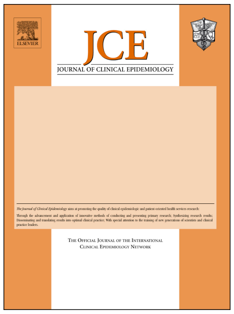


Treatment effects may remain the same even when trial participants differed from the target population

Mike J. Bradburn, Ellen Cecilia Lee, David Alexander White, Daniel Hind, Norman R. Waugh, Deborah Denise Cooke, David Hopkins, Peter Mansell, Simon Richard Heller

PII: S0895-4356(19)30818-2

DOI: https://doi.org/10.1016/j.jclinepi.2020.05.001

Reference: JCE 10139

To appear in: *Journal of Clinical Epidemiology*

Received Date: 22 October 2019

Revised Date: 15 April 2020

Accepted Date: 4 May 2020

Please cite this article as: Bradburn MJ, Lee EC, White DA, Hind D, Waugh NR, Cooke DD, Hopkins D, Mansell P, Heller SR, Treatment effects may remain the same even when trial participants differed from the target population, *Journal of Clinical Epidemiology* (2020), doi: https://doi.org/10.1016/j.jclinepi.2020.05.001.

Treatment effects may remain the same even when trial participants differed from the target population

Bradburn, Mike J [1]*

Lee, Ellen Cecilia [1]

White, David Alexander [1]

Hind, Daniel [1]

Waugh, Norman R [2]

Cooke, Deborah Denise [3]

Hopkins, David [4]

Mansell, Peter [5]

Heller, Simon Richard [6]

Author affiliations

[1] Clinical Trials Research Unit, University of Sheffield, Sheffield, UK

[2[Warwick Medical School, University of Warwick, UK

[3] School of Health Sciences, University of Surrey, UK

[4] Institute of Diabetes, Endocrinology & Obesity, King's Health Partners, London, UK

[5] Department of Diabetes and Endocrinology, Nottingham University Hospitals NHS Trust, UK

[6] Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK

*Corresponding author

Clinical Trials Research Unit

School of Health and Related Research (ScHARR)

University of Sheffield

30 Regent Street

Sheffield S1 4DA

m.bradburn@sheffield.ac.uk

+44 114 222 0706

Abstract

Objective

RCTs have been criticised for lacking external validity. We assessed whether a trial in people with type I diabetes mellitus (T1DM) mirrored the wider population, and applied sample-weighting methods to assess the impact of differences on our trial's findings.

Study design and setting

The REPOSE trial was nested within a large UK cohort capturing demographic, clinical and quality of life (QoL) data for people with T1DM undergoing structured diabetes-specific education. We firstly assessed whether our RCT participants were comparable to this cohort using propensity score modelling. Following this we re-weighted the trial population to better match the wider cohort and re-estimated the treatment effect.

Results

Trial participants differed from the cohort in regards to sex, weight, HbA1c and also QoL and satisfaction with current treatment. Nevertheless, the treatment effects derived from alternative model weightings were similar to that of the original RCT.

Conclusions

Our RCT participants differed in composition to the wider population but the original findings were unaffected by sampling adjustments. We encourage investigators take steps to address criticisms of generalisability, but doing so is problematic: external data, even if available, may contain limited information and analyses can be susceptible to model misspecification.

Keywords

Randomized controlled trials; External validity; generalizability; sample reweighting.

Clinical trial registration:

ISRCTN: 61215213

ClinicalTrials.gov: NCT01616784

EudraCT: 2010-023198-21

1 Introduction

The randomised controlled trial (RCT) is considered the gold standard for assessing health interventions. RCTs nevertheless have drawn a number of criticisms, one of which is their generalisability[1]. Trial participants are assumed to be a random sample of a wider target population in order to make valid inference but, as several authors have previously claimed, this is an optimistic (and probably unrealistic) assumption[1,2]. RCTs recruit volunteers who consent to be studied; those who participate may differ in some important characteristics to those who do not, a phenomenon referred to as volunteer bias[3]. Furthermore, the institutions involved in recruiting may themselves not be a representative of the wider population[4,5]. Therefore, whilst the randomised groups within the trial are comparable to each other ("internally valid"), inferences drawn from such studies are questionable if those who enter an RCT are in some way different to those who do not ("externally invalid")[6,7].

An important caveat is that even where selected and non-selected participants differ in their characteristics , the treatment comparison remains unbiased unless the reasons for non-uptake are also associated with the treatment effect (i.e. effect modifiers)[8,9]. These conditions are seldom known for any given situation: trials generally collect very limited data on individuals who do not participate [10,11], and studies typically have little power to detect differential effects across patient characteristics[12]. Whilst there is compelling meta-epidemiological evidence that RCT cohorts differ from wider patient populations[13–23], and the potential for bias in the treatment effect has been demonstrated[24,25], less is known as to whether this impacts on the overall estimated treatment difference. Proponents of RCTs claim that randomisation, even on selected participants, is less biased than alternative study designs[26]. Nevertheless, the credibility and uptake of a trial's findings are at least questionable if its participants do not mirror those that a healthcare professional sees in routine practice.

Methodology to adjust for non-random sampling are commonly used when analysing surveys[27,28] but seldom employed for RCTs[8,29]. We describe a cohort of people with type 1 diabetes mellitus (T1DM), a subset of whom entered into a RCT. This allows us to explore whether trial participants were representative and, if not, whether this impacted on the trial findings. We also describe some practicalities and limitations of methods which can be used to assess these.

2 Methods

2.1 Case study

2.1.1 The REPOSE trial

The REPOSE (Relative Effectiveness of Pumps over MDI and Structured Education) trial was a randomised trial evaluating two modalities for delivering insulin to people with type 1 diabetes (T1DM)[30]. Participants were randomised to continuous subcutaneous insulin infusion (CSII, also known as insulin pump) or to the conventional multiple daily injections (MDI). As part of the trial, all participants also attended a DAFNE (Dose Adjustment for Normal Eating) diabetes educational course [31] shortly after randomisation.

The REPOSE trial recruited participants from 8 UK centres between March 2012 and June 2013. The primary endpoint was the reduction in HbA1c at 2 years among the subset of participants with HbA1c ≥58mmol/mol (7.5%) at baseline, and demonstrated only marginal clinical benefit of CSII (mean difference 2.7mmol/mol (0.24%), 95% confidence interval -0.5 to 5.8mmol/mol, p=0.10).

2.1.2 The DAFNE UK database

The DAFNE programme is a five-day course which aims to improve self-management of T1DM by equipping people with the skills necessary to review and respond to results of blood glucose monitoring, estimate their carbohydrate intake and administer accurate insulin doses dependent on a range of variables through adjustment of insulin on a meal by meal basis [31]. The UK National Institute for Health and Care Excellence recommends people with T1DM undergo a structured educational programme and specifically endorse DAFNE[32]. In 2008, a nationwide DAFNE database was set up to collect details of individuals who had undergone DAFNE training in the UK and consented to their data being used for research purposes[33]. 72 centres collected pre-course demographic and diabetes-specific characteristics, ten of which were designated research sites in which individuals were given the option to consent to additional data collection including insulin dosage, laboratory measures and health utility and quality of life (QoL) inventories. A summary of the data collected is presented in box 1.

2.2 Comparison of cohort and nested RCT populations

We used the DAFNE database to identify people undergoing DAFNE training as part of the REPOSE trial and compared these against people from the wider T1DM population who partook in the DAFNE course and who would have been eligible for REPOSE. We compared the characteristics of the REPOSE trial participants against three subgroups who had attended DAFNE training:

1. Those undertaking DAFNE at the same eight centres during the REPOSE RCT recruitment period but were not recruited into the trial.
2. Those undertaking DAFNE in any of the 72 centres between December 2008 and March 2015.
3. Those undertaking DAFNE in the 10 DAFNE research centres and consenting to additional data collection between December 2008 and March 2015.

The three comparisons assess different facets of generalisability. The rationale for comparison 1 is to assess whether any participant characteristics are associated with participation in the REPOSE RCT and which therefore may indicate a selection effect in the recruitment process. Comparisons 2 and 3 assess how well the trial population represents the wider cohort of people with T1DM that undergo DAFNE training and any differences here may also reflect temporal effects or case-mix differences between recruiting and non-recruiting centres. Comparison 2 includes more individuals but is limited to less detailed data.

2.3 Statistical methods

2.3.1 Assessing (dis)similarity between the RCT and cohort populations

The dissimilarity between the RCT and cohort was assessed using a logistic regression model for which the probability of an individual i entering the RCT is dependent on measurable characteristics $x_{1i}, \ldots, x_k$

$$log\left\{\frac{p_i(x)}{1-p_i(x)}\right\} = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{ki}$$

Where $p_i(x)$ denotes their predicted probability of inclusion. In the context of assessing and adjust for differences between two populations, this probability is referred to as their propensity score. Several (related) measures of dissimilarity of RCT and non-RCT participants have been proposed including the c-statistic, the Somer's D coefficient of concordance, the standardised mean difference and the Tipton index[8,9,34].

2.3.2 Re-weighted analysis using the propensity score

The similarity of the RCT and cohort characteristics was modelled using a two-stage approach based on propensity score reweighting[35]. In the first stage, the aforementioned logistic regression model was used to model the probability that an individual undergoing a DAFNE educational course would also partake in the RCT, using their characteristics at the time of the course as explanatory covariates. The second stage repeats the efficacy analysis but with different weights allocated to the observations. More specifically, each individual's contribution to the analysis was up-weighted or down-weighted dependent on their value of $p_i(x)$. Used in this way, propensity scores give more weight to individuals whose characteristics are underrepresented in the RCT compared with the wider cohort, therefore producing an unbiased estimate of the population average treatment effect (PATE) – assuming the regression model is correct. We will return to the development of the model, and to this assumption in particular, in the next section.

Several approaches have been suggested for this reweighting[8,9,36–39], but here we limit our analysis to two common methods. The first method stratifies the entire population (both randomised and non-randomised) into a small number of equally sized subgroups based on their fitted value $p_i(x)$. A common choice is to split based on population quintiles [40], and this is the approach taken here. Next, the treatment effect is calculated within each subgroup among the trial participants. Finally, the overall treatment effect is calculated as the simple average of the subgroups. Although each of the five subgroup estimates are weighted equally, the trial population has now been reweighted - unless its composition matches that of the entire population (i.e. 20% of the trial population are in each subgroup), more weight is given to subgroups with fewer trial participants and vice versa. Further details, including model assumptions and diagnostic checks, are provided by O'Muircheartaigh & Hedges [9].

The second approach directly weights each individual inversely to their probability of inclusion, thereby giving more weight to individuals who are under-represented on the basis of their characteristics. The weights are standardised to ensure they sum to the total sample size:

$$w_i = \widetilde{w}_i \times (N/\widetilde{W})$$

Where

$$\widetilde{w}_i = \frac{1}{p_i(x)}$$

and

$$\widetilde{W} = \sum_{i \in RCT} \widetilde{w}_i(x)$$

The RCT is then re-analysed with these weights applied. The second approach has the advantage of not having to arbitrarily divide the trial population into subgroups. However, its findings may become unstable where weights are very large or very small – in particular, very low probabilities lead to overly influential individual data points. A heuristic but effective method to mitigate is to truncate the lowest and highest weights [41]; in this analysis the upper and lower 1% weights were reset to equal the 1st and 99th centiles respectively.

2.3.3 Derivation of the propensity score model

The logistic regression model was built sequentially, starting with the most readily available data. Firstly, demographic and routinely collected diabetes characteristics were entered into the model. Five variables (age, sex, baseline HbA1c, any previous severe hypoglycaemic episode and time since diagnosis) were included in all models, irrespective of statistical significance; otherwise, those demonstrating significance at the 0.10 statistical significance level were retained. Covariates entered in the next phase comprised laboratory values and vital signs whose impact on inclusion were considered less direct; those associated with a significant improvement in the model log-likelihood were retained. The model was then re-assessed for covariates which could be removed: in particular, covariates containing missing data were replaced with other related variables to assess whether a similar fit could be achieved using the Akaike Information Criterion (AIC). Finally, the model produced was compared with those produced by backwards and forwards stepwise regression, with additional investigations made where covariates differed. Continuous covariates were assessed for non-linearity using fractional polynomial models of 2 and 3 degrees of freedom[42] and model improvement was assessed by the likelihood ratio test. The propensity score model was reviewed and approved by the trial's chief investigator prior to any re-weighted efficacy analysis being undertaken.

2.3.4 Model Assessment

Three types of model assessments were undertaken. Firstly, the stability and goodness of fit of the propensity model were assessed by bootstrap methods and the Hosmer-Lemeshow test respectively[43]. Second, the residual imbalance following each reweighting method described in 2.3.2 was assessed graphically by distributional plots and by the Somer's D statistic[44,45]. Finally, the stability of the reweighted treatment effect was assessed by re-fitting the propensity model with each covariate sequentially left out and also by bootstrap methods as above: in both cases, the re-weighted treatment effect was estimated on the modified propensity model and compared with the original re-weighted estimate. The Stata code used to undertake these is provided in Appendix 2.

3 Results

3.1 Participant flow

267 people were randomised into REPOSE. After excluding participants who were known to be ineligible for REPOSE a total of 22020 DAFNE records were available, 2961 of whom were from DAFNE research centres and 339 attended one of the eight REPOSE centres during the

recruitment period (Figure 1). The characteristics of these and the RCT participants are summarised in table 1. Although data completeness was generally high, there were considerable missing data for laboratory parameters and QoL, which is described in more detail below. These were combined with the 267 RCT participants to create the propensity score models used to re-weight the RCT population.

As previously noted, the primary outcome was HbA1c change amongst participants with a baseline HbA1c ≥7.5% (58mmol/mol). The reweighted analysis was undertaken on the 224 participants meeting this criteria and who had 24 month outcome data.

3.2 Comparisons between RCT and cohort participants

3.2.1 Comparison with non-recruited individuals from the same centres and during the recruitment period

RCT participants were slightly older, had greater weight and BMI, and had been living with T1DM longer compared to non-randomised individuals attending the recruiting centres during recruitment. The randomised individuals also comprised proportionately fewer females and fewer individuals with a prior severe hypoglycaemic episode (table 1). These characteristics were used to derive the first propensity model. Weight was used in preference to BMI as this gave a better model fit (in this and other comparisons). The relationship with disease duration and study entry was non-linear with the probability of entry highest among individuals between 25-30 years post-diagnosis. Only two of the six covariates (weight and years with T1DM) were statistically significant in the multivariable propensity score model, but nevertheless there was modest dissimilarity between RCT and cohort individuals (table 2).

3.2.2 Comparison with non-recruited individuals from all centres

Compared to the entire cohort, RCT participants were similar in terms of age and years since T1DM diagnosis but otherwise yielded findings and indices of dissimilarity comparable to those seen in the previous model.

3.2.3 Comparison with non-recruited individuals from research centres

The final models incorporated additional factors collected in the research centres only. The two QoL inventories (DSQoL and SF-12) were collected only up to 2012 in the cohort, and many laboratory parameters also contained missing data. To deal with this, two models were fitted: health utility, QoL and laboratory data were omitted from the first and added to the second, with the likelihood ratio testing their importance.

Both models showed greater separation than previous models (table 2). RCT participation was lower among nulliparous females than females with a previous pregnancy and males, and presence of retinopathy was positively associated with study entry. Adding QoL and laboratory data substantially improved the model (likelihood ratio test $p<0.0001$ among those with data). The two SF-12 QoL measures (physical and mental health) were both highly correlated with EQ-5D health utility: SF-12 was preferred on the basis of lower AIC. Both physical and mental health components were higher among RCT participants, though mental health was non-linearly associated with RCT inclusion and was highest among individuals scoring between 40-60 which corresponds to a +/-1 standard deviation of the population average[46]. Two of the diabetes specific QoL domains were significantly associated with RCT inclusion: RCT participants had tighter treatment goals and greater satisfaction with current treatment compared to non- participants. Of the laboratory parameters, LDL cholesterol and total cholesterol were higher among RCT participants. The two were heavily correlated, and total cholesterol alone was retained as this had far less – although still substantial- missing data.

3.3 Re-analysis of the RCT based on propensity score weights

The methods performed similarly in terms of reducing the residual imbalance between the RCT and cohort (see appendix 1 for diagnostic plots) and all three methods were used to re-weight the treatment effect. The ensuing point estimates varied between 1.0 to 2.8 mmol/mol (0.09% to 0.26%) with the greatest reduction when reweighting against the detailed research cohort individuals. These estimates were in keeping with the 2.4 mmol/mol (0.22%) difference estimated on complete cases in the original RCT and the 5.5 mmol/mol (0.5%) defined as the clinically relevant difference in the sample size calculation. All 95% confidence intervals overlapped zero, with notably wider confidence intervals for the stratified method (figure 2).

3.4 Sensitivity analyses

In order to assess the sensitivity of the analyses to the propensity model adjustment, alternative models were fitted to the comparison within research centres using the stratified model, since these were most likely to be affected due to overfitting and small strata sizes respectively. The effect of removing, substituting or adding covariates on the effect sizes are shown graphically in figure 3. Although no material differences were evident, alternative adjustments generally led to lower effect sizes; in particular omitting total cholesterol from the model gave an effect size close to zero.

4 Discussion

4.1 Findings

We re-analysed a randomised trial conducted in people with type I diabetes mellitus to better match the characteristics of the wider population. We had access to a large external database containing extensive data for individuals who did not participate in the RCT, thus allowing us to explore participant characteristics that are not routinely collected in the general population. Although RCT participants differed in some ways from those in the external cohort, the original RCT findings were broadly in keeping with the effect sizes derived from re-weighting the RCT population.

4.2 Strengths and limitations

4.2.1 Strengths and limitation of the methodology

The methods described here need external data containing factors that affect both study enrolment and treatment response. This is likely to be the exception rather than the rule; even where such data exist the model building will inevitably fail to explain all the subtleties which affect inclusion into a trial. In addition, RCTs generally employ extensive ongoing data cleaning and would be therefore expected to be of a higher standard than data collected as routine practice. Differences between the RCT and wider populations may therefore arise due to differences in the data collection processes [47,48]. Most importantly however, these methods do not assess the fidelity of the intervention, the baseline standard of care and the transferability of the outcome(s) assessed, which are largely non-measurable entities. We therefore take the more pragmatic approach noted by O'Muircheartaigh & Hedges that a partial reduction in the mismatch between trial and target populations is better than no adjustment at all [9].

4.2.2 Strengths and limitation of the case study

A strength of this study is the extensive data collected for the reference population. Previous research has shown that some people volunteer for an RCT with the perception it may improve their treatment[49,50]: here, a person that considers CSII is willing to (and hopeful of) changing how they manage their diabetes. The inclusion of QoL and (in particular) DSQoL treatment satisfaction inventories went some way to quantifying this for participants in the research database. Nevertheless, analytical methods cannot fully incorporate the underlying subtleties of why, for example, younger people were less represented.

A limitation is that entry to the DAFNE research cohort was itself voluntary thereby raising possible volunteer bias, although the overall consent rate was over 80%[33]. A further limitation is that some DAFNE participants would inevitably have been ineligible for the RCT for reasons that were not captured by the database. Had this exclusion criteria been applied more fully to the DAFNE cohort we would almost certainly have found it more comparable to the trial population, but also less comparable to the wider population who may still be candidates for the intervention irrespective of the RCT inclusion criteria. The missing data, whilst not substantial, is another limitation which we have not addressed in this analysis.

Despite these limitations, the DAFNE cohort may still represent a "best case" of external data. The database was set up for research activities and is likely to be more extensive and of better quality than that which is generally available on non-recruited individuals. The "core" data collected at DAFNE sites is comparable to that which would be available routinely in clinic, and the differences between RCT and cohort were modest on these characteristics. By having access to a large cohort containing quality of life, diabetic complications and treatment goals/satisfaction we were able to better discriminate those in the RCT from those who were not. Although our reference data is undoubtedly a strength of the current research project, this is unlikely to be available to other researchers wanting to recalibrate their own findings to the wider population.

Finally, the specific factors associated with uptake will likely differ across interventions, settings, and trials. One feature of T1DM is the emphasis on self-management, which does not apply (or apply as strongly) to all conditions. The REPOSE trial also had relatively broad inclusion criteria and people were screened routinely rather than selectively with high uptake among those eligible. We do not advocate any attempt to extrapolate our relationships for trial inclusion outside of this study.

4.3 Other literature

Several reviews have shown RCTs to be non-random subsets of general patient populations (sometimes by design) with findings including the under-representation of older aged patients in trials of cancer [13], osteoarthritis[14] and lower back pain[18]; concomitant chronic morbidity in trials of coronary heart disease[17]; smoking habits and motivation in smoking cessation[19]; lung function in asthma[16]; unemployment and lower educational qualification in drug abuse[22]; and both low and high disease severity in pharmacological and psychotherapeutic depression trials [21]. Whilst considerations and adjustments for non-random sampling are commonplace in electoral polls and market research [27,28], equivalent developments for RCTs are rarely quantified[29] and have only recently received serious attention. Those which have done are often limited to basic demographic data in the general population [51–54], and the relatively small attenuations to the revised effect size may be because (non-)recruitment was not adequately characterised by these reasons alone.

A recent attempt to address these issues at the design stage is the PRECIS-2 tool [55,56], which incorporates the restrictiveness of exclusion criteria among its criteria. Similarly, the Innovative Medicines Initiative (IMI) described considerations for obtaining “real-world evidence” from RCTs which encompassed criteria for site selection and patient inclusion [5,10,57] – specifically, to make these more broad . Raymond [58] further advocated RCTs should be embedded within clinical care in order to evaluate standard practices on a larger scale. Nevertheless, even the most pragmatic and inclusive trials are reliant on the willingness of individuals to participate and therefore subject to volunteer bias[8,59]. Under-representation is therefore unlikely to be addressed by design considerations alone.

This raises the final, fundamental question regarding how the wider world itself is to be characterised. Some authors advocate RCTs nested within cohorts (akin to REPOSE) [60,61], although this requires the cohort also contains a random subsample of the population being studied. Makady et al[62] reported a literature review and stakeholder interviews to attempt to reach consensus on what constitutes “real world data”; whilst non-RCT settings were unsurprisingly favoured, there was inconsistency among those interviewed on which sources (e.g. general practice data, disease specific registries, claim databases) appropriately represent the wider population. The quality of such data is a further concern with high levels of miscoding, misclassification and misdiagnosis reported for patients with diabetes in routinely collected data across several western countries [63–65]. The impact of this is difficult to quantify, but there appears inherent inaccuracy in the characterisation (and indeed identification) of the wider disease populations.

4.4 Recommendations

Whilst we encourage researchers to compare their data against a reference population, there are barriers to doing so. Obtaining reference data is often difficult, increases the trial cost and may lead to unwelcome findings. Investigating representativeness is necessarily a part-exploratory analysis, and an investigator faced with inconsistent findings from competing models will be conflicted as to how to report these if the validity of the RCT is called into question. Lastly, whilst non-overlap (e.g. a c-statistic of 0.7) suggests a non-random RCT sample, the converse is not necessarily true; a c-statistic of 0.5 may arise because the RCT is truly a random (and therefore unpredictable) sample, but may also arise because the model is incomplete. All in all, it is easier not to look.

We nevertheless encourage more consideration be given to the generalisation of RCTs, the simple reason being that RCT evidence may be downplayed (or even overlooked) because it appears distant from a readers own practice and experience[66,67]. A sensitivity analysis that re-weights the findings to the wider population (however basic) may help bridge this gap. Such analyses will ideally be planned in advance as far as possible and consider the sample size implications of the modelling at the design stage [68,69]

5 Conclusions

We found our RCT differed from the wider population but nevertheless gave the same results on sampling reweighting. We argue that clinical investigators and statisticians should apply methods such as these more often. There is an increasing belief that RCTs and real-world evidence are divergent, and prospectively planned analyses that compare RCT participants against an external population can help build confidence in its findings. Obtaining and modelling external data are not straightforward, but we believe that any step, however incomplete, will better align RCTs to routine practice.

Acknowledgements

The authors gratefully acknowledge the participants in the DAFNE database cohort and the REPOSE RCT, and to Susan Moore at North Tyneside General Hospital for her help in providing the DAFNE database.

Funding sources

The REPOSE trial was funded by the UK National Institute for Health Research Health Technology Assessment Programme (project number 08/107/01) and the DAFNE research database was funded by the UK National Institute for Health Research Programme Grant for Applied Research (project number PG-0606-1184). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HTA, the National Institute for Health Research, the NHS or the Department of Health.

References

[1] Rothwell PM. External validity of randomised controlled trials: “To whom do the results of

this trial apply?" Lancet 2005;365:82–93. doi:10.1016/S0140-6736(04)17670-8.

[2] Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to Applicability of Randomised Trials: Exclusions and Selective Participation. J Health Serv Res Policy 1999;4:112–21. doi:10.1177/135581969900400210.

[3] Boughner RL. Volunteer Bias. Encycl. Res. Des., Thousand Oaks California United States: SAGE Publications, Inc.; 2010. doi:10.4135/9781412961288.n492.

[4] Califf RM, Harrington RA. American Industry and the U.S. Cardiovascular Clinical Research Enterprise. J Am Coll Cardiol 2011;58:677–80. doi:10.1016/j.jacc.2011.03.048.

[5] Worsley SD, Oude Rengerink K, Irving E, Lejeune S, Mol K, Collier S, et al. Series: Pragmatic trials and real world evidence: Paper 2. Setting, sites, and investigator selection. J Clin Epidemiol 2017;88:14–20. doi:10.1016/j.jclinepi.2017.05.003.

[6] Rothwell PM. Factors That Can Affect the External Validity of Randomised Controlled Trials. PLoS Clin Trials 2006;1:e9. doi:10.1371/journal.pctr.0010009.

[7] Zwarenstein M, Oxman A. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006;59:1125–6. doi:10.1016/j.jclinepi.2006.05.010.

[8] Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A (Statistics Soc 2011;174:369–86. doi:10.1111/j.1467-985X.2010.00673.x.

[9] O'Muircheartaigh C, Hedges L V. Generalizing from unrepresentative experiments: a stratified propensity score approach. J R Stat Soc Ser C (Applied Stat 2014;63:195–210. doi:10.1111/rssc.12037.

[10] Malmivaara A. Generalizability of findings from randomized controlled trials is limited in the leading general medical journals. J Clin Epidemiol 2019;107:36–41. doi:10.1016/j.jclinepi.2018.11.014.

[11] Humphreys K, Maisel NC, Blodgett JC, Fuh IL, Finney JW. Extent and Reporting of Patient Nonenrollment in Influential Randomized Clinical Trials, 2002 to 2010. JAMA Intern Med 2013;173:1029. doi:10.1001/jamainternmed.2013.496.

[12] Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. Epidemiol Perspect Innov 2007;4:4. doi:10.1186/1742-5573-4-4.

[13] Al-Refaie WB, Vickers SM, Zhong W, Parsons H, Rothenberger D, Habermann EB. Cancer trials versus the real world in the United States. Ann Surg 2011;254:438–42; discussion 442-3. doi:10.1097/SLA.0b013e31822a7047.

[14] Liberopoulos G, Trikalinos NA, Ioannidis JPA. The elderly were under-represented in osteoarthritis clinical trials. J Clin Epidemiol 2009;62:1218–23. doi:10.1016/j.jclinepi.2008.12.009.

[15] Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16:495. doi:10.1186/s13063-015-1023-4.

[16] Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, Shirtcliffe P, et al. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? Thorax 2007;62:219–23. doi:10.1136/thx.2006.066837.

[17] Buffel du Vaure C, Dechartres A, Battin C, Ravaud P, Boutron I. Exclusion of patients with concomitant chronic conditions in ongoing randomised controlled trials targeting 10 common chronic conditions and registered at ClinicalTrials.gov: a systematic review of registration

details. BMJ Open 2016;6:e012265. doi:10.1136/bmjopen-2016-012265.

[18] Paeck T, Ferreira ML, Sun C, Lin C-WC, Tiedemann A, Maher CG. Are older adults missing from low back pain clinical trials? A systematic review and meta-analysis. Arthritis Care Res (Hoboken) 2014;66:1220–6. doi:10.1002/acr.22261.

[19] Motschman CA, Gass JC, Wray JM, Germeroth LJ, Schlienz NJ, Munoz DA, et al. Selection criteria limit generalizability of smoking pharmacotherapy studies differentially across clinical trials and laboratory studies: A systematic review on varenicline. Drug Alcohol Depend 2016;169:180–9. doi:10.1016/j.drugalcdep.2016.10.018.

[20] Gollop ND, Ford J, Mackeith P, Thurlow C, Wakelin R, Steel N, et al. Are patients in heart failure trials representative of primary care populations? A systematic review. BJGP Open 2018;2:bjgpopen18X101337. doi:10.3399/bjgpopen18X101337.

[21] von Wolff A, Jansen M, Hölzel LP, Westphal A, Härter M, Kriston L. Generalizability of findings from efficacy trials for chronic depression: an analysis of eligibility criteria. Psychiatr Serv 2014;65:897–904. doi:10.1176/appi.ps.201300309.

[22] Susukida R, Crum RM, Stuart EA, Ebnesajjad C, Mojtabai R. Assessing sample representativeness in randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. Addiction 2016;111:1226–34. doi:10.1111/add.13327.

[23] Braslow JT, Duan N, Starks SL, Polo A, Bromley E, Wells KB. Generalizability of Studies on Mental Health Treatment and Outcomes, 1981 to 1996. Psychiatr Serv 2005;56:1261–8. doi:10.1176/appi.ps.56.10.1261.

[24] Wang W, Ma Y, Huang Y, Chen H. Generalizability analysis for clinical trials: a simulation study. Stat Med 2017;36:1523–31. doi:10.1002/sim.7238.

[25] Santacatterina M, Bottai M. Inferences and conjectures in clinical trials: a systematic review of generalizability of study findings. J Intern Med 2016;279:123–6. doi:10.1111/joim.12389.

[26] Fuller J. Rationality and the generalization of randomized controlled trial evidence. J Eval Clin Pract 2013;19:644–7. doi:10.1111/jep.12021.

[27] Armstrong JS, Overton TS. Estimating Nonresponse Bias in Mail Surveys. J Mark Res 1977;14:396. doi:10.2307/3150783.

[28] Gideon (ed) L. Handbook of Survey Methodology for the Social Sciences. New York, NY: Springer New York; 2012. doi:10.1007/978-1-4614-3876-2.

[29] Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. Am J Epidemiol 2018. doi:10.1093/aje/kwy228.

[30] REPOSE Study Group. Relative effectiveness of insulin pump treatment over multiple daily injections and structured education during flexible intensive insulin treatment for type 1 diabetes: cluster randomised trial (REPOSE). Bmj 2017;356:j1285. doi:10.1136/bmj.j1285.

[31] DAFNE Study Group. Training in flexible, intensive insulin management to enable dietary freedom in people with type 1 diabetes: dose adjustment for normal eating (DAFNE) randomised controlled trial. BMJ 2002;325:746.

[32] NICE. Type 1 diabetes in adults: diagnosis and management (NICE guideline CG17) 2015.

[33] Heller S, Lawton J, Amiel S, Cooke D, Mansell P, Brennan A, et al. Improving management of type 1 diabetes in the UK: the Dose Adjustment For Normal Eating (DAFNE) programme as a research test-bed. A mixed-method analysis of the barriers to and facilitators of successful diabetes self-management, a health economic analys. Program Grants Appl Res 2014;2:1–188. doi:10.3310/pgfar02050.

[34] Tipton E. How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. J Educ Behav Stat 2014;39:478–501. doi:10.3102/1076998614558486.

[35] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55. doi:10.1093/biomet/70.1.41.

[36] Hansen BB. The prognostic analogue of the propensity score. Biometrika 2008;95:481–8. doi:10.1093/biomet/asn004.

[37] Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. Am J Epidemiol 2017;186:1010–4. doi:10.1093/aje/kwx164.

[38] Tipton E. Stratified Sampling Using Cluster Analysis. Eval Rev 2013;37:109–39. doi:10.1177/0193841X13516324.

[39] Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. J R Stat Soc Ser A (Statistics Soc 2015;178:757–78. doi:10.1111/rssa.12094.

[40] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937–60. doi:10.1002/sim.1903.

[41] Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLoS One 2011;6:e18174. doi:10.1371/journal.pone.0018174.

[42] Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Appl Stat 1994;43:429. doi:10.2307/2986270.

[43] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Wiley; 2013.

[44] Newson RB. The role of Somers' D in propensity modelling. 22nd UK Stata Users' Gr Meet 2016. https://www.stata.com/meeting/uk16/slides/newson_uk16.pdf (accessed September 21, 2017).

[45] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083–107. doi:10.1002/sim.3697.

[46] Jenkinson C, Layte R, Jenkinson D, Lawrence K, Petersen S, Paice C, et al. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? J Public Health Med 1997;19:179–86.

[47] Najafzadeh M, Schneeweiss S. From Trial to Target Populations ? Calibrating Real-World Data. N Engl J Med 2017;376:1203–5. doi:doi:10.1056/NEJMp1614720.

[48] Brakenhoff TB, van Smeden M, Visseren FLJ, Groenwold RHH. Random measurement error: Why worry? An example of cardiovascular risk factors. PLoS One 2018;13:e0192298. doi:10.1371/journal.pone.0192298.

[49] Estcourt S, Epton J, Epton T, Vaidya B, Daly M. Exploring the motivations of patients with type 2 diabetes to participate in clinical trials: a qualitative analysis. Res Involv Engagem 2016;2:34. doi:10.1186/s40900-016-0050-y.

[50] McCann SK, Campbell MK, Entwistle VA. Reasons for participating in randomised controlled trials: conditional altruism and considerations for self. Trials 2010;11:31. doi:10.1186/1745-6215-11-31.

[51] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol 2010;172:107–15. doi:10.1093/aje/kwq084.

[52] Robinson D, Woerner MG, Pollack S, Lerner G. Subject selection biases in clinical trials: data from a multicenter schizophrenia treatment study. J Clin Psychopharmacol 1996;16:170–6.

[53] Jordan S, Watkins A, Storey M, Allen SJ, Brooks CJ, Garaiova I, et al. Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis. PLoS One 2013;8:e67912. doi:10.1371/journal.pone.0067912.

[54] Rovers MM, Straatman H, Ingels K, van der Wilt G-J, van den Broek P, Zielhuis GA. Generalizability of trial results based on randomized versus nonrandomized allocation of OME infants to ventilation tubes or watchful waiting. J Clin Epidemiol 2001;54:789–94. doi:10.1016/S0895-4356(01)00340-7.

[55] Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;62:464–75. doi:10.1016/j.jclinepi.2008.12.011.

[56] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015;350:h2147–h2147. doi:10.1136/bmj.h2147.

[57] Oude Rengerink K, Kalkman S, Collier S, Ciaglia A, Worsley SD, Lightbourne A, et al. Series: Pragmatic trials and real world evidence: Paper 3. Patient selection challenges and consequences. J Clin Epidemiol 2017;89:173–80. doi:10.1016/j.jclinepi.2016.12.021.

[58] Raymond J. Reflections on the TEAM Trial: Why Clinical Care and Research Should be Reconciled. Can J Neurol Sci / J Can Des Sci Neurol 2011;38:198–202. doi:10.1017/S0317167100011343.

[59] Kaptchuk TJ. The double-blind, randomized, placebo-controlled trial. J Clin Epidemiol 2001;54:541–9. doi:10.1016/S0895-4356(00)00347-4.

[60] Ioannidis JPA, Adami H-O. Nested randomized trials in large cohorts and biobanks: studying the health effects of lifestyle factors. Epidemiology 2008;19:75–82. doi:10.1097/EDE.0b013e31815be01c.

[61] Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. BMJ 2010;340:c1066. doi:10.1136/bmj.c1066.

[62] Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value Heal 2017;20:858–65. doi:10.1016/j.jval.2017.03.008.

[63] de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabet Med 2010;27:203–9. doi:10.1111/j.1464-5491.2009.02917.x.

[64] Hinton W, Liyanage H, McGovern A, Liaw S-T, Kuziemsky C, Munro N, et al. Measuring Quality of Healthcare Outcomes in Type 2 Diabetes from Routine Data: a Seven-nation Survey Conducted by the IMIA Primary Health Care Working Group. Yearb Med Inform 2017;26:201–8. doi:10.15265/IY-2017-005.

[65] Liaw S-T, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records - existing tools may be unreliable and potentially unsafe. Aust Fam Physician 2013;42:820–3.

[66] Greenfield S. Making Real-World Evidence More Useful for Decision Making. Value Heal

2017;20:1023–4. doi:10.1016/j.jval.2017.08.3012.

[67] SIMON SD. Is the Randomized Clinical Trial the Gold Standard of Research? J Androl 2001;22:938–43. doi:10.1002/j.1939-4640.2001.tb03433.x.

[68] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276–96. doi:10.1002/sim.7992.

[69] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res 2019;28:2455–74. doi:10.1177/0962280218784726.

Box 1 Data items collected for the REPOSE RCT and the DAFNE database

Demographic information:

Age, sex, ethnicity, weight, BMI, smoking status*, previous pregnancy*

Diabetes characteristics:

Years since diagnosis, HbA1c, previous severe hypoglycaemic episode, previous DKA, diabetic complications*, years since last reported complication*, insulin dose*

Laboratory and vital signs:

Creatinine*, lipids*, blood pressure*

Quality of life:

Health utility (EQ-5D*), generic QoL (SF-12+), Diabetes Specific QoL (DSQoL)+

* Information collected in DAFNE research centres and RCT only.

+ information collected in DAFNE research centres up to December 2012 and RCT only.

List of Appendices

Appendix 1: Assessment of balance between RCT and non-RCT participants prior to re-weighting and under different re-weighting models.

Appendix 2. Stata statistical code

List of tables

List of figures

Table 1 Baseline characteristics of RCT participants and the external cohort

	RCT (N=267)	RCT centres during recruitment (N=339)	Research centres (N=2961)	Entire cohort (N=22020)
Demographics				
Age (years)	41 [28,50]	39 [27,51]	39 [28,49]	40 [28,51]
Missing	0	0	0	26 (0.1%)
Sex				
Male	160 (59.9%)	174 (51.3%)	1560 (52.7%)	10832 (49.2%)
Female	107 (40.1%)	165 (48.7%)	1397 (47.2%)	11097 (50.4%)
Missing	0	0	4 (0.1%)	91 (0.4%)
Previous pregnancy (in females) *				
Yes	48 (44.9%)	-	411 (29.4%)	
No	56 (52.3%)	-	941 (67.2%)	
Missing	3 (2.8%)	-	48 (3.4%)	
Ethnicity				
White British	244 (91.4%)	317 (93.5%)	2687 (90.7%)	20083 (91.2%)
Other white	8 (3.0%)	13 (3.8%)	214 (7.2%)	1331 (6.0%)
Other non-white	7 (2.6%)	7 (2.1%)	27 (0.9%)	294 (1.3%)
Missing/prefer not to say	8 (3.0%)	2 (0.6%)	33 (1.1%)	312 (1.4%)

Weight (Kg)	80 [68,91]	75 [64,86]	76 [65,87]	76 [66,87]
Missing	0	14 (4.1%)	93 (3.1%)	1064 (4.8%)
BMI (Kg/m2)	27 [24,30]	25 [23,28]	26 [23,29]	26 [23,29]
Missing	0	25 (7.4%)	190 (6.4%)	1782 (8.1%)
Smoking status				
Smoker	53 (19.9%)	26 (7.7%)	570 (19.3%)	570 (2.6%)
Ex-smoker	69 (25.8%)	32 (9.4%)	735 (24.8%)	735 (3.3%)
Never smoker	145 (54.3%)	63 (18.6%)	1616 (54.6%)	1616 (7.3%)
Missing	0 (0.0%)	218 (64.3%)	40 (1.4%)	19099 (86.7%)
Diabetes characteristics and complications				
Years since diagnosis	16 [8,27]	13 [5,23]	14 [5,25]	15 [7,26]
Missing	0	4 (1.2%)	22 (0.7%)	424 (1.9%)
Any previous severe hypoglycaemic episode	30 (11.2%)	47 (13.9%)	637 (21.5%)	4583 (20.8%)
Missing	0	0	0	0
Any previous DKA	20 (7.5%)	27 (8.0%)	226 (7.6%)	1593 (7.2%)
Missing	0	30 (8.8%)	203 (6.9%)	1413 (6.4%)

Daily insulin dose (IU/Kg weight) *	0.69 [0.53,0.89]	-	0.66 [0.50,0.85]	-
Missing	0	-	197 (6.7%)	-
Diabetic complications				
Any complication *	148 (55.4%)	-	1137 (38.4%)	-
Neuropathy *	19 (7.1%)	-	145 (4.9%)	-
Retinopathy *	115 (43.1%)	-	690 (23.3%)	-
Other complication *	101 (37.8%)	-	818 (27.6%)	-
Laboratory measures				
HbA1c (mmol/mol) *	72 [65,85]	72 [60,81]	70 [60,83]	70 [61,81]
Missing	0	18 (5.3%)	146 (4.9%)	1988 (9.0%)
Creatinine(umol/L) *	73 [64,86]	-	74 [65,85]	-
Missing	4 (1.5%)	-	302 (10.2%)	-
Cholesterol (mmol/mol) *	4.90 [4.30,5.60]	-	4.60 [4.00,5.20]	-
Missing	1 (0.4%)	-	274 (9.3%)	-
Triglycerides(mmol/L) *	1.10 [0.80,1.60]	-	1.00 [0.70,1.50]	-
Missing	0	-	429 (14.5%)	-
Generic QoL and health utility				
SF-12 Physical composite QoL*	53.1 [48.3,56.7]	-	52.7 [45.3,56.1]	-
Missing	10 (3.7%)	-	582 (19.7%)	-

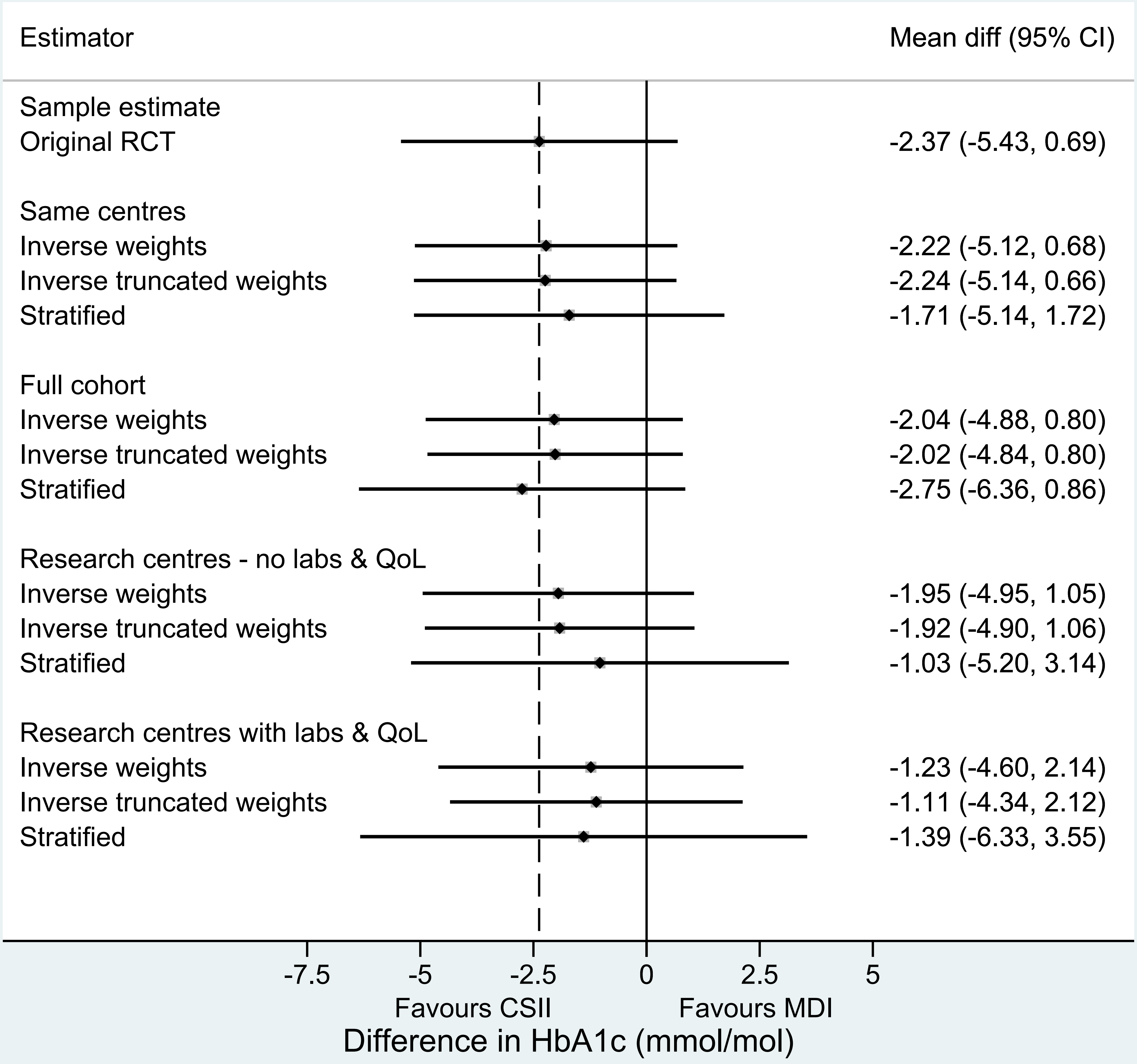
SF-12 Mental composite QoL*	49.4 [42.0,54.4]	-	48.9 [38.7,55.0]	-
Missing	10 (3.7%)	-	582 (19.7%)	-
EQ-5D health utility *	0.848 [0.760,1.000]	-	1.000 [0.779,1.000]	-
Missing	4 (1.5%)	-	168 (5.7%)	-
Diabetes specific QoL				
Treatment goals*	74.0 [68.0,82.0]	-	78.0 [70.0,84.0]	-
Missing	4 (1.5%)	-	561 (18.9%)	-
Treatment satisfaction *	62.0 [48.9,72.0]	-	62.0 [51.1,74.0]	-
Missing	4 (1.5%)	-	563 (19.0%)	-
Preference weighted treatment satisfaction *	57.4 [47.3,66.3]	-	59.3 [49.3,68.7]	-
Missing	4 (1.5%)	-	577 (19.5%)	-
Social relations *	14.5 [5.5,29.1]	-	18.2 [7.3,32.7]	-
Missing	4 (1.5%)	-	563 (19.0%)	-
Leisure time restrictions and flexibility *	20.0 [10.0,36.7]	-	20.0 [6.7,40.0]	-
Missing	4 (1.5%)	-	578 (19.5%)	-
Physical complaints *	25.0 [12.5,40.0]	-	27.5 [12.5,45.0]	-
Missing	4 (1.5%)	-	577 (19.5%)	-
Worries about the future *	56.0 [36.0,76.0]	-	60.0 [36.0,80.0]	-
Missing	4 (1.5%)	-	579 (19.6%)	-
Daily hassle of functions *	30.0 [15.0,50.0]	-	35.0 [20.0,53.3]	-

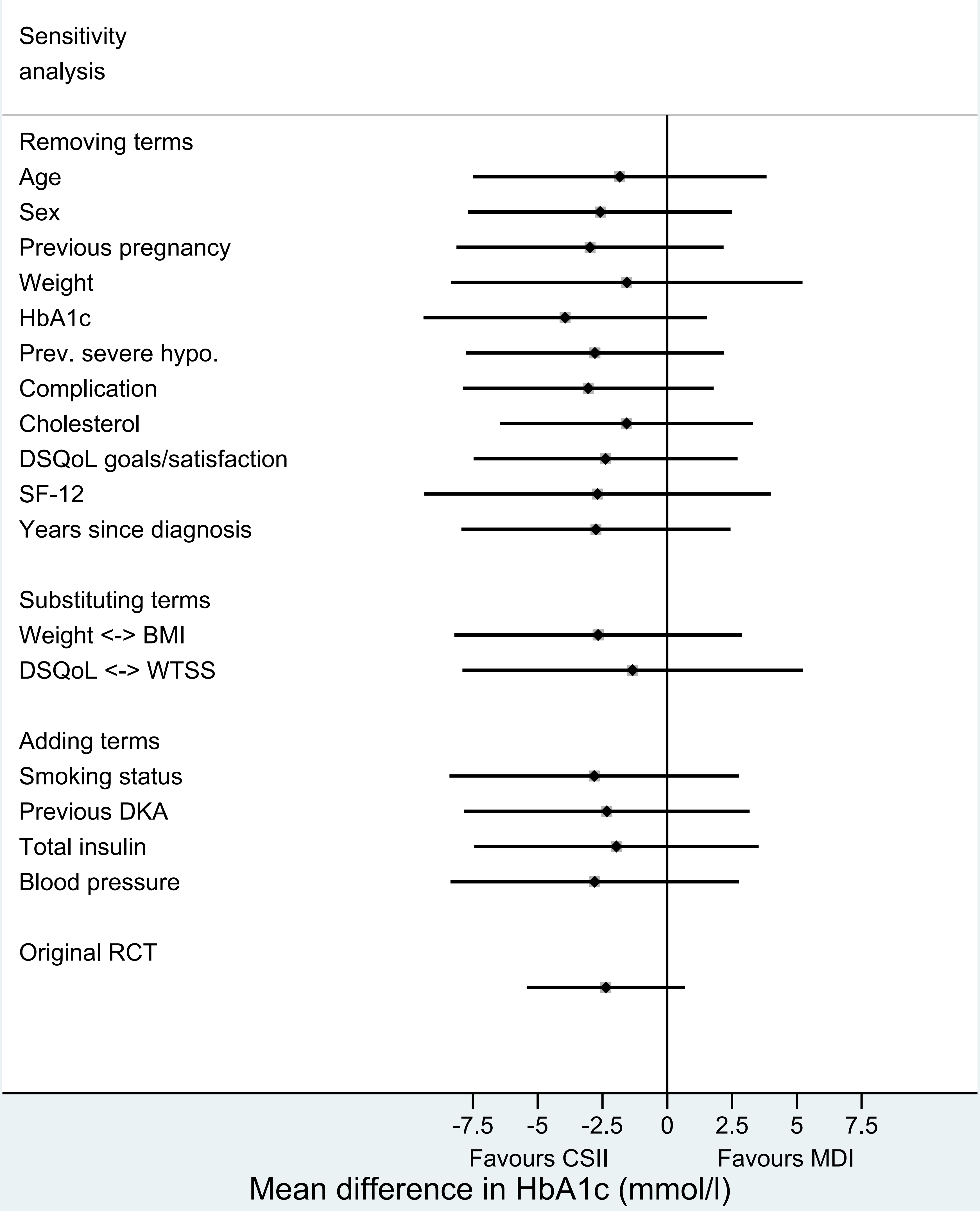
Missing	4 (1.5%)	-	565 (19.1%)	-

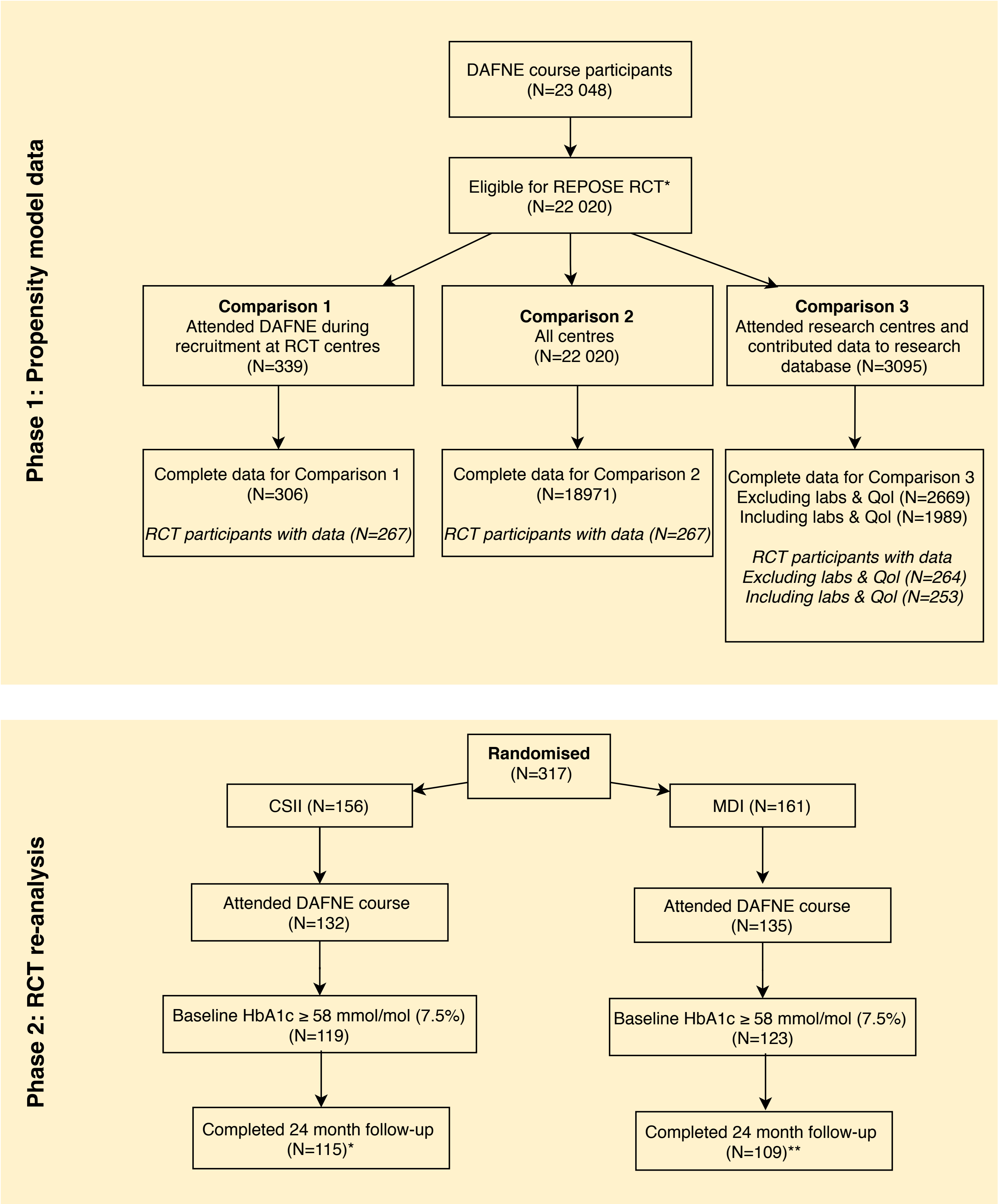
Figures are median [iqr] or number (%) *Collected only in research centres.

Table 2 Indices of comparability of RCT and cohort

		Indices of generalisability		
Estimator	N	c-statistic	SMD	Tipton index
Same centres	573	0.62	0.45	0.96
Full cohort	19238	0.63	0.45	0.95
Research centres				
No labs & QoL	2933	0.68	0.63	0.93
Including labs & QoL	2242	0.72	0.89	0.91


Estimator
Mean diff (95% CI)
Sample estimate
Original RCT
-2.37 (-5.43, 0.69)
Same centres
Inverse weights
-2.22 (-5.12, 0.68)
Inverse truncated weights
-2.24 (-5.14, 0.66)
Stratified
-1.71 (-5.14, 1.72)
Full cohort
Inverse weights
-2.04 (-4.88, 0.80)
Inverse truncated weights
-2.02 (-4.84, 0.80)
Stratified
-2.75 (-6.36, 0.86)
Research centres - no labs & QoL
Inverse weights
-1.95 (-4.95, 1.05)
Inverse truncated weights
-1.92 (-4.90, 1.06)
Stratified
-1.03 (-5.20, 3.14)
Research centres with labs & QoL
Inverse weights
-1.23 (-4.60, 2.14)
Inverse truncated weights
-1.11 (-4.34, 2.12)
Stratified
-1.39 (-6.33, 3.55)
-7.5
-5
-2.5
0
2.5
5
Favours CSII
Favours MDI
Difference in HbA1c (mmol/mol)


Sensitivity
analysis
Removing terms
Age
Sex
Previous pregnancy
Weight
HbA1c
Prev. severe hypo.
Complication
Cholesterol
DSQoL goals/satisfaction
SF-12
Years since diagnosis
Substituting terms
Weight <-> BMI
DSQoL <-> WTSS
Adding terms
Smoking status
Previous DKA
Total insulin
Blood pressure
Original RCT
-7.5
-5
-2.5
0
2.5
5
7.5
Favours CSII
Favours MDI
Mean difference in HbA1c (mmol/l)


Phase 1: Propensity model data
DAFNE course participants (N=23 048)
Eligible for REPOSE RCT* (N=22 020)
Comparison 1
Attended DAFNE during recruitment at RCT centres (N=339)
Comparison 2
All centres (N=22 020)
Comparison 3
Attended research centres and contributed data to research database (N=3095)
Complete data for Comparison 1 (N=306)
RCT participants with data (N=267)
Complete data for Comparison 2 (N=18971)
RCT participants with data (N=267)
Complete data for Comparison 3
Excluding labs & Qol (N=2669)
Including labs & Qol (N=1989)
RCT participants with data
Excluding labs & Qol (N=264)
Including labs & Qol (N=253)
Phase 2: RCT re-analysis
Randomised (N=317)
CSII (N=156)
MDI (N=161)
Attended DAFNE course (N=132)
Attended DAFNE course (N=135)
Baseline HbA1c ≥ 58 mmol/mol (7.5%) (N=119)
Baseline HbA1c ≥ 58 mmol/mol (7.5%) (N=123)
Completed 24 month follow-up (N=115)*
Completed 24 month follow-up (N=109)**

* 3 participants had no data for comparison 3 (no labs/QoL); 8 participants had no data for comparison 3 (incl. labs/QoL)
** 6 participants had no data for comparison 3 (incl. labs/QoL)

CRediT author statement

Mike Bradburn: Conceptualization; Data curation; Formal analysis; Visualization; Writing – original draft; Writing – review & editing; Investigation; Funding acqusition; Supervision.

Ellen Lee: Data curation; Formal analysis; Visualization; Writing – review & editing.

David White: Writing – review & editing; Project administration.

Daniel Hind: Investigation; Writing – review & editing.

Norman Waugh: Funding acqusition; Writing – Review & Editing

Deborah Cooke: Funding acqusition; Writing – Review & Editing

David Hopkins: Funding acqusition; Writing – Review & Editing

Peter Mansell: Funding acqusition; Writing – Review & Editing

Simon Heller: Funding acqusition; Writing – review & editing.

Additional information:

Simon Heller was the chief investigator on the grant which funded DAFNE (NIHR PGfAR project number PG-0606-1184). Peter Mansell, Debbie Cooke and David Hopkins members were members of the DAFNE study team.

Simon Heller was the chief investigator on the grant which funded REPOSE (NIHR HTA project No 08/107/01). Mike Bradburn, Ellen Lee, David White, Norman Waugh and Peter Mansell were members of the REPOSE study team.

Mike Bradburn and Ellen Lee undertook the statistical analysis of the present manuscript. Daniel Hind supplied additional references and information which were used in the background and discussion. All authors contributed to the writing of the manuscript.

What’s new

Key findings

- RCTs have been criticised for failing to adequately represent the wider population
- We re-analysed an RCT in type I diabetes mellitus using re-weighting methods to better match a large national cohort with regards to disease characteristics, QoL and treatment satisfaction
- Our trial participants differed from those of the cohort, but the revised treatment effect was similar to that seen in the original RCT

What this adds to what is known;

- Our case study is consistent with claims that RCTs are selective populations, but emphasises this does not necessarily invalidate the RCT findings

What is the implication and what should change now

- We advocate trialists consider comparisons such as those employed here to explore the potential impact of external generalisability