ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Comparative-Effectiveness Research/HTA

Emulating Target Trials With Real-World Data to Inform Health Technology Assessment: Findings and Lessons From an Application to Emergency Surgery

Silvia Moler-Zapata, MSc, Andrew Hutchings, MSc, Stephen O'Neill, PhD, Richard J. Silverwood, PhD, Richard Grieve, PhD

ABSTRACT

Objectives: International health technology assessment (HTA) agencies recommend that real-world data (RWD) are used in some circumstances to add to the evidence base about the effectiveness and cost-effectiveness of health interventions. The target trial framework applies the design principles of randomized-controlled trials to RWD and can help alleviate inevitable concerns about bias and design flaws with nonrandomized studies. This article aimed to tackle the lack of guidance and exemplar applications on how this methodology can be applied to RWD to inform HTA decision making.

Methods: We use Hospital Episode Statistics data from England on emergency hospital admissions from 2010 to 2019 to evaluate the cost-effectiveness of emergency surgery for 2 acute gastrointestinal conditions. We draw on the case study to describe the main challenges in applying the target trial framework alongside RWD and provide recommendations for how these can be addressed in practice.

Results: The 4 main challenges when applying the target trial framework to RWD are (1) defining the study population, (2) defining the treatment strategies, (3) establishing time zero (baseline), and (4) adjusting for unmeasured confounding. The recommendations for how to address these challenges, mainly around the incorporation of expert judgment and use of appropriate methods for handling unmeasured confounding, are illustrated within the case study.

Conclusions: The recommendations outlined in this study could help future studies seeking to inform HTA decision processes. These recommendations can complement checklists for economic evaluations and design tools for estimating treatment effectiveness in nonrandomized studies.

Keywords: comparative effectiveness, emergency surgery, health technology assessment, real-world data, target trial framework.

VALUE HEALTH. 2023; ■(■):■–■

Introduction

Health technology assessment (HTA) agencies require robust effectiveness and cost-effectiveness evidence to support decision making in healthcare. Studies using real-world data (RWD) such as disease registry data or electronic health records can help build an evidence base, given their ability to include patients from large, heterogenous populations, and assess the effectiveness and cost-effectiveness for settings of direct decision-making relevance.[1,2] Nevertheless, the risk of bias from confounding and other design flaws in these studies constitute a major barrier to a more widespread adoption of real-world evidence in HTA decision making.[3,4]

Good research practices recommendations by HTA agencies such as the UK's National Institute for Health and Care Excellence (NICE) include the use of economic evaluation checklists and other quality assessment tools or the reporting health economic analysis plans, but these offer limited guidance on how to address fundamental issues pertaining to study design of studies using RWD.[5,6] Recently, NICE's latest manual of methods and processes for technology evaluation formally recognized the importance of RWD in informing decision making and emphasized the need for studies that consider how the principles of the "target trial" framework could be applied to HTA.[7,8]

The target trial framework can help mitigate concerns about the study design in observational (nonexperimental) studies by applying the design principles of randomized-controlled trials (RCTs).[8,9] This approach requires the definition of a target (hypothetical) pragmatic trial protocol, which is then emulated using observational data. The target trial framework can help to better identify and minimize the risk of bias in the study and make methodological assumptions and design choices transparent for evidence users. In settings with high-quality observational data analyses, emulating the target trial principles has been found to help replicate the results of published RCTs.[10-12] More recently,

Gomes et al[13] described the potential uses of the target trial framework in HTA, but did not use the methods in an application. In general, there is a lack of guidance on applying the target trial framework in the HTA context, which raises major challenges, in particular around the interrelated issues of defining from RWD the study population, time zero (baseline), and the specification of the intervention and comparators.

This article aimed to critically examine the application of the target trial framework principles to the HTA context when assessing the effectiveness of health interventions from RWD. We draw on a case study, the "Emergency Surgery or Not" (ESORT) study, to describe common challenges in applying the target trial framework to assess comparative effectiveness from routine data and offer a series of recommendations for future studies.[14] Previous publications from the ESORT study have reported the relative effectiveness[15] and cost-effectiveness[16] of emergency surgery (ES) for 5 acute conditions and detailed an advanced quantitative approach to address confounding.[15-17] The unique contribution of this article is to define and emulate the key elements of the target trial protocol, in evaluating the cost-effectiveness of ES for patients admitted to hospital with acute gastrointestinal conditions (section 2). We also report the accompanying results of the cost-effectiveness analysis (section 3). In section 4, we draw on these findings to offer general recommendations for future studies.

Methods

Overview

The ESORT study exemplifies the challenges that arise for HTA when there is little evidence from RCTs to inform routine clinical practice. In this particular setting, there were few published RCTs and economic evaluations that evaluated ES versus alternative non-ES (NES) strategies for common acute conditions.[14] The ESORT study helped address this gap in the literature by using information on 2010 to 2019 hospital admissions from the Hospital Episode Statistics (HES) database, linked to Office for National Statistics mortality data, to assess the cost-effectiveness of ES for 5 acute gastrointestinal conditions. In this article, we focus on 2 of the conditions, acute appendicitis and acute gallstone disease, which have the highest prevalence. The evaluation of costs and outcomes was from a hospital perspective, over a 1-year time horizon, and applied the key principles of the target trial framework as described in the following sections.[16]

Target Population for the Decision Problem

The application of the target trial framework to the HTA context requires that eligibility criteria for the study population are defined to represent the target population of interest and that it only includes those subgroups for whom there is equipoise about the choice of intervention versus comparator strategies. If subgroups of patients are included for whom there is a lack of equipoise between the treatment choices, it is likely that there will be differences in unobserved and observed baseline prognostic factors between the comparison groups, leading to biased estimates of treatment effectiveness because of confounding by indication.[18]

In the ESORT study, these considerations informed the choice of inclusion and exclusion criteria. Table 1 lists those criteria and describes how they were emulated using HES data. Some inclusion criteria, such as the patient's age and the requirement to be assessed by a surgeon, were intended to ensure equipoise and were emulated directly from the HES data. For example, by specifying an inclusion criterion that the patient must be at some point under the care of a consultant surgeon, it was anticipated that the study would exclude some patients whose prognosis was so "severe" or "mild" according to unobserved, as well as observed, characteristics that they would not be considered for ES. For other criteria, such as the reason for admission, the information from the routine data and the available evidence were insufficient to define which patient subgroups to include. In particular, although there was information on the patients' diagnosis according to International Classification of Diseases, Tenth Revision (ICD-10) codes, it was unclear which of the subcategories of ICD-10 corresponded to patient subgroups that would be eligible for ES in routine practice and for whom there was equipoise between the comparison strategies.

The ESORT study addressed the challenge of defining those elements of the target trial protocol that could not be specified from the routine data, by convening 2 panels of 12 clinicians with relevant expertise that followed a modified Delphi process (see ESORT Study Group[19] for details). The panelists were required to judge which inclusion and exclusion criteria were appropriate, given the requirement for equipoise between the comparison groups, and to define the interventions of interest (see next section). Views of the members of the clinical panels were crucial to ensure these definitions would reflect the heterogeneity across NHS hospitals in volume of patients or teaching status. The consensus of the panel required at least 9 from 12 responses in favor of the inclusion of the category, and 5 (appendicitis) and 3 ICD subcategories (gallstone disease) were designated for inclusion (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for full list). The panel's consensus also implied that for patients with acute appendicitis those with ICD-10 codes corresponding to appendiceal cancer and pregnancy should be excluded because of lack of equipoise, but for patients with gallstone disease none of the subcategories should be excluded.

Definition of Treatment Strategies

The main challenge in defining the treatment strategies from the RWD is to ensure they reflect the use of the comparators in routine clinical practice. In the ESORT study, the treatment strategies under assessment were complex, combining different surgical and nonsurgical procedures, and both were already in use within England's National Health Service. ES involves operative management that is immediate, urgent, or expedited.[20] To operationalize the ES definition, the expert panel were asked to consider which of the Office of Population Censuses and Surveys (OPCS) procedure codes listed within the HES data met the definition of ES and to define the appropriate time window. The panel's consensus was that 21 (appendicitis) and 45 procedure codes (gallstones) (see Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for full list), respectively, met the definition for ES and that for both conditions the time window for ES should be within 7 days of assessment (baseline/time zero, see below).

The definition of the comparator strategy should consider whether the information in the RWD is sufficient to ensure the comparator strategy is defined in enough detail to evaluate the causal contrast of interest.[21,22] In the ESORT study, any patient who did not receive one of the designated procedures within the 7-day period was assigned to the NES strategy. This definition includes management with antibiotic therapy and either no surgery within the 1-year time horizon or surgery that does not meet the ES criteria (ie, either an OPCS procedure code not considered to be ES within the designated ES window or an OPCS procedure code considered to be ES but outside the window). The proposed

Table 1. Protocol for target trial of emergency surgery versus non-emergency surgery for acute appendicitis and acute gallstone disease

Protocol component	Description of target trial of ES	How was the protocol element emulated in the ESORT?
Eligibility criteria	Inclusion criteria: - Patient was at least 18 years old at admission. - Emergency admission, via emergency department or primary care - The condition was the reason for admission into hospital. - The diagnosis was confirmed by consultant surgeon. Exclusion criteria: - According to clinical condition-specific exclusion criteria - Emergency admission for the condition in the previous year - Surgery for the condition within the previous 90 days - Patient transferred between hospitals before surgical assessment.	Inclusion criteria: - Emulated directly from HES data. - Emulated directly from HES data. - Expert panel defined diagnostic (ICD-10) codes with equipoise between comparator strategies.* - Emulation directly from HES data Exclusion criteria: - Expert panel designated exclusion criteria with (ICD-10) codes.† - Emulated directly from HES data - Emulated directly from HES data (using definitions of treatment strategies below) - Emulated directly from HES data Additional criteria according to data availability: - Patient was admitted to an ineligible hospital for ESORT.‡ - Admission lacked information on admission or discharge status or date.
Treatment strategies	- ES defined as urgent, expedited, or immediate surgery for the condition.[1,9] - NES: (1) medical management with no surgery for the condition and (2) surgery that did not meet the criteria for ES, either because not relevant procedure or after the 7-day time window, possibly preceded by medical management	- Expert panel defined the 2 criteria for ES: (1) the procedure constituted "surgery for the condition" according to selected OPCS codes,§ and (2) to be considered "emergency," the panel designated a time window of 7 days from the date of assessment (see below). - Emulation assumed patient assigned NES if they did not meet ES criteria
Time zero and follow-up period	- Time zero is analogous to the time of randomization and is when all the eligibility criteria are met, the assignment to ES or NES occurs, and follow-up starts. - Follow-up ends at the earliest of 1 year, death, or end of study period.	- Emulation assumed time zero was the start date of the first FCE for the first admission, in which the specialty code was general surgery, colorectal surgery, or upper gastrointestinal surgery. - Emulation censored patients at the date of death, if that was within 1 year from day zero. Complete follow-up data were available for all patients.
Treatment assignment	- Individuals are randomly assigned to a strategy at baseline.	- Treatments groups were assumed to be balanced after adjustment for differences in measured and unmeasured prognostic factors in the statistical analysis.
Outcomes	- Life-years at 1 year from randomization - QALYs at 1 year from randomization - Total costs at 1 year from randomization - Net monetary benefit at 1 year from randomization	- Emulated directly from HES data (linked to ONS death data) - Emulation required adjusting life-years using published age- and sex-adjusted HRQoL scores from similar populations. - Emulation required calculating resource use for categories considered to be main drivers of total costs (length of stay, including critical care; operative and diagnostic procedures and readmissions up to 1 year) and valuing resource use data using relevant estimates of unit costs taken from national unit cost databases. - Emulated combining cost and QALY data
Causal contrast of interest	- ITT effect (effect of assignment of patients to interventions at baseline) - PP effect (effect of complying with the trial protocol)	- ITT effect could not be emulated because information on the initial treatment assignment was not available from HES. - Emulation of the PP effect required taking differences between the treatment groups in estimated total costs, life-years, QALYs, and net monetary benefits at 1 year.

continued on next page

Table 1. Continued

Protocol component	Description of target trial of ES	How was the protocol element emulated in the ESORT?
Analysis plan	- ITT analysis and PP analysis with adjustment for baseline prognostic factors	- Emulation of the PP analysis required using a LIV approach to mitigate the risk of confounding because of unmeasured prognostic factors associated with ES receipt. The IV was the hospital's tendency to operate. Models were adjusted for a wide range of case-mix measures (age, sex, frailty level, comorbidity profile, ethnicity, index of multiple deprivation), fixed effects for each financial year, and proxies of quality of acute care (rates of emergency admission and mortality for each hospital and acute condition in 2009-2010 and in the year before the admission).
	- Subgroup analyses by baseline age, sex, frailty, and number of comorbidities	- Emulated directly from HES data

ES indicates emergency surgery; ESORT, Emergency Surgery or Not; FCE, finished consultant episode; HES, Hospital Episode Statistics; HRQoL, health-related quality of life; ICD, International Classification of Diseases; ITT, intention-to-treat; IV, instrumental variable; LIV, local instrumental variables; NES, non-emergency surgery; ONS, Office for National Statistics; OPCS, Office of Population Censuses and Surveys; PP, per-protocol; QALY, quality-adjusted life-year.
*See Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for full list ICD-10 codes for the 2 conditions.
†ICD-10 codes for acute appendicitis: pregnancy (O00-O9A; Z00-Z99) and appendiceal cancer (C00-D49). ICD-10 codes for acute gallstone disease: none.
‡Of all eligible acute general hospitals with at least 200 emergency general surgery admissions per year, those that ceased activity in 5 years before December 31, 2019, were excluded.
§See Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for full list of procedure codes included in definition of ES the 2 conditions.

definition of the comparator strategy in ESORT reflects the variation in the provision NES strategies in routine clinical practice, but also the limited availability of granular information within the HES data on specific NES treatments (eg, duration or dosage for antibiotic therapy). This lack of granular information meant that although the study could evaluate the cost-effectiveness of ES versus a general "basket" of NES strategies, it could not contrast with any one specific NES strategy.

Definition of Time Zero and Follow-Up

The careful definition of the emulated target trial's "point of randomization" or "time zero" can help minimize the risk of bias in the study.[8,23] In an RCT, time zero is defined as the time when eligibility is met, the alternative treatment strategies commence, and the follow-up begins. In RWD studies, it is often impossible to establish temporality from events recorded in the data, and if eligibility and treatment assignment are not aligned with the start of follow-up, then selection bias (if patients are excluded according to events that occurred after the onset of treatment) and immortal time bias (if there is a period of the follow-up over which outcomes of interest cannot occur) can emerge.[24,25] The criteria for time zero are as follows: (1) it does not precede the time when the eligibility criteria are met, (2) it must be identified for all patients regardless of the assigned treatment arm strategy, and (3) it should minimize the time window used to define treatment initiation to reduce the possibility of immortal time bias.

In ESORT, emulating time zero was not straightforward. The study considered using the date of hospital admission or the date either strategy was initiated, but both were deemed inadequate. The date of admission preceded the date diagnosis was confirmed by a surgeon, which was an inclusion criterion, for some patients (violation of 1). The date of treatment initiation was not available for the NES comparator (violation of 2). A third alternative, the date that the patient was first under the care of a consultant surgeon, was judged to be the most appropriate definition of time zero. After this initial surgical assessment, patients with these acute conditions would be assigned to either treatment strategy, without delay. Given the study's eligibility criteria, once the patient had the surgical assessment, all the eligibility criteria were met. This definition of time zero could still lead to bias, if during the 7-day time window for defining receipt of ES (rather than NES), the risk of the outcomes of interest differed between the comparison groups. For patients with acute appendicitis and acute gallstone disease, this would seem unlikely because patients are at very low risk of adverse outcomes, such as death, over that period.[26] When assessing ES for other conditions with higher rates of in-hospital mortality, methods such as "cloning, censoring, and weighting" could help to reduce the risk of immortal time bias.[7]

Outcomes

The nature of RWD might pose additional challenges for the emulation of the target trial given that data on outcome measures are often unavailable or available with insufficient detail. Nevertheless, through HES, the ESORT study had access to rich resource use data including the total duration of hospital stay (including readmissions) and survival time from HES linked to Office for National Statistics mortality data, which was used to derive life-years. Although information on health-related quality of life after ES and NES for the conditions was not available from HES, it could be obtained from available published studies reporting health-related quality of life weights for comparable populations. These weights were combined with information on key events (eg, emergency admissions) and survival time to derive 1-year quality-adjusted life-years (QALYs).[16] The main cost-effectiveness outcome was the incremental net monetary benefit (INB) at 1 year, using NICE's recommended threshold of £20 000 per QALY.[27]

Causal Contrast

RCTs are typically concerned with estimating the intention-to-treat effect, that is, the effect of being assigned to a particular treatment strategy, and the per-protocol (PP) effect, that is, the effect of receiving the treatment as prescribed in the protocol. In observational studies, where treatment received is observed but

Figure 1. Flowchart of eligibility for a target trial of emergency surgery versus non-emergency surgery for acute appendicitis emulated using Hospital Episodes Statistics data.

ESORT indicates Emergency Surgery or Not; TTO, tendency to operate.

treatment assignment is not, a PP analysis is generally favored. In the ESORT study, the broad protocol definition of both the ES and the NES strategies enabled the study to estimate a PP effect. Here, the assumption that patients in either group adhered to their treatment assignment was plausible and consistent with routine practice.

Analysis Plan

The risk of confounding bias poses a major threat to the validity of observational studies, and alternative methods make different assumptions, which need to be carefully considered.[28] The ESORT study used a local instrumental variable (LIV) approach to mitigate the concerns about unmeasured confounding. Briefly, LIV allows for treatment selection according to measured and unmeasured prognostic factors and can report consistent estimates of the overall effect for the population (ie, the average treatment effect [ATE]) and subpopulations of interest (ie, conditional ATEs) provided a series of assumptions hold.[16]

The instrument in the ESORT study was the hospital's tendency to operate (TTO), which was a proxy for their preference for ES, calculated from historic data. The assumptions underlying LIV were as follows: (1) TTO only influenced the outcome through its effect on treatment assignment (exclusion restriction), (2) TTO was associated with treatment assignment (relevance assumption), (3) TTO was independent of unmeasured confounders (exchangeability condition), and (4) TTO had the same direction of effect on the probability of treatment receipt, irrespective of the level of the instrumental variable (IV) (monotonicity assumption). These assumptions were judged plausible, given the findings that the IV was sufficiently strong (assumption 2) and balanced the observed covariates (3) and by implication and a priori reasoning

Table 2. Patient characteristics of the 2 cohorts of patients by emergency surgery and non-emergency surgery groups.

	Acute appendicitis (N = 268 144)		Acute gallstone disease (N = 240 977)	
	ES (n = 247 506)	NES (n = 20 638)	ES (n = 52 004)	NES (n = 188 973)
Sex: n (%)				
Male	134 270 (54)	10 409 (50)	15 140 (29)	63 046 (33)
Female	113 224 (46)	10 228 (50)	36 864 (71)	125 927 (67)
Age: mean	38 (16)	47 (20)	51 (18)	56 (19)
IMD quintile: n (%)				
1 – most deprived	49 495 (20)	4319 (21)	11 774 (23)	44 650 (24)
2	47 818 (20)	3898 (19)	9586 (19)	34 792 (19)
3	49 203 (20)	4128 (20)	10 641 (21)	37 561 (20)
4	50 337 (21)	4024 (20)	10 881 (21)	39 759 (21)
5 – least deprived	46 636 (19)	3907 (20)	8686 (17)	30 285 (16)
SCARF index: n (%)				
Fit	206 796 (84)	15 015 (73)	34 056 (66)	114 973 (61)
Mild frailty	34 544 (14)	4052 (20)	13 608 (26)	52 629 (28)
Moderate frailty	5041 (2)	1155 (6)	3385 (6)	16 175 (9)
Severe frailty	1125 (0)	416 (2)	955 (2)	5196 (3)
Ethnicity: n (%)				
Black/Black mixed	5771 (2)	627 (3)	827 (2)	3923 (2)
Asian/Asian mixed	11 592 (5)	1122 (5)	2204 (4)	9124 (5)
White	194 968 (79)	16 371 (79)	44 396 (85)	162 727 (86)
Chinese and other	9054 (4)	708 (3)	997 (2)	4092 (2)
Charlson index: n (%)				
0 – comorbidities	207 525 (84)	15 321 (74)	36 737 (71)	120 748 (64)
1	35 721 (14)	3989 (19)	12 287 (24)	49 863 (26)
2	3715 (2)	1035 (5)	2544 (5)	14 503 (8)
3+ – comorbidities	545 (0)	293 (1)	436 (1)	3859 (2)

Note. N indicates the cohort size and n indicates the group size.
ES indicates emergency surgery; IMD, index of multiple deprivation; NES, non-emergency surgery; SCARF, Secondary Care Administrative Records Frailty.

also unobserved covariates (1) and was unlikely to have a differential effect on the probability of ES receipt at different levels of TTO (4) (see Appendix Fig. 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010).

We conducted analyses that made alternative assumptions as sensitivity analyses. We undertook conventional risk-adjustment (using generalized linear model [GLM] regression) approaches, adjusting for the same baseline measures as in the LIV analysis, but making the alternative assumption that all the requisite confounders were adjusted for. For completeness we also included a naive comparison, which assumed there were no confounders. For each approach, we reported the INB for the overall target population of interest (ATE) and for LIV the INB according to prespecified subgroups of prime policy relevance (defined by age, sex, frailty level, and number of comorbidities.

Results

Cohort Description

We identified 268 144 patients with acute appendicitis and 240 977 with gallstone disease who met the target trial eligibility (see Fig. 1 and Appendix Fig. 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010). Of these patients, 92% (appendicitis) and 22% (gallstone disease) met the definition of ES, and the baseline characteristics of the comparison groups are presented in Table 2. In each cohort, those patients who had ES were on average younger and fitter and had fewer comorbidities (see Table 2).

Cost-Effectiveness Results

The LIV approach reports overall INB estimates for ES versus NES of −£86.2 (95% CI −1163 to 991) and £221 (−450 to 892) for appendicitis and gallstone disease, respectively (a 1-year time horizon was deemed appropriate to capture changes in patient outcomes and differential costs from major clinical events associated with the provision of ES and NES from ultimate discharge from the index hospital admission up to 1 year. In Moler-Zapata et al,[16] we show that the results reported in the study are not sensitive to extending the time horizon to 5 years) (Table 3). The regression adjustment reported similar estimates for the INB of −£223 (95% CI −342 to −104) for acute appendicitis and −£220 (95% CI −316 to 124) for gallstone disease (see also Appendix Table 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for results for costs, life-years, and QALYs). By contrast, the unadjusted INB estimates were £1431 (95% CI 1259-1603) and £1002 (95% CI 832-1171) for acute appendicitis

Table 3. Estimated group means and incremental costs (£GBP 2019/2020), QALYs, and net monetary benefit (£GBP 2019/2020) at 1 year of emergency surgery vs non-emergency surgery strategies using the local instrumental variable approach.

	Emergency surgery	Non-emergency surgery	Mean differences (95% CI)
		Acute appendicitis (N = 268 144)	
Costs	3366	3475	−109 (−1130 to 913)
Life-years	0.996	0.999	−0.003 (−0.006 to −0.001)
QALYs	0.942	0.952	−0.010 (−0.024 to 0.003)
Net benefit	15 475	15 561	−86.2 (−1163 to 991)
		Acute gallstone disease (N = 240 977)	
Costs	5477	5554	−76.8 (−702 to 548)
Life-years	0.970	0.978	−0.009 (−0.022 to 0.005)
QALYs	0.877	0.870	0.007 (−0.001 to 0.015)
Net benefit	12 059	11 838	221 (−450 to 892)

Note. Variables used for adjustment in models: age (years), sex, ethnicity, index of multiple deprivation (quintiles), number of comorbidities (Charlson index), frailty level (as measured by the SCARF index), method of admission, year fixed effects, and proxies for the quality of acute care within the hospital.
GDP indicates gross domestic product; QALY, quality-adjusted life-year; SCARF, Secondary Care Administrative Records Frailty.

and gallstone disease, respectively (see also Appendix Table 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010 for results for costs, life-years, and QALYs). When considering population subgroups, the LIV analysis suggests that ES was not cost-effective for patients with severe frailty (for both conditions) and patients with 2, 3, or more comorbidities (acute appendicitis) (see Fig. 2A and B, and Appendix Fig. 3A and 3B in Supplemental Materials found at https://doi.org/10.1016/j.jval.2023.04.010).

Discussion

International HTA agencies are expanding their use of comparative-effectiveness evidence from RWD studies.[1,2] NICE's new real-world evidence framework sets out recommendations to help RWD studies provide trustworthy evidence to inform decision making, which include using the target trial framework to inform study design choices.[29] This article illustrates how this framework can be applied to HTA in a study evaluating the cost-effectiveness of ES for 2 common acute gastrointestinal conditions, which exemplifies common challenges in applying the target trial alongside RWD to inform HTA. In Table 4, we draw on the findings from this study to outline some recommendations for future studies that aim to assess comparative effectiveness from RWD.

This article makes 3 important contributions to the literature. First, it contributes to methods for informing HTA decision making using effectiveness evidence from RWD and can help reduce barriers to the appropriate use of this evidence in decision making. NICE describes 3 main barriers to the adoption of real-world evidence in their evaluations: (1) the risk of bias, (2) the quality and relevance of the data, and (3) concerns about the trustworthiness of the evidence.[29] To tackle concerns about the trustworthiness of evidence, study design choices need to be made traceable and transparent for decision makers. Current good-practice recommendations, including the reporting of checklists for economic evaluations, provide, in general, insufficient basis for judging study design choices outside of RCTs.[3,30] The target trial framework allows users of the evidence generated from RWD to assess its rigorousness and trustworthiness according to how closely the study design mimics that of an RCT. Published RCTs estimates can be used as "benchmarks" in HTA to assess choices about aspects of the study design, including the plausibility of the assumptions underlying the different statistical approaches.[12] A further step would be to use the target trial framework in the design of systematic reviews and network meta-analyses of RCTs.[31] Nevertheless, in many settings, RCT evidence for benchmarking is unavailable or unsuitable as it fails to include the target populations, comparators, or endpoints of decision-making relevance. This study shows that RWD can still be used to support HTA decision making in those settings. Although applying the notions of target trial framework helps ensure that groups are comparable, thereby reducing the potential for confounding, this study highlights the importance of considering statistical methods that make alternative underlying assumptions about residual confounding. In ESORT, the unadjusted comparison of means that makes the implausible assumption of no confounding at all leads to a different conclusion to the GLM regression and LIV approaches that make more plausible assumptions about confounding and lead to similar results.

Second, the article tackles the lack of guidance on how to apply the principles of the target trial framework in RWD studies to ensure they meet the main requirements of HTA. We identify a series of challenges that are raised when using routine data for emulating target trials pertaining to (1) defining the study population, (2) defining the intervention and all relevant comparator strategies, (3) establishing time zero, and (4) using appropriate methods to adjust for confounding. Table 4 offers point-by-point recommendations for how to address these challenges. The first challenge relates to the inability to emulate the target trial's eligibility criteria, which can result in bias because of imbalances in the distribution of patient characteristics. To inform HTA, applying the target trial framework would require RWD studies to emulate trials with active comparators (the "standard of care").[32] Then, to minimize the risk of confounding from imbalances in prognostic factors, the eligibility criteria need to ensure that only patients for whom there is likely to be equipoise between treatment strategies are included. In the ESORT study, the criterion that the patient must be "under the care of a surgeon" helped exclude patients whose prognosis was so poor according to unobserved, as well as observed, characteristics that they would not be considered for ES (eg, patients in advanced stages of the disease). When defining the eligibility criteria, another important consideration is that the population needs to include all patient subgroups of relevance for HTA decision making. When

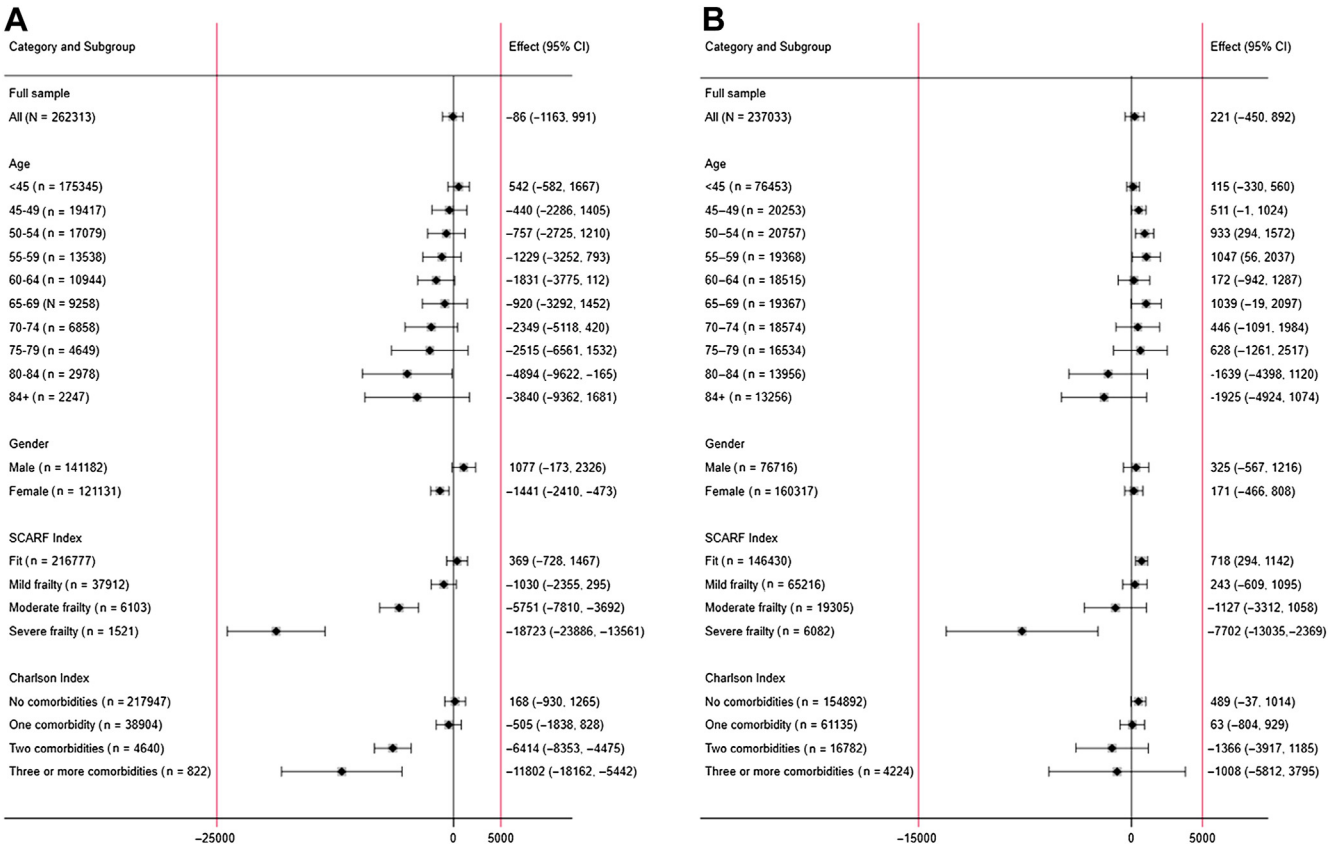
Table 4. Challenges and recommendations for studies applying the target trial framework alongside real-world data to inform health technology assessment.

Protocol component	Challenge for studies using RWD to inform HTA	Implications	Example from target trial of ES	Recommendation
Eligibility criteria	Data might be insufficient to emulate the trial's eligibility criteria.	Estimates of comparative effectiveness could be subject to selection bias/ confounding if the distributions of patient characteristics are not balanced.	Unclear which ICD-10 diagnostic subcategories describe patients with diagnoses of acute appendicitis and acute gallstone disease.	Use expert opinion to adapt the trial's eligibility criteria to the data available.
	Population selected for study might include patients for whom there is no equipoise between treatment strategies.	Estimates of comparative effectiveness could be subject to confounding bias.	No equipoise for some patients with designated diagnostic codes for the condition (eg, pregnant patients with designated codes of appendicitis).	Use clinical guidelines and/or expert opinion to define and exclude patients subgroups for whom there is no equipoise.
	Population selected for study might fail to include subgroups of interest for decision making.	Findings could fail to inform HTA decision making if they are not generalizable to the target population or omit relevant subgroup analyses.	Unclear which patient subgroups are eligible and in equipoise for ES and NES strategies in routine practice.	Use clinical guidelines and/or expert opinion to define subgroups of interest.
Treatment strategies	The definition of the intervention (eg, its timing) might differ from the intervention of interest.	Findings could fail to inform decision making if they do not reflect routine clinical practice.	Unclear which OPCS-4 procedure codes and timings describe ES.	Use clinical guidelines and/or expert opinion to define the intervention and comparators.
	The comparator strategy might not be defined with sufficient level of detail.	Findings could fail to inform decision making because of the interventions involved in the causal contrast not being well defined.	The study could not inform the comparative effectiveness of ES vs specific NES treatments, but could do so against not receiving ES.	Carefully assess whether the causal contrast can be estimated given the data available.
Time zero	Start of follow-up might predate the assessment of the eligibility criteria.	Findings could be subject to selection bias.	Using the date of admission as day zero could result in bias because of postbaseline events being used to exclude patients.	Consider the likely bias arising from alternative candidates for day zero.
	Time of treatment assignment might not be aligned with that of eligibility assessment and start of follow-up.	Findings could be subject to immortal time bias.	Using the date of admission as day zero could result in bias if, during time until treatment initiation, the risk of event of interest differed between the groups.	Include as a criteria for day zero that it should minimize time to treatment initiation.
Statistical analysis	Residual confounding might exist after emulating the main components of the target trial, from both measured and unmeasured prognostic factors.	Estimates of comparative effectiveness could be biased by residual confounding.	Naive comparisons are unlikely to provide robust estimates, whereas adjustment in LIV and GLM regression resulted in similar findings.	Consider appropriate methods for tackling confounding and, where possible, assess the underlying assumptions in the method used.
	Not all statistical methods might be appropriate for studying the causal contrast(s) of interest.	Findings might not be generalizable to the target population.	Estimates of traditional IV methods usually pertain to narrow populations, but LIV can retrieve an overall effect.	Carefully assess the plausibility of the assumptions required for the estimation of the causal contrast.

ES indicates emergency surgery; GLM, generalized linear model; HTA, health technology assessment; IV, instrumental variable; ICD, International Classification of Diseases; LIV, local instrumental variable; NES, non-emergency surgery; OPCS, Office of Population Censuses and Surveys; RWD, real-world data.

published clinical guidance is insufficient to identify these populations, expert judgment should be used to adapt the target trial's eligibility criteria to the data available and to the requirements of HTA (see Table 4). Sensitivity analyses around the different eligibility criteria could help assess the implications of these decisions and should be adopted (in Hutchings et al,[15] we describe how alternative definitions of the ES window could be considered in sensitivity analyses to accommodate the effects of seasonality or capacity constraints in decision-making processes about ES).[33] In the ESORT study, the clinical panel exercise provided a basis for this. The study could define alternative more/less strict definitions of the eligibility criteria by varying the threshold for required

Figure 2. Forest plots of estimated incremental net monetary benefit (INB) of emergency surgery (ES) versus non-emergency surgery (NES) for acute appendicitis (panel A) and acute gallstonedisease (panel B) across population subgroups.



Values to the left (right) of the 0 line denote that NES (ES) is cost-effective for the subgroup. ES indicates emergency surgery; INB, incremental net monetary benefit; NES, nonemergency surgery; SCARF, Secondary Care Administrative Records Frailty.

number of responses favoring inclusion of an ICD-10 code subcategory.

The main challenge in defining the intervention and comparator strategies is to specify the treatment(s), dosage(s), and timing(s) that characterize their provision in routine clinical practice (second challenge, see Table 4). The definition could be informed by clinical guidelines for management of the condition, but as in the ESORT study, these are often unavailable. Unless the treatments of interest are specified within the RWD, the study will be of limited use for informing HTA decision making.[21] Further to this, the study should carefully consider whether the comparators are defined in sufficient detail to evaluate the causal contrast of interest.[34,35] We recommend drawing on expert opinion to define the interventions and comparators of interest from those recorded within the routine data.

The ESORT study highlights the challenges in defining time zero (baseline) from the RWD (third challenge), which cannot precede eligibility, and must minimize any delay before treatment initiation (see Table 4). In studies such as ESORT, where treatment initiation for one or all treatment strategies is not observed in the data, the choice of time zero should be carefully evaluated. The ESORT study defined time zero as the date when the patient was first under the care of a surgeon. This definition is expected to carry low risk of bias given that (1) it is does not precede the time of eligibility assessment and, (2) although it may not coincide with the time of treatment initiation, the probability of events until treatment initiation is small for these conditions. To help ensure the definition of time zero meets the requirement mentioned earlier, tools that help establish temporality from RWD, such as design diagrams,[36] and approaches such as cloning, censoring, and weighting should be adopted in settings where immortal time bias is suspected.[7]

In relation to the fourth challenge, our article builds on precedent work on the use of IV methods for confounding adjustment and in particular the combination of the target trial framework with IV methods to reduce the risk of bias from unmeasured confounding, which is a major concern in RWD.[37] ESORT uses a LIV approach which, unlike traditional IV methods such as 2-stage least squares, can provide estimates of the ATE and conditional ATEs that apply directly to the target population.

We argue that the application of principles of the target trial framework should not replace the use of available design tools such as the Strengthening the Reporting of Observational Studies in Epidemiology checklist pertaining to the choice of statistical approaches for estimating treatment effects in observational studies.[38] As this article illustrates, a fundamental element of this type of research is that the plausibility of the underlying assumptions is assessed, and alternative approaches that make contrasting assumptions are considered (see Table 4).

The third contribution of this article is to illustrate how the target trial framework can be used to inform recommendations about technologies in settings where appropriate RCT evidence is not available. The assessment of the relative cost-effectiveness of ES for acute appendicitis and acute gallstone disease in ESORT

contributes to the scarce evidence on the effects of providing ES versus alternative strategies for patients with acute gastrointestinal conditions. This article extends the limited available evidence on the relative cost-effectiveness of ES versus NES for treating acute appendicitis and acute gallstone disease. Previous cost-effectiveness analyses produced conflicting results,[39-42] and very few included formal subgroup analyses. The ESORT study finds that, overall, it is uncertain whether ES is cost-effective for patients with these conditions and that there is substantial heterogeneity in the cost-effectiveness of ES across frailty and comorbidity subgroups. Our findings suggest that redirecting surgical resources could result in improved population outcomes, and emphasize the importance of perioperative assessments in routine practice according to patient characteristics such as frailty.

Although the ESORT study exemplifies key issues in emulating target trials for HTA using individual patient data from routine sources, it cannot consider all the issues that may arise when using RWD in HTA. In ESORT, the completeness and accuracy of HES data[43] meant there were minimal concerns about the risk of attrition bias or reporting bias, which can result from imbalances in the duration of follow-up and reporting of outcome data, but these could be present when using other sources of RWD. In addition, the target trial framework was applied to the endpoints available within the HES data, namely survival time and health service utilization. In other settings, the requirement to access RWD on other outcome measures (eg, patient-reported outcomes) could add another layer of complexity to the study. The ESORT study directly addresses the use of RWD for HTA purposes when individual patient data are available from a single study. More generally, greater consideration is needed on how the principles may expand to settings where individual patient data are not available for any or all the comparators of interest (eg, creating external controls in single arms trials). Finally, there is scope for incorporating public and patient involvement and engagement processes within the framework described in this article. In Grieve et al,[17] we discuss how insights from discussions with public and patient involvement and engagement representatives were used in ESORT to guide decisions about the study design such as the choice of outcomes. Further exemplar applications and guidance on how this could be done more formally in studies using RWD are required.

In conclusion, this article addresses common challenges that arise when applying the target trial framework to assess comparative effectiveness and cost-effectiveness for the purposes of HTA, when using RWD. The article provides recommendations for improving the study design pertaining to the definition of the study population, comparators, and analytical approaches to help address concerns about the use of RWD in decision making.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2023.04.010.

Article and Author Information

Accepted for Publication: April 26, 2023

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2023.04.010

Author Affiliations: Department of Health Services Research and Policy, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, England, UK (Moler-Zapata, Hutchings, O'Neill, Grieve); Centre for Longitudinal Studies, UCL Social Research Institute, University College London, London, England, UK (Silverwood).

Correspondence: Silvia Moler-Zapata, MSc, Department of Health Services Research and Policy, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, England, United Kingdom. Email: silvia.moler@lshtm.ac.uk

Author Contributions: *Concept and design*: Moler-Zapata, Hutchings, O'Neill, Silverwood, Grieve
Acquisition of data: Hutchings, Silverwood, Grieve
Analysis and interpretation of data: Moler-Zapata, Hutchings, O'Neill, Silverwood, Grieve
Drafting of the article: Moler-Zapata, Hutchings, O'Neill, Grieve
Critical revision of the article for important intellectual content: Moler-Zapata, Hutchings, O'Neill, Silverwood
Statistical analysis: Moler-Zapata, O'Neill
Obtaining funding: Moler-Zapata, Hutchings, O'Neill, Silverwood, Grieve
Supervision: Moler-Zapata, Hutchings, O'Neill, Grieve

Conflict of Interest Disclosures: Dr Hutchings reported receiving grants from the National Institute for Health and Care Research during the conduct of the study. No other disclosures were reported.

Funding/Support: The ESORT study was funded by the National Institute for Health Research Health Services and Delivery Research (grant number 18/02/25). This report is independent research supported by the National Institute for Health Research ARC North Thames. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; and decision to submit the article for publication.

Acknowledgment: The authors gratefully acknowledge the valuable input and support from ESORT team members at London School of Hygiene and Tropical Medicine, especially to David Lugo-Palacios (assistant professor) and Beth Silver (project manager).

REFERENCES

1. Garrison Jr LP, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR real-world data Task Force report. *Value Health*. 2007;10(5):326–335.
2. Makady A, Ham RT, de Boer A, et al. Policies for use of real-world data in Health Technology Assessment (HTA): a comparative study of six HTA agencies. *Value Health*. 2017;20(4):520–532.
3. Faria R, Hernández Alava M, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The Use of Observational Data to Inform Estimates of Treatment Effectiveness in Technology Appraisal: Methods for Comparative Individual Patient Data. National Institute for Health and Care Excellence. https://www.sheffield.ac.uk/sites/default/files/2022-02/TSD17-DSU-Observational-data-FINAL.pdf. Accessed September 25, 2022.
4. Bell H, Wailoo AJ, Hernandez M, et al. The use of real world data for the estimation of treatment effects in NICE decision making: report by the decision support unit. https://www.sheffield.ac.uk/sites/default/files/2022-02/RWD-DSU-REPORT-Updated-DECEMBER-2016.pdf. Accessed September 25, 2022.
5. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)—explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. *Value Health*. 2013;16(2):231–250.
6. Thorn J, Ridyard C, Hughes D, et al. Health economics analysis plans: where are we now? *Value Health*. 2016;19(7):A397.
7. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–764.
8. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70–75.
9. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat Med*. 2019;25(10):1601–1606.
10. Caniglia EC, Zash R, Jacobson DL, et al. Emulating a target trial of antiretroviral therapy regimens started before conception and risk of adverse birth outcomes. *AIDS*. 2018;32(1):113–120.
11. Petito LC, García-Albéniz X, Logan RW, et al. Estimates of overall survival in patients with cancer receiving different treatment regimens: emulating

hypothetical target trials in the surveillance, epidemiology, and end results (SEER)-Medicare linked database [published correction appears in JAMA Netw Open. 2020 Apr 1;3(4):e204966] *JAMA Netw Open*. 2020;3(3):e200452.
12. Franklin JM, Patorno E, Desai RJ, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation*. 2021;143(10):1002–1013.
13. Gomes M, Latimer N, Soares M, et al. Target trial emulation for transparent and robust estimation of treatment effects for health technology assessment using real-world data: opportunities and challenges. *Pharmacoeconomics*. 2022;40(6):577–586.
14. Emergency Surgery or NoT (ESORT) study. Study protocol. ESORT Study Group. https://www.lshtm.ac.uk/media/38711. Accessed August 4, 2022.
15. Hutchings A, O'Neill S, Lugo-Palacios D, et al. Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. *Anaesthesia*. 2022;77(8):865–881.
16. Moler-Zapata S, Grieve R, Lugo-Palacios D, et al. Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. *Med Decis Making*. 2022;42(8):1010–1026.
17. Grieve R, Hutchings A, Moler Zapata S, et al. *Clinical effectiveness and cost-effectiveness of emergency surgery for adult emergency hospital admissions with common acute gastrointestinal conditions: the ESORT study [Internet]*. Southampton (UK): National Institute for Health and Care Research; 2023 (Health and Social Care Delivery Research, No. 11.1.).
18. Yoshida K, Solomon DH, Haneuse S, et al. A tool for empirical equipoise assessment in multigroup comparative effectiveness research. *Pharmacoepidemiol Drug Saf*. 2019;28(7):934–941.
19. Protocol for Emergency Surgery or NoT (ESORT) study. Clinical panel. ESORT Study Group. https://www.lshtm.ac.uk/media/39151. Accessed August 4, 2022.
20. National Confidential Enquiry into patient outcomes and death. The NCEPOD Classification of Intervention. www.ncepod.org/uk/classification.html. Accessed July 3, 2022.
21. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health*. 2004;58(4):265–271.
22. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008;32(suppl 3):S8–S14.
23. Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining bias in studies of statin treatment and survival in patients with cancer [published correction appears in *JAMA Oncol.* 2018;4(1):133] *JAMA Oncol*. 2018;4(1):63–70.
24. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010;340:b5087.
25. Maringe C, Benitez Majano S, Exarchakou A, et al. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *Int J Epidemiol*. 2020;49(5):1719–1729.
26. Di Saverio S, Podda M, De Simone B, et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. *World J Emerg Surg*. 2020;15(1):27.
27. Guide to the methods of technology appraisal. National Institute for Health and Care. https://www.nice.org.uk/process/pmg9/chapter/foreword. Accessed June 2, 2023.
28. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*. 2013;347:f6409.
29. NICE real-world evidence framework. National Institute for Health and Care Excellence. www.nice.org.uk/corporate/ecd9. Accessed August 4, 2022.
30. Orsini LS, Monz B, Mullins CD, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1504–1513.
31. Zhao SS, Lyu H, Solomon DH, Yoshida K. Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework. *Ann Rheum Dis*. 2020;79(7):883–890.
32. Developing NICE guidelines: the manual. Process and methods guide. National Institute for Health and Care Excellence. http://www.nice.org.uk/article/pmg20. Accessed July 6, 2022.
33. Lodi S, Phillips A, Lundgren J, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. *Am J Epidemiol*. 2019;188(8):1569–1577.
34. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
35. Hernán MA, Robins JM. *Causal Inference: What If*. New York, NY: Chapman & Hall/CRC; 2020.
36. Patorno E, Schneeweiss S, Wang SV. Transparency in real-world evidence (RWE) studies to build confidence for decision-making: reporting RWE research in diabetes. *Diabetes Obes Metab*. 2020;22(suppl 3):45–59.
37. Swanson SA. Instrumental variable analyses in pharmacoepidemiology: what target trials do we emulate? *Curr Epidemiol Rep*. 2017;4(4):281–287.
38. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Ann Intern Med*. 2007;147(8):W163–W194.
39. Javanmard-Emamghissi H, Boyd-Carson H, Hollyman M, et al. The management of adult appendicitis during the COVID-19 pandemic: an interim analysis of a UK cohort study. *Tech Coloproctol*. 2021;25(4):401–411.
40. O'Leary DP, Walsh SM, Bolger J, et al. A randomized clinical trial evaluating the efficacy and quality of life of antibiotic-only treatment of acute uncomplicated appendicitis: results of the COMMA trial. *Ann Surg*. 2021;274(2):240–247.
41. Morris S, Gurusamy KS, Patel N, Davidson BR. Cost-effectiveness of early laparoscopic cholecystectomy for mild acute gallstone pancreatitis. *Br J Surg*. 2014;101(7):828–835.
42. Macafee DA, Humes DJ, Bouliotis G, Beckingham IJ, Whynes DK, Lobo DN. Prospective randomized trial using cost–utility analysis of early versus delayed laparoscopic cholecystectomy for acute gallbladder disease. *Br J Surg*. 2009;96(9):1031–1040.
43. Emergency Surgery or NoT (ESORT) study. Analysis plan. ESORT Study Group. https://www.lshtm.ac.uk/media/51011. Accessed June 15, 2021.