The findings of a surgical hip fracture trial were generalisable to the UK national hip fracture database

Hopin Lee[1,2]*	Email: hopin.lee@ndorms.ox.ac.uk
Jonathan A Cook[1]	Email: jonathan.cook@ndorms.ox.ac.uk
Sarah E Lamb[1]	Email: sarah.lamb@ndorms.ox.ac.uk
Nick Parsons[3]	Email: nick.parsons@warwick.ac.uk
David J Keene[4]	Email: david.keene@ndorms.ox.ac.uk
Alex L Sims[5]	Email: asims@nhs.net
Matthew L Costa[4]	Email: matthew.costa@ndorms.ox.ac.uk
Xavier L Griffin[6,7]	Email: x.griffin@qmul.ac.uk

1. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
2. School of Medicine and Public Health, University of Newcastle, Newcastle, Australia
3. Statistics and Epidemiology Unit, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK
4. Kadoorie Centre, John Radcliffe Hospital, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
5. Northumbria NHS Foundation Trust, Northumberland, UK
6. Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Newark Street, London, UK
7. Barts Health NHS Trust, London, UK

***Corresponding Author:**

Dr Hopin Lee
Botnar Research Centre
Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences, University of Oxford, Windmill Road, Headington, Oxford, UK OX3 7LD
Tel: +44 018 652 26493
Email: hopin.lee@ndorms.ox.ac.uk

Declarations of interest: none

Abstract

Objective: To estimate the generalisability of treatment effects observed in a randomised trial of hip fracture surgery implants, to a broader population of people undergoing hip surgery in the United Kingdom.

Study Design and Setting: In 2018, the WHiTE-3 trial (n=958) demonstrated that a modular hemiarthroplasty implant conferred no additional benefit over the traditional monoblock implant for quality of life and length of hospital stay. We compared and weighted the trial sample against two target populations: WHiTE-cohort (n=2,457) and UK-National Hip Fracture Database (NHFD, n=190,894), and re-estimate expected treatment effects for the target populations.

Results: Despite differences in baseline characteristics of the trial sample and target populations, the re-estimated treatment effects were comparable. For quality of life, the differences between the trial estimate and WHiTE-cohort and NHFD estimates were 0.01 points on the EuroQol (EQ5D). For length of stay, the difference between the trial estimate and WHiTE-cohort was 0.50 days; and the difference between the trial estimate and NHFD estimate was -0.47 days.

Conclusion: This generalisability analysis of the WHiTE-3 trial found that the inferences from the trial can be generalised to the group of individuals in the UK NHFD and the WHiTE-cohort who met the inclusion criteria for WHiTE-3.

Key words: External validity, generalizability, hip fracture, surgery

Running title: Generalisability of treatment effects from RCTs

Word Count: Abstract: 198; main text: 3512

What is new?

Key findings

- Participants recruited to an RCT of hip fracture surgery implants differed from individuals captured in the wider UK national hip fracture database who would have been eligible for the trial.

- By using inverse probability weighting, this generalisability analysis shows that the treatment effect estimates derived from the hip surgery RCT would be generalisable to the wider UK population with intracapsular hip fractures, based on the selected covariates that were modelled.

What this adds to what was known

- It was unclear whether the findings of the hip surgery RCT would be externally valid to a wider target population of patients with similar hip fractures in the UK

- This generalisability analysis provides some assurance (limited by the selected covariates modelled) that the findings of the hip surgery RCT are generalisable to the wider UK population despite some differences in key patient characteristics.

What is the implication and what should change now?

- The findings of this hip surgery RCT should be implemented to the wider UK population; and where possible, generalisability analyses should be routinely conducted to assess the external validity of RCTs.

Introduction

Randomised controlled trials (RCTs) generate internally valid evidence about the effectiveness of health interventions. However for RCTs to inform policy and practice, their findings should also be generalisable to wider populations that comprise of patients in routine care [1–4]. The generalisability of RCTs can be compromised when eligibility criteria are restrictive, or if enrolment systematically limits the sample characteristics, either by clinicians recruiting selectively or by patients declining to take part [5,6]. If the characteristics that vary between the sample and the population also modify the treatment effect, the treatment effect estimates from an RCT may not be directly generalisable to this wider populations [5,7].

Many studies have explored the external validity of RCTs by documenting differences based upon summary demographics between RCT participants and target populations [21–25]. Although these descriptive comparisons can be useful for assessing the representativeness of RCT samples, they do not inform clinicians and policy makers about whether the treatment effects derived from RCTs can be generalised to a target population.

One way of formally assessing the generalisability of treatment effects from RCTs is to compare the target population average treatment effect (TATE) to the sample average treatment effect (SATE). The TATE is the expected treatment effect estimate in the broader population of individuals who would have been eligible but were not recruited into an RCT; whereas the SATE is the estimate that is directly derived from participants in the RCT [1]. The extent to which an RCT can be considered generalisable to a target population is reflected by how closely the SATE approximates the TATE. Differences between SATE and TATE can arise when baseline factors that modify the treatment effect are unequally distributed between the RCT sample and the target population [1,2]. A recent study by Bradburn et al. (2020) used a sample-weighting method to assess the generalisability of a diabetes trial [8]. We used a similar approach but with application to a different clinical setting with larger nested datasets (surgical management of hip fracture) and compared alternative estimators for assessing generalisability.

In a trial of 964 patients with intracapsular hip fractures, the World Hip Trauma Evaluation (WHiTE)-3 trial showed no clinically meaningful difference in health-related quality of life and length of hospital stay when the modern cemented modular polished-taper stemmed hemiarthroplasty was compared against the traditional cemented monoblock [9]. Although these findings are considered internally valid, it is unclear whether they can be generalised to the wider UK population with this injury. This is

particularly an issue because surgeons and patients may have preferences about surgical interventions which could induce selective sampling in surgical RCTs. The aim of this study is to assess the generalisability of the WHiTE-3 trial by estimating TATEs in the WHiTE-cohort and UK National Hip Fracture Database (NHFD) and comparing them against the SATE from WHiTE-3.

Methods

To assess the generalisability of the WHiTE-3 trial to the UK population, we used patient-level data from the WHiTE-3 trial, WHiTE-cohort and NHFD. By using a weighting approach to balance possible effect modifiers between the WHiTE-3 trial and target populations, we estimated the TATE for health-related quality of life (HRQoL) at four months post-fracture (primary outcome in WHiTE-3) and length of hospital stay (secondary outcome). We then compared the TATEs to their respective SATEs from the trial. The methods used to estimate the TATE are described below. All analyses were complete-case analyses.

Data sources and preparation

WHiTE-3 trial

WHiTE-3 is a pragmatic, multi-centre, two-arm, parallel group randomised controlled trial of patients 60 years and over with a hip fracture, nested within the WHiTE-cohort. The trial compared the modern Exeter-Unitrax cemented modular polished-taper stemmed hemiarthroplasty (modular arm, n=482) against the traditional Thompson cemented monoblock (monoblock arm, n=482) in patients with displaced intracapsular fractures of the hip. All patients >60 years with intracapsular hip fractures considered suitable for a hemiarthroplasty were eligible. Patients with pre-existing symptomatic hip arthritis were excluded [9]. In this generalisability analysis, we used baseline covariate data, the randomised treatment group assignment indicator, and the following outcome measures: HRQoL using the EuroQol (EQ-5D-5L) [10] at four months post-fracture and length of hospital stay (LOS) in days. As recommended by Parsons et al. (2018), we used death-adjusted EQ-5D-5L scores that assume that an individual's EQ-5D-5L score becomes 0 at death, and this value was carried forward to the endpoint [11]. More detail about the WHiTE-3 trial is provided in **Appendix A1** and Sims et al. (2018) [9].

Target populations: WHiTE-Cohort and NHFD

The WHiTE-cohort was established in 2011 to measure outcomes in a comprehensive cohort of UK patients with hip fracture. In the WHiTE-cohort, all patients are treated under a single comprehensive treatment pathway based on the National Institute for Health and Care Excellence (NICE) Hip Fracture

Guidelines [12]. The WHiTE-cohort also nests a series of observational studies and RCTs such as the WHiTE-3 trial [13]. Between 23-04-2011 and 16-11-2017, data on 8,673 patients were available.

The NHFD is a UK national registry that was established in 2007 by the British Geriatrics Society and British Orthopaedic Association. The primary function of the NHFD is to audit clinical care according to UK national standards. The NHFD collects demographic, operative and peri-operative and outcome data (not HRQoL) on patients treated for almost all patients who fracture their hip in England, Wales and Northern Ireland. We had access to anonymised data collected between 01-01-2008 and 31-12-2018, including 614,398 patients admitted to an English hospital.

We restricted both WHiTE-cohort and NHFD datasets to only include individuals who would have been eligible for the WHiTE-3 trial based on its inclusion criteria (patients over the age of 60 years, receiving a hemiarthroplasty for an intracapsular fracture of the hip).

Statistical analysis

Covariate selection

In this generalisability analysis, we identified baseline covariates from the WHiTE-cohort and NHFD that were also collected in the WHiTE-3 trial. These covariates were selected based on their plausibility of modifying the effects of treatments tested in WHiTE-3. From the WHiTE-cohort, these covariates were: age, sex, pre-fracture place of residence, pre-op ASA score, pre-fracture mobility, pathological fracture, pre-op AMTS score, diagnosed diabetes pre-fracture, regular smoker pre-fracture, alcohol units per week pre-fracture, diagnosed chronic renal failure pre-fracture, pre-fracture EQ-5D - Mobility subscale, EQ-5D - Self Care subscale, EQ-5D - Usual Activities subscale, EQ-5D - Pain subscale, EQ-5D - Anxiety subscale, and EQ-5D - VAS. For the NHFD, the available covariates were age, sex, place of residence, pre-op ASA score, pre-fracture mobility, pathological fracture, and pre-op AMTS score.

Data harmonisation

When the response levels of categorical covariates in the trial sample did not match the response levels in the target populations, we grouped the levels to the lowest categorisation so that the measures would be comparable across the trial and target populations. The target population datasets were then combined with the trial dataset so that the combined dataset included an indicator variable for trial/target population membership, the outcome variables from the trial (indicated as missing for the target population), randomised treatment group assignment indicator from the trial (indicated as

missing for the target population), and the common set of covariates across both the trial and target populations. In effect, we created two combined datasets: WHiTE-3:WHiTE-cohort and WHiTE-3:NHFD.

Modelling the probability of trial participation

We used logistic regression to model the probability of trial participation based on a linear combination of baseline covariates. We used the estimated probabilities from the logistic model to assess differences between the WHiTE-3 trial and target populations (WHiTE-cohort and NHFD) and to calculate inverse probability weight (IPW)s for the estimation of TATE [2,14].

Comparison of the trial and target population

To assess how well the WHiTE-3 trial sample represented the target populations (WHiTE-cohort and NHFD) with respect to the overall combination of observed covariates, we calculated Tipton's generalisability index [15]. Tipton's index ranges from 0 (no overlap between trial sample and target population) to 1 (the trial sample is equivalent to a random sample drawn from the target population based on selected covariates). Indices greater than 0.8 are usually considered to represent trial samples that are very similar to the target population [15].

We also assessed how similar the WHiTE-3 trial sample were to those in the target populations (WHiTE-cohort and NHFD) by assessing each covariate independently. To do this, we calculated the standardised mean difference (SMD) between the trial and target population (WHiTE-cohort/NHFD) for each covariate. The SMD was calculated as the difference in means between groups (WHiTE-3 trial sample versus WHiTE-Cohort/NHFD) divided by the standard deviation of the pooled values. We calculated these SMDs before and after weighting the WHiTE-3 trial sample by the IPWs.

Estimation of the treatment effect in the target population (TATE)

To estimate the TATE, we primarily used the inverse probability of trial participation method [14]. By using the estimated probabilities derived from the trial participation model, we weighted the WHiTE-3 trial sample. This involved assigning an IPW to each participant in the trial sample. Participants with a high probability of being in the target population (based on the joint distribution of observed covariates) were assigned a larger weight than those who had a lower probability of being in the target population. This produces in essence a weighted pseudo-trial sample that is more representative of the target population, based on the joint distribution of observed covariates. To estimate the TATE for HRQoL and LOS outcomes, we used a weighted linear regression to the pseudo-trial sample.

We applied the above strategy to estimate TATEs for HRQoL and LOS outcomes in the WHITE-Cohort and NHFD. We then compared the TATE to the original SATE derived from the WHITE-3 trial. We used the 'generalize' R package to compare the trial and target populations and to estimate TATEs [16]

Sensitivity analyses

Restricted set of covariates

To generalise the WHiTE-3 trial to the WHiTE-Cohort, we included 17 covariates that could plausibly modify the treatment effect. However, between WHiTE-3 trial and NHFD, we were only able to include 7 covariates because the remaining 10 covariates (diagnosed diabetes, regular smoker, alcohol units per week, diagnosed chronic renal failure, EQ-5D - Mobility subscale, EQ-5D - Self Care subscale, EQ-5D - Usual Activities subscale, EQ-5D - Pain subscale, EQ-5D - Anxiety subscale, and EQ-5D - VAS) were not measured in the NHFD. To explore the effect of omitting these covariates in the estimation of TATE, we generalised the WHiTE-3 trial to the WHiTE-Cohort by including a restricted set of covariates that were used to generalise WHiTE-3 to NHFD (age, sex, place of residence, ASA score, pre-fracture mobility, pathological fracture, and pre-op AMTS score).

Exclusion of trial participants from WHiTE-cohort

Because of the hierarchical partially-nested structure of the datasets (NHFD:WHiTE-cohort:WHiTE-3 trial), it was plausible for some WHiTE-3 trial participants to be captured in the NHFD and WHiTE-cohort datasets. Although we were able to uniquely identify and exclude WHiTE-3 participants from the WHiTE-cohort dataset, we had limited pseudo-anonymised NHFD data and thus were unable to identify WHiTE-3 participants from NHFD. To explore the effects of including/excluding WHiTE-3 trial participants from the target population datasets, we estimated the TATE in WHiTE-cohort with and without WHiTE-3 participants.

Results

Patient flow

Figure 1 outlines the patient flow in the WHiTE-cohort and NHFD. We excluded individuals who were ineligible for WHiTE-3, those recruited to WHiTE-3, individuals with missing covariates, and individuals with covariate values outside the bounds covered by the WHiTE-3 trial (individuals aged above 104 and with an ASA score of V). We excluded individuals in WHiTE-cohort and NHFD who had covariate levels

outside the bounds of covariate levels covered by the WHiTE-3 trial to avoid extrapolating outside the trial data and to avoid violation of the positivity assumption (explained in the discussion). This resulted in 190,894 individuals from the NHFD and 2,457 individuals from WHiTE-cohort. The total WHiTE-3 trial sample was 964. We excluded 6 individuals who had non-intracapsular fractures, resulting in 958 individuals in the trial sample. The baseline characteristics of individuals in the WHiTE-3 trial and individuals in target populations are presented in **Table 1**. **Appendices A2 and A3** describe baseline characteristics of individuals in the NHFD and WHiTE-Cohort who would have been eligible for the WHiTE-3 trial, and those excluded from the analysis.

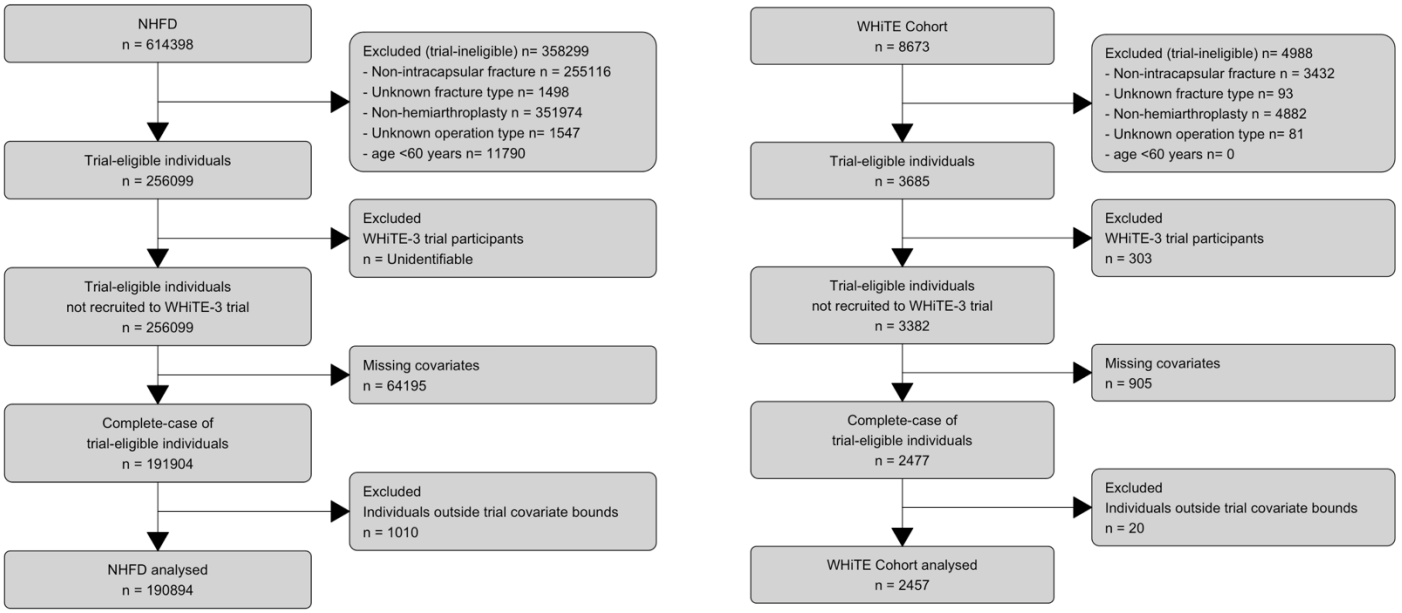

Figure 1. Patient flow in NHFD and WHiTE-cohort

Table 1: Baseline covariates of individuals in the NHFD, WHiTE-cohort, and WHiTE-3 trial

				WHiTE-3 Trial	
	NHFD (N=190894)	**WHiTE Cohort (N=2457)**	**WHiTE-3 Trial (N=958)**	**Thompson arm (N=481)**	**Exeter arm (N=477)**
Age					
Mean (SD)	83.8 (7.43)	84.0 (7.38)	83.4 (7.66)	83.3 (7.34)	83.5 (7.97)
Median [Min, Max]	85.0 [61.0, 104]	85.0 [61.0, 103]	84.0 [60.0, 104]	84.0 [62.0, 99.0]	85.0 [60.0, 104]
Sex					
Female	135886 (71.2%)	1777 (72.3%)	649 (67.7%)	326 (67.8%)	323 (67.7%)
Male	55008 (28.8%)	680 (27.7%)	309 (32.3%)	155 (32.2%)	154 (32.3%)
Residence					
Own home/sheltered housing	147736 (77.4%)	2037 (82.9%)	548 (73.3%)	272 (73.5%)	276 (73.0%)
Residential care	24708 (12.9%)	228 (9.28%)	113 (15.1%)	57 (15.4%)	56 (14.8%)
Nursing care	13213 (6.92%)	145 (5.90%)	63 (8.42%)	33 (8.92%)	30 (7.94%)
Rehabilitation unit	631 (0.331%)	3 (0.122%)	3 (0.401%)	2 (0.541%)	1 (0.265%)
Hospital	4060 (2.13%)	38 (1.55%)	19 (2.54%)	4 (1.08%)	15 (3.97%)
Other	546 (0.286%)	6 (0.244%)	2 (0.267%)	2 (0.541%)	0 (0%)
Missing	0 (0%)	0 (0%)	210 (21.9%)	111 (23.1%)	99 (20.8%)
ASA score					
I	2070 (1.08%)	26 (1.06%)	3 (0.403%)	1 (0.272%)	2 (0.532%)
II	45561 (23.9%)	604 (24.6%)	161 (21.6%)	78 (21.2%)	83 (22.1%)
III	114855 (60.2%)	1469 (59.8%)	468 (62.9%)	240 (65.2%)	228 (60.6%)
IV	28408 (14.9%)	358 (14.6%)	112 (15.1%)	49 (13.3%)	63 (16.8%)
Missing	0 (0%)	0 (0%)	214 (22.3%)	113 (23.5%)	101 (21.2%)
Prefracture mobility					
Freely mobile without aids	58847 (30.8%)	874 (35.6%)	297 (40.5%)	142 (39.2%)	155 (41.7%)
Mobile outdoors with one aid	41748 (21.9%)	690 (28.1%)	172 (23.4%)	82 (22.7%)	90 (24.2%)
Mobile outdoors with two aids or frame	23816 (12.5%)	442 (18.0%)	120 (16.3%)	61 (16.9%)	59 (15.9%)
Some indoor mobility but never goes outside without help	63497 (33.3%)	410 (16.7%)	128 (17.4%)	68 (18.8%)	60 (16.1%)
No functional mobility	2986 (1.56%)	41 (1.67%)	17 (2.32%)	9 (2.49%)	8 (2.15%)
Missing	0 (0%)	0 (0%)	224 (23.4%)	119 (24.7%)	105 (22.0%)
Pathological fracture					
Non-pathological	187691 (98.3%)	2425 (98.7%)	735 (99.5%)	363 (99.2%)	372 (99.7%)
Pathological	3203 (1.68%)	32 (1.30%)	4 (0.541%)	3 (0.820%)	1 (0.268%)
Missing	0 (0%)	0 (0%)	219 (22.9%)	115 (23.9%)	104 (21.8%)
Pre-op AMTS score					
Mean (SD)	6.83 (3.69)	7.44 (3.41)	6.55 (3.72)	6.44 (3.76)	6.66 (3.67)
Median [Min, Max]	9.00 [0, 10.0]	9.00 [0, 10.0]	8.00 [0, 10.0]	8.00 [0, 10.0]	8.00 [0, 10.0]

	NHFD (N=190894)	WHiTE Cohort (N=2457)	WHiTE-3 Trial (N=958)	WHiTE-3 Trial Thompson arm (N=481)	WHiTE-3 Trial Exeter arm (N=477)
Missing	0 (0%)	0 (0%)	219 (22.9%)	117 (24.3%)	102 (21.4%)
Diabetes					
No	NA	2083 (84.8%)	510 (83.5%)	253 (83.0%)	257 (84.0%)
Yes	NA	374 (15.2%)	101 (16.5%)	52 (17.0%)	49 (16.0%)
Missing	190894 (100%)	0 (0%)	347 (36.2%)	176 (36.6%)	171 (35.8%)
Regular smoker					
No	NA	2245 (91.4%)	566 (92.6%)	283 (92.8%)	283 (92.5%)
Yes	NA	212 (8.63%)	45 (7.36%)	22 (7.21%)	23 (7.52%)
Missing	190894 (100%)	0 (0%)	347 (36.2%)	176 (36.6%)	171 (35.8%)
Alcohol (units/week)					
0-7 units	NA	2245 (91.4%)	549 (89.9%)	274 (89.8%)	275 (89.9%)
8-14 units	NA	113 (4.60%)	32 (5.24%)	17 (5.57%)	15 (4.90%)
15-21 units	NA	47 (1.91%)	19 (3.11%)	9 (2.95%)	10 (3.27%)
> 21 units	NA	52 (2.12%)	11 (1.80%)	5 (1.64%)	6 (1.96%)
Missing	190894 (100%)	0 (0%)	347 (36.2%)	176 (36.6%)	171 (35.8%)
Chronic renal failure					
No	NA	2278 (92.7%)	589 (96.4%)	299 (98.0%)	290 (94.8%)
Yes	NA	179 (7.29%)	22 (3.60%)	6 (1.97%)	16 (5.23%)
Missing	190894 (100%)	0 (0%)	347 (36.2%)	176 (36.6%)	171 (35.8%)
EQ5D - Mobility					
No problem	NA	756 (30.8%)	161 (26.4%)	72 (23.6%)	89 (29.2%)
Slight problem	NA	604 (24.6%)	162 (26.6%)	74 (24.3%)	88 (28.9%)
Moderate problem	NA	714 (29.1%)	187 (30.7%)	110 (36.1%)	77 (25.2%)
Severe problem	NA	357 (14.5%)	91 (14.9%)	42 (13.8%)	49 (16.1%)
Unable	NA	26 (1.06%)	9 (1.48%)	7 (2.30%)	2 (0.656%)
Missing	190894 (100%)	0 (0%)	348 (36.3%)	176 (36.6%)	172 (36.1%)
EQ5D - Self Care					
No problem	NA	1478 (60.2%)	298 (48.9%)	139 (45.6%)	159 (52.1%)
Slight problem	NA	349 (14.2%)	96 (15.7%)	44 (14.4%)	52 (17.0%)
Moderate problem	NA	294 (12.0%)	98 (16.1%)	56 (18.4%)	42 (13.8%)
Severe problem	NA	127 (5.17%)	55 (9.02%)	32 (10.5%)	23 (7.54%)
Unable	NA	209 (8.51%)	63 (10.3%)	34 (11.1%)	29 (9.51%)
Missing	190894 (100%)	0 (0%)	348 (36.3%)	176 (36.6%)	172 (36.1%)
EQ5D - Usual Activities					
No problem	NA	1007 (41.0%)	205 (33.8%)	98 (32.3%)	107 (35.3%)
Slight problem	NA	435 (17.7%)	110 (18.2%)	55 (18.2%)	55 (18.2%)
Moderate problem	NA	441 (17.9%)	117 (19.3%)	51 (16.8%)	66 (21.8%)
Severe problem	NA	233 (9.48%)	84 (13.9%)	48 (15.8%)	36 (11.9%)
Unable	NA	341 (13.9%)	90 (14.9%)	51 (16.8%)	39 (12.9%)
Missing	190894 (100%)	0 (0%)	352 (36.7%)	178 (37.0%)	174 (36.5%)

	NHFD (N=190894)	WHiTE Cohort (N=2457)	WHiTE-3 Trial (N=958)	WHiTE-3 Trial Thompson arm (N=481)	WHiTE-3 Trial Exeter arm (N=477)
EQ5D - Pain					
No problem	NA	1217 (49.5%)	264 (43.3%)	130 (42.6%)	134 (43.9%)
Slight problem	NA	464 (18.9%)	137 (22.5%)	64 (21.0%)	73 (23.9%)
Moderate problem	NA	514 (20.9%)	130 (21.3%)	68 (22.3%)	62 (20.3%)
Severe problem	NA	228 (9.28%)	73 (12.0%)	38 (12.5%)	35 (11.5%)
Unable	NA	34 (1.38%)	6 (0.984%)	5 (1.64%)	1 (0.328%)
Missing	190894 (100%)	0 (0%)	348 (36.3%)	176 (36.6%)	172 (36.1%)
EQ5D - Anxiety					
No problem	NA	1392 (56.7%)	305 (50.1%)	147 (48.2%)	158 (52.0%)
Slight problem	NA	483 (19.7%)	137 (22.5%)	65 (21.3%)	72 (23.7%)
Moderate problem	NA	402 (16.4%)	106 (17.4%)	58 (19.0%)	48 (15.8%)
Severe problem	NA	145 (5.90%)	53 (8.70%)	30 (9.84%)	23 (7.57%)
Unable	NA	35 (1.42%)	8 (1.31%)	5 (1.64%)	3 (0.987%)
Missing	190894 (100%)	0 (0%)	349 (36.4%)	176 (36.6%)	173 (36.3%)
EQ5D - VAS					
Mean (SD)	NA (NA)	65.0 (21.1)	59.6 (22.0)	58.5 (22.3)	60.8 (21.7)
Median [Min, Max]	NA [NA, NA]	70.0 [2.00, 100]	60.0 [2.00, 100]	60.0 [2.00, 100]	60.0 [5.00, 100]
Missing	190894 (100%)	0 (0%)	349 (36.4%)	176 (36.6%)	173 (36.3%)

Results presented as mean (SD) for continuous variables and count (%) for discrete variables.

Probability of trial participation and inverse probability weights

Figure 2 presents the densities of the estimated probabilities of trial participation based on the observed covariates. These plots show overlap between the estimated probabilities of trial participation in the WHiTE-3 trial sample and the WHiTE-Cohort (upper panel), and NHFD (lower panel). The overlap between these probability densities suggest that it is appropriate to attempt generalisation to both target populations. These probabilities were used to calculate IPWs.

For the generalisation of WHiTE-3 to the WHiTE-Cohort, the sample average of the estimated IPWs was 5.3, and the maximum IPW was 22.8. For the generalisation of WHiTE-3 to the NHFD, the sample average of the estimated IPWs was 268.8, and the maximum IPW was 987.7. Intuitively, the individuals with larger weights (higher probability of being in the target population) have greater influence in the estimation of the TATE than the individuals with smaller weights (lower probability of being in the target population). To limit the influence of extreme weights, we truncated the weights at the 99th percentile (13.7 for WHiTE-cohort and 740.1 for NHFD). This removed 6 individuals for the generalisation to the

WHiTE-cohort, and 8 individuals for the generalisation to NHFD. The distributions of the weights before and after truncation are presented in **Appendix A4**.

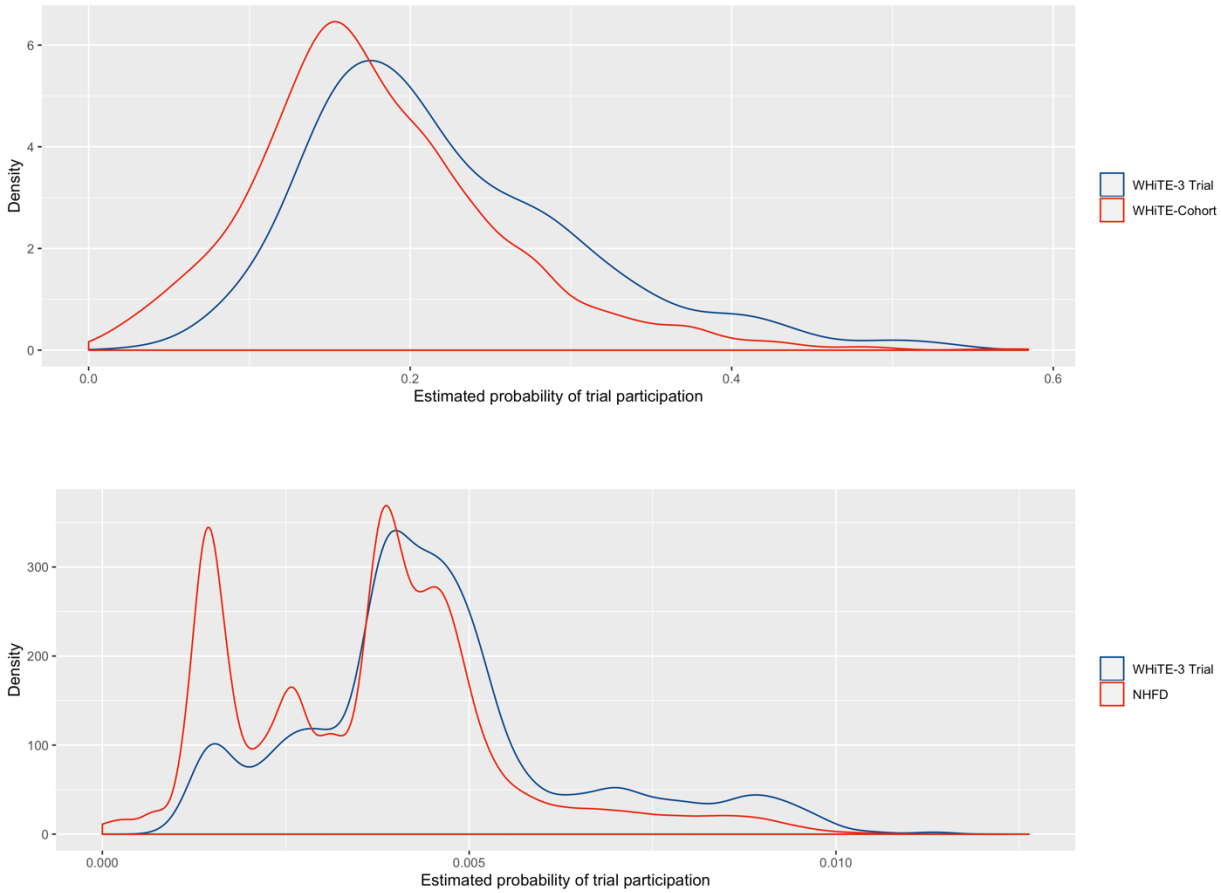

Figure 2. Densities of the estimated probabilities of trial participation

Comparisons between trial sample and target populations

For most covariates, the SMDs between the trial sample and target populations were smaller after weighting (**Figure 3**). This indicates that the weighted WHiTE-3 trial sample better resembled the WHiTE Cohort and NHFD after weighting. When the covariates were jointly considered, Tipton's generalisability index was 0.97 for generalisation to WHiTE-Cohort and 0.96 for generalisation to NHFD. These indices indicate that the weighted WHiTE-3 trial sample was highly representative of the WHiTE-Cohort and NHFD.

Target Population Average Treatment Effect (TATE)

Estimates of the TATE are presented in **Figure 4** and **Table 2**. The HRQoL SATE estimate from the WHiTE-3 trial was closely resembled by the TATE estimates for the NHFD and WHiTE-cohort. The difference between the WHiTE-cohort-TATE and WHiTE-3 SATE estimate was 0.01 points; and the difference between the NHFD-TATE and WHiTE-3 SATE point estimate was 0.01 points.

The LOS SATE estimate from the WHiTE-3 trial was also closely resembled by the TATE estimates for the WHiTE-cohort and NHFD. The mean difference between the WHiTE-cohort-TATE and WHiTE-3 SATE estimate was 0.50 days; and the mean difference between the NHFD-TATE and WHiTE-3 SATE estimate was -0.47 days.

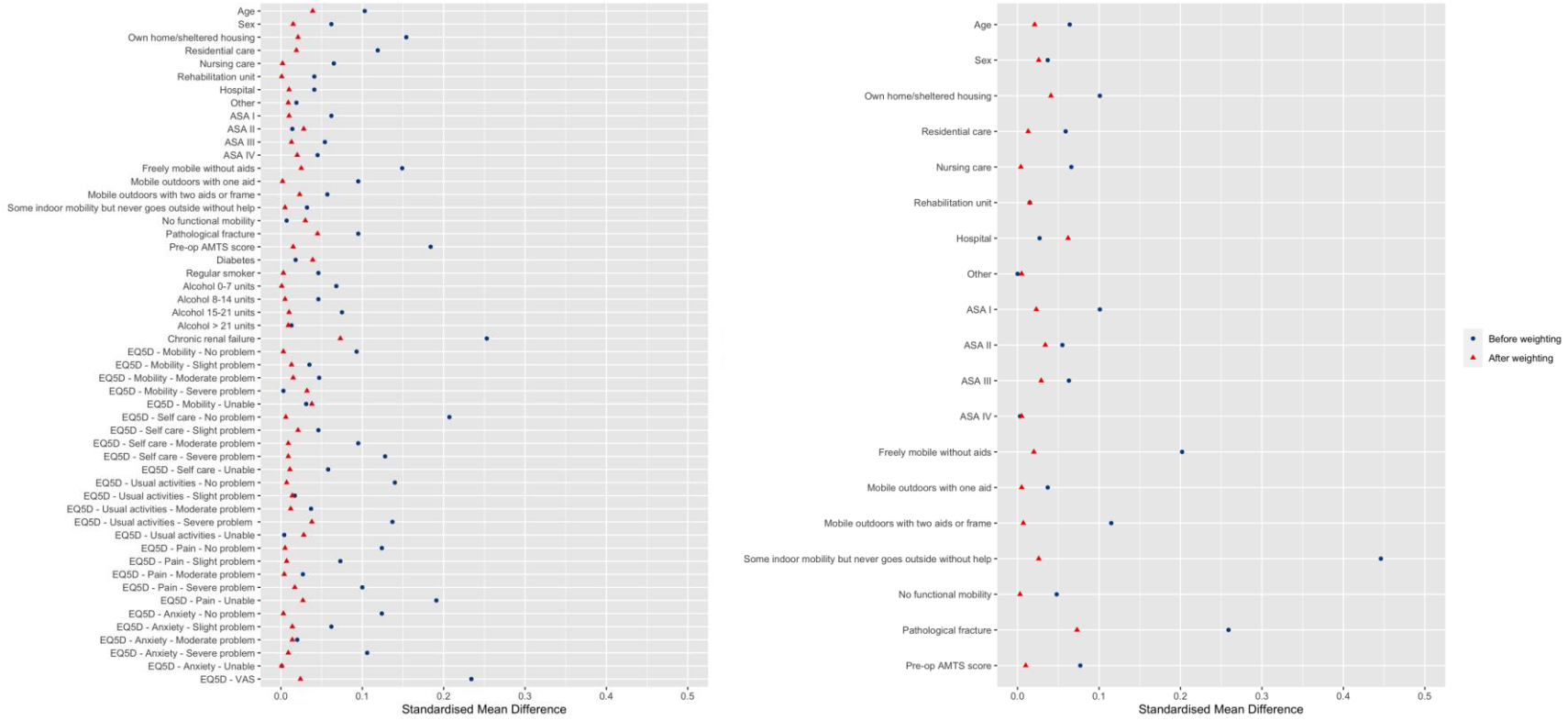

Figure 3. Standardised mean differences between trial and target populations before and after weighting. Left panel displays covariates used to generalise to WHiTE-Cohort and right panel displays covariates used to generalise to the UK National Hip Fracture Database.

Table 2. Target population Average Treatment Effect (TATE) and Sample Average Treatment Effect (SATE) estimates for Health-related Quality of Life (EQ5D) at 4 months and Length of Stay (LOS)

Outcome	Sample	Estimand	Point estimate	Lower 95% CI	Upper 95% CI	Tipton's generalisability index*
HRQoL	WHiTE Trial	SATE	0.06	-0.01	0.12	NA
HRQoL	WHiTE Cohort	TATE	0.05	-0.01	0.11	0.97
HRQoL	NHFD	TATE	0.05	-0.01	0.11	0.96
LOS	WHiTE Trial	SATE	-0.70	-1.90	0.51	NA
LOS	WHiTE Cohort	TATE	-0.17	-1.46	1.13	0.97
LOS	NHFD	TATE	-1.14	-2.35	0.08	0.96

* Tipton's generalisability index ranges from 0 (no overlap between trial sample and target population) to 1 (the trial sample is equivalent to a random sample drawn from the target population based on selected covariates). Indices greater than 0.8 are usually considered to represent trial samples that are very similar to the target population

Sensitivity analysis results

Restricted set of covariates

A sensitivity analysis using the restricted set of covariates to estimate TATE for the WHiTE-cohort did not influence the effect on HRQoL. There was a small influence on the TATE for LOS. In the main analysis with the complete set of covariates, the TATE for LOS was -0.17 (95% CI; -1.46 to 1.13), whereas the TATE estimated with the restricted set of covariates was -1.02 (95% CI; -2.23 to 0.18). The difference in point estimate between the SATE and restricted covariate TATE was 0.35 days; and the difference between the SATE and complete covariate TATE was -0.50. Results are presented in **Appendix A5**.

Identifying and excluding trial participants from WHiTE-cohort

In the main analysis, we excluded 303 participants from the WHiTE-cohort who were recruited to the WHiTE-3 trial. A sensitivity analysis that retained the 303 trial participants in WHiTE-cohort showed that the deviation from the primary result was negligible; -0.02 points for HRQoL and 0.31 days for LOS. Because we only had limited access to anonymised NHFD data with few covariates for matching, we were unable to uniquely identify WHiTE-3 participants from NHFD. Results are presented in **Appendix A6**.

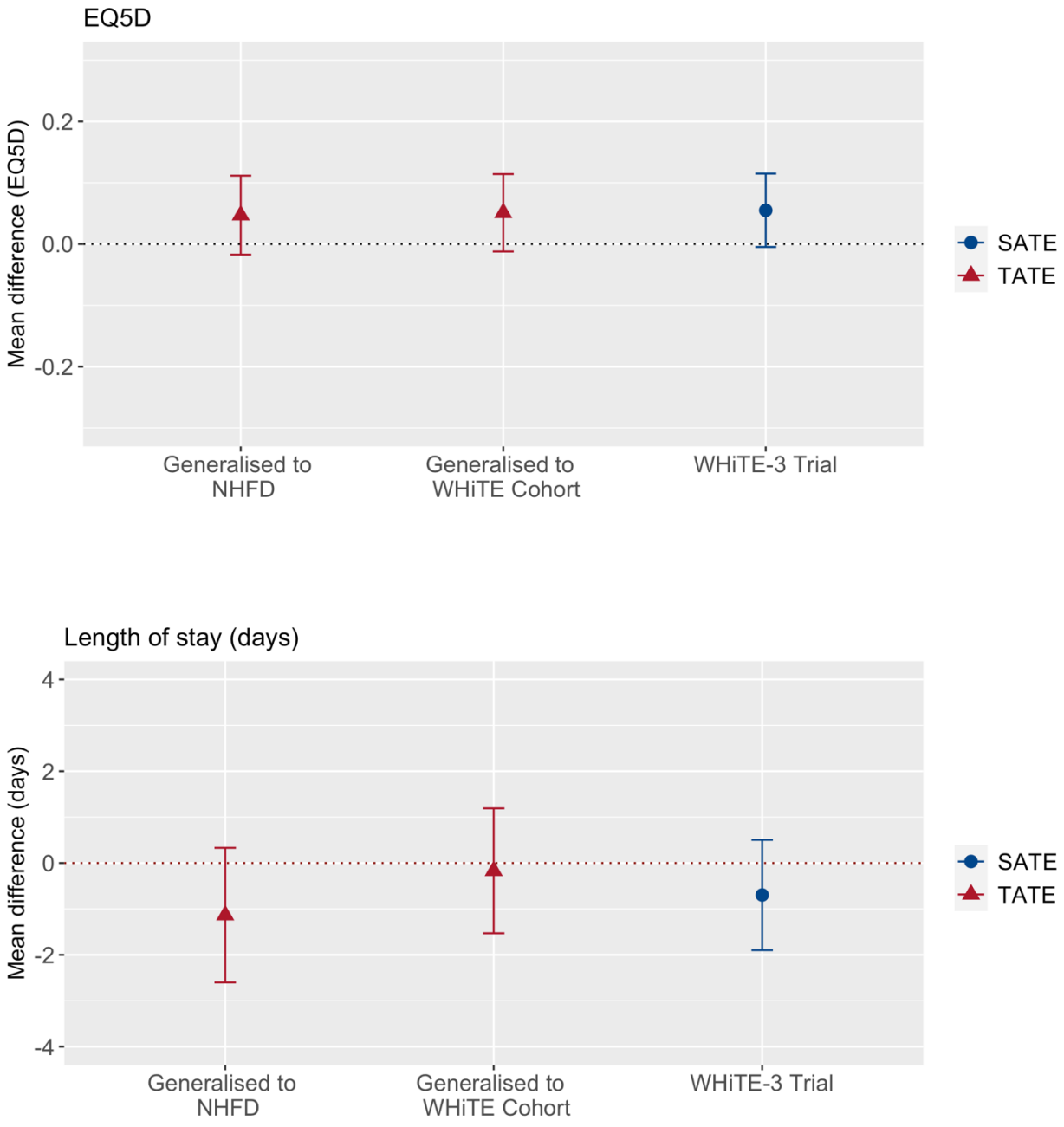

Figure 4. Target population Average Treatment Effect (TATE) and Sample Average Treatment Effect (SATE) estimates for Health-related Quality of Life (EQ5D) at 4 months and Length of Stay (LOS)

Discussion

This generalisability analysis suggests that the findings of the WHiTE-3 trial are generalisable to the wider UK population with intracapsular hip fractures undergoing hemiarthroplasty, based on the selected covariates that were modelled. Although there were small differences between the SATE and TATEs for LOS, they did not reach minimally importance difference levels (0.074 points) [17]. In summary, these findings suggest that the modern modular hemiarthroplasty will not provide additional benefit over the traditional monoblock, in terms of quality of life and length of hospital stay, if it were to be implemented nationwide in the UK.

The sensitivity analyses suggest that the main findings were largely robust to alternative modelling approaches, restricted covariate sets, and the inclusion/exclusion of trial participants in the target population dataset. We also did not detect clinically meaningful differences between the TATEs across the WHiTE-cohort and NHFD, suggesting that the WHiTE-cohort might be a good representative cohort of the NHFD, as supported by Metcalfe et al. [18].

The validity of the TATE estimates relies on three key structural assumptions. Firstly, given the joint distribution of observed covariates, every individual in the target population must have a non-zero probability of being a trial participant. We sought to preclude the violation of this assumption by limiting the target population datasets to individuals who were within the bounds of the covariates observed in the WHiTE-3 trial and checked that the estimated probabilities were greater than zero. Secondly, we assumed that the treatment and control groups in the WHiTE-3 trial were exchangeable. This assumption is met because participants in the WHITE-3 trial were randomly allocated. Finally, there should not be any unobserved covariates that drive selection into the trial and modify the treatment effect. Although we used expert clinical judgement to select a set of covariates that were thought to drive selection and modify the treatment effect, it is possible that there are other unobserved covariates that were omitted. The TATE estimates presented in this study are only as valid to the extent that these assumptions are sufficiently satisfied.

This study has some limitations. We cannot be certain that we included all relevant covariates that drive selection and modify treatment effect. This is difficult to verify because there is limited data on effect modifiers for the interventions tested in WHiTE-3. Although our sensitivity analysis with the restricted set of covariates provides some assurance of the stability of TATEs, future work could attempt to use richer datasets of individual or aggregate level data to balance a broader range of covariates [19,20].

Across all covariates included in this complete case analysis, there were between 0 to 14.7% individuals with at least one missing covariate from the NHFD; 0 to 15.7% individuals with at least one missing covariate from the WHiTE-Cohort; and 0 to 36.7% individuals with at least one missing covariate from the WHiTE-3 trial. If covariate data were not missing completely at random, selection bias might have influenced the weights that were used to re-weight the trial participants [26]. This in turn could have biased the TATEs that were estimated in this complete-case analysis. There were also missing data for the outcome data (HRQoL and LOS) from the trial. Missingness in the outcome data could also introduce bias because the outcome data are required to estimate both the unweighted SATE and weighted TATEs. Missing outcome data would also decrease the precision of SATE and TATE estimates. Thus, missing covariate data will always affect the TATE estimates but will only affect the SATE analysis if the model adjusts for these covariates; in contrast missing outcome data would affect both the SATE and TATE estimates. Methodological research and corresponding simulation studies considering various missing data patterns could help improve understanding of how missing covariate and outcome data influence the bias and precision of TATE estimates in generalisability analyses.

Future work could extend this generalisability analysis to assess the transportability of WHiTE-3 findings to a wider target population of patients who may not have been eligible for WHiTE-3. This seems particularly important in context of recent findings of a multi-national trial which demonstrated that hemiarthroplasty does not confer greater benefit than total hip arthroplasty for patients with displaced femoral neck fractures.[21] Future methodological work could also investigate the role of relative sample sizes between the trial and target population that would optimise the precision of TATE estimators. Although the external validity of trials have been largely assessed through descriptive comparisons to date, the increasing availability of representative datasets [22] coupled with methodological advances will allow for more opportunities to routinely assess the external validity of randomised trials.

Conclusion

Taken together, in a generalisability analysis of the WHiTE-3 trial, this study found that the inferences from the trial can be generalised to the group of individuals in the UK NHFD and its nesting WHiTE-cohort who met the inclusion criteria for WHiTE-3.

Acknowledgements

The authors would like to thank Dr May Ee Png, WHiTE-3 and WHiTE investigators for sharing the WHiTE data; the Healthcare Quality Improvement Partnership (HQIP), Royal College of Physicians (RCP), Falls and Fragility Fracture Audit Programme (FFFAP) and Crown Informatics for granting access to the UK National Hip Fracture Database; and Dr Benjamin Ackerman for assisting in the implementation of the 'generalize' package.

Ethical Approval

WHiTE-3 trial approved on 14 October 2014 by NRES Committee West Midlands under REC reference number 14/WM/1098NIHR Study ID: 17502

Sources of funding

The WHiTE-3 trial (but not this generalisability analysis) was funded by Stryker, USA with support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. Hopin Lee is supported by the Australian National Health and Medical Research Council (APP1126767).

References

[1] Dahabreh IJ, Robertson SE, Tchetgen EJT, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 2018;75:685–694. https://doi.org/10.1111/biom.13009.

[2] Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. Prev Sci 2015;16:475–85. https://doi.org/10.1007/s11121-014-0513-z.

[3] Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. Arch Intern Med 2008;168:133–5. https://doi.org/10.1001/archinternmed.2007.30.

[4] 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;16:1–14. https://doi.org/10.1186/s13063-015-1023-4.

[5] Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. Epidemiology 2017;28:553–61. https://doi.org/10.1097/EDE.0000000000000664.

[6] Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: Suggested framework and systematic review. Br Med J 2006;333:346–9. https://doi.org/10.1136/bmj.333.7563.346.

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

[8] Bradburn MJ, Lee EC, White DA, Hind D, Norman R, Cooke DD, et al. Treatment effects may remain the same even when trial participants differed from the target population. J Clin Epidemiol 2020;124:126–38. https://doi.org/10.1016/j.jclinepi.2020.05.001.

[9] Sims AL, Parsons N, Achten J, Griffin XL, Costa ML, Reed MR. A randomized controlled trial comparing the Thompson hemiarthroplasty with the Exeter polished tapered stem and Unitrax modular head in the treatment of displaced intracapsular fractures of the hip. Bone Jt J 2018;100B:352–60. https://doi.org/10.1302/0301-620X.100B3.BJJ-2017-0872.R2.

[10] van Hout B, Janssen MF, Feng Y-SS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Heal J Int Soc

Pharmacoeconomics Outcomes Res 2012;15:708–15. https://doi.org/10.1016/j.jval.2012.02.008.

[11] Parsons N, Griffin XL, Achten J, Chesser TJ, Lamb SE, Costa ML. Modelling and estimation of healthrelated quality of life after hip fracture: A re-analysis of data from a prospective cohort study. Bone Jt Res 2018;7:1–5. https://doi.org/10.1302/2046-3758.71.BJR-2017-0199.

[12] National Institute for Health and Care Excellence (NICE). Hip fracture: the management of hip fracture in adults [CG124]. Secondary Hip fracture: the management of hip fracture in adults [CG124]. 2015.

[13] Costa ML, Griffin XL, Achten J, Metcalfe D, Judge A, Pinedo-Villanueva R, et al. World Hip Trauma Evaluation (WHiTE): Framework for embedded comprehensive cohort studies. BMJ Open 2016;6:1–6. https://doi.org/10.1136/bmjopen-2016-011679.

[14] Ackerman B, Schmid I, Rudolph KE, Seamans MJ, Susukida R, Mojtabai R, et al. Implementing statistical methods for generalizing randomized trial findings to a target population. Addict Behav 2019;94:124–32. https://doi.org/10.1016/j.addbeh.2018.10.033.

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

[16] Ackerman B. generalize: Generalizing Average Treatment Effects from RCTs to Target Populations. R package 2019.

[17] Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32. https://doi.org/10.1007/s11136-004-7713-0.

[18] Metcalfe D, Costa ML, Parsons NR, Achten J, Masters J, Png ME, et al. Validation of a prospective cohort study of older adults with hip fractures. Bone Joint J 2019;101-B:708–14. https://doi.org/10.1302/0301-620X.101B6.BJJ-2018-1623.R1.

[19] Hong JL, Webster-Clark M, Jonsson Funk M, Stürmer T, Dempster SE, Cole SR, et al. Comparison of Methods to Generalize Randomized Clinical Trial Results Without Individual-Level Data for the Target Population. Am J Epidemiol 2019;188:426–37. https://doi.org/10.1093/aje/kwy233.

[20] Egami N, Hartman E. Covariate Selection for Generalizing Experimental Results. ArXiv 2019.

[21] HEALTH investigators. Total Hip Arthroplasty or Hemiarthroplasty for Hip Fracture. N Engl J Med 2019:1–10. https://doi.org/10.1056/nejmoa1906190.

[22] Knottnerus JA, Tugwell P. Trials embedded in cohorts , registries , and health care databases are gaining ground. J Clin Epidemiol 2020;120:5–6. https://doi.org/10.1016/j.jclinepi.2020.02.007.