.g. research) – it may never be enough for some analyses or commentators, but it may be sufficient to guide decision-making if used appropriately and built on for future, better research.

Even when RWD are used to supplement primary data collection in RCTs, a decision has to be made about which data should be based on questionnaires (e.g. person or proxy-reported) or EHRs, with some information needing perhaps to come from surveys/questionnaires (e.g. related to informal care) as such information is never/rarely recorded in EHRs (60). RCTs and RWD studies embedded within the confines of any given single or linked datasets are restricted to the data available, which requires making assumptions to utilise the data and/or having a restricted analysis perspective (e.g. restricted time horizon and costing perspective) (61). For example, for the ADAPT economic evaluation our costing perspective is that of the TTad-service only, whereas data linkage with other datasets (e.g. hospital data) could have expanded this perspective; however, it still wouldn't be sufficient to cover a whole healthcare and social care perspective desirable to NICE which would require extensive data linkage/cleaning/structuring, but such limited perspective analyses can still be informative. Modelling-based CEAs (e.g. decision-analytic models) allows evidence synthesis of a range of estimates which can overcome some of the limitations of within-study CEAs (e.g. account for broader outcome/cost implications over a longer-time horizons) (62, 63). However, causal within-study CEAs alongside non-randomised RWD studies still have a place to statistically identify the potential causal effect of treatments/exposures on costs and health-outcomes, rather than relying on mathematical models using existing estimates which may or may not have external validity/generalisability.

5.4. Limitations of causal analyses for non-randomised studies including ADAPT

Quantitative research constantly requires making (untestable) assumptions about the analyses we conduct and often requires making inferences within the limitations of the study, data, and analyses conducted. For example, here we stated unmeasured and time-varying confounding was not adequately accounted for and is a key limitation which may impact the inferences we can make from our estimates. Although, it is worth recognising that the vast majority of non-randomised studies making causal claims are potentially subject to some sort of confounding bias. There is still value in such evidence to guide decision-making, if sufficient structural and analytical processes have been used to account for such bias and limitations are transparent. It is also worth recognising that attempting to undertake every recommended 'ideal' structural and analytical method, including all possible sensitivity analyses, is a substantial undertaking for any researcher (as alluded to in this article given the substantial additional supplementary appendices available online). As such, pre-specified analysis plans should balance off the ideal analyses within time and budget restrictions, while making sure any time and budget is sufficient for an adequate analysis if not the ideal; subsequent additional sensitivity analyses can always be run if time/funding permits (see Appendix S8), rather than overreaching to do the 'perfect' analyses.

For the ADAPT study, it is justifiable to suggest there is evidence that the enhanced TTad-service has the potential to provide cost-savings (albeit the QALY gains are small and particularly uncertain) compared to TAU-only from the perspective of TTad-services, within the confines and limitations of the analysis conducted. As always, future research is required which can build from the strengths and limitations of this study.

6. Conclusion

When treatment allocation is non-randomised, the TT framework alongside doubly-robust analyses should be used to reduce selection and confounding bias. Although the consideration and use of such methods is more complex and time-consuming than the analytical stage of many (perhaps not all) traditionally conducted within-RCT CEAs, the approach and methods described in this article can aid people conduct better and more appropriate causal economic evaluation alongside non-randomised studies including those using RWD.

Declarations

Author Contributions: *Concept and design*: MF, FDV, JK. *Acquisition of data*: MF, AP, FDV, JK. *Analysis and interpretation of data*: MF. *Drafting of the manuscript*: MF. *Critical revision of the paper for important intellectual content*: MF, AP, FDV, BK, NL, MHA, TY, JK. *Statistical analysis*: MF. *Obtaining funding*: JK, MF, FDV, TY.

Conflict of Interest Disclosures: MF, AP, FDV, BK, JK reports funding from NIHR SPHR, MF and TY report funding from NIHR ARC-YH, during the conduct of the study. No other disclosures were reported.

Funding/Support: The ADAPT study was funded by the NIHR School for Public Health Research (SPHR; grant reference number: SPHR-PHPES004-IAPT) The writing of the manuscript and analysis was part-funded by the NIHR SPHR and part-funded by the NIHR Applied Research Collaboration Yorkshire and Humber (NIHR ARC YH; NIHR award identifier: 200166).

Role of the Funder/Sponsor: The NIHR SPHR and ARC-YH had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The funding agreement ensured the authors' independence in developing the purview of the manuscript, writing, and publishing the manuscript.

Acknowledgment: We thank the Talking Therapies for anxiety and depression service providers for providing functionally anonymised electronic health record data and information about the data used in our analyses. We thank the rest of the ADAPT study team, including Katrina d'Apice and Patricia N Albers (University of Bristol, UK), and Esther Curtin (London School of Hygiene and Tropical Medicine, UK). MF acknowledges the use of Stata code (or conversion from R code) from the following courses: 'Introduction to Causal Inference', University of Leeds, England; 'Causal Inference for Assessing Effectiveness in Real World Data and Clinical Trials', UMIT TIROL, Austria.

References

1. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Social science & medicine. 2018; 210: 2–21.
2. Naimi AI, Whitcomb BW. Defining and Identifying Average Treatment Effects. American Journal of Epidemiology. 2023; 192: 685–87.
3. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes. Oxford university press, 2015.
4. NICE. NICE health technology evaluations: the manual. In: NICE, ed., 2022.
5. Rowen D, Azzabi Zouraq I, Chevrou-Severac H, et al. International regulations and recommendations for utility data for health technology assessment. Pharmacoeconomics. 2017; 35: 11–19.
6. Franklin M, Lomas J, Walker S, et al. An educational review about using cost data for the purpose of cost-effectiveness analysis. PharmacoEconomics. 2019: 1–13.
7. Hunter RM, Baio G, Butt T, et al. An educational review of the statistical issues in analysing utility data for cost-utility analysis. Pharmacoeconomics. 2015; 33: 355–66.
8. NICE. NICE real-world evidence framework In: NICE, ed., 2022.
9. Gomes M, Latimer N, Soares M, et al. Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. Pharmacoeconomics. 2022; 40: 577–86.
10. Frieden TR. Evidence for health decision making—beyond randomized, controlled trials. New England Journal of Medicine. 2017; 377: 465–75.
11. Franklin M, Lomas J, Richardson G. Conducting value for money analyses for non-randomised interventional studies including service evaluations: an educational review with recommendations. Pharmacoeconomics. 2020; 38: 665–81.
12. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. American journal of epidemiology. 2016; 183: 758–64.
13. NICE. National Institute for Health and Care Excellence (NICE). National Institute for Health and Care Excellence (NICE),, 2023.
14. National Health Service (NHS). The NHS Long Term Plan. online, 2019.
15. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020.
16. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. bmj. 2016; 355.
17. Ali MS, Prieto-Alhambra D, Lopes LC, et al. Propensity score methods in health technology assessment: principles, extended applications, and recent advances. Frontiers in pharmacology. 2019: 973.
18. Funk MJ, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. American journal of epidemiology. 2011; 173: 761–67.

19. Faria R, Alava MH, Manca A, et al. The use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data: NICE DSU technical support document. NICE DSU Technical Support Document (TSD), 2015.
20. El Alili M, van Dongen JM, Esser JL, et al. A scoping review of statistical methods for trial-based economic evaluations: The current state of play. Health Economics. 2022; 31: 2680–99.
21. Bowrin K, Briere J-B, Levy P, et al. Cost-effectiveness analyses using real-world data: an overview of the literature. Journal of Medical Economics. 2019; 22: 545–53.
22. Guertin JR, Conombo B, Langevin R, et al. A systematic review of methods used for confounding adjustment in observational economic evaluations in cardiology conducted between 2013 and 2017. Medical Decision Making. 2020; 40: 582–95.
23. Kreif N, Grieve R, Sadique MZ. Statistical methods for cost-effectiveness analyses that use observational data: A critical appraisal tool and review of current practice. Health economics. 2013; 22: 486–500.
24. Kennedy-Martin M, Slaap B, Herdman M, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. The European Journal of Health Economics. 2020; 21: 1245–57.
25. NHS Digital. Psychological Therapies, Annual report on the use of IAPT services, 2021-22. Psychological Therapies, Annual report on the use of IAPT services, 2022.
26. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine. 2022; 54.
27. National Institute for Health and Care Excellence (NICE). Common mental health problems: identification and pathways to care. NICE clinical guidelines, 2011.
28. NHS Digital. Improving Access to Psychological Therapies (IAPT) Data Set. NHS Digital,, 2023.
29. The National Collaborating Centre for Mental Health (NCCMH). The Improving Access to Psychological Therapies Manual. 2023.
30. Curtin EL, d'Apice K, Porter A, et al. Perspectives on an enhanced 'Improving Access to Psychological Therapies'(IAPT) service addressing the wider determinants of mental health: a qualitative study. BMC health services research. 2023; 23: 536.
31. Microsoft Corporation. Microsoft Excel 2016. 2016.
32. Faria R, Gomes M, Epstein D, et al. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. Pharmacoeconomics. 2014; 32: 1157–70.
33. Leurent B, Gomes M, Faria R, et al. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. PharmacoEconomics. 2018; 36: 889–901.
34. NICE. Guide to the methods of technology appraisal. In: National Institute for Health and Care Excellence (NICE), ed. London, 2013.
35. NICE. Position statement on use of the EQ-5D-5L valuation set for England (updated November 2018). London: National Institute for Health and Care Excellence (NICE),, 2018.
36. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC,, 2017.
37. Ramsey S, Willke R, Briggs A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. Value in health. 2005; 8: 521–33.
38. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II—an ISPOR Good Research Practices Task Force report. Value in Health. 2015; 18: 161–72.
39. Husereau D, Drummond M, Petrou S, et al. Consolidated health economic evaluation reporting standards (CHEERS)—explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. Value in health. 2013; 16: 231–50.
40. Smith MJ, Mansournia MA, Maringe C, et al. Introduction to computational causal inference using reproducible Stata, R, and Python code: A tutorial. Statistics in medicine. 2022; 41: 407–32.
41. Fewell Z, Hernán MA, Wolfe F, et al. Controlling for time-dependent confounding using marginal structural models. The Stata Journal. 2004; 4: 402–20.
42. Gabrio A, Plumpton C, Banerjee S, et al. Linear mixed models to handle missing at random data in trial-based economic evaluations. Health Economics. 2022; 31: 1276–87.
43. Morga A, Latimer NR, Scott M, et al. Is Intention to Treat Still the Gold Standard or Should Health Technology Assessment Agencies Embrace a Broader Estimands Framework?: Insights and Perspectives From the National Institute for Health and Care Excellence and Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen on the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use E9 (R1) Addendum. Value in Health. 2023; 26: 234–42.
44. International Council for Harmonisation (ICH). ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998.
45. International Council for Harmonisation (ICH). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1). International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2019.
46. Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? Journal of Patient-Reported Outcomes. 2020; 4: 1–8.
47. Franklin M, Hernández Alava M. Enabling QALY estimation in mental health trials and care settings: mapping from the PHQ-9 and GAD-7 to the ReQoL-UI or EQ-5D-5L using mixture models. Quality of Life Research. 2023: 1–16.
48. Tennant PW, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. International journal of epidemiology. 2021; 50: 620–32.

49. Rodrigues D, Kreif N, Lawrence-Jones A, et al. Reflection on modern methods: constructing directed acyclic graphs (DAGs) with domain experts for health services research. International Journal of Epidemiology. 2022; 51: 1339–48.
50. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. Clinical Kidney Journal. 2022; 15: 14–20.
51. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine. 2015; 34: 3661–79.
52. Thoemmes F, Ong AD. A primer on inverse probability of treatment weighting and marginal structural models. Emerging Adulthood. 2016; 4: 40–59.
53. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. Statistical models in epidemiology, the environment, and clinical trials: Springer, 2000.
54. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association. 2001; 96: 440–48.
55. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health economics. 2005; 14: 487–96.
56. Franklin M, Hunter RM, Enrique A, et al. Estimating cost-effectiveness using alternative preference-based scores and within-trial methods: exploring the dynamics of the quality-adjusted life-year using the EQ-5D 5-level version and recovering quality of life utility index. Value in Health. 2022; 25: 1018–29.
57. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. International journal of epidemiology. 2014; 43: 1969–85.
58. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Annals of internal medicine. 2017; 167: 268–74.
59. Morris TP, Walker AS, Williamson EJ, et al. Planning a method for covariate adjustment in individually randomised trials: a practical guide. Trials. 2022; 23: 328.
60. Franklin M, Thorn J. Self-reported and routinely collected electronic healthcare resource-use data for trial-based economic evaluations: the current state of play in England and considerations for the future. BMC medical research methodology. 2019; 19: 1–13.
61. Franklin M, Davis S, Horspool M, et al. Economic evaluations alongside efficient study designs using large observational datasets: the PLEASANT trial case study. Pharmacoeconomics. 2017; 35: 561–73.
62. Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Oup Oxford, 2006.
63. Brennan A, Akehurst R. Modelling in health economic evaluation: What is its place? What is its value? Pharmacoeconomics. 2000; 17: 445 – 59.

Tables

Table 1
Target Trial protocol overview for the ADAPT study economic evaluation

Component	Description
Eligibility criteria	**New referrals to TTad-services**: no attendance at the TTad site in the previous 6-months since the new referral **Newly referred during**: March 2021 to March 2022 (intervention & geographical-control) or March 2018 and March 2019 (historical-control) **Baseline data**: recorded PHQ-9 (depression severity) and GAD-7 (anxiety severity) score at baseline – necessary for '**Condition caseness at baseline**' **Baseline condition caseness**: classified as having depression caseness (PHQ-9 ≥ 10) or anxiety caseness (i.e. GAD-7 ≥ 8) at index assessment (i.e. baseline) **As-started treatment**: service-users had attended at least one treatment session to be defined as 'as-started' treatment **Sufficient TTad data for follow-up period**: available data time-horizon must be at least that of the analysis follow-up period (e.g. primary analysis: 16-weeks)
Treatment strategies	**Intervention**: Enhanced TTad-service (South-West, England), as TAU and/or 'Healthy Living Healthy Minds' programme and/or 1–1 Wellbeing Navigator sessions. **Geographical control**: TAU TTad-service in South-East, England; **Historical control**: TAU TTad-service in the intervention area but before the enhanced service had been implemented.
Assignment	**Non-randomised and unblinded**: all intervention site referrals are offered the enhanced TTad-service, with uptake based on service-user preference
Follow-up period(s)	**Baseline (time zero)**: index appointment to assess condition caseness and allocate people to the waiting-list before first treatment session **Primary follow-up**: starts at baseline and ends at 16-weeks after baseline, regardless of TAU received and service discharge **Secondary follow-up**: starts at baseline and ends at 12-, 20-, or 24-weeks after baseline, regardless of TAU received and service discharge
Outcome(s)	**Primary**: PHQ-9 and GAD-7 scores, sex, and age mapped to EQ-5D-5L UK crosswalk utility scores for QALY estimation **Secondary**: PHQ-9 and GAD-7 scores, sex, and age mapped to EQ-5D-5L VSE OR ReQoL-UI utility scores for QALY estimation **Resource-use/costs**: TTad-service EHR recorded resources-use with costs applied for, or inflated to, the year 2020/21
Estimand(s) (causal contrasts)	**Primary ITT**: In new referrals to TTad-services, what is the between-group difference in mean TTad-service costs and QALYs accumulated since index assessment (i.e. baseline), with QALYs based on EQ-5D-5L crosswalk utilities predicted from PHQ-9 and GAD-7 scores, age, and sex, for those referred to the enhanced TTad-service compared to treatment-as-usual (TAU) for those within a geographical-control-site up to 16-weeks after baseline, regardless of TAU received and service discharge? **Primary PP**: same as 'as-started', but intervention group participants must have had at least one enhanced TTad-service treatment session
Analysis plan	**Overall**: pre-specified SHEAP with any deviations to be reported at publication stage; below are highlighted examples. **Selection bias**: a Target Trial approach is used to aid reduce/avoid selection bias among other biases in the analysis. **Confounding bias**: IPTWs used within MSMs alongside baseline confounder adjustment as a double robust approach, informed by bespoke DAGs as a visual presentation of hypothesised causal relationship between treatment strategies and outcomes(s) and other covariates to identify confounders for conditioning **Information bias**: informative censoring is a known issue as PHQ-9 and GAD-7 data collection stops at the point of patient discharge (administrative censoring); however, people discharged due to completing treatment or feeling well enough to be discharged were hypothesised to have better mental health outcomes post-discharge than those who left the service early (e.g. due to service dissatisfaction). As such, IPCWs are used alongside IPTWs in the MSMs to account for this informative censoring. **Sensitivity analyses**: alternative utility scores, time-horizons, and weighting assumptions/specifications (e.g. for IPTWs) will be used to assess result robustness
Acronyms & abbreviations	**ADAPT**, **A**ssessing a **D**istinct Improving **A**ccess to **P**sychological **T**herapies (researchregistry7322); **EHR**, electronic health record; **GAD-7**, Generalized Anxiety Disorder-7 anxiety-scale; **HRQoL**, health-related quality-of-life; **IAPT**, Improving Access to Psychological Therapies; **IPCW**, inverse probability of censoring weights; **IPTW**, inverse probability of treatment weights; **ITT**, intention-to-treat; **MSM**, marginal structural model; **PHQ-9**, Patient Health Questionnaire-9 depression-scale; **PP**, per protocol; **QALY**, quality-adjusted life-year; **ReQoL-UI**, Recovering Quality of Life Utility Index; **SHEAP**, statistical and health economic analysis plan; **TAU**, treatment-as-usual; **TTad**, NHS Talking Therapies for anxiety and depression services; **VSE**, Value Set for England.

Table 2
Baseline descriptive statistics of sample populations for the primary 16-week time-horizon analyses

	Intervention **ITT**	**Intervention** **PP**	**Geographical** **control**	**Historical** **control**
N	5,441	549	2,149	4,001
Socio-demographics				
Age, mean (SD, range)	35 (13.4, 16 to 92)	39 (14.1, 17 to 89)	39 (15.6, 15 to 107)	37 (14.0, 16 to 93)
Sex, Female %N	69.7%	67.9%	69.6%	63.9%
Ethnicity, %N				
- White	90.4%	85.8%	90.1%	90.5%
- Mixed	3.3%	3.8%	2.4%	3.1%
- Asian/Asian British	2.7%	4.4%	3.0%	3.1%
- Black/Black British	2.3%	4.2%	3.7%	2.9%
- Other	1.2%	1.8%	0.8%	0.9%
IMD decile^{a}, mean (SD, range)	5.5 (1.9, 2.4 to 9.3)	5.3 (1.9, 2.4 to 9.3)	5.6 (2.3, 2.3 to 9.3)	5.1 (1.8, 2.5 to 9.1)
Mental health scores				
PHQ-9, mean (SD, range)	14.4 (5.2, 0 to 27)	16.6 (5.1, 1 to 27)	15.7 (5.5, 0 to 27)	15.8 (5.4, 5.4, 0 to 27)
GAD-7, mean (SD, range)	13.4 (3.5, 0 to 21)	14.1 (4.7, 0 to 21)	14.1 (4.4, 0 to 21)	13.9 (4.4, 0 to 21)
Depression caseness (PHQ-9 ≥ 10), %N	83.7%	91.6%	86.7%	87.3%
Anxiety caseness (GAD-7 ≥ 8), %N	89.9%	91.7%	92.8%	92.1%
Depression & anxiety caseness, %N	73.6%	82.5%	79.5%	79.4%
Predicted health-related quality-of-life				
Predicted EQ-5D-5L crosswalk, mean (SD, range)	0.634 (0.093, 0.359 to 0.828)	0.596 (0.105, 0.359 to 0.821)	0.609 (0.104, 0.359 to 0.809)	0.609 (0.104, 0.359 to 0.840)
Predicted EQ-5D-5L VSE, mean (SD, range)	0.712 (0.077, 0.472 to 0.874)	0.682 (0.088, 0.472 to 0.873)	0.691 (0.086, 0.471 to 0.865)	0.693 (0.086, 0.472 to 0.882)
Predicted ReQoL-UI, mean (SD, range)	0.757 (0.057, 0.605 to 0.898)	0.734 (0.059, 0.605 to 0.894)	0.743 (0.061, 0.605 to 0.890)	0.743 (0.061, 0.605 to 0.902)

Acronyms & abbreviations. **ADAPT**, **A**ssessing a **D**istinct Improving **A**ccess to **P**sychological **T**herapies (researchregistry7322); **EHR**, electronic health record; **GAD-7**, Generalized Anxiety Disorder-7 anxiety-scale; **HRQoL**, health-related quality-of-life; **IAPT**, Improving Access to Psychological Therapies; **IMD**, Index of Multiple Deprivation; **ITT**, intention-to-treat; **LSOAs**, Lower Layer Super Output Areas; **PHQ-9**, Patient Health Questionnaire-9 depression-scale; **PP**, per protocol; **ReQoL-UI**, Recovering Quality of Life Utility Index; **SD**, standard deviation; **TTad**, NHS Talking Therapies for anxiety and depression services; **VSE**, Value Set for England.

[a] IMD decile description: LSOAs in decile 1 fall within the most deprived 10% of LSOAs nationally and LSOAs in decile 10 fall within the least deprived 10% of LSOAs nationally

Table 3

Cost-effectiveness results for ITT and PP-intervention Vs control over 16-weeks with bootstrapped bias-corrected confidence intervals

Comparison	Method[a]	QALYs (ATE)				Costs, £ (ATE)				ICER[b]	Prob. CE < λ per QALY[c]		
		Mean	bSEM	BC 95% bCIs		Mean	bSEM	BC 95% bCIs			λ = £0	λ = £20k	λ = £30k
Intervention Vs Geographical control (16-weeks)													
ITT (N = 5,441) Vs.	Naïve regression	0.00035	0.00058	-0.00075	0.00152	-£30.64	£4.21	-£38.90	-£22.26	Dominates	100%	100%	99.3%
Geog (N = 2,149)	Doubly robust	0.00052	0.00102	-0.00105	0.00277	-£29.64	£4.26	-£37.99	-£20.69	Dominates	100%	98.1%	95.0%
PP (N = 549) Vs.	Naïve regression	-0.00478	0.00130	-0.00729	-0.00233	-£34.09	£5.80	-£45.95	-£22.70	< Q & < £	100%	49.2%	31.3%
Geog (N = 2,149)	Doubly robust	-0.00157	0.00261	-0.00736	0.00322	-£33.33	£5.84	-£45.24	-£21.82	< Q & < £	100%	57.8%	46.2%
Intervention Vs Historical control (16-weeks)													
ITT (N = 5,441) Vs.	Naïve regression	0.00364	0.00053	0.00261	0.00471	-£84.54	£4.10	-£92.78	-£76.61	Dominates	100%	100%	100%
Hist (N = 4,001)	Doubly robust	0.00153	0.00069	0.00025	0.00298	-£84.86	£4.05	-£92.52	-£76.50	Dominates	100%	100%	100%
PP (N = 549) Vs.	Naïve regression	0.00008	0.00134	-0.00278	0.00262	-£85.50	£5.80	-£97.60	-£75.07	Dominates	100%	99.9%	95.9%
Hist (N = 4,001)	Doubly robust	-0.00103	0.00170	-0.00535	0.00167	-£88.15	£5.79	-£99.83	-£77.06	< Q & < £	100%	96.9%	89.4%

Acronyms, short-hand text, and symbols. ATE, average treatment effect; **BC 95% bCIs**, bias-corrected 95% bootstrapped confidence intervals; **ICER**, incremental cost-effectiveness ratio; **ITT**, intention-to-treat; **PP**, per-protocol; **QALYs / Q**, quality-adjusted life-years; **bSEM**, bootstrapped standard error of the mean; **w/**, with; **£**, cost in Great British Pounds (2020/21); **λ**, cost-effectiveness threshold per QALY.

[a] Naïve regression involved using a generalised linear model (GLM) with baseline covariate adjustment only. The doubly robust method used censoring-related and treatment-related inverse probability weights within marginal structural models (MSM) followed by g-computation. The family and link functions were the same for GLMs and MSMs: utility (for QALYs), Normal-distribution with identity-link; costs, Gamma-distribution with log-link.

[b] A numerical ICER is only presented if the intervention produces higher mean incremental QALYs (> Q) and costs (>£) than control. If > Q & <£, then intervention dominates control. If < Q & >£, the intervention is dominated by control. If < Q & <£, then the intervention produces less QALYs but is cost-saving, thus this is stated rather than presenting a numerical ICER.

[c] Cost-effectiveness thresholds (**λ**) are set at £20,000 (£20k) and £30,000 (£30k) as the NICE approval norms; when λ=£0, this represents the probability of cost-savings given QALYs have no designated £ value.

Table 4

Cost-effectiveness results for ITT or PP intervention vs geographical control over 16-weeks – sensitivity analyses using different met

No.	Method[a]	Outcome	Mean	SEM	95% CIs		Mean	SEM	95% CIs		Mean	SEM	95% C
ITT			Intervention (ITT), N = 5,441				Geographical-control, N = 2,149				ATE		
1	Naïve regression	QALYs	0.20532	0.00029	0.20475	0.20589	0.20498	0.00049	0.20402	0.20593	0.00035	0.00058	-0.000
	Naïve regression	Costs (£)	£173.85	£1.79	£170.35	£177.36	£204.49	£3.68	£197.29	£211.70	-£30.64	£4.10	-£38.6
2	MSM w/ sIPTW	QALYs	0.20555	0.00030	0.20497	0.20613	0.20488	0.00047	0.20395	0.20580	0.00067	0.00056	-0.000
	MSM w/ sIPTW	Costs (£)	£173.78	£1.79	£170.28	£177.28	£203.42	£3.78	£196.01	£210.84	-£29.64	£4.17	-£37.8
3	MSM w/ sIPCW	QALYs	0.20706	0.00028	0.20651	0.20761	0.20685	0.00047	0.20593	0.20778	0.00020	0.00056	-0.000
	No. 2	Costs (£)	£173.78	£1.79	£170.28	£177.28	£203.42	£3.78	£196.01	£210.84	-£29.64	£4.17	-£37.8
4	MSM w/ sIPTCW	QALYs	0.20723	0.00028	0.20668	0.20779	0.20676	0.00048	0.20582	0.20769	0.00048	0.00056	-0.000
	No. 2	Costs (£)	£173.78	£1.79	£170.28	£177.28	£203.42	£3.78	£196.01	£210.84	-£29.64	£4.17	-£37.8
5	No. 4 & g-comp[c]	QALYs	0.20519	N/A	N/A	N/A	0.20467	N/A	N/A	N/A	0.00052	N/A	N/A
	No. 2 & g-comp[c]	Costs (£)	£175.13	N/A	N/A	N/A	£204.77	N/A	N/A	N/A	-£29.64	N/A	N/A
PP			Intervention (PP), N = 549				Geographical-control, N = 2,149				ATE		
6	Naïve regression	QALYs	0.19803	0.00102	0.19604	0.20003	0.19980	0.00049	0.19883	0.20076	-0.00176	0.00113	-0.003
	Naïve regression	Costs (£)	£171.52	£4.22	£163.25	£179.79	£205.62	£3.64	£198.49	£212.74	-£34.09	£5.56	-£45.0
7	MSM w/ sIPTW	QALYs	0.19782	0.00101	0.19584	0.19979	0.19982	0.00049	0.19885	0.20079	-0.00200	0.00112	-0.004
	MSM w/ sIPTW	Costs (£)	£172.49	£4.23	£164.20	£180.79	£205.94	£3.64	£198.80	£213.09	-£33.45	£5.54	-£44.3
8	MSM w/ sIPCW	QALYs	0.20304	0.00107	0.20094	0.20514	0.20425	0.00057	0.20314	0.20537	-0.00122	0.00124	-0.003
	No. 2	Costs (£)	£172.49	£4.23	£164.20	£180.79	£205.94	£3.64	£198.80	£213.09	-£33.45	£5.54	-£44.3
9	MSM w/sIPTCW	QALYs	0.20267	0.00108	0.20056	0.20479	0.20428	0.00058	0.20315	0.20541	-0.00161	0.00121	-0.003
	No. 2	Costs (£)	£172.49	£4.23	£164.20	£180.79	£205.94	£3.64	£198.80	£213.09	-£33.45	£5.54	-£44.3
10	No. 4 & g-comp[c]	QALYs	0.19800	N/A	N/A	N/A	0.19958	N/A	N/A	N/A	-0.00157	N/A	N/A
	No. 2 & g-comp[c]	Costs (£)	£172.23	N/A	N/A	N/A	£205.56	N/A	N/A	N/A	-£33.33	N/A	N/A

Acronyms, short-hand text, and symbols. ATE, average treatment effect; **95% CIs**, confidence intervals; **g-comp**, g-computation; **ICER**, incremental cost-effectiveness ratio; **IPCW**, inverse probability of censoring weight; **IPTW**, inverse probability of treatment weight; **IPTCW**, inverse probability of treatment and censoring weight (i.e. **IPTW*IPCW**); **ITT**, intention-to-treat; **MSM**, marginal structural model; **No.**, number; **PP**, per-protocol; **QALYs / Q**, quality-adjusted life-years; **SEM**, standard error of the mean; **w/**, with; **£**, cost in Great British Pounds (2020/21)

[a] Naïve regression involved using a generalised linear model (GLM) with baseline covariate adjustment only. The family and link functions were the same for GLMs and MSMs: utility (for QALYs), Normal-distribution with identity-link; costs, Gamma-distribution with log-link.

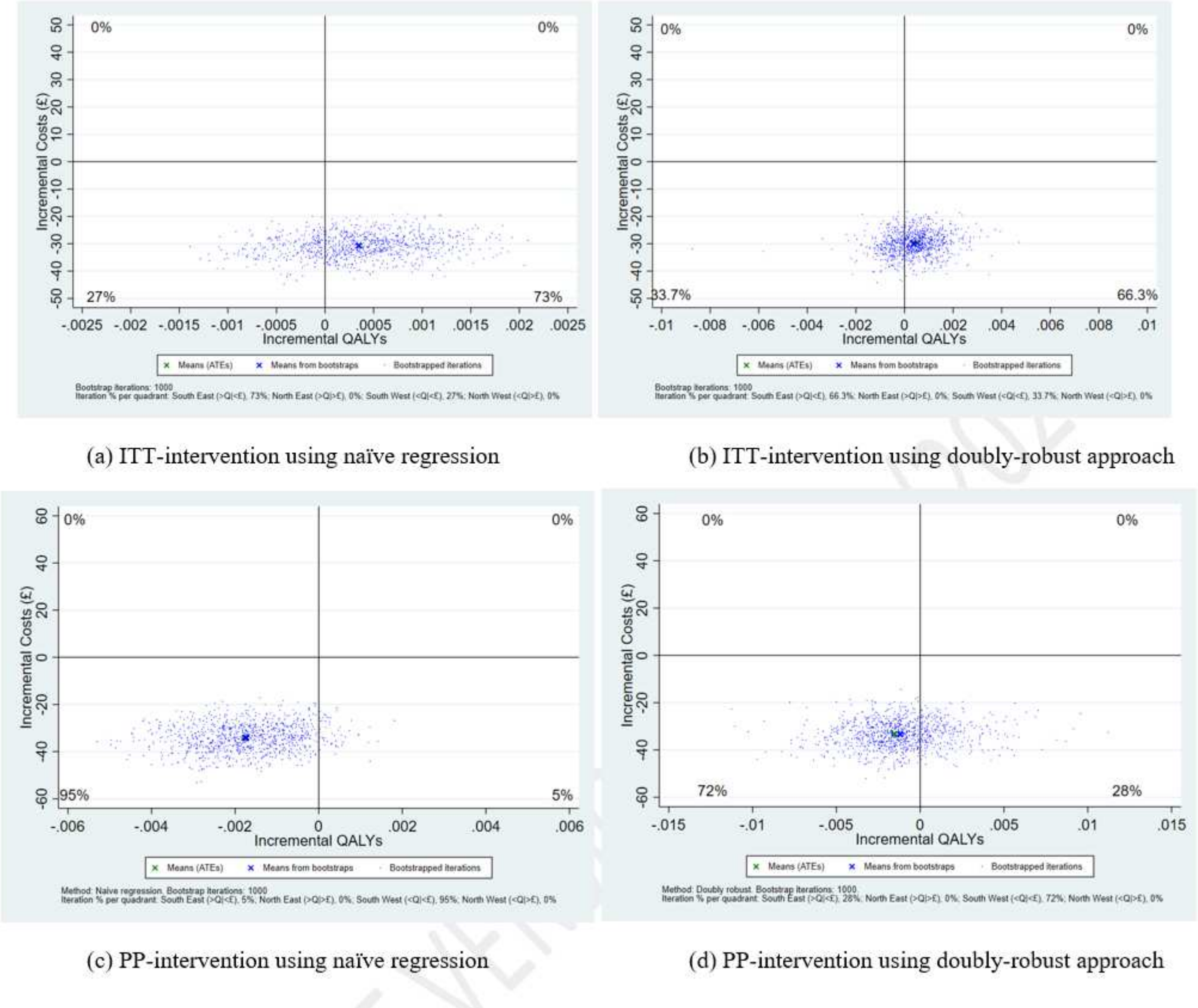
[b] A numerical ICER is only presented if the intervention produces higher mean incremental QALYs (> Q) and costs (>£) than control. If > Q & <£, then intervention dominates control. If < Q & >£, the intervention is dominated by control. If < Q & <£, then the intervention produces less QALYs but is cost-saving, thus this is stated rather than presenting a numerical ICER.

[c] Due to the two-step process associated with g-computation, distributional statistics are not produced and therefore not presented.

Appendix

The Appendix is not available with this version

Figures



Bootstrap iterations: 1000
Iteration % per quadrant: South East (>Q|<£), 73%; North East (>Q|>£), 0%; South West (<Q|<£), 27%; North West (<Q|>£), 0%

(a) ITT-intervention using naïve regression

Bootstrap iterations: 1000
Iteration % per quadrant: South East (>Q|<£), 66.3%; North East (>Q|>£), 0%; South West (<Q|<£), 33.7%; North West (<Q|>£), 0%

(b) ITT-intervention using doubly-robust approach

Method: Naive regression. Bootstrap iterations: 1000
Iteration % per quadrant: South East (>Q|<£), 5%; North East (>Q|>£), 0%; South West (<Q|<£), 95%; North West (<Q|>£), 0%

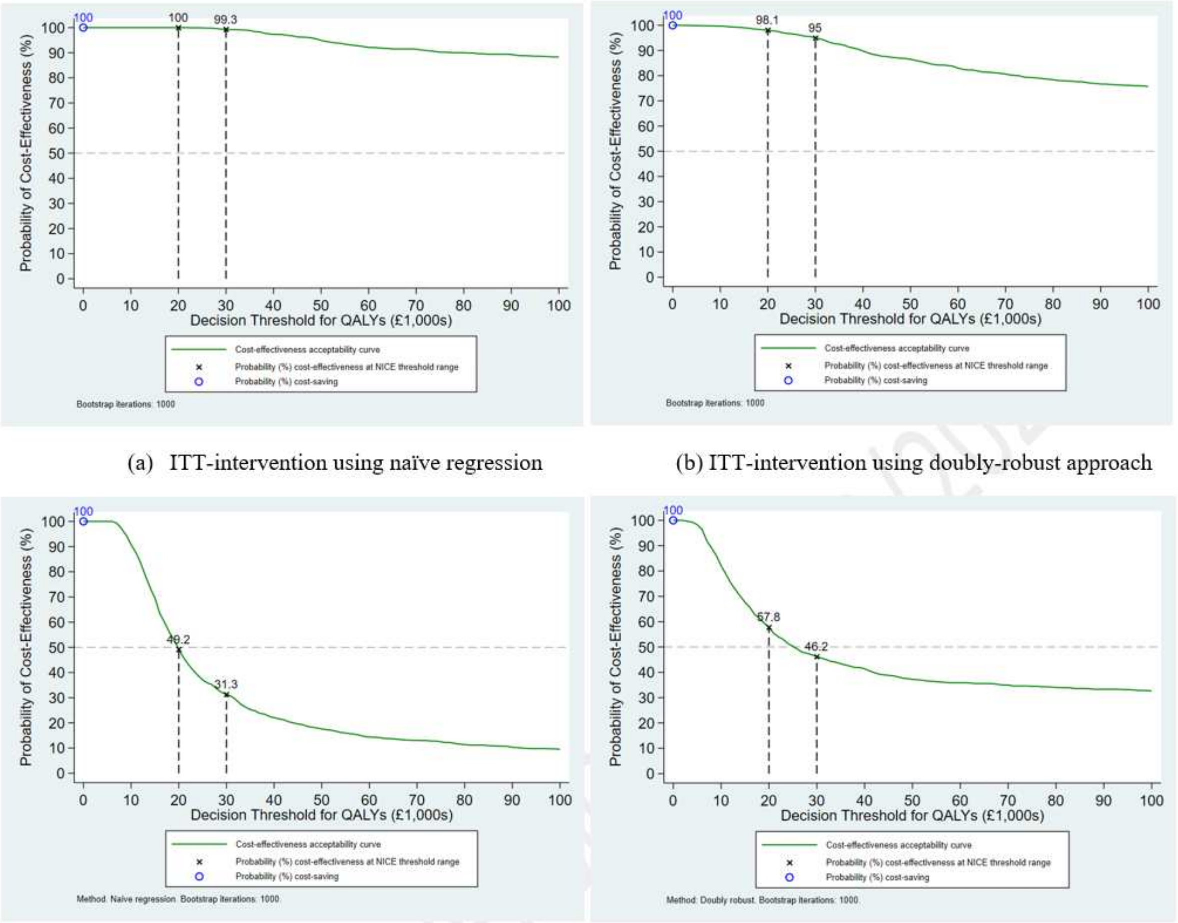
(c) PP-intervention using naïve regression

Method: Doubly robust. Bootstrap iterations: 1000.
Iteration % per quadrant: South East (>Q|<£), 28%; North East (>Q|>£), 0%; South West (<Q|<£), 72%; North West (<Q|>£), 0%

(d) PP-intervention using doubly-robust approach

Figure 1

Cost-effectiveness planes up to 16-weeks – ITT or PP Vs Geographical control, using naïve regression or doubly-robust approach



(a) ITT-intervention using naïve regression

(b) ITT-intervention using doubly-robust approach

(c) PP-intervention using naïve regression

(d) PP-intervention using doubly-robust approach

Figure 2

Cost-effectiveness acceptability curves up to 16-weeks – ITT or PP Vs Geographical control, using naïve regression or doubly-robust approach