

A protocol for emulating a published randomised controlled trial using registry data: effects of azithromycin in young adults with cystic fibrosis

Emily Granger, Jonathan Todd, Susan Charman, Elizabeth Cromwell, Gwyneth Davies, Freddy Frost, Alex Gifford, Bin Huang, Nicole Mayer-Hamblett, Lutz Naehrlich, Josh Ostrenga, Sanja Stanojevic, Rhonda Szczesniak, Ruth Keogh, on behalf of the CF-Trial Emulation Network (CF-TEN)

Corresponding author:

Dr Emily Granger, Department of Medical Statistics, Faculty of Epidemiology and Population Health, LSHTM, Keppel Street, London, WC1E 7HT

emily.granger@lshtm.ac.uk

Abstract

Introduction Target trial emulation is a framework for evaluating the effects of treatments using observational data. The trial emulation approach involves specifying key elements of a protocol for a target trial (a randomised controlled trial designed to address the question of interest) and then describing how best to emulate the trial using observational data. Recent years have seen an uptake of target trial emulation in several disease areas, although there are limited examples in cystic fibrosis (CF). This protocol describes a study which aims to assess the applicability of target trial emulation in CF. We aim to emulate an existing trial in CF and assess to what extent the results from the trial can be replicated using registry data.

Methods and analysis The target trial is a published randomised controlled trial which found evidence for beneficial effects of azithromycin use on lung function in young adults with CF. Two emulated trials are planned: one using data from the UK CF Registry and one using data from the US CF Registry. The inclusion and exclusion criteria, treatment and outcome definitions, follow-up period, and estimand of interest are all designed to match the published trial as closely as possible. The analysis step of the trial emulations will use causal inference methods to control for confounding. Results obtained in the emulated trials using registry data will be compared to those from the target trial.

Ethics and dissemination Ethical approval has been granted by the London School of Hygiene and Tropical Medicine Ethics Committee (Ref: 29609). This study has also been approved by the UK CF Registry Research Committee and the North Star Review Board. The results of this study will be published in a peer-reviewed journal and presented at relevant scientific conferences.

Keywords cystic fibrosis; registries; observational study; statistics & research methods

Word count: 4651

Article summary

Strengths and limitations of this study

- We use data from the UK and US CF Registries. These are the two largest national CF registries, and the UK CF Registry is cited as an exemplar patient registry in the NICE real-world evidence framework.
- We use the target trial emulation approach. This approach helps to clearly articulate the study design and to avoid certain biases. We provide an example of target trial emulation in a disease area where there are limited applications.
- The CF registries do not contain data on treatment doses or adherence, which limits our ability to match the treatment strategies in the target trial precisely.
- The CF registries do not contain data for all secondary outcomes used in the target trial.

1. Rationale and background

Randomised controlled trials (RCTs) are the gold standard approach for evaluating the effects of treatments. However, RCTs are costly, and sufficiently large trials are not always feasible, particularly in patient populations with a rare disease, such as cystic fibrosis (CF). When an RCT is not feasible, an alternative is to use observational data to 'emulate' a trial [1]. The trial emulation approach involves specifying key elements of a protocol for a target trial (an RCT we would like to conduct, if it were feasible) and then describing how best to emulate the target trial using the observational data at hand. This approach combines the study design principles of RCTs, with an analysis appropriate for observational data.

Recent years have seen an uptake of target trial emulation in several disease areas [2-7]. There is also rising interest in emulating existing RCTs in an attempt to replicate the results from the existing RCTs using observational data. The RCT DUPLICATE initiative recently published the results of 32 trial emulations using insurance claims data to replicate existing trials [8]. They found that emulated trials based on insurance claims data can obtain similar results to the original RCTs. Matthews et al [9] used Swedish registry data to emulate the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) randomised trial [10]. Admon et al [11] used target trial emulation to predict results of the PreVent (Preventing Hypoxemia with Manual Ventilation during Endotracheal Intubation) Trial [12] before they were published.

Despite the widespread use of target trial emulation across other areas of medicine, there are limited applications within the CF literature [13,14]; thus, its applicability to CF remains unclear. We aim to assess the applicability of target trial emulation in CF using data from the UK and US CF patient registries by emulating a published RCT within CF, and assessing the extent to which the RCT findings could be replicated. Here we set out the trial emulation protocol, including the statistical analysis plan. We follow the reporting guidelines recommended in the HARmonised Protocol Template to Enhance Reproducibility (HARPER) [15].

2. Research question and objectives

The primary objective is to emulate a published RCT of the effects of azithromycin in young adults with CF by Clement et al. [16] using observational data from two patient registries, and to assess the extent to which the RCT results can be replicated. The RCT of Clement et al [16] provides the target trial that this study aims to emulate. This trial was selected as we anticipate that it will be possible to replicate using the UK and US CF Registry data, based on our knowledge of the treatment and outcome data recorded. Table 1 summarises the research question addressed in the target trial.

3. Data sources

UK CF Registry

The UK CF Registry was established in 1995 and is a national database sponsored and managed by the Cystic Fibrosis Trust, with UK National Health Service research ethics approval. It records longitudinal data on approximately 99% of people with CF in the United Kingdom [17,18].

Data are collected on time-invariant variables, such as sex at birth, cystic fibrosis transmembrane conductance regulator (*CFTR*) genotype, date of birth, diagnosis data, and longitudinal variables that change over time. Longitudinal data are collected at approximately annual review clinic visits on over 250 variables covering several domains. These include clinical measurements taken on the day, and other variables covering the previous 12 months period such as: hospital admissions, treatments prescribed, culture and microbiology, health complications, nutrition, physiotherapy, smoking, and outcomes (death and transplants). From 2016, the UK CF Registry started collecting treatment prescription start and stop dates.

US CF Registry

The US CF Registry began collecting data on people with CF in the United States in 1986 and is managed by the Cystic Fibrosis Foundation. It contains longitudinal information on approximately 80% of people with CF in the United States [19-21].

Data are collected on demographic characteristics and on longitudinal variables that change over time. Data collection takes place at 'encounter visits' at CF care centres, with data also being abstracted annually. This study will use the encounter visit data which includes relevant information regarding hospitalizations, clinical measurements, medication usage, culture and microbiology, health complications. The encounter visits include routine clinical visits and visits in a hospital or the individual's home. The non-routine hospital and home IV visits may be due to an individual experiencing worsening of their respiratory symptoms and therefore their lung function may be unstable at this time. These are referred to as "unstable" visits, whereas routine clinical visits are referred to as "stable" visits. The analyses in this study will use data from the stable visits only.

4. Research Methods

4.1 Study design

We will conduct two studies nested within existing longitudinal data sets (one using UK CF Registry data, one using US CF Registry data), designed using the target trial emulation framework. Table 2 summarises the key components of the protocols for the target trial and the emulated trials.

4.2 Setting

4.2.1 Time-periods

The target trial was conducted from 2001 to 2003, with results published in 2006. Within the data collection period for the target trial, results from other azithromycin RCTs were published [22-24], after which there was uptake of this treatment in routine clinical practice. We plan to emulate the trial using data from three time periods of three years' duration: 2003-2005, 2007-2009, 2016-2018). Period 1 is close to the timing of the target trial, while allowing time for the treatment to have come into use. Periods 2 and 3 were chosen based on features of the data and ending the timeframe in 2018 means that we only use data from the time before CFTR modulators became widespread in clinical practice.

Table 3 provides further details and justification about the three time periods. For each 3-year time period, the first two years are used as the "recruitment-period", defined as the period during which individuals are considered for inclusion in the emulated trial. The target trial recruited participants over two years. Individuals are included in the emulated trial data if they meet the inclusion and exclusion criteria in at least one of their visits during the recruitment-period.

4.2.2 Definition of time 0 and the index visit

Time 0 is defined as the time at which individuals meet the eligibility criteria and "enter" the emulated trial, analogous to the time of randomization in the target trial. Time 1 is 12 months post-baseline, and the outcome of interest is FEV₁% at time 1.

For the emulated trials conducted within period p ($p = 1,2,3$) we define an 'index visit' for each individual who meets the eligibility criteria during the recruitment period, such that the date of the index visit is time 0. Follow-up visits take place approximately 12 months after the index visit and an eligibility criterion is that individuals are required to have a follow-up visit. Index visits and follow-up visits are defined differently in the UK and US emulated trials, due to differences in data collection between the two registries. Moreover, the visit we use for treatment, outcome and covariate data at time 0 and time 1 differs between the UK as US emulated trials; Figure 1 summarises these differences.

UK CF Registry Data

1. Using annual review data only

Within period p , the index visit for a given individual is the first annual review visit at which they meet the eligibility criteria in the 2-year recruitment period. The follow-up visit is the date of the next annual review which is closest to 12 months after time 0, but falls within 9-15 months after time 0. The outcome is FEV₁% measured on the day of the follow-up visit. Individuals are included in the treated group if they are recorded as being prescribed azithromycin at the follow-up visit, because the information recorded on treatment use at the annual review refers to treatment use over the past year. Individuals not recorded as being prescribed azithromycin at the follow-up visit are included in the control group.

2. *Using annual review data and prescription dates data*

In the UK Registry, prescription dates data are available from 2016 onwards. Therefore, in period 3 (2016-2018), we can conduct a second analysis making use of the dates data for a more precise time 0. In the second analysis, the index and follow-up visits are defined as above for individuals in the control group. For treated individuals, the index visit is defined similarly, however time 0 is defined as the first date post-index visit at which azithromycin is prescribed. The follow-up visit is defined as the next annual review visit which is closest to 12 months after time 0, but falls within 9-15 months after time 0.

US CF Registry Data

Within period p , the index visit is defined as the first stable encounter visit at which an individual meets the eligibility criteria in the 2-year recruitment period. The follow-up visit is defined as the date of the stable encounter visit that is closest to 12 months after the index visit, but which falls within 9-15 months after the index visit. Individuals are included in the treatment group if they are recorded as being prescribed treatment at the index visit, as this is often assumed to be the treatment start date. Once an individual starts treatment, we assume they remain on treatment until the follow-up visit. Individuals are included in the control group if they are recorded as not being prescribed azithromycin at the index visit. Controls who are recorded as starting treatment at an encounter visit between index visit and the follow-up visit are censored at the date of that encounter visit. For the US emulated trials, we also define the prior visit to be the most recent stable encounter visit prior to the index visit.

4.2.3 Inclusion and exclusion criteria

Table 2 summarises the inclusion and exclusion criteria for the target and emulated trials. The data on liver function test results and serum creatinine levels (exclusion criteria 3 and 5) may have large amounts of missingness and therefore unusable. An alternative is to use indicator variables for any recorded non cystic fibrosis-related liver disease (for exclusion criteria 3), or chronic kidney disease (for exclusion criteria 5).

4.3 Variables

4.4.1 Treatment strategies

The active and comparator treatment strategies used in the target trial are provided in Table 2. We aim to match these strategies as closely as possible; however, the target trial specifies doses and frequency of treatment, and this information is not available in the UK or US CF registries.

For both the UK and US emulated trials, the active treatment is prescription of prophylactic oral or chronic oral azithromycin and the comparator is no prescription of prophylactic oral or chronic oral azithromycin.

4.4.2 Outcomes

Where possible, the emulated trials will replicate outcomes studied in the target trial; however, data are not available in the registries for all secondary outcomes. The primary outcome in the target trial, and both emulated trials, is absolute FEV₁% at time 1. Secondary outcomes in the target trial include: number of pulmonary exacerbations, forced vital capacity, nutritional status with body mass index

(BMI) z-score, the use of antibiotics, modifications of microbiological analysis of sputum or throat cultures, and quality of life. Secondary outcomes in the emulated trials include:

- Prescription of IV antibiotics at time 1 (as a proxy for pulmonary exacerbations)
- Percent predicted forced vital capacity at time 1
- BMI z-score at time 1

For the emulated trials, FEV₁% will be calculated using the Global Lung Initiative 2012 equations [25] and BMI z-scores will be calculated using the WHO reference distribution [26].

4.4.3 Covariates

In the target trial, individuals were randomly allocated to the treatment or placebo strategy. In the emulated trials, there is no randomisation. The data on treatment use within the UK and US CF Registries reflects treatment decisions made based on clinical indication and on clinician and patient preference. Being prescribed azithromycin is therefore assumed to be informed by a number of factors, many of which are also associated with the outcomes of interest. The association between prescription of azithromycin and FEV₁% (and secondary outcomes) is therefore believed to be confounded by the following factors: age, number of days on intravenous antibiotics (IV days), non-IV hospital admissions, presence of *Pseudomonas aeruginosa*, *Staphylococcus aureus* or Nontuberculous Mycobacteria, cystic fibrosis-related diabetes (CFRD), use of hypertonic saline, inhaled antibiotics or DNase, and rate of decline in FEV₁%, BMI z-score and FEV₁%. These are depicted in a directed acyclic graph (DAG) in Figure 2.

Age, FEV₁% and BMI z-score are continuous. Rate of decline in FEV₁% is calculated as difference between the absolute FEV₁% measured at the index visit and the prior visit. Data on treatment prescription, presence of infections, CFRD diagnosis, pancreatic insufficiency and non-IV hospital admissions will be binary indicators. Indicators for pancreatic insufficiency and non-IV hospital admissions are created using existing variables in the data (see Appendix Tables A.1 and A.2). Registry data provide dates for treatment with IV antibiotics (at home or hospital). These data will be used to create a variable indicating number of days on IV antibiotics since last annual review (including treatment administered at home and hospital). IV days will then be treated as a categorical variable with four categories: 0, 1-14, 15-28, 28+.

4.5 Data analysis

The following data analysis plan will be implemented in both the UK and US emulated trials.

4.5.1 Notation

Let A denote an indicator variable for treatment strategy ($A = 0$ indicates no prescription of azithromycin and $A = 1$ indicates prescription of azithromycin). Let $Y_t(0)$ denote the potential outcome under treatment $A = 0$ at time t , $t \in \{0,1\}$ where $t = 1$ is 12 months after $t = 0$. Similarly, let $Y_t(1)$ denote the potential outcome under treatment $A = 1$ at time t . Finally, C denotes the confounding factors listed in section 4.4.3 and in Figure 2. Sections 4.5.2 and 4.5.3 describe the causal estimand of interest and main analysis plan with a focus on the primary outcome. Section 4.5.4 summarises key differences in the analysis plan for the secondary outcomes.

4.5.2 Causal estimand of interest

The target trial reported the difference in mean changes (between month 0 and month 12) in FEV₁% between treatment groups in the total population:

$$ATE = E\{Y_1(1) - Y_0(1)\} - E\{Y_1(0) - Y_0(0)\} \quad (1)$$

where the expectations refer to the population of individuals meeting the criteria for the target trial. This is equivalent to the difference in means at the end of follow-up as the observed value of Y at time 0 is unaffected by treatment, i.e. $Y_0(a) = Y_0$, i. e.

$$ATE = E\{Y_1(1)\} - E\{Y_1(0)\} \quad (2)$$

Our causal estimand is interpreted as the expected difference in FEV₁% at month 12 if everyone had taken azithromycin for 12 months, compared to a scenario where no-one took azithromycin for 12 months.

4.5.3 Main analysis

In the target trial, the authors investigated the change in FEV₁% from baseline, with adjustment for baseline FEV₁%. This is equivalent to a regression of the mean FEV₁% at follow-up with adjustment for baseline FEV₁%. We note that in the absence of imbalance between treatment groups in baseline FEV₁%, the adjustment of baseline FEV₁% is not required, but may result in gains in efficiency. Both approaches result in estimates of the ATE as defined in equations (1) and (2).

In the emulated trials, we need to account for differences between treatment groups at baseline, including baseline FEV₁%. We also require assumptions of positivity, no interference, consistency and conditional exchangeability (conditional on \mathbf{C}). We use augmented inverse-probability-of-treatment weighting (AIPTW) to control for potential confounding by \mathbf{C} . The first step in this method is to estimate the propensity scores, i.e., the probability of treatment conditional on baseline covariates:

$$PS = P(A = 1|\mathbf{C}) \quad (3)$$

Propensity scores will be estimated using logistic regression. Secondly, we specify an outcome model conditional on treatment and covariates:

$$Y_1 = \gamma_0 + \gamma_A A + \boldsymbol{\gamma}_C \mathbf{C} + \varepsilon \quad (4)$$

This is used to estimate expected outcomes under the two treatments. The AIPTW estimator for $E[Y_1(a)]$, ($a \in \{0,1\}$) is given by:

$$\hat{E}[Y_1(a)] = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{I(A_i = a_i) Y_i}{\hat{P}(A_i = a_i | \mathbf{C}_i)} - \frac{I(A_i = a_i) - \hat{P}(A_i = a_i | \mathbf{C}_i)}{\hat{P}(A_i = a_i | \mathbf{C}_i)} \hat{E}(Y_i | A_i = a_i, \mathbf{C}_i) \right\} \quad (5)$$

where $\hat{P}(A_i = a_i | \mathbf{C}_i)$ can be obtained using predictions from the propensity score model. The estimator in equation 5 can be used to estimate $E[Y_1(0)]$ and $E[Y_1(1)]$, and the difference between these two expectations is an estimate of the average treatment effect in the population. The AIPTW estimator is doubly robust, in that it provides consistent estimates if either the propensity score model or the outcome model is correctly specified, under the stated assumptions [27]. Standard errors can be obtained based on the efficient influence function [28].

Additional analysis in the UK Emulated Trial making use of data on treatment prescription dates

A limitation of this approach is that it assumes individuals with $A = 1$ at time 1 have been taking azithromycin for the past 12 months. Realistically, individuals may initiate treatment with azithromycin at any time between time 0 and time 1. For the most recent time-period, we can conduct a second analysis using the UK data and making use of data on treatment prescription dates.

In this second analysis, AIPTW is used as above, but the outcome model used previously (equation 3) is modified to include a variable indicating time and interaction term between time and treatment. We define a new time variable, t^* , which measures time in months. For treated individuals, $t^* = 0$ on the first date they are prescribed azithromycin after the index visit. For control individuals, $t^* = 0$ for the date of the index visit. We let $Y_{t^*=12}$ denote FEV₁% measured on the day of the annual review after $t^* = 0$ and closest in time to $t^* = 12$. Note that this accommodates the fact that annual review visits do not always take place exactly 12 months apart. The outcome model will then be defined as:

$$Y_{t^*=12} = \delta_0 + \delta_A A + \boldsymbol{\gamma}_C \mathbf{C} + \delta_{t^*} t^* + \delta_{At^*} A t^* + \varepsilon \quad (6)$$

After fitting this model, we set $t^* = 12$ to obtain the relevant expected outcomes.

Diagnostics

The distribution of weights will be assessed using summary statistics and plots. Methods such as trimming or truncating will be considered to deal with extreme weights. Standardised mean differences will be used to compare the balance in the distribution of confounders between treatment and control groups in the original and weighted samples.

4.5.4 Secondary outcomes

The secondary outcomes are: prescription of IV antibiotics, percent predicted forced vital capacity (FVC%) and BMI z-score. Analysis of the continuous outcomes (FVC% and BMI z-score) can be implemented as described above. When the outcome is prescription of IV antibiotics, Y is binary and the causal estimand of interest is the risk difference. The analysis plan for our binary outcome is similar, except logistic regression will be used for the outcome models (equations 4 and 6).

4.5.5 Sensitivity analyses

Sensitivity to the no unmeasured confounders assumption

Our analysis relies on the assumption that there are no unmeasured confounders. Unfortunately, there may exist some factors that are associated with both treatment prescription and outcome, which are not captured in the registries (denoted by U in Figure 2). Sensitivity to unmeasured confounders will be summarised using E-values [29].

Allowing individuals to enter the emulated trials more than once

For the main analysis, individuals will be included in the emulated trial once. Individuals “enter” the trial at time 0, which is defined as the earliest year within the recruitment period that they meet the inclusion and exclusion criteria. This approach restricts the analysis to using information from everyone at one time point only and may be inefficient. Alternatively, we can allow individuals to “enter” the trial twice if they meet the inclusion and exclusion criteria in both years during the recruitment period. Standard errors will need to take into account that individuals are included multiple times.

4.5.6 Missing data

The amount of missing data in each variable will be summarised in tables by treatment group. Where there is missing data in binary time-varying variables that are usually static for long time periods, we will use a simple imputation approach. For missing visits where the prior visit and subsequent visit are equal, we will assume the missing value is also equal and impute accordingly. This approach will be used for the following variables: pancreatic insufficiency, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, NTM, CFRD, inhaled antibiotics, inhaled steroids, hypertonic saline and DNase. Missingness patterns in the remaining missing data will be explored. If there are missing outcomes that are missing at random conditional on C , then a complete case analysis is appropriate [28]. If a complete case analysis is not appropriate, more complex missing data methods such multiple imputation by chained equations [30] may be considered.

Comparison of results against the target trial

We will compare our results with those from the target trial with the aim of determining whether results from the emulated trials are compatible with the target trial. The following criteria will be considered, as were used in the RCT DUPLICATE Project [8]:

- (1) do the estimated ATEs from the emulated trials replicate the direction and statistical significance of the estimated ATE in the target trial?
- (2) do the estimated ATEs from the emulated trials lie within the 95% confidence intervals for the ATE estimates reported in the target trial?
- (3) is there evidence against the null hypothesis of no difference between the ATE estimates from the emulated trials and that from target trial? To assess this, we calculate the standardised mean

difference between the effect estimate obtained in the target trial and that obtained in the emulated trial. Evidence against the null hypothesis at the 5% level is indicated by a standardised mean difference greater than 1.96.

5 Limitations

There are a number of limitations to this trial emulation, which are sources of potential bias in our results, and which may explain any differences in findings between the target trial and the emulated trial. Here, we identify a number of sources of bias and/or potential reasons we may observe differences in the results, due to either limitations regarding data availability in the registries, differences in sample size, or differences in the study populations.

Data availability

The target trial specified a particular dose of azithromycin depending on an individual's weight. The trial also reported a high adherence, estimated at 95% for azithromycin and placebo. Neither the UK nor US Registry provides reliable data on treatment doses and it is possible that individuals in the registry will take different doses to those given in the target trial (see Table 2). There is also no data on adherence and our emulated trial relies on data on treatment prescription, which may differ from actual treatment use.

Some of the exclusion criteria of the target trial cannot be replicated exactly in the emulated trial. For example, the target trial included a criterion based on liver function tests, with individuals excluded if they had liver disease with liver function tests more than twice the laboratory upper limit. In the UK Registry, the closest variable to this criterion is an indicator for acute liver failure with liver function tests greater than three times the laboratory upper limit. The US Registry has a similar variable, but data collection for this variable began in 2015, so it can only be used for the most recent time period.

The target trial calculated the outcome, FEV₁%, using the Knudson equations [31]; we plan to use the Global Lung Initiative (GLI) equations in the emulated trial [25], as these are now more commonly used. Previous research suggests that results will be minimally affected by choice of reference equations [32].

The main analyses in the UK emulated trials will use data from consecutive annual review visits. We assume that the annual review visits are 12 months apart and that individuals in the treatment group were taking azithromycin for the 12 months in between visits. In practice, the annual review visits are not always exactly one year apart, and individuals may begin treatment with azithromycin at any time during the time between visits. We address this limitation to some extent in an additional analysis for the UK Registry data, in which we incorporate prescription dates data.

Finally, our analysis relies on the assumption that all confounding of the treatment-outcome association is accounted for in the analysis. It is possible that there are some factors associated with both azithromycin prescription and the outcome, that are not collected in the registry. We plan a sensitivity analysis to assess how sensitive our results are to unmeasured confounders.

Sample size

The target trial included 82 individuals (40 in the treated group and 42 in the placebo group). The authors note in their discussion that it is possible the study was not adequately powered to detect significant differences in FEV₁%.

We have not performed sample size calculations for the emulated trials and there is some debate as to whether sample size calculations are needed in studies using observational data [33-35]. We plan to use all the available data in the UK CF Registry, or US CF Registry, and expect much larger sample sizes than were used in the target trial.

Differences in the study populations

Ideally, we would conduct the emulated trials using data from a similar time-period as the target trial, to ensure homogeneity in the clinical settings. The target trial was conducted from 2001 to 2003, but azithromycin was not commonly used in clinical practice at this time. Therefore, for the emulated trials, we need to wait for the treatment to uptake in clinical practice, i.e., after the earlier azithromycin trials were published in 2002 [22,23]. Additionally, in 2007, the UK CF Registry introduced a new web-based data collection system which improved data collection and data quality. Restricting the emulated trials to years prior to 2007 would therefore not make use of the years with higher data quality. On the other hand, using later years could result in differences in the clinical setting between the emulated and target trials. For this reason, we have suggested multiple time periods for the emulated trials and will compare results between time periods. Finally, due to the way the data are collected in the two registries, we require different definitions of time 0 for the UK and US emulated trials. The different definitions may lead to slightly different populations of interest and therefore the estimands between the UK and US emulated trials would be based on different populations. This could lead to different results between the emulated trials.

6 Ethics and dissemination

This project will use anonymised data from the UK Cystic Fibrosis Registry, which has Research Ethics Approval (ref: 24/EE/0012) and from the US Cystic Fibrosis Registry. This protocol was reviewed by Advarra IRB and it was confirmed that no study specific IRB approval was required to use the US Cystic Fibrosis Registry data. No additional data beyond that contained in the registries will be collected for the project. Ethical approval has been granted by the London School of Hygiene and Tropical Medicine Ethics Committee (Ref: 29609). The study has also been approved by the UK CF Registry Research Committee and the North Star Review Board.

We plan to publish the results of this study in a high-ranking peer-reviewed journal. Findings will also be presented at relevant scientific conferences such as the European Cystic Fibrosis Conference, the North American Cystic Fibrosis Conference, and the International Society for Clinical Biostatistics.

7 Conclusions

This work is being undertaken by the CF Trial Emulation Network (CF-TEN), a new multidisciplinary international collaborative network. We plan to replicate multiple trials on the effects of the use of azithromycin on health outcomes for people with cystic fibrosis.

This work will contribute to the evidence base for the target trial emulation approach in CF. If the trial emulations are a success, we could extend the research to study questions beyond the trial. For example, the longer-term effects of azithromycin, effects of azithromycin use on other outcomes such as risk of NTM infection, or combination effects of multiple treatments. Such questions are often difficult to study in RCTs due to additional costs or lack of statistical power.

If we can consistently replicate the results of existing trials, then we can gain confidence in future studies using registry data to investigate questions that may not be answered in a RCT.

Acknowledgements

The authors would like to thank the Cystic Fibrosis Foundation and the Cystic Fibrosis Trust for the use of patient registry data to conduct this study. The authors also thank the patients, care providers and clinic coordinators at CF centres throughout the USA and the UK for their contributions to the patient registries.

Data Statement

UK Registry Data are available following application to the UK CF Registry Research Committee. <https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry/apply-for-data-from-the-uk-cf-registry>.

Author Contributions

All authors contributed to the study question and design. EG wrote the first draft of the protocol manuscript. All authors contributed to further drafts and approved the final version.

Funding

GD is supported by a UKRI Future Leaders fellowship (MR/T041285/1). RHK and EG are supported by a UKRI Future Leaders fellowship (MR/S017968/1) awarded to RHK. BH reports funding support from CF Statistical Network and Expertise Award (StatNet SZCZES18Y7). AG, BH and RS are supported by a CFF Leadership and Design grant (SZCZES22A0-LAD, and RS is supported by an NIH/NHLBI grant (R01 HL141286). The funders had no role in the writing, decision to publish, or any other aspect of this manuscript.

Competing Interests

GD reports speaker honoraria from Chiesi Ltd and Vertex Pharmaceuticals, and advisory board and clinical trial leadership roles with Vertex Pharmaceuticals unrelated to the current work. RHK reports a speaker honorarium from Vertex Pharmaceuticals. FF reports speaker honoraria from Chiesi Ltd, Vertex Pharmaceuticals. SS reports speaker fees from Vyaire Medical and consulting fees from Chiesi Pharmaceuticals. RS reports personal fees from CFF.

References

- [1] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol.* 2016 Apr 15;183(8):758-64.
- [2] Wing K, Williamson E, Carpenter JR, et al. Real world effects of COPD medications: a cohort study with validation against results from randomised controlled trials. *Eur Respir J.* 2021 Mar; 57(3): 2001586. doi: [10.1183/13993003.01586-2020](https://doi.org/10.1183/13993003.01586-2020)
- [3] Fu EL, Evans M, Clase CM, et al. Stopping Renin-Angiotensin System Inhibitors in Patients with Advanced CKD and Risk of Adverse Outcomes: A Nationwide Study. *J Am Soc Nephrol* 2021 Feb;32(2):424-435
- [4] Cain SE, Saag MS, Petersen M, et al. Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. *Int J Epidemiol.* 2016 Dec 1;45(6):2038-2049
- [5] García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol.* 2017;32:495–500.
- [6] Trevisan M, Fu EL, Yang X, et al. Stopping mineralocorticoid receptor antagonists after hyperkalaemia: trial emulation in data from routine care. *European Journal of Heart Failure* 2021; 10: 1698-1707. doi.org/10.1002/ejhf.2287
- [7] Ioannou GN, Locke ER, O'Hare AM, Bohnert ASB, Boyko EJ, Hynes DM, Berry K. COVID-19 Vaccination Effectiveness Against Infection or Death in a National U.S. Health Care System: A Target Trial Emulation Study. *Ann Intern Med.* 2022;175(3):352-361.
- [8] Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA.* 2023 Apr 25;329(16):1376–85.
- [9] Matthews AA, Dahabreh IJ, Fröbert O, Lindahl B, James S, Feychting M, Jernberg T, Berglund A, Hernán MA. Benchmarking Observational Analyses Before Using Them to Address Questions Trials Do Not Answer: An Application to Coronary Thrombus Aspiration. *Am J Epidemiol.* 2022 Aug 22;191(9):1652-1665.
- [10] Fröbert O, Lagerqvist B, Gudnason T, Thuesen L, Svensson R, Olivecrona GK, James SK. Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A

multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and rationale. *Am Heart J*. 2010 Dec;160(6):1042-8.

[11] Admon AJ, Donnelly JP, Casey JD, Janz DR, Russell DW, Joffe AM, Vonderhaar DJ, Dischert KM, Stempel SB, Dargin JM, Rice TW, Iwashyna TJ, Semler MW. Emulating a Novel Clinical Trial Using Existing Observational Data. Predicting Results of the PreVent Study. *Ann Am Thorac Soc*. 2019 Aug;16(8):998-1007.

[12] Casey JD, Janz DR, Russell DW, Vonderhaar DJ, Joffe AM, Dischert KM, Brown RM, Zouk AN, Gulati S, Heideman BE, Lester MG, Toporek AH, Bentov I, Self WH, Rice TW, Semler MW; PreVent Investigators and the Pragmatic Critical Care Research Group. Bag-Mask Ventilation during Tracheal Intubation of Critically Ill Adults. *N Engl J Med*. 2019 Feb 28;380(9):811-821

[13] Granger E, Keogh RH, Frost F. The long-term effects of insulin use in incident cystic fibrosis-related diabetes: a target trial emulated using longitudinal national registry data. *ERJ Open Res*. 2022 Nov 7;8(4):00170-2022.

[14] Granger E, Davies G, Keogh RH. Emulated trial investigating effects of multiple treatments: estimating combined effects of mucoactive nebulisers in cystic fibrosis using registry data. *Thorax*. 2023 Oct;78(10):1011-1018.

[15] Wang SV, Pottegård A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, Berger ML, Crane G, Goettsch W, Hua W, Kabadi S, Kern DM, Kurz X, Langan S, Nonaka T, Orsini L, Perez-Gutthann S, Pinheiro S, Pratt N, Schneeweiss S, Toussi M, Williams RJ. HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. *Pharmacoepidemiol Drug Saf*. 2023 Jan;32(1):44-55

[16] Clement A, Tamalet A, Leroux E, Ravilly S, Fauroux B, Jais JP. Long term effects of azithromycin in patients with cystic fibrosis: A double blind, placebo controlled trial. *Thorax*. 2006 Oct;61(10):895-902

[17] Taylor-Robinson D, Archangelidi O, Carr SB, Cosgriff R, Gunn E, Keogh RH, MacDougall A, Newsome S, Schlüter DK, Stanojevic S, Bilton D; CF-EpinNet collaboration. Data Resource Profile: The UK Cystic Fibrosis Registry. *Int J Epidemiol*. 2018 Feb 1;47(1):9-10e.

[18] Cystic Fibrosis Trust. Cystic Fibrosis Trust Homepage. Available at: <https://www.cysticfibrosis.org.uk/> (Accessed: 05/07/2023).

[19] Knapp EA, Fink AK, Goss CH, Sewall A, Ostrenga J, Dowd C, Elbert A, Petren KM, Marshall BC. The Cystic Fibrosis Foundation Patient Registry. Design and Methods of a National Observational Disease Registry. *Ann Am Thorac Soc*. 2016 Jul;13(7):1173-9.

[20] Cromwell EA, Ostrenga JS, Todd JV, Elbert A, Brown AW, Faro A, Goss CH, Marshall BC. Cystic fibrosis prevalence in the United States and participation in the Cystic Fibrosis Foundation Patient Registry in 2020. *J Cyst Fibros*. 2023 May;22(3):436-442.

[21] Cystic Fibrosis Foundation. Cystic Fibrosis Foundation Homepage. Available at: <https://www.cff.org/> (Accessed: 05/07/2023).

[22] Wolter J, Seeney S, Bell S, Bowler S, Masel P, McCormack J. Effect of long term treatment with azithromycin on disease parameters in cystic fibrosis: a randomised trial. *Thorax* 2002;57(3):212-6.

[23] Equi A, Balfour-Lynn IM, Bush A, Rosenthal M. Long term azithromycin in children with cystic fibrosis: a randomised, placebo-controlled crossover trial. *Lancet* 2002;360(9338):978-84.

[24] Saiman L, Marshall BC, Mayer-Hamblett N, Burns JL, Quittner AL, Cibene DA, Coquillette S, Fieberg AY, Accurso FJ, Campbell PW 3rd; Macrolide Study Group. Azithromycin in patients with

cystic fibrosis chronically infected with *Pseudomonas aeruginosa*: a randomized controlled trial. *JAMA*. 2003 Oct 1;290(13):1749-56.

[25] Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MS, Zheng J, Stocks J; ERS Global Lung Function Initiative. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J*. 2012 Dec;40(6):1324-43

[26] World Health Organisation. BMI-for-age (5-19 years). 2021. Available at: www.who.int. <https://www.who.int/toolkits/growth-reference-data-for-5to19-years/indicators/bmi-for-age>. [Accessed: 06/07/2023]

[27] Kurz CF. Augmented Inverse Probability Weighting and the Double Robustness Property. *Med Decis Making*. 2022 Feb;42(2):156-167

[28] Zhong Y, Kennedy EH, Bodnar LM, Naimi AI. AIPW: An R Package for Augmented Inverse Probability-Weighted Estimation of Average Causal Effects. *American Journal of Epidemiology*. 2021 Dec; 190(12): 2690-2699.

[29] VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*. 2017 Aug 15;167(4):268-274

[30] White, Ian R., Patrick Royston, and Angela M. Wood. "Multiple imputation using chained equations: issues and guidance for practice." *Statistics in medicine* 30.4 (2011): 377-399.

[31] Knudson RJ, Lebowitz MD, Holberg CJ, Burrows B. Changes in the normal maximal expiratory flow-volume curve with growth and aging. *Am Rev Respir Dis*. 1983; **127**: 725– 734.

[32] Konstan MW, Wagener JS, VanDevanter DR, Pasta DJ, Millar SJ, Morgan WJ; Scientific Advisory Group and the Investigators and Coordinators of the Epidemiologic Study of Cystic Fibrosis. Comparison of FEV₁ reference equations for evaluating a cystic fibrosis therapeutic intervention. *Pediatr Pulmonol*. 2017 Aug;52(8):1013-1019.

[33] Hernán MA. Causal analyses of existing databases: no power calculations required. *J Clin Epidemiol*. 2022;144:203-205.

[34] Morris TP, van Smeden M. Causal analyses of existing databases: the importance of understanding what can be achieved with your data before analysis (commentary on Hernán). *J Clin Epidemiol*. 2022;142:261-263.

[35] Mansournia MA. Sample size considerations are needed for the causal analyses of existing databases. *J Clin Epidemiol*. 2022;141:212.

Table 1: Description of the primary research question addressed in the target trial (Clement et al, Thorax 2006;61:895-902)

Objective:	Investigate whether long-term use of azithromycin is associated with respiratory benefits in young people with CF
Population:	CF patients aged older than 6 years and forced expiratory volume in 1 second (FEV ₁ %) of 40% or more.
Exposure:	Oral azithromycin
Comparator:	Placebo pills
Primary Outcome:	Change in FEV ₁ %
Time:	12 months
Setting:	Patients recruited from 18 CF accredited care centres in France
Main measure of effect:	Difference in change in FEV ₁ % from baseline between treatment groups

Table 2: Summary of key components of the protocol for the target trial and emulated trials

Protocol Component	Target trial based on Clement et al, Thorax 2006;61:895-902	Emulation of the target trial using UK CF Registry data	Emulation of the target trial using US CF Registry data
Eligibility Criteria	<p>Include: French individuals diagnosed with cystic fibrosis (sweat chloride >60mmol/l or a genotype known to cause the disease), aged 6-21 years, with the ability to perform pulmonary function tests with FEV₁%>40, and the ability to swallow tablets.</p> <p>Individuals were excluded if they had the following:</p> <ol style="list-style-type: none"> 1. Allergy to macrolide antibiotics 2. Long-term (>3 months) with macrolides during the 12-month period before study entry 3. Liver disease with liver function tests >2 times the laboratory upper limit 4. History of portal hypertension 5. Kidney disease with serum creatinine > 150 µmol/l and/or creatinine clearance <50ml/min 6. Use of any of the following in the 3 months before study entry: DNase, inhaled tobramycin, inhaled steroids 	<p>Individuals will be considered for inclusion if they have a clinically confirmed diagnosis of cystic fibrosis (i.e. are present in the UK CF Registry) and have an observation date within the recruitment-periods defined in section 4.3.1, aged between 6 and 21 years, and obtained FEV₁%>40 on their pulmonary function test (taken on the day of the annual review). It is assumed that all individuals have the ability to swallow tablets.</p> <p>Exclusion criteria are as follows:</p> <ol style="list-style-type: none"> 1. Intolerance to macrolide antibiotics recorded at any time during study period. 2. Prescription of chronic oral or prophylactic oral macrolides (including azithromycin) recorded at time 0. 3. Acute liver failure with > 3 x the upper laboratory limit, INR>2, or not responsive to vitamin K at time 0 4. Recorded cirrhosis with portal hypertension at time 0. 5. Serum creatinine levels > 150 µmol/l at time 0. 6. Prescription of DNase, inhaled tobramycin or inhaled corticosteroids recorded at time 0. 7. No follow-up visit for time 1. 	<p>Individuals will be considered for inclusion if they have a clinically confirmed diagnosis of cystic fibrosis (i.e. are present in the US CF Registry) and have an observation or encounter date within the time-periods defined in section 4.3.1, aged between 6 and 21 years, and obtained FEV₁%>40 on their pulmonary function test (taken on the day of the encounter visit). It is assumed that all individuals have the ability to swallow tablets.</p> <p>Exclusion criteria are as follows:</p> <ol style="list-style-type: none"> 1. As in UK Emulated Trial 2. As in UK Emulated Trial 3. Non-cystic fibrosis related liver disease recorded at time 0. Laboratory results from liver tests are not available in the US registry. 4. Recorded cirrhosis at time 0. Portal hypertension is not available in the US registry. 5. As in the UK Emulated Trial 6. As in the UK Emulated Trial 7. As in UK Emulated Trial

Treatment strategies	<p>The active intervention was azithromycin supplied as 250mg tablets and the comparator was placebo pills.</p> <p>Individuals weighing less than 40kg took one tablet 3 days per week and individuals weighing more than 40kg took two tablets 3 days per week.</p>	<p>The active intervention is prescription of oral azithromycin and the comparator is no prescription of oral azithromycin.</p> <p>Further details on the treatment strategies are provided in section 4.4.1</p>
Assignment procedures	<p>Individuals were randomised to treatment strategy. Randomisation was stratified according to centre and <i>P. aeruginosa</i> infection status. The patients and all study investigators remained blinded to the treatment assignment until study completion.</p>	<p>In the emulated trials individuals are not randomly assigned to the treatment strategy. This is accounted for in the analysis.</p>
Follow-up period	12 months	As in the target trial
Outcome	<p>Primary outcome: mean change in FEV₁% between month 0 and month 12.</p> <p>Secondary outcomes included: evaluation of the number of pulmonary exacerbations, the use of antibiotics, modifications of microbiological analysis of sputum or throat cultures, changes in FVC, nutritional status with measurement of body mass index (BMI), and quality of life.</p>	<p>Primary outcome: absolute FEV₁% at the end of follow-up.</p> <p>Secondary outcomes include: prescription of IV antibiotics, FVC, BMI z-score.</p> <p>Further details on the outcomes are provided in section 4.4.2.</p>
Causal contrasts of interest	Intention-to-treat	Per-protocol
Analysis plan	<p>For continuous outcomes, mean differences between treatment groups were estimated using mixed models; for binary outcomes, logistic regression was used; for count outcomes, Poisson</p>	<p>A direct acyclic graph is used to inform which variables need to be controlled for (see section 4.4.3).</p> <p>Confounding by measured variables will be accounted for using inverse-probability-of-treatment weighting.</p>

	regression was used. Models were adjusted for interactions taking into account effects due to centre and baseline characteristics (including <i>P. aeruginosa</i> infection status)	Further details on the analysis plan are provided in section 4.5.
--	---	---

Table 3: Description of the time-periods considered in the UK and US registry data

Time period	Justification for time-period	UK data	US data
2003-2005	Closest time-period to the target trial, allowing a couple of years for use of azithromycin to uptake in clinical practice. This will only be conducted using UK data as the US registry did not collect data on azithromycin use during this time.	✓	×
2007-2009	US registry started collecting data on azithromycin in 2006 and one year wash-out period is required to select individuals who were not taking azithromycin prior to study entry. Therefore, this time-period is as close to the time-period used in the target trial as is possible for the US registry.	✓	✓
2016-2018	UK registry started collecting data on treatment prescription dates in 2016. Therefore, this time-period is a more recent period that pre-dates widespread use of CFTR modulators, but also allows use of treatment dates data.	✓	✓

Figure Captions

Figure 1: Study Design Diagram illustrating from which visits data are extracted for time 0 and time 1 in the UK and US Emulated Trials. Subscripts denote time.

Figure 2: Directed Acyclic Graph depicting assumed confounding relationships for the association between azithromycin at time 0 ($azithromycin_0$) and FEV₁% at time 1 ($FEV_1\%_1$). U represents any unmeasured confounders. Subscripts denote time; 0 indicates pre-baseline.*