University of Sheffield

Sheffield Centre
For Health &
Related Research

HEALTH ECONOMICS & DECISION SCIENCE

Discussion Paper Series

HEDS Discussion Paper 24.02

Title: Protocol for emulating Non-Small Cell Lung and Triple Negative Breast Cancer Target Trials from England's Cancer Registry Data

Authors: Saleema Rex

Corresponding Author: Saleema Rex
Sheffield Centre for health and related research
(SCHARR), Division of Population Health, School of
Medicine and Population Health, University of Sheffield,
Regent Court, 30 Regent Street, Sheffield, S1 4DA UK
Email: ssrex1@sheffield.ac.uk

Disclaimer:
This series is intended to promote discussion and to provide information about work in progress.
The views expressed in this series are those of the authors.
Comments are welcome, and should be sent to the corresponding author.

RECReATE

Researching the use of England's Cancer Registry
data for Assessing Treatment Effectiveness

University of Sheffield

Protocol for emulating Non-Small Cell Lung and Triple Negative Breast Cancer Target Trials from England's Cancer Registry Data.

Authors

PhD student: Saleema Rex

Supervisors: Nicholas Latimer, Ron Akehurst, Mike Bradburn

Institute: Health Economics and Decision Science,

School of Medicine and Population Health,

University of Sheffield

Project summary

RECReATE: Researching the use of England's Cancer Registry data for Assessing Treatment Effectiveness.	
Protocol tile	Protocol for emulating Non-Small Cell Lung and Triple Negative Breast Cancer Target Trials Using England's Cancer Registry Data.
Project Objective	The project aims to evaluate whether England’s cancer registry data can be reliably used to compare the effectiveness of different cancer treatments provided in the NHS using the Target Trial framework by replicating two existing lung cancer trials and one breast cancer trial.
PhD Student	Mrs Saleema Rex
Project team members	Prof Nicholas Latimer Prof Ron Akehurst Mr Mike Bradburn
Clinical advisors	Dr Robin Young, Consultant, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield Dr Matthew Winter, Consultant, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield
Project institution	School of Medicine and Population Health, University of Sheffield
Project Funder	Yorkshire Cancer Research, Lumanity
Study protocol version	Version 1

Study protocol version history

Version number	Date of approval	Summary of changes
Version 1	7th May 2024	First version

Contents

List of tables

List of figures

Abbreviations

A&E	Accident and Emergency
ALK	Anaplastic Lymphoma Kinase
BMI	Body Mass Index
BRCA	BReast CAncer gene
CDF	Cancer Drug Fund
CfG	Centre for Guidelines
DAG	Directed Acyclic Graph
ECOG	Eastern Cooperative Oncology Group
EGFR	Epidermal Growth Factor Receptor Mutation
EGFR-TK	EGFR tyrosine kinase
ER	Estrogen Receptor
FDA	Food and Drug Administration
HER2	Human Epidermal Growth Factor Receptor 2
HES	Hospital Episodes Statistics
HTA	Health Technology Assessment
ICD	The International Classification of Diseases
IMD	Index of Multiple Deprivation
IPW	Inverse-Probability Weights
ITT	Intention-To-Treat
KM	Kaplan Meier
NCRAS	National Cancer Registration and Analysis Service
NCRD	National Cancer Registration Dataset
NDRS	National Disease Registration Service
NHS	National Health Services
NICE	National Institute for Health and Care Excellence
NSCLC	Non-small cell lung cancer

OLDW	OptumLabs Data Warehouse
OPERAND	Observational Patient Evidence for Regulatory Approval and uNderstanding Disease
OS	Overall survival
PD-L1	Programmed cell Death Ligand 1
PgR	Progesterone Receptor
PP	Per-Protocol
RCT	Randomised Controlled Trial
RECReATE	Researching the use of England's Cancer Registry data for Assessing Treatment Effectiveness
RTDS	Radiotherapy dataset
RWD	Real-World Data
RWE	Real-World Evidence
SACT	Systemic Anti-Cancer therapy
SAP	Statistical Analysis Plan
SDE	Secure Data Environment
TA	Technology Appraisal
TAC	Technology Appraisal Committees
TKI	Tyrosine Kinase Inhibitors
TMLE	Targeted Maximum Likelihood Estimation
TNBC	Triple Negative Breast Cancer
TNM	Tumour Node Metastasis
TRE	Trusted Research Environments
TT	Target Trial
UK	United Kingdom

Abstract

Background

England’s cancer registry data in theory offer a wealth of data for comparing the effects of cancer treatments [1, 2]. Linking England’s cancer registry data with other healthcare datasets allows researchers to estimate how well treatments perform in real-world settings, benefiting patients, clinicians, and policymakers. However, drawing reliable conclusions from observational data can be challenging as biases inherent to observational studies and data quality issues can lead to misleading results [2, 3].

Miguel Hernan and James Robins (2016) proposed the Target Trial (TT) framework as a more robust approach to estimate comparative treatment effectiveness using observational data [4]. This framework leverages counterfactual theories and principles from randomised controlled trials (RCTs), considered the gold standard for evaluating treatments [5]. However, the reliability of the estimates obtained from applying the TT framework to analyse England's cancer registry data remains unknown.

Addressing the Uncertainty: A Benchmarking Approach

This PhD project will address this uncertainty through a benchmarking process. Benchmarking involves comparing estimates derived from observational analyses against established gold standard evidence from RCTs [5]. By designing and emulating three TTs leveraged on three carefully selected existing RCTs (LUX-Lung 7 [6], KEYNOTE-024 [7], and TNT [8]) using England’s cancer registry data within the TT framework, this project will assess the reliability of treatment effectiveness estimates using this approach.

Expected Impact

This study has the potential to provide valuable insights into the reliability of comparative effectiveness estimates derived from England's cancer registry data. These findings will inform clinicians, policymakers, and patients on the appropriate use and interpretation of results from future comparative effectiveness studies utilising this data source.

1. Background

1.1. The RECReATE project

The case studies described in this protocol are part of a larger project titled ‘RECReATE’ (**R**esearching the use of **E**ngland's **C**ancer **Re**gistry data for **A**ssessing **T**reatment **E**ffectiveness) led by Professor Nicholas Latimer. The overarching objective of the RECReATE project is to evaluate whether Real World Data (RWD), especially England’s cancer registry data, can be used to produce reliable estimates of comparative treatment effectiveness that can aid vital decision making by key stakeholders [9]. One way to assess the reliability of observational evidence is by comparing observational evidence against reliable external evidence. This process of evaluating reliability is often called 'Benchmarking' [10, 11]. In the case of comparative treatment effectiveness research, ‘Benchmarking’ can be done by comparing RWE against existing RCT evidence, as RCT evidence is considered the ‘Gold Standard’ [12].

The RCT DUPLICATE project, funded by the United States Food and Drug Administration (FDA) agency, emulated Target Trials (TTs) designed based on existing RCTs using insurance claims data to evaluate whether the evidence from observational data can lead to the same conclusion as an RCT and the circumstances under which this is possible [13]. Similarly, the Observational Patient Evidence for Regulatory Approval and uNderstanding Disease (OPERAND) project evaluates the use of evidence from observational data for regulatory decision-making by replicating existing clinical trials using administrative claims and Electronic Health Records (EHR) data from the OptumLabs Data Warehouse (OLDW) [10].

As a component of the RECReATE initiative, three existing RCTs have been selected for replication using England’s cancer registry data, in a manner similar to the RCT DUPLICATE and OPERAND projects. These three RCTs were used to design the three case studies described in this protocol. Other case studies utilising England’s cancer registry data and other data sources are also being conducted concurrently at the University of Sheffield as part of the RECReATE project.

1.2. Project rationale

Due to the extensive development of cancer drugs, a considerable number of NICE (The National Institute for Health and Care Excellence) Technology Appraisals (TAs) are conducted on cancer treatments to determine whether they should be provided by the National Health Service (NHS) [14]. Of the total number of Technology Appraisals (TAs) conducted between March 2000 and June 2021, about 48.3% (340 out of 701) were specifically focused on cancer treatments [15]. Randomised controlled trials (RCTs) and systematic reviews of RCTs serve as the primary sources of empirical evidence about the efficacy of treatments within the context of both Health Technology Assessment (HTA) and clinical guidelines procedures.

Using random treatment allocation in RCTs effectively eliminates or mitigates the presence of confounding by indication bias and certain types of selection bias (e.g., immortal-time bias). The randomisation process leads to treatment groups that are more evenly distributed, ensuring comparability across measured and unmeasured factors as the sample size increases. However, RCT evidence may not be available for all necessary treatment comparisons due to several factors, such as the prohibitively high costs and limited resources associated with conducting RCTs or ethical considerations [16, 17]. In certain instances, RCT studies can also be subjected to biases including selection bias. The Cochrane Risk of Bias tool 2 for RCTs categorises the biases in an RCT into five main domains. These domains are inappropriate randomisation process, deviations from intended interventions, missing data, poor outcome measurement and selective reporting [18, 19]. These biases can affect an RCT's internal and external validity. In general, RCTs are considered to have high internal validity but poor external validity due to dissimilarities between the people recruited within the RCT and the real-world population [16]. When RCT evidence is either unavailable or deemed inadequate, NICE decision making can be informed by Real-World Evidence (RWE) on comparative treatment effectiveness [20].

1.3. England's Cancer Registry data

Cancer is a disease area where extensive data are collected routinely. In England, the National Disease Registration Service (NDRS), including the National Cancer Registration and Analysis Service (NCRAS), is responsible for systematically collecting data on cancer patients. NCRAS collect data from multiple sources to provide real-world data covering the entire cancer pathway [1]. NHS England is the custodian of these datasets [21]. NHS England can provide anonymised linked datasets from multiple sources, including patient characteristics such as demographics, tumour details, planned treatment regimens, treatment cycles, medications, and outcome information. All this information can, in theory, be used to estimate the comparative effectiveness of cancer treatments in the NHS [1]. These datasets are collectively referred to in this study as England's cancer registry data for convenience and consistency.

England's cancer registry data is a potentially viable data source for estimating RWE of comparative effectiveness that could aid decision-making. However, data quality issues are a big concern for routinely collected data, as the primary purpose of data collection is patient care rather than clinical research. In contrast to experimental research studies, which employ specifically developed data collection forms to gather data at regular intervals and implement numerous data quality checks, routinely collected data are inputted into hospital systems with few quality checks by clinical staff. These data-collecting conditions can affect the data accuracy, completeness and consistency [22]. Therefore, evaluating the data quality of England’s cancer registry data for obtaining comparative effectiveness estimates is vital.

1.4. The Target Trial framework

The TT framework has been proposed by Miguel Hernan and James Robins (2016) as an approach for obtaining comparative effectiveness estimates using real-world data such as England’s cancer registry data [4, 23]. The primary sources of bias in RWE of comparative treatment effectiveness often stem from data inadequacy or poor quality, unsuitable or erroneous application of study design, substandard study conduct, or flawed statistical analysis. The TT framework incorporates a counterfactual framework and clinical trial principles to reduce these associated biases.

The underlying concept of the TT framework is to emulate the RCT that would have been conducted if it were feasible and ethical to carry out such a trial. This hypothetical trial is referred to as the "Target Trial". The explicit design of the TTs are an essential component of this framework, which enables the observational study to be carried out with a similar rigour and transparency as experimental studies.

All fundamental components of a clinical trial are incorporated into the design of the TTs. The key components covered in the TT framework are:

- Eligibility criteria
- Time zero/Baseline
- Treatment strategies
- Treatment group assignment
- Follow-up period
- Outcome(s) of interest
- Estimand(s) of interest
- Analysis plan

The observational data is then used to emulate the specified TT using a systematic process, which aims to reduce biases at multiple stages to obtain estimates comparable to RCT estimates.

1.5. Causal Inference methods

In the real world, treatment decisions are made due to factors that could also affect the outcome. When these factors are not adequately handled by study design or data analysis method, they can result in confounding by indication bias. However, reducing this confounding bias requires knowledge of the causal pathway of treatment choice and factors affecting the outcome. The choice of statistical methods depends on several factors, including the presumed causal treatment pathway.

Cancer treatment pathways are especially complex as treatments are provided over a long duration, resulting in time-varying treatments, as treatment decisions are made at multiple time points. Depending on individual response to the treatments, treatments are often modified, changed or discontinued. As treatment decisions are made at multiple time points, confounding bias can occur at multiple times points. It is imperative to consider the presence of time-dependent or time-varying confounding factors in the statistical analysis to reduce these biases.

A traditional statistical analysis method, such as a regression analysis, including or excluding time-dependent variables, can lead to biased estimates. Time-dependent confounding requires specialised statistical methods to reduce bias.

The causal inference framework is an area of statistics which aims to estimate causal effects from nonrandomised data. Within the causal inference paradigm, time-dependent confounders can be effectively addressed using approaches such as "g-methods" when adequate data are available.

1.6. Case studies selection

The three RCTs addressed in this protocol were selected in a systematic manner, taking into consideration their feasibility to be replicated using England's cancer registry data. Two of the three selected RCTs examined interventions for Non-Small Cell Lung Cancer (NSCLC), while the other RCT investigated treatments for Triple Negative Breast Cancer (TNBC). Three TTs will be designed based on these chosen RCTs and will be emulated using England's cancer registry data. Furthermore, four causal inference analysis methods have been selected following a systematic process, for the purpose of analysing the designated TTs, with the aim of increasing the likelihood of producing treatment effect estimates that are less biased. The three trial emulations will be deigned and conducted by Ms. Saleema Rex, a doctoral candidate under the supervision of Prof. Nicholas Latimer, Prof. Ron Akehurst, and Mr. Mike Bradburn.

2. Data management

England's cancer registry data can be linked to other routinely collected health datasets using pseudonymised codes at patient or tumour levels. The key datasets identified as necessary for this project work are:

- National Cancer Registrations dataset (NCRD) (audited and rapid registrations)
- Systematic Anti-Cancer Therapy (SACT) dataset
- Radiotherapy dataset (RTDS)
- Hospital Episode Statistics (HES) admitted care data
- HES outpatients' data

- HES accident and emergency data
- Cancer Waiting Time data

England's cancer registry data can be obtained directly from NHS England and other data access providers for research purposes. For this project, DATA-CAN, a Health Data Research Hub specifically dedicated to cancer-related data in the UK, was chosen as the most appropriate collaborator to facilitate data access to enable conduct of these case studies.

2.1. DATA-CAN: a Health Data Research Hub

DATA-CAN's objective is to enable accessibility of health data for cancer researchers and healthcare practitioners, aiming to improve cancer services and optimise patient outcomes. The fundamental objective of this project work is to improve cancer care by identifying whether England's cancer registry data can be used to obtain reliable estimates of comparative treatment effectiveness, which could help patients and other vital decision-makers. These shared goals enabled the collaboration with DATA-CAN to conduct the case studies described in this protocol.

2.2. NHS England's Secure Data Environment (SDE)

NHS England's Secure Data Environments (SDE), previously referred to as Trusted Research Environments (TREs), are designed to offer authorised researchers a secure and protected platform for conducting data analysis and research activities. DATA-CAN is a crucial collaborator in having a data-sharing agreement with NHS England to access England's cancer registry data and other important healthcare datasets via the SDE to improve patient care. A collaborative agreement has been made with DATA-CAN to facilitate this research work. As an integral component of the SDE, NHS England will review any output from the system to ensure it complies with the data security and data-sharing agreement before it can be downloaded for research reporting.

3. Target Trial Emulations

Three RCTs were identified as feasible to be replicated from England's cancer registry data using Hernan and Robins' TT framework [4]. These trials were used to design the TTs described below.

3.1. Target Trial 1: LUX-Lung 7 trial

The first TT is designed based on the LUX-Lung 7 trial. LUX-Lung 7 was an open-label, randomised, controlled, phase 2B trial comparing afatinib with gefitinib as a first-line treatment for patients with biomarker epidermal growth factor receptor (EGFR) mutation-positive non-small-cell lung cancer (NSCLC) [6]. The EGFR is a transmembrane receptor found on the epithelial cell surface. The most common mutations in EGFR are sensitive to EGFR tyrosine kinase inhibitors (TKIs). Osimertinib, erlotinib, gefitinib, afatinib and dacomitinib are medications used as EGFR TKIs [6]. Within these, afatinib and gefitinib were compared in this trial. Gefitinib is the first-generation TKI, and afatinib is a second-generation TKI; both are administrated orally [6].

Table 3-3 provides an overview of the characteristics of the original RCT and outlines the process in which this trial will be emulated from England's cancer registry data.

Table 3-1: Target Trial 1 study design

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England's cancer registry data
Patient eligibility inclusion criteria	Age 18 or over	Age 18 or over
	NSCLC diagnosis	NSCLC diagnosis identified using ICD code 'C34' and morphology codes (Full list: **Table 9-1** Lung cancer morphology codes listed in the appendix).
	EGFR mutation	Patients treated with gefitinib or afatinib (used as a proxy for EGFR mutation).
	Patients with stage IIIB or IV cancer	NSCLC stage IIIB or IV
	Eastern Cooperative Oncology Group (ECOG) Performance status of 0 or 1	ECOG performance status of 0 or 1

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England's cancer registry data
	At least one lung tumour lesion	Not identifiable from the data but the treatment received itself can be taken as evidence of having at least one lung tumour lesion.
Patient eligibility exclusion criteria	Surgical treatments in the four months before randomisation	Surgical treatments in the four months prior to and including day of study treatment initiation.
	Previously treated with Chemotherapy or other targeted therapies	Chemotherapy or anti-cancer treatments after NSCLC diagnosis but on or before study treatment initiation date.
	Active brain metastases	Diagnosis of brain tumour metastases will be identified using ICD code C79.31 (Secondary malignant neoplasm of brain) after NSCLC diagnosis but on or before study treatment initiation date. If treatment for brain metastases is ongoing at the time of baseline / time zero, then the tumour will be considered as active, and will be excluded.
	Previous or concomitant malignancies at other sites, except effectively treated non-melanoma skin cancers, carcinoma in situ of the cervix, ductal carcinoma in situ or effectively treated malignancy that has been in remission for more	It is not possible to determine whether a cancer has been effectively treated due to lack of data availability, therefore, all patients with a cancer diagnosis (excluding lung cancer) identified using ICD codes (C00-C96 malignant

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England's cancer registry data
	than 3 years and is considered to be cured in the opinion of investigator.	neoplasms but excluding C34 lung cancer) and C97 malignant neoplasms of independent primary multiple sites) prior to and including NCSLC diagnosis date will be excluded.
	Previously treated with an investigational drug within four weeks	Exclusion criteria not applicable for trial emulation as only patients with afatinib and gefitinib as first line therapy will be included.
	Leptomeningeal disease	Diagnosis of Leptomeningeal disease identified using ICD 10 code C79.3 (Secondary malignant neoplasm of brain and cerebral meninges) and C79.4 (Secondary malignant neoplasm of other and unspecified parts of nervous system) in the two years prior to and including the date of treatment initiation date [24].
	Pre-existing interstitial lung disease	Diagnosis of interstitial lung disease identified using ICD 10 code J84 (Other interstitial pulmonary diseases) in the two years prior to and including treatment initiation date.
	Any history or presence of poorly controlled gastrointestinal disorders	It is not possible to determine whether a gastrointestinal disorder is poorly controlled, therefore,

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England’s cancer registry data
		patients with a diagnosis of gastrointestinal disorders identified using ICD 10 codes K20-K31 (Diseases of oesophagus, stomach and duodenum) in the two years prior to and including treatment initiation date will be excluded.
	Clinically relevant cardiovascular abnormalities	Diagnosis of myocardial infarction, congestive heart failure, and peripheral vascular disease identified using ICD codes (ICD10 codes listed in appendix 9.1 Codes for identification) in the two years prior to and including treatment initiation date.
	Cardiac left ventricular function with resting ejection fraction of less than the institutional lower limit of normal	It is not possible to identify patients who had cardiac left ventricular function with resting ejection fraction of less than the institutional lower limit of normal. However, patients with left ventricular failure could be identified using ICD 10 code I50.1. Therefore, as a pragmatic approach, patients who had a diagnosis of cardiac left ventricular function failure using ICD 10 code I50.1 (Left ventricular failure) in the

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England's cancer registry data
		two years prior to and including treatment initiation date.
	Active hepatitis B infection, active hepatitis C infection, or known HIV infection	Diagnosis of hepatitis B (identified using ICD10 codes B16 (Acute hepatitis B, B17.0 (Acute delta-super infection of hepatitis B carrier), B18.0 (Chronic viral hepatitis B with delta-agent), B18.1 (Chronic viral hepatitis B without delta-agent), and B19.1 (Unspecified viral hepatitis B)), hepatitis C (identified using ICD10 codes B17.1 (Acute hepatitis C), B18.2 (Chronic viral hepatitis C), and B19.2 (Unspecified viral hepatitis C)) or HIV (identified using ICD10 codes B20-B22 and B24) in the two years prior to and including treatment initiation date.
Time zero/Baseline	Randomisation date	Date of study treatment initiation
Treatment strategies	**Afatinib arm** Afatinib 40 mg orally once daily Dose escalation to 50 mg was allowed after four weeks and can be reduced to 20 mg. A treatment gap of 14 days was allowed. **Gefitinib arm** Gefitinib daily dose of 250 mg	**Afatinib arm** Patients treated with afatinib as first-line therapy for NSCLC. **Gefitinib arm** Patients treated with gefitinib as first-line therapy for NSCLC.

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England’s cancer registry data
	Modifications in the administration of gefitinib were allowed. A treatment gap of 14 days was allowed. **Treatment changes:** Patient were allowed to switch to other treatment upon treatment failure. A full list of treatments patients received are in appendix section 9.1 Codes for identification)	**Per-protocol treatment strategy:** Patients who received treatments following the first line afatinib and gefitinib that were also received in LUX-Lung 7 were considered to be consistent with a per-protocol treatment strategy.
Assignment procedures	Patients were randomised to afatinib or gefitinib on a 1:1 ratio, and EGFR mutation type exon 19 deletions vs Leu858Arg, brain metastases present vs absent were used as stratification factors.	To emulate the random treatment allocation, measured variables that could reduce confounding will be included in the analysis.
Follow-up period	The study’s primary analysis was planned at a follow-up period of at least 32 months for patients still alive. However, overall survival estimates were reported for a maximum follow-up of 50 months.	Patients whose time zero/baseline was at least 32 months and up to 50 months before the cut-off date of England’s cancer registry data availability will be included.
Outcome	Overall survival is measured as the time from randomisation until death from any cause.	Overall survival is measured as the time from date of study treatment initiation until death from any cause.
Causal contrasts of interest	Intention-to-treat (ITT) effect: the effect of the intervention based on assigned treatment strategies	Analogues of the ITT effect and PP effect will be estimated. (Further information provided below)

Key component	LUX-Lung 7 trial	Target Trial 1: Trial emulation from England's cancer registry data
	irrespective of treatment compliance. Per-protocol (PP) effect: not reported.	
Analysis plan	Intention-to-treat (ITT) analysis: A Cox proportional hazards model was used to calculate HRs and 95% CIs for overall survival.	An analogue of ITT and Per-protocol analysis will be carried out using Inverse Probability Weighting (IPW), and G-formula. Sensitivity analysis to evaluate the impact of unmeasured confounding will be carried out using E-value. Depending on resource availability and time availability, primary outcome analysis will be repeated using G-estimation and Targeted Maximum Likelihood Estimation (TMLE).

Eligibility criteria

All lung patients aged 18 or over will be identified from the NCRD using the ICD diagnosis code 'C34' and age information. The morphology codes associated with NSCLC (a complete list of morphology codes can be found in the appendix section 9.1 Codes for identification) will be used to identify NSCLC patients. These NSCLC patient records will be linked to SACT data to identify patients who received gefitinib or afatinib as first-line therapy. SACT data is available only for patients diagnosed from April 2012; therefore, all patients with diagnosis date prior to April 2012 will be excluded.

NCRD linked to the SACT dataset will be further linked to HES datasets, the RTDS and Cancer Waiting Time data using pseudo identifiers to incorporate additional variables needed to evaluate further eligibility conditions detailed in **Table 3-1**: Target Trial 1 design, and confounder adjustment. Patients who fail one or more of these eligibility criteria will be marked as not eligible.

Two sets of analysis datasets will be created, one aiming to mimic the LUX-Lung 7 trial participants and another representing the real-world population.

Trial matching population dataset

The trial matching population dataset will include NSCLC patients aged 18 or over, who received gefitinib or afatinib as first-line therapy; and satisfy further eligibility conditions detailed in **Table 3-1**.

Real-world population dataset

In addition to the benchmarking analysis, an analysis dataset that includes all NSCLC lung cancer patients aged 18 or over, who received gefitinib or afatinib as first-line therapy, will be created. The results from this real-world population will help compare the comparative treatment effectiveness estimates derived from wider patient population and RCT eligible population.

Time Zero/ Baseline

In the original trial, the baseline is the randomisation date, but in the emulated trial, the baseline is the study treatment initiation date. It is hypothesised that most patients will start their treatment soon after the diagnosis date as the treatments compared are offered as first-line therapy. The delay between the cancer diagnosis and study treatment start date will be included in the analysis as it could be a potential confounder.

Treatment strategy and assignment procedures

In the UK, NICE recommended gefitinib and afatinib as the first-line treatment for people with locally advanced or metastatic NSCLC if they tested positive for the EGFR tyrosine kinase (EGFR-TK) mutation in July 2010 and April 2014, respectively [14]. As the year of NICE recommendation is different for both these treatments, and therefore it is likely that patients who received these treatments in the NHS will differ by calendar year, the year of diagnosis will be included in the analysis to reduce the impact of the period effect.

Follow-up period and outcome selection

The original study reported overall survival outcomes at 32 months and up to a maximum follow-up of 50 months. In order to mimic the original trial's follow-up, patients whose time zero/Baseline was at least 32 months and up to 50 months before the cut-off date of England's cancer registry data availability will be included for analysis.

Time fixed and time variable confounder selection

Baseline time fixed factors: sex, age, ethnicity, age at diagnosis, year of diagnosis, time to treatment from cancer diagnosis, geographical region, Charlson co-morbidity index (derived using HES hospital admissions records based on the methodology published by Quan et. al. The full list of relevant ICD10 codes are listed in **Table 9-6** in the appendix) with a two years look back period starting from three months prior to cancer diagnosis [25], route of diagnosis, cancer morphology, cancer stage, history of brain metastasis, performance status at baseline, body mass index (BMI) at baseline, number of hospital admission days, number of outpatient visits, and number of A&E attendances in the last two years.

Time-dependent post-baseline confounders: Performance status, BMI, treatment line, rate of hospital admission, rate of outpatient visits, rate of A&E attendances, and hospital admission indicator will be included in the analysis to reduce bias.

The above listed baseline and post baseline time-dependent confounding factors have been identified as relevant through discussions with clinical experts and will be included in the analysis irrespective of whether they are statistically associated with the treatment or outcome.

Causal contrasts of interest(s)

Two causal contrasts are of interest: 1) an analogue of an intention to treat analysis, and 2) an analogue of a per-protocol analysis. These causal effects of interests will be estimated both for the real-world population and for the trial matching population.

Analogue of intention to treat population dataset

All patients who initiated gefitinib as first-line therapy will be assigned to the gefitinib group, and patients who initiated afatinib will be assigned to the afatinib group. All subsequent treatment changes will be ignored. Patients will only be censored at the end of the study period.

Analogue of per-protocol population dataset

To determine the causal effect of adhering to the treatment regimen closely resembling that of trial participants, patients initiating gefitinib as first-line therapy will be categorised into the gefitinib group, while patients initiating afatinib will be categorised into the afatinib group. Furthermore, patients who deviate from treatments that were permitted in the LUX-Lung 7 trial after study treatment initiation will be censored at the point of deviation. A full list of treatments that were allowed in LUX-Lung 7 trial are listed in appendix section 9.1 Codes for identification.

Analysis Plan

The average treatment effect from the above mentioned analyses population will be estimated using multiple analysis methods. These methods are described in more detail in section 44. Statistical Analysis Plan.

3.2. Target Trial 2: KEYNOTE-024 trial

KEYNOTE-024 was an international, open-label, phase 3 trial comparing pembrolizumab with platinum-based chemotherapy for treating patients with programmed cell death ligand 1 (PD-L1) positive metastatic NSCLC as first-line therapy [26]. PD-L1 is a coregulatory molecule that acts as an inhibitor for T-cell-mediated cell deaths, which allows cancer cells to thrive. Monoclonal antibodies such as nivolumab, pembrolizumab, atezolizumab and durvalumab can treat patients with PD-L1-positive NSCLC. The KEYNOTE-024 trial compared the effect of pembrolizumab with chemotherapy.

Table 3-2 provides an overview of the characteristics of the original RCT and outlines the process in which this trial will be emulated from England’s cancer registry data.

Table 3-2: Target Trial 2 study design

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England's cancer registry data
Patient eligibility inclusion criteria	Aged 18 and over	Aged 18 and over
	Patients with untreated metastatic NSCLC	NSCLC diagnosis identified using ICD code 'C34' and morphology codes (Full list: **Table 9-1** Lung cancer morphology codes listed in the appendix).
	ECOG performance score of 0 or 1	ECOG performance score of 0 or 1
	No previous history of systemic therapy for metastatic disease	No previous history of systemic therapy (identified from the SACT dataset) given after NSCLC diagnosis but on or before study treatment initiation date.
	Histologically or cytologically confirmed stage IV NSCLC with no sensitising EGFR mutations or ALK translocations	People with cancer stage IV. EGFR mutations or ALK translocations status cannot be determined from data.
	Life expectancy of at least three months	Life expectancy cannot be determined from the data. However, treatments compared are unlikely to be given to patients with less than three months life expectancy.
	PD-L1 proportion score of 50% or over	PD-L1 proportion cannot be determined from the data.
Patient eligibility exclusion criteria	Patients receiving systemic glucocorticoids or other immunosuppressive treatments were excluded.	It will not be possible to identify patients who were given glucocorticoids or other

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England’s cancer registry data
		immunosuppressive treatments from the available data.
	Untreated brain metastases	Diagnosis of brain tumour metastases will be identified using ICD code C79.31 (Secondary malignant neoplasm of brain) after NSCLC diagnosis but on or before study treatment initiation date. If treatment for brain metastases is ongoing at the time of baseline / time zero, then the tumour will be considered as active, and will be excluded.
	Active autoimmune disease received treatment in the previous two years	Diagnosis of autoimmune disease identified using ICD 10 codes [27] listed in **Table 9-2** found in the appendix) in the 2 years prior to and including day of study treatment initiation. (Note: It is not possible to determine whether any treatment was offered for autoimmune disease, so any patients with a history of autoimmune disease in the 2 years prior to and including day of study treatment initiation are excluded).

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England's cancer registry data
	Active interstitial lung disease	Diagnosis of interstitial lung disease identified using ICD 10 code J84 (Other interstitial pulmonary diseases) in the two years prior to and including treatment initiation date.
	History of pneumonitis treated with glucocorticoids	Diagnosis of pneumonitis identified using ICD-10 codes recommended by Neibart, Shane S., et al. 2021 [28] (listed in *Table* **9-3** found in appendix 9.1 Codes for identification) in the 2 years prior to and including day of study treatment initiation.
Time zero/Baseline	Randomisation date	Date of study treatment initiation.
Treatment strategies	**Pembrolizumab arm** Pembrolizumab was administered at a fixed dose of 200-mg every three weeks for 35 cycles. **Chemotherapy arm** Platinum-based chemotherapy (carboplatin plus pemetrexed, cisplatin plus pemetrexed, carboplatin plus gemcitabine, cisplatin plus gemcitabine, or	**Pembrolizumab arm** Patients who received pembrolizumab treatment of any dosage **Chemotherapy arm** Patients who received chemotherapy treatment (carboplatin plus pemetrexed, cisplatin plus pemetrexed, carboplatin plus gemcitabine, cisplatin plus gemcitabine, or

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England's cancer registry data
	carboplatin plus paclitaxel) for 4 to 6 cycles. Pemetrexed-based chemotherapy was used for non-squamous tumours, and pemetrexed was continued as maintenance. **Treatment changes:** Chemotherapy arm patients were allowed to switch from chemotherapy to pembrolizumab after disease progression.	carboplatin plus paclitaxel) of any dosage. Patients who met the eligibility criteria but did not initiate pembrolizumab or chemotherapy treatment will be excluded from the analysis. **Per-protocol treatment strategy:** Patients who received chemotherapy as first line treatment and pembrolizumab as second line treatment will be considered to be consistent with a per-protocol treatment strategy.
Assignment procedures	Patients were randomly assigned, in a 1:1 ratio, to receive treatment with either pembrolizumab or the investigator's choice of one of the following five platinum-based chemotherapy regimens. Randomisation was stratified by ECOG performance-status score (0 vs 1), tumour histologic type (squamous vs non-squamous), and region of enrolment (East Asia vs non–East Asia) and did not include any provisions regarding equal	To emulate the random treatment allocation, measured variables that could reduce confounding will be included in the analysis.

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England's cancer registry data
	distribution of enrolment across participating sites or stratification by the site.	
Follow-up period	The overall survival estimates were reported with a follow-up of 20 months and an updated OS estimates reported with five year follow-up.	Patients whose time zero/Baseline was at least 20 months and up to a maximum of 60 months before the cut-off date of England's cancer registry data availability will be included.
Outcome	Overall survival is measured as the time from randomisation until death from any cause.	Overall survival is measured as the time from date of study treatment initiation until death from any cause.
Causal contrasts of interest	Efficacy analyses and adverse events were reported in the intention-to-treat population. The control patients were allowed to crossover to the pembrolizumab after disease progression. An updated analysis results adjusting for control participants' crossing over to pembrolizumab using two-stage method, rank-preserving structural failure time model and inverse probability censoring weights approach was also reported [29].	Analogues of the ITT effect, PP effect and an additional analysis adjusting for crossover effect will be estimated. (Further information provided below)

Key component	KEYNOTE-024 trial	Target Trial 2: Trial emulation from England's cancer registry data
Analysis plan	The Kaplan–Meier method was used to estimate overall survival curves. Data for patients who were alive or lost to follow-up were censored at the time of the last contact. Between-group differences in overall survival were assessed using a stratified log-rank test. Hazard ratios and associated 95% confidence intervals were assessed using a stratified Cox proportional-hazards model with Efron's method of handling ties. The same stratification factors used for randomisation were applied to the stratified log-rank and Cox models.	An analogue of ITT, PP and crossover adjusted analysis will be carried out using IPW, and G-formula. Sensitivity analysis to evaluate the impact of unmeasured confounding will be carried out using E-value. Depending on resource availability and time availability, primary outcome analysis will be repeated using G-estimation and TMLE.

Eligibility criteria

All lung patients aged 18 and over will be identified from the NCRD using the ICD10 diagnosis code 'C34' and age information. The morphology codes associated with NSCLC (a complete list of morphology codes can be found in the appendix) will be used to identify NSCLC patients. These NSCLC patient records will be linked to SACT data to identify patients who received either pembrolizumab or a platinum-based chemotherapy as their first-line therapy. NCRD linked to SACT dataset will be further linked to HES datasets, RTDS and Cancer Waiting Time data using pseudo identifiers to incorporate additional variables needed to evaluate further eligibility conditions detailed in **Table 3-2**, and confounder adjustment. Patients who fail one or more of these eligibility criteria will be marked as not eligible.

Two sets of analysis datasets will be created, one aiming to mimic the KEYNOTE-24 trial participants and another representing the real-world population.

Trial matching population dataset

The trial matching population dataset will include NSCLC lung cancer patients aged 18 or over, who received pembrolizumab or a platinum-based chemotherapy as first-line therapy; and satisfy further eligibility conditions detailed in **Table 3-2**.

Real-world population dataset

In addition to benchmarking analysis, an analysis dataset that includes all NSCLC lung cancer patients aged 18 or over, who received gefitinib or afatinib as first-line therapy, will be created. The results from this real-world population will help compare the comparative treatment effectiveness estimates derived from wider patient population and RCT eligible population.

Time zero/Baseline

In the original trial, the baseline is the randomisation date, but in the emulated trial, the baseline is the treatment initiation date. It is hypothesised that most patients will start their treatment soon after the diagnosis date as the treatments compared are offered as first-line therapy. The delay between the cancer diagnosis and treatment start date will be included in the analysis as it could be a potential confounder.

Treatment strategy and assignment procedures

In England, NICE recommends pembrolizumab for treating untreated PD-L1-positive metastatic NSCLC in adults whose tumours express PD-L1 (with at least a 50% tumour proportion score) and have no EGFR- or ALK-positive mutation (Ref: TA531, July 2018). The control group received chemotherapy in the KEYNOTE-024 trial; since chemotherapy was not a recommended for first-line treatment at the time pembrolizumab was approved, patients who received chemotherapy as first-line therapy are likely to be prior to July 2018. The analysis will include the year of diagnosis to reduce the potential impact of this period effect due to using historical control patients.

Follow-up period and outcome selection

The original study reported an overall survival outcome over 20 months. Additionally, overall survival estimates at five year follow-up was reported. In order to mimic the original trial's follow-up results, patients whose time zero/Baseline was at least 20 months and up to a maximum of 60 months before the cut-off date of England's Cancer Registry data availability will be included for analysis.

Time fixed and time variable confounder selection

Baseline time fixed factors: sex, age, ethnicity, age at diagnosis, year of diagnosis, time to treatment from cancer diagnosis, geographical region, Charlson co-morbidity (derived using HES hospital admissions records based on the methodology published by Quan et. al. The full list of relevant ICD10 codes are listed in **Table 9-6** in the appendix) with a two years look back period starting from three months prior to cancer diagnosis [25], route of diagnosis, cancer morphology, cancer stage, history of brain metastasis, performance status at baseline, BMI at baseline, number of hospital admission days, number of outpatient visits, and number of A&E attendances in the last two years.

Time-dependent post-baseline confounders: Performance status, BMI, treatment line, rate of hospital admission days, rate of outpatient visits, rate of A&E attendances, hospital admission indicator) will be included in the analysis to reduce bias.

The above listed baseline and post baseline time-dependent confounding factors have been identified as relevant through discussions with clinical experts and will be included in the analysis irrespective of whether they are statistically associated with the treatment or outcome.

Causal contrasts of interest(s)

Three causal contrasts are of interest: 1) an analogue of intention to treat analysis, 2) an analogue of per-protocol analysis, and 3) an analogue of crossover adjusted analysis. These causal effects of interests will be estimated by creating and analysing three datasets each for real-world population and trial matching population.

Analogue of intention to treat population

All patients who initiated pembrolizumab as first-line therapy will be assigned to the pembrolizumab group, and patients who initiated a platinum-based chemotherapy will be assigned to the chemotherapy group. All subsequent treatment changes will be ignored. Patients will only be censored at the end of the study period.

Analogue of per-protocol population

To determine the causal effect of adhering to the treatment regimen closely resembling that of trial participants, patients initiating pembrolizumab as first-line therapy will be categorised into the pembrolizumab group, while patients initiating a platinum-based chemotherapy will be categorised into the chemotherapy group. The KEYNOTE-024 trial allowed chemotherapy patients to receive pembrolizumab after disease progression. In order to match the trial participants, the designed TT will allow patients in chemotherapy group receiving pembrolizumab as second line therapy, but censor patients who received other treatments.

Analogue of crossover adjusted population

The KEYNOTE-24 trial reported results of an updated analysis adjusting for crossover effect due to control participants' crossing over to intervention treatment i.e., to pembrolizumab upon disease progression. To estimate the causal effect analogue to the trial's crossover adjusted analysis estimates, patients initiating pembrolizumab as first-line therapy will be categorised into the pembrolizumab group, while patients initiating a platinum-based chemotherapy will be categorised into the chemotherapy group. To match the crossover adjusted population, patients in chemotherapy group receiving pembrolizumab as second line therapy will be censored at the point of starting pembrolizumab [29].

Analysis Plan

The average treatment effect from the above mentioned analyses population will be estimated using multiple analysis methods. These methods are described in more detail in section 4 Statistical Analysis Plan.

3.3. Target Trial 3: TNT trial

TNT was a phase III RCT comparing carboplatin with docetaxel for treating people with triple-negative breast cancer [8]. Both the compared treatments are types of chemotherapy.

Carboplatin is platinum-based, and docetaxel is a taxane. The majority of patients had surgical treatment for breast cancer at baseline.

Table 3-3 provides an overview of the characteristics of the original RCT and outlines the process in which this trial will be emulated from England's cancer registry data.

Table 3-3: Target trial 3 study design

Key component	**TNT trial**	**Target Trial 3: Trial emulation from England's cancer registry data**
Patient eligibility inclusion criteria	Histologically confirmed Estrogen Receptor (ER) negative, Progesterone Receptor (PgR) negative, Human Epidermal growth factor Receptor 2 (HER2) negative, primary invasive breast cancer (or) PgR unknown but ER negative and HER2 negative, and otherwise eligible (or) Confirmed BReast CAncer gene *(BRCA) 1* or *BRCA2* mutation carrier, with any ER, PgR and HER2 status	Patients with breast cancer identified using ICD10 code 'C50' and, negative Estrogen Receptor status, negative Progesterone Receptor status and negative or unknown HER2 status. *BRCA1* or *BRCA2* mutation status cannot be determined from the data.
	Aged 18 and over	Aged 18 and over
	Sex: female	Sex: female
	Measurable confirmed metastatic or recurrent locally advanced disease unsuitable for local therapy but suitable for taxane chemotherapy	Positive metastatic status or locally advanced disease identified using TNM staging information. Note: It is not possible to determine whether the disease was unsuitable for local therapy but suitable for taxane chemotherapy.

Key component	TNT trial	Target Trial 3: Trial emulation from England's cancer registry data
	Patients with stable, treated brain metastases will be eligible, providing informed consent can be given and that other sites of measurable disease are present.	Diagnosis of brain tumour metastases will be identified using ICD code C79.31 (Secondary malignant neoplasm of brain) after NSCLC diagnosis but on or before study treatment initiation date. If treatment for brain metastases is ongoing at the time of baseline / time zero, then the tumour will be considered as untreated, and will be excluded.
	Patients with bone metastases receiving bisphosphonates for palliation will be eligible, providing other sites of measurable disease are present.	It will not be possible to identify patients with bone metastases receiving bisphosphonates, therefore, this eligibility criteria will be ignored.
	ECOG performance Status 0, 1 or 2	ECOG performance Status 0, 1 or 2
	Adequate haematology biochemical indices.	It will not be possible to determine the adequacy of haematology biochemical indices from the data.
	Adequate renal function	It will not be possible to determine the adequacy of renal function from the data.
Patient eligibility exclusion criteria	Patients unfit for chemotherapy or those with neuropathy >grade 1 (sensory or motor). Known allergy to platinum compounds or mannitol.	It will not be possible to determine the patient's fitness to receive the interventions compared to the data. However, receipt of the study treatments indicates that the

Key component	TNT trial	Target Trial 3: Trial emulation from England's cancer registry data
	Known sensitivity to taxanes. Patients with inoperable locally advanced disease suitable for local radiotherapy or an anthracycline-containing regimen.	patients were considered fit to receive them.
	Previous chemotherapy for metastatic disease other than anthracycline as in inclusion criteria above.	Patients who were given chemotherapy after breast cancer diagnosis date but on or before study treatment initiation date.
	Previous exposure to a taxane in adjuvant chemotherapy within 12 months of trial entry.	Patients who were given taxane in the 12 months prior to or on study treatment initiation date.
	Previous treatment with a taxane for a recurrent locally advanced disease, which was not completely excised.	Diagnosis of more than one locally advanced disease prior to breast cancer diagnosis, which has been treated with a taxane in 1 year prior to or on study treatment initiation date.
	Previous treatment with a platinum chemotherapy drug	Patients who were given a platinum chemotherapy drug after breast cancer diagnosis date but on or before study treatment initiation date.
	Patients with a life expectancy of fewer than three months	Life expectancy cannot be determined from the data. However, only people with at least three months of life expectancy

Key component	TNT trial	Target Trial 3: Trial emulation from England's cancer registry data
		will likely be offered the treatments compared.
	Previous malignancies other than adequately treated in situ carcinoma of the uterine cervix or basal or squamous cell carcinoma of the skin, unless there has been a disease-free interval of at least ten years.	It is not possible to determine whether a cancer has been effectively treated due to lack of data availability, therefore, all patients with a cancer diagnosis (excluding breast cancer) identified using ICD codes (C00-C96 Malignant neoplasms but excluding C50) and C97 malignant neoplasms of independent primary multiple sites) prior to and including breast cancer diagnosis date will be excluded.
	Previous or synchronous second breast cancer (unless also confirmed ER-, PgR-/unknown and HER2-)	Patients with a diagnosis of another breast cancer identified using ICD10 code 'C50' but not triple negative i.e, negative ER status, negative PgR status and negative or unknown HER2 status, prior to or on study treatment initiation date.
	Patients with bone-limiting disease	Patients diagnosed with bone-limiting disease identified using ICD 10 codes M80-M85 (Disorders of bone density and structure) in two

Key component	TNT trial	Target Trial 3: Trial emulation from England’s cancer registry data
		years prior to or on study treatment initiation date.
	Other uncontrolled severe medical conditions or concurrent medical illnesses are likely to compromise life expectancy or the completion of trial therapy.	It is not possible to determine this from the data.
	Pregnant, lactating or potentially childbearing women not using adequate contraception.	Pregnant, lactating or potentially childbearing women may be given treatment drugs in routine practice and these factors are not considered clinically prognostic so will not be excluded. In addition, it will not be possible to determine the contraception use from the data.
Time zero/Baseline	Randomisation date	Date of study treatment initiation.
Treatment strategies	**Carboplatin** Dose depending on patient factors every three weeks for six cycles (18 weeks) 100mg/m2 every three weeks for six cycles (18 weeks) **Docetaxel** 100mg/m2 every three weeks for six cycles (18 weeks) **Treatment changes:**	**Carboplatin** Patients who initiated carboplatin treatment of any dosage **Docetaxel** Patients who initiated docetaxel treatment of any dosage Patients who met the eligibility criteria but did not initiate carboplatin or docetaxel are excluded from the analysis.

Key component	TNT trial	Target Trial 3: Trial emulation from England’s cancer registry data
	Patients were allowed to crossover i.e., patients randomised to carboplatin were offered docetaxel and patients randomised to docetaxel were offered carboplatin at the time of treatment failure.	**Per-protocol treatment strategy:** Patients were allowed to crossover i.e., patients who initiated carboplatin but later changed to docetaxel and patients initiated docetaxel but later changed to carboplatin were considered as being compliant to per-protocol treatment strategy.
Assignment procedures	Patients allocated to carboplatin or docetaxel (1:1 ratio) utilising a computerised minimisation algorithm with a random element. Balancing factors were centre, previous adjuvant taxane chemotherapy, presence of liver or lung metastasis, performance status (0/1 vs 2) and recurrent locally advanced vs metastatic carcinoma.	To emulate the random treatment allocation, measured variables that could reduce confounding will be included in the analysis.
Follow-up period	The study’s primary analysis was planned at a follow-up period of at least 15 months for patients still alive.	Patients whose study treatment initiation date was at least 15 months before the cut-off date of England’s cancer registry data availability will be included.
Outcome	Overall survival was measured as the time from randomisation until death from any cause.	Overall survival was measured as the time from study treatment

Key component	TNT trial	Target Trial 3: Trial emulation from England’s cancer registry data
		initiation date until death from any cause.
Causal contrasts of interest	Efficacy analyses were done in the intention-to-treat population, and safety analyses were done in patients who received at least one dose of the study drug. Patients were offered alternative treatment upon progression or discontinuation due to toxicity. The per-protocol effect was not published, but that study reported no evidence of a difference for crossover treatments.	Analogues of the ITT effect and PP effect will be estimated. (Further information provided below)
Analysis plan	Survival endpoints were displayed using Kaplan Meier plots, and survival analysis modelling utilised restricted mean survival methodology given that the proportionality of hazards assumption required for Cox survival analysis did not hold.	An analogue of ITT and Per-protocol analysis will be carried out using IPW, and G-formula. Sensitivity analysis to evaluate the impact of unmeasured confounding will be carried out using E-value. Depending on resource availability and time availability, primary outcome analysis will be repeated using G-estimation and TMLE.

Eligibility criteria

All NSCLC patients aged 18 and over will be identified from the NCRD using diagnosis ICD10 code C50 and progesterone, estrogen and HER2 statuses. These breast cancer patient records will be linked to SACT to identify patients who received carboplatin or docetaxel as first-line therapy. NCRD linked to the SACT dataset will be further linked to HES datasets, RTDS and Cancer Waiting Time data using pseudo identifiers to incorporate additional variables needed to evaluate further eligibility conditions detailed in **Table 3-3** Target Trial 3 design, and confounder adjustment. Patients who fail one or more of these eligibility criteria will be marked as not eligible.

Two sets of analysis datasets will be created, one aiming to mimic the TNT trial participants and another representing the real-world population.

Trial matching population dataset

The trial matching population dataset will include patients with a diagnosis of breast cancer identified using ICD10 code 'C50' and, negative estrogen receptor status, negative progesterone receptor status and negative or unknown HER2 status; received carboplatin or docetaxel as first-line therapy; and satisfy further eligibility conditions detailed in **Table 3-3**.

Real-world population dataset

The real-world population dataset will include all patients with a diagnosis of breast cancer identified using ICD10 code 'C50' and, negative estrogen receptor status, negative progesterone receptor status and negative or unknown HER2 status; and received carboplatin or docetaxel as first-line therapy

Time Zero/ Baseline

In the original trial, the baseline is the randomisation date, but in the emulated trial, the baseline is the study treatment initiation date. It is hypothesised that most patients will start their treatment soon after the diagnosis date as the treatments compared are offered as first-line therapy. The delay between the cancer diagnosis and study treatment start date will be included in the analysis as it could be a potential confounder.

Treatment strategy and assignment procedures

NICE recommended atezolizumab with nab-paclitaxel for untreated PD-L1-positive, locally advanced or metastatic, triple-negative breast cancer in July 2020 (Ref: TA639). Chemotherapy is also a treatment option for triple-negative breast cancer patients in England [14]. Therefore, intervention and control patients are likely to have been diagnosed with Triple Negative Breast Cancer before July 2020 or switched to atezolizumab with nab-paclitaxel upon policy change. These potential period and treatment crossover effects will be accounted for in the analysis.

Follow-up period and outcome selection

The original study reported an overall survival outcome at 15 months. In order to mimic the original trial's follow-up, patients whose date of study treatment initiation was at least 15 months before the cut-off date of England's cancer registry data availability will be included for analysis.

Time fixed and time variable confounder selection

> **Baseline time fixed factors:** age, ethnicity, age at diagnosis, year of diagnosis, time to treatment from cancer diagnosis, geographical region, Charlson co-morbidity index (derived using HES hospital admissions records based on the methodology published by Quan et. al. The full list of relevant ICD10 codes are listed in **Table 9-6** in the appendix) with a two years look back period starting from three months prior to cancer diagnosis [25], route of diagnosis, cancer morphology, cancer stage, history of brain metastasis, performance status at baseline, BMI at baseline, number of hospital admission days, number of outpatient visits, and number of A&E attendances in the last two years.
>
> **Time-dependent post-baseline confounders:** Performance status, BMI, treatment line, rate of hospital admission days, rate of outpatient visits, rate of A&E attendances, hospital admission indicator) will be included in the analysis to reduce bias.

The above listed baseline and post baseline time-dependent confounding factors have been identified as relevant through discussions with clinical experts and will be included in the analysis irrespective of whether they are statistically associated with the treatment or outcome.

Causal contrasts of interest(s)

Two causal contrasts are of interest: 1) an analogue of intention to treat analysis, and 2) an analogue of per-protocol analysis. These causal effects of interests will be estimated by creating and analysing two datasets each for real-world population and trial matching population.

Analogue of intention to treat population

All patients who initiated carboplatin as first-line therapy will be assigned to the carboplatin group, and patients who initiated docetaxel will be assigned to the docetaxel group. All subsequent treatment changes will be ignored. Patients will only be censored at the end of the study period.

Analogue of per-protocol population

To determine the causal effect of adhering to the treatment regimen closely resembling that of trial participants, patients initiating carboplatin as first-line therapy will be categorised into the carboplatin group, while patients initiating docetaxel will be categorised into the docetaxel group. The TNT trial allowed cross over between groups at disease progression (and in some cases pre-progression). In order to match the trial participants, the designed TT will allow patients in carboplatin group receiving docetaxel as second line therapy, and patients in docetaxel group receiving carboplatin as second line therapy but patients who received treatments that were not received in the TNT trial will be censored.

Analysis Plan

The average treatment effect from the above mentioned analyses population will be estimated using multiple analysis methods. These methods are described in more detail in next section 4 Statistical Analysis Plan.

4. Statistical Analysis Plan

4.1. General considerations

All analysis results will report the point estimates and their corresponding 95% confidence interval (CI). When reporting counts or percentages of patients by groups, if the number of patients in a particular group is considered small (e.g., less than 7), they will be masked to protect the patients' privacy.

4.2. Missing data

Missing post-baseline factors will be imputed by using the last observation carried forward approach. Patients with missing baseline variables that are continuous will be excluded from the analysis. For missing values in baseline categorical variables, a missing category will be created to indicate whether the value is missing, and will be included in the analysis [30].

Directed Acyclic Graphs (DAGs)

A simplified Directed Acyclic Graph (DAG) will be reported to represent the causal assumptions for each of the designed TTs (reference: Appendix **Figure 1**).

Positivity assumption checks

Overlap of patient characteristics considered important to check the conditional exchangeability assumption will be evaluated by graphically plotting the distribution of each of these characteristics by treatment group (e.g., age, weight). In addition, propensity scores calculated at baseline will be plotted by treatment groups to evaluate the overall overlap of propensity scores at baseline (reference: Appendix **Figure 2**).

Sample size

Original RCT sample sizes are taken as the required sample sizes for the designed TTs. If the patient population available for the TTs is either smaller or larger than the number of patients included in the original RCT, the analyses of the TTs will be conducted using all available patient data, but potential impact of changes in sample size will be considered when comparing the results.

Descriptive statistics

In the descriptive statistics, all continuous measures will be reported as means, standard deviations, and medians, interquartile ranges, minimum and maximum. Categorical data will be reported as counts and percentages.

Data flow diagram

For each TT study, a data flow diagram will be used to display the flow of participants. This diagram will report the number of patients evaluated for eligibility, the number of patients eligible to be included, the number in each treatment arm and the number included in the final analysis. The template data flow diagram is in the appendix 9.2 Data flow template.

4.3. Summary statistics

The completeness of key data items used for eligibility assessment and potential confounders will be reported for each TT study (reference Appendix: **Table 9-7**). The summary statistics of key characteristics of eligible patients and those who are not eligible will be reported (reference Appendix **Table 9-8**). Comprehensive summary statistics of important baseline characteristics of patients will be reported by treatment arm and overall (reference Appendix: **Table 9-9**). A summary of the treatment cycles received by patients and treatment changes will be reported (reference Appendix: **Table 9-10**). The duration of follow-up between the baseline and end of the study will be reported by treatment group and overall (reference Appendix: **Table 9-11**).

4.4. Impact of COVID-19 pandemic

The Covid-19 pandemic has brought changes to cancer care provided within the NHS in the UK. It has also changed the mortality risk of people with cancer due to the direct risk of the Covid-19 disease. An additional exploratory analysis to evaluate the impact of Covid-19 will be carried out on all NSCLC patients and triple negative breast cancer patients diagnosed from 2015 onwards.

4.5. Overall survival outcome

The primary outcome is overall survival. Kaplan-Meier survival curves will be used to illustrate the survival analysis data.

4.6. Primary outcome analysis

The primary outcome of overall survival will be analysed using four analysis methods described below.

Inverse probability weighting analysis

Inverse probability weighting will be used to estimate average treatment effects from all the analysis datasets created. Inverse probability treatment weighting at baseline and inverse probability censoring weights will be calculated. Inverse probability treatment weighting will include all potential baseline confounders and inverse probability censoring weighting will include both baseline and time-varying confounding factors. These weights will be combined and applied to the data to create a pseudo-population in which each treatment arm has greater similarity in their characteristics following re-weighting. The weighted data will be used to compare the treatment effect. An appropriate outcome model will be used to analyse the weighted data. The model-based standard errors may be inaccurate due to using weighted data for analysis, so bootstrapping or robust standard errors will be used to estimate the 95% confidence intervals.

G-formula analysis

The G-formula will be used to estimate average treatment effects from all the analysis datasets created. In addition to the baseline and time-varying confounding variables, the censoring flag will be included in the analysis to account for informative censoring. Parametric models will be used to model the outcome on treatments to estimate log hazard ratios. Bootstrapping standard errors will be used to estimate the 95% confidence intervals.

G-estimation analysis

G-estimation will be used to estimate conditional treatment effects from all the analysis datasets created. G-estimation itself is not considered sufficient to handle informative censoring, therefore, as recommended by Hernan et al. a pseudo-population will be created by inversely weighting the data by the inverse probability of remaining uncensored during follow-up using IPCW. The IPCW analysis will include all the baseline and post-baseline time-varying confounding variables to reduce the bias due to time-varying confounding. The inverse probability weighted data will be analysed using G-estimation to estimate the treatment effect.

Targeted Maximum Likelihood Estimation (TMLE) analysis

Targeted Maximum Likelihood Estimation (TMLE) will be used to estimate the average treatment effect on overall survival adjusting for both baseline and post baseline time-varying confounding factors. TMLE is a doubly robust method that can yield a valid estimate when either the outcome model or the treatment and censoring models are correctly specified. The causal interpretation of the estimates obtained depends on additional causal assumptions being true. TMLE is considered a semi parametric approach, as it utilises machine learning techniques that can help avoid making statistical distributional assumptions required for parametric analysis methods when they unlikely to be true.

4.7. Sensitivity analysis

E-value

To evaluate the potential impact of unmeasured confounding, the E-value will be used to estimate the required strength of a potential confounder to change the interpretation of the treatment effect [31].

5. Comparison with original RCT results

Similar to the RCT DUPLICATE project, the following criteria will be applied to evaluate whether the estimates generated by the TTs correspond to the original RCT trial results.

First, the consensus on statistical significance will be used to compare the results. The LUX-Lung 7 trial and TNT trial both reported null treatment effects, therefore, for TT1 and TT3, a null result will be considered as having a consensus, and other results will be considered as not. The KEYNOTE-024 trial reported a statistically significant treatment effect favourable to the pembrolizumab arm, so a similar statistically significant result favouring the pembrolizumab arm will be considered as having a consensus. When the number of records included in the trial emulations are larger or smaller than the original trial sample size, then the potential of having a statistical significance or not due to the differences in sample sizes will be considered when reporting the results.

Second, the consistency of the point estimates between the TT and the original RCTs will be compared. If the point estimates from the TTs are within the 95% confidence interval of the respective RCT treatment effect reported, then they will be considered as being consistent.

Third, the standardised effect sizes estimated from the TT and the original RCT will be compared. If the differences in standardised effect sizes between the TT and its respective RCT is within -1.96 and +1.96, then they will be considered as having an agreement.

In addition, survival curves TT emulations will be plotted and will be compared against the Kaplan Meier (KMs) from the trials to check wither the emulated survival curves lie within the 95% CIs of the trial survival curves.

6. Ethics and dissemination

This project work has the ethical approval of the University of Sheffield (Reference number: 049343). The results obtained from this research will be disseminated through conference presentations and journal publications.

7. Protocol summary

The TT studies described in this protocol will be emulated using England's Cancer Registry data. Data analysis will be undertaken only after the protocol is approved by all reviewers and uploaded online. The estimates obtained from these TT emulations will be compared against the original study results using a set of pre-specified criteria to determine the suitability of England’s cancer registry data for estimating reliable comparative effectiveness.

8. References

1. *NCRAS*. Feb 2021; Available from: http://www.ncin.org.uk/home.
2. Bright, C.J., et al., *Data resource profile: the systemic anti-cancer therapy (SACT) dataset.* International journal of epidemiology, 2020. **49**(1): p. 15-15l.
3. Mansournia, M.A., et al., *Handling time varying confounding in observational research.* Bmj, 2017. **359**.
4. Hernán, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* Am J Epidemiol, 2016. **183**(8): p. 758-64.
5. Dahabreh, I.J., et al., *Using trial and observational data to assess effectiveness: Trial emulation, transportability, benchmarking, and joint analysis.* Epidemiologic Reviews, 2023: p. mxac011.
6. Park, K., et al., *Afatinib versus gefitinib as first-line treatment of patients with EGFR mutation-positive non-small-cell lung cancer (LUX-Lung 7): a phase 2B, open-label, randomised controlled trial.* Lancet Oncol, 2016. **17**(5): p. 577-89.
7. Herbst, R.S., et al., *Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial.* Lancet, 2016. **387**(10027): p. 1540-1550.
8. Tutt, A., et al., *Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial.* Nat Med, 2018. **24**(5): p. 628-637.
9. Merriel, S.W.D., et al., *Cross-sectional study evaluating data quality of the National Cancer Registration and Analysis Service (NCRAS) prostate cancer registry data using the Cluster randomised trial of PSA testing for Prostate cancer (CAP).* BMJ Open, 2017. **7**(11): p. e015994.
10. Sheffield, K.M., et al., *Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions.* Journal of Comparative Effectiveness Research, 2020. **9**(15): p. 1043-1050.
11. Franklin, J.M., et al., *Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative.* Circulation, 2021. **143**(10): p. 1002-1013.
12. Cuschieri, S., *The CONSORT statement.* Saudi journal of anaesthesia, 2019. **13**(Suppl 1): p. S27-S30.
13. Franklin, J.M., et al., *Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project.* Clin Pharmacol Ther, 2020. **107**(4): p. 817-826.
14. *NICE appraisal*. 2020 [cited 2020 01/11/2020]; Available from: https://www.nice.org.uk/process/pmg9/chapter/the-appraisal-of-the-evidence-and-structured-decision-making.
15. *Technology appraisal data: cancer appraisal recommendations* 2021; Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance/data/cancer-appraisal-recommendations.
16. Kennedy-Martin, T., et al., *A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results.* Trials, 2015. **16**: p. 495.
17. Freedman, B.J.T.N.E.J.o.M., https://www. nejm. org/doi/full/10./NEJM198707163170304, *Equipoise and the ethics of clinical research.* 1987.
18. Kahan, B.C., S. Rehal, and S. Cro, *Risk of selection bias in randomised trials.* Trials, 2015. **16**: p. 405.
19. Mansournia, M.A., et al., *Biases in randomized trials: a conversation between trialists and epidemiologists.* Epidemiology (Cambridge, Mass.), 2017. **28**(1): p. 54.
20. *NICE real-world evidence framework guidance document*. 2022 [cited 2022; Available from: https://www.nice.org.uk/corporate/ecd9/chapter/overview.

21. NHS-Digital, Dec 2020.

22. Cai, L. and Y. Zhu, *The challenges of data quality and data quality assessment in the big data era.* Data science journal, 2015. **14**.

23. Hernán, M.A. and J.M. Robins, *Causal inference: what if*. 2020, Boca Raton: Chapman & Hall/CRC.

24. systems, H.d.s.a. *Leptomeningeal disease*. 2024; Available from: http://remote.health.vic.gov.au/viccdb/view.asp?Query_Number=3570.

25. Quan, H., et al., *Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.* Medical care, 2005. **43**(11): p. 1130-1139.

26. Reck, M., et al., *Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer.* N engl J med, 2016. **375**: p. 1823-1833.

27. Harpsøe, M.C., et al., *Body mass index and risk of autoimmune diseases: a study within the Danish National Birth Cohort.* International journal of epidemiology, 2014. **43**(3): p. 843-855.

28. Neibart, S.S., et al., *Validation of a claims‐based algorithm for identifying non‐infectious pneumonitis in patients diagnosed with lung cancer.* Pharmacoepidemiology and drug safety, 2021. **30**(12): p. 1624-1629.

29. Reck, M., et al., *Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater.* J Clin Oncol, 2019. **37**(7): p. 537-546.

30. Song, M., et al., *The missing covariate indicator method is nearly valid almost always.* arXiv preprint arXiv:2111.00138, 2021.

31. VanderWeele, T.J. and P. Ding, *Sensitivity Analysis in Observational Research: Introducing the E-Value.* Ann Intern Med, 2017. **167**(4): p. 268-274.

9. Appendix

9.1. Codes for identification

Lung cancer morphology codes

Table 9-1: Lung cancer morphology codes to classify NSCLC

Group	Specific group	Morphology code	Description
SCLC	Small cell carcinoma	8041	Small cell carcinoma NOS
SCLC	Small cell carcinoma	8042	Oat cell carcinoma
SCLC	Small cell carcinoma	8045	Small cell-large cell carcinoma
SCLC	Small cell carcinoma	8044	Small cell carcinoma, intermediate cell
SCLC	Small cell carcinoma	8043	Small cell carcinoma, fusiform cell
SCLC	Small cell carcinoma	8002	Malignant tumour, small cell type
NSCLC	Adenocarcinoma	8140	Adenocarcinoma NOS
NSCLC	Adenocarcinoma	8250	Bronchiolo-alveolar adenocarcinoma
NSCLC	Adenocarcinoma	8480	Mucinous adenocarcinoma
NSCLC	Adenocarcinoma	8481	Mucin-producing adenocarcinoma
NSCLC	Adenocarcinoma	8260	Papillary adenocarcinoma NOS
NSCLC	Adenocarcinoma	8550	Acinar cell carcinoma
NSCLC	Adenocarcinoma	8251	Alveolar adenocarcinoma
NSCLC	Adenocarcinoma	8310	Clear cell adenocarcinoma NOS
NSCLC	Adenocarcinoma	8490	Signet ring cell carcinoma
NSCLC	Adenocarcinoma	8323	Mixed cell adenocarcinoma
NSCLC	Adenocarcinoma	8570	Adenocarcinoma with squamous metaplasia
NSCLC	Adenocarcinoma	8211	Tubular adenocarcinoma
NSCLC	Adenocarcinoma	8141	Scirrhous adenocarcinoma
NSCLC	Adenocarcinoma	8440	Cystadenocarcinoma NOS
NSCLC	Adenocarcinoma	8143	Superficial spreading adenocarcinoma
NSCLC	Adenocarcinoma	8144	Adenocarcinoma, intestinal type
NSCLC	Adenocarcinoma	8147	Basal cell adenocarcinoma
NSCLC	Adenocarcinoma	8190	Trabecular adenocarcinoma
NSCLC	Adenocarcinoma	8201	Cribriform carcinoma
NSCLC	Adenocarcinoma	8252	Bronchioalvelolar carcinoma non mucinous
NSCLC	Adenocarcinoma	8320	Granular cell carcinoma
NSCLC	Adenocarcinoma	8401	Apocrine adenocarcinoma
NSCLC	Adenocarcinoma	8470	Mucinous cystadenocarcinoma NOS
NSCLC	Adenocarcinoma	8572	Adenocarcinoma with spindle cell metaplasia

Group	Specific group	Morphology code	Description
NSCLC	Adenocarcinoma	8574	Adenocarcinoma with neuroendocrine differentiation
NSCLC	Squamous cell carcinoma	8070	Squamous cell carcinoma NOS
NSCLC	Squamous cell carcinoma	8071	Squamous cell carcinoma, keratinising NOS
NSCLC	Squamous cell carcinoma	8072	Squamous cell carcinoma, large cell, non-keratinisi
NSCLC	Squamous cell carcinoma	8073	Squamous cell carcinoma, small cell, non-keratinisi
NSCLC	Squamous cell carcinoma	8074	Squamous cell carcinoma, spindle cell
NSCLC	Squamous cell carcinoma	8052	Papillary squamous cell carcinoma
NSCLC	Squamous cell carcinoma	8083	Basaloid Squamous Cell carcinoma
NSCLC	Squamous cell carcinoma	8076	Squamous cell carcinoma, microinvasive
NSCLC	Carcinoid	8240	Carcinoid tumour NOS
NSCLC	Carcinoid	8249	Atypical carcinoid tumour
NSCLC	Carcinoid	8241	Carcinoid tumour, argentaffin, malignant
NSCLC	Carcinoid	8244	Composite carcinoid
NSCLC	Carcinoid	8243	Goblet cell carcinoid
NSCLC	Carcinoid	8245	Adenocarcinoma tumour
NSCLC	Other NSCLC	8046	Non-small cell carcinoma, NOS
NSCLC	Other NSCLC	8012	Large cell carcinoma NOS
NSCLC	Other NSCLC	8246	Neuroendocrine carcinoma
NSCLC	Other NSCLC	8560	Adenosquamous carcinoma
NSCLC	Other NSCLC	8032	Spindle cell carcinoma
NSCLC	Other NSCLC	8022	Pleomorphic carcinoma
NSCLC	Other NSCLC	8033	Pseudosarcomatous carcinoma
NSCLC	Other NSCLC	8200	Adenoid cystic carcinoma
NSCLC	Other NSCLC	8980	Carcinosarcoma NOS
NSCLC	Other NSCLC	8430	Mucoepidermoid carcinoma
NSCLC	Other NSCLC	8031	Giant cell carcinoma
NSCLC	Other NSCLC	8050	Papillary carcinoma NOS
NSCLC	Other NSCLC	8972	Pulmonary blastoma
NSCLC	Other NSCLC	8013	Large cell neuroendocrine
NSCLC	Other NSCLC	8123	Basaloid carcinoma
NSCLC	Other NSCLC	8940	Mixed tumour, malignant NOS
NSCLC	Other NSCLC	8075	Adenoid squamous cell carcinoma
NSCLC	Other NSCLC	8230	Solid carcinoma NOS
NSCLC	Other NSCLC	8255	Adenocarcinoma with mixed cell types

Group	Specific group	Morphology code	Description
NSCLC	Other NSCLC	8030	Giant cell and spindle cell carcinoma
NSCLC	Other NSCLC	8034	Polygonal cell carcinoma
NSCLC	Other NSCLC	8082	Lymphoepithelial carcinoma
NSCLC	Other NSCLC	8145	Carcinoma, diffuse type
NSCLC	Other NSCLC	8562	Epithelial-myoepithelial carcinoma
NSCLC	Unspecified lung	8010	Carcinoma NOS
NSCLC	Unspecified lung	8000	Neoplasm, malignant
NSCLC	Unspecified lung	8020	Carcinoma, undifferentiated NOS
NSCLC	Unspecified lung	8021	Carcinoma, anaplastic type NOS
NSCLC	Unspecified lung	8001	Tumour cells, malignant
NSCLC	Unspecified lung	8004	Malignant tumour, fusiform cell type
NSCLC	Unspecified lung	8003	Malignant tumour, giant cell type
Exclude	Excluded - sarcoma	8800	Sarcoma NOS
Exclude	Excluded - sarcoma	8801	Spindle cell sarcoma
Exclude	Excluded - sarcoma	8890	Leiomyosarcoma NOS
Exclude	Excluded - sarcoma	8810	Fibrosarcoma NOS
Exclude	Excluded - sarcoma	9120	Haemangiosarcoma
Exclude	Excluded - sarcoma	9040	Synovial sarcoma NOS
Exclude	Excluded - sarcoma	8803	Small cell sarcoma
Exclude	Excluded - sarcoma	8900	Rhabdomyosarcoma NOS
Exclude	Excluded - sarcoma	8802	Giant cell sarcoma (except of bone M9250/3)
Exclude	Excluded - sarcoma	8830	Fibrous histiocytoma, malignant
Exclude	Excluded - sarcoma	8804	Epithelioid sarcoma
Exclude	Excluded - sarcoma	8811	Fibromyxosarcoma
Exclude	Excluded - sarcoma	8850	Liposarcoma NOS
Exclude	Excluded - sarcoma	8851	Liposarcoma, well-differentiated
Exclude	Excluded - sarcoma	8852	Myxoid liposarcoma
Exclude	Excluded - sarcoma	8854	Pleomorphic liposarcoma
Exclude	Excluded - sarcoma	8894	Angiomyosarcoma
Exclude	Excluded - sarcoma	8901	Pleomorphic rhabdomyosarcoma
Exclude	Excluded - sarcoma	8910	Embryonal rhabdomyosarcoma
Exclude	Excluded - sarcoma	8920	Alveolar rhabdomyosarcoma
Exclude	Excluded - sarcoma	8933	Adenosarcoma
Exclude	Excluded - sarcoma	8963	Rhabdoid sarcoma
Exclude	Excluded - sarcoma	9170	Lymphangiosarcoma
Exclude	Excluded - sarcoma	9180	Osteosarcoma NOS
Exclude	Excluded - sarcoma	9220	Chondrosarcoma NOS
Exclude	Excluded - sarcoma	9260	Ewing's sarcoma
Exclude	Excluded - unusual	9133	Epithelioid haemangioendothelioma, malignant

Group	Specific group	Morphology code	Description
Exclude	Excluded - unusual	8720	Malignant melanoma NOS
Exclude	Excluded - unusual	8040	Tumourlet (uncertain malignancy)
Exclude	Excluded - unusual	8011	Epithelioma, malignant
Exclude	Excluded - unusual	8146	Monomorphic adenoma
Exclude	Excluded - benign	8333	Microfollicular adenoma
Exclude	Excluded - unusual	9364	Peripheral neuroectodermal tumour
Exclude	Excluded - unusual	8120	Transitional cell carcinoma NOS
Exclude	Excluded - unusual	8263	Adenocarcinoma in tubulovillous adenoma
Exclude	Excluded - unusual	8520	Lobular carcinoma NOS
Exclude	Excluded - unusual	8815	Solitary fibrous tumour
Exclude	Excluded - unusual	8982	Myoepithelioma
Exclude	Excluded - unusual	9050	Mesothelioma, malignant
Exclude	Excluded - unusual	9071	Endodermal sinus tumour
Exclude	Excluded - unusual	9080	Teratoma, malignant NOS
Exclude	Excluded - unusual	9130	Haemangioendothelioma, malignant
Exclude	Excluded - unusual	9473	Primitive neuroectodermal tumour

ICD-10 codes to identify Autoimmune diseases

Table 9-2: ICD-10 codes to identify Autoimmune diseases (Harpsøe, Maria C., et al. 2014 [27])

Autoimmune disease	ICD-10 codes
Addison’s disease	E27.1, E27.2
Ankylosing spondylitis	M45, M08.1
Behcet’s disease	M35.2
Buerger’ s syndrome	M31.1B, DI7.31
Celiac disease	K90.0
Crohn’ s disease	K50
Dermatitis herpetiformis	L13.0
Diabetes mellitus type 1	E10
Dupuytren’ s disease	M72.0
Erythema nodosum	L52
Goodpasture’ s syndrome	M31.0
Graves’ disease	E05.0
Guillain-Barré syndrome	G61.0

Autoimmune disease	**ICD-10 codes**
Haemolytic anaemia	D59.0, D59.1
Hashimoto’s thyroiditis	E06.3
Henoch-Schönlein purpura	D69.0
ITP	D69.3
Kawasaki syndrome	M30.3
Localized lupus erythematosus	L93
Localized scleroderma	L94.0, L94.1, L94.3
Myasthenia gravis	G70.0
Multiple sclerosis	G35
Pemphigoid	L12
Pernicious anaemia	D51.0
Pemphigus foliacus	L10.2
Pemphigus vulgaris	L10.0
Polyarteritis nodosa	M30.0
Polymyositis/dermatomyositis	M33
Primary biliary cirrhosis	K74.3
Psoriasis	L40
Rheumatic fever	I00, I01
Rheumatoid arthritis	M05, M06
Raynaud’s phenomenon	DI73.0
Reiter’s disease	M02.3
Sarcoidosis	D86
Sjögren’s syndrome	M35.0
Sympathetic ophthalmia	H44.1B
Systemic lupus erythematosus	M32
Systemic scleroderma	M34
Temporal arteritis	M31.5, M31.6, M35.3
Ulcerative colitis	K51
Vitiligo	L80
Wegener’s granulomatosis	M31.3

Codes for Non-Infectious Pneumonitis

Table 9-3: **ICD-10-CM Codes for Non-Infectious Pneumonitis** (validated by Neibart, Shane S., et al. 2021 [28])

Code	Description
J70.0	Acute pulmonary manifestations due to radiation
J70.1	Chronic and other pulmonary manifestations due to radiation
J70.2	Acute drug-induced interstitial lung disorders
J70.3	Chronic drug-induced interstitial lung disorders
J70.4	Drug-induced interstitial lung disorders, unspecified
J70.8	Respiratory conditions due to other specified external agents
J70.9	Respiratory conditions due to unspecified external agent
J84.1	Other interstitial pulmonary diseases with fibrosis
J84.111	Idiopathic interstitial pneumonia not otherwise specified
J84.113	Idiopathic non-specific interstitial pneumonitis
J84.114	Acute interstitial pneumonitis

Target Trial 1 emulation treatments

Table 9-4: Target Trial 1: LUX Lung 7 trial post progression treatments

Treatment	Afatinib	Gefitinib
Systemic anti-cancer therapy	106 (72.6)	116 (76.8)
Chemotherapy or chemotherapy-based combination	84 (57.5)	91 (60.3)
Platinum based	70 (47.9)	71 (47.0)
EGFR TKI	67 (45.9)	84 (55.6)
EGFR TKI monotherapy	63 (43.2)	74 (49.0)
First-generation gefitinib	22 (15.1)	27 (17.9)

Treatment	Afatinib	Gefitinib
First-generation erlotinib	23 (15.8)	30 (19.9)
Second-generation afatinib	6 (4.1)	12 (7.9)
Second-generation poziotinib	0 (0.0)	4 (2.6)
Third-generation osimertinib	15 (10.3)	17 (11.3)
Third-generation olmutinib	5 (3.4)	5 (3.3)
EGFR TKI-containing combination Including gefitinib (afatinib arm, n = 7; gefitinib arm, n = 11), erlotinib (n = 0; n = 5) and osimertinib (n = 0; n = 1).	7 (4.8)	15 (9.9)
Immune checkpoint inhibitor	3 (2.1)	4 (2.6)
Radiotherapy	26 (17.8)	34 (22.5)

Codes to identify cardiovascular abnormalities

Table 9-5: ICD10 codes to identify cardiovascular abnormalities

Condition	ICD 10 codes
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9

Coding Algorithms for Charlson comorbidities

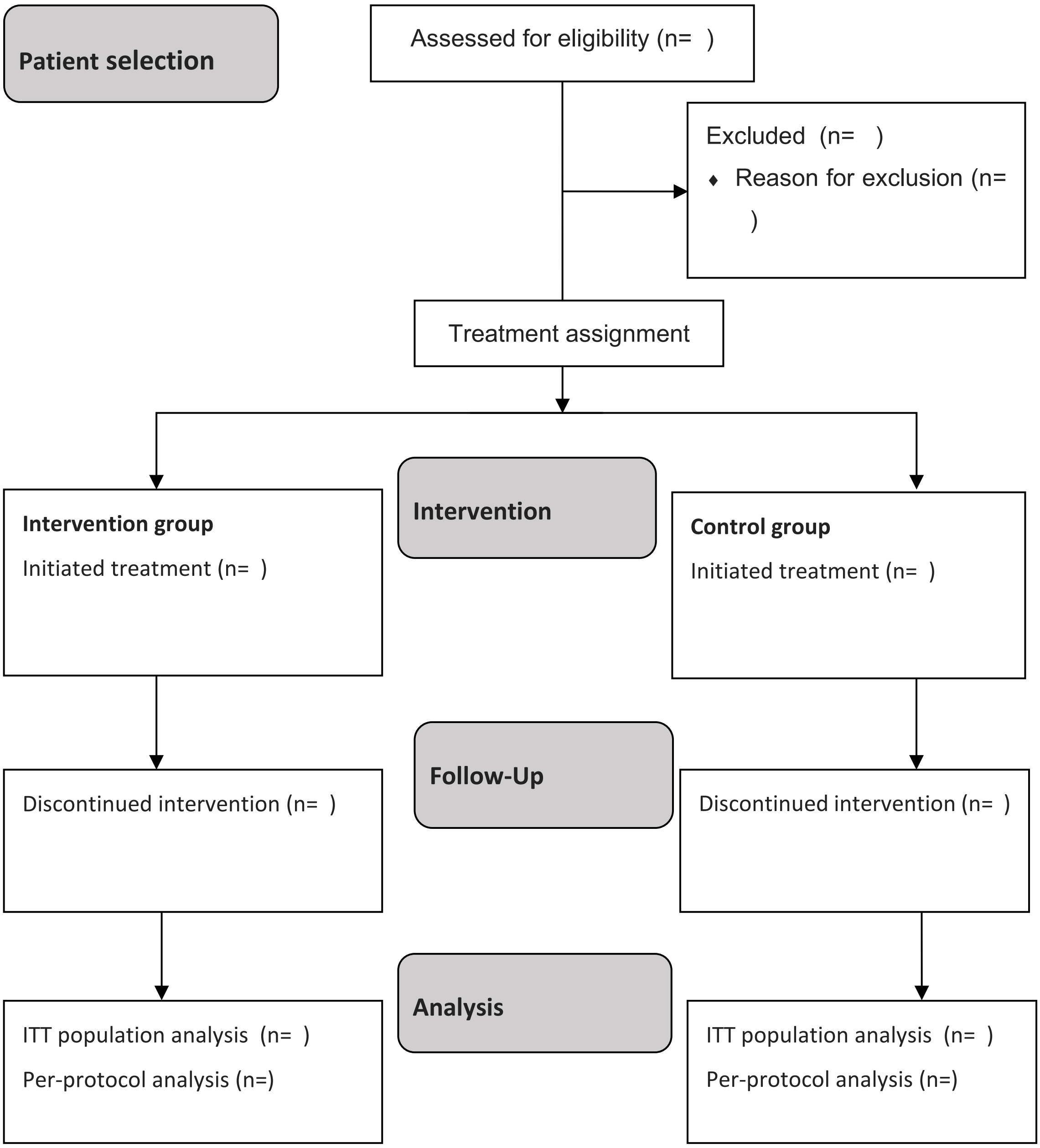
Table 9-6: Coding Algorithms for Defining Comorbidities in ICD-10 Administrative Data [25]

Condition	ICD 10 codes
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular Disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic pulmonary disease	I27.8, I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Peptic ulcer disease	K25.x–K28.x
Mild liver disease	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.2–E10.5, E10.7, E11.2–E11.5, E11.7, E12.2–E12.5, E12.7, E13.2– E13.5, E13.7, E14.2–E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9

Condition	ICD 10 codes
Renal disease	I12.0, I13.1, N03.2–N03.7, N05.2– N05.7, N18.x, N19.x, N25.0, Z49.0– Z49.2, Z94.0, Z99.2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	C00.x–C26.x, C30.x–C34.x, C37.x– C41.x, C43.x, C45.x–C58.x, C60.x– C76.x, C81.x–C85.x, C88.x, C90.x–C97.x
Moderate or severe liver disease	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
Metastatic solid tumor	C77.x–C80.x
AIDS/HIV	B20.x–B22.x, B24.x

9.2. Data flow template

Target Trial Data Flow Diagram


Patient selection
Assessed for eligibility (n=)
Excluded (n=)
♦ Reason for exclusion (n=)
Treatment assignment
Intervention group
Initiated treatment (n=)
Intervention
Control group
Initiated treatment (n=)
Discontinued intervention (n=)
Follow-Up
Discontinued intervention (n=)
ITT population analysis (n=)
Per-protocol analysis (n=)
Analysis
ITT population analysis (n=)
Per-protocol analysis (n=)

9.3. Example graphs and figures

Example DAG

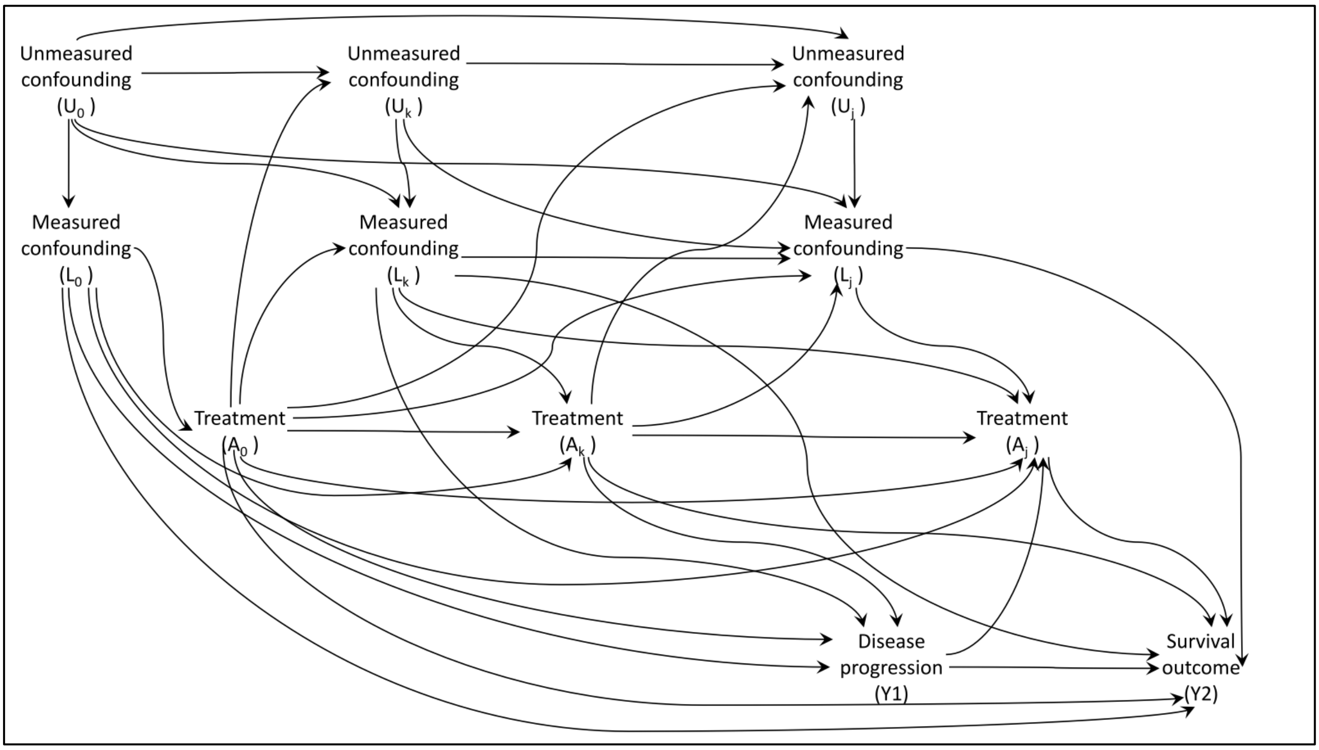

Unmeasured confounding (U0)
Unmeasured confounding (Uk)
Unmeasured confounding (Uj)
Measured confounding (L0)
Measured confounding (Lk)
Measured confounding (Lj)
Treatment (A0)
Treatment (Ak)
Treatment (Aj)
Disease progression (Y1)
Survival outcome (Y2)

Figure 1: Example DAG

U_0 = Unmeasured confounding at baseline (lifestyle factors (e.g., smoking, alcohol use), health status, genetic factors, socio-economic background, educational level)

L_0 = Measured confounding at baseline (cancer staging, age at diagnosis, ethnicity, measure of deprivation, co-morbidities, PSA level, Gleason score, BMI / height and weight, ECOG performance status, year of diagnosis, geographic region)

A_0 = Treatment strategy at baseline

U_k = Unmeasured confounding after baseline and before disease progression (e.g., changes to lifestyle factors)

L_k = Measured confounding after baseline and before disease progression (e.g. performance status, co-morbidities)

A_k = Treatment changes (e.g., first line to second line treatment)

U_j = Unmeasured confounding after disease progression and before survival outcome/end of follow-up (changes to lifestyle factors)

A_j = Measured confounding after disease progression and before survival outcome/end of follow-up (e.g. performance status, co-morbidities)

T_j = Treatment changes after disease progression (e.g., start of a new treatment)

Y1 = Disease progression outcome

Y2 = Survival outcome

Propensity score overlap

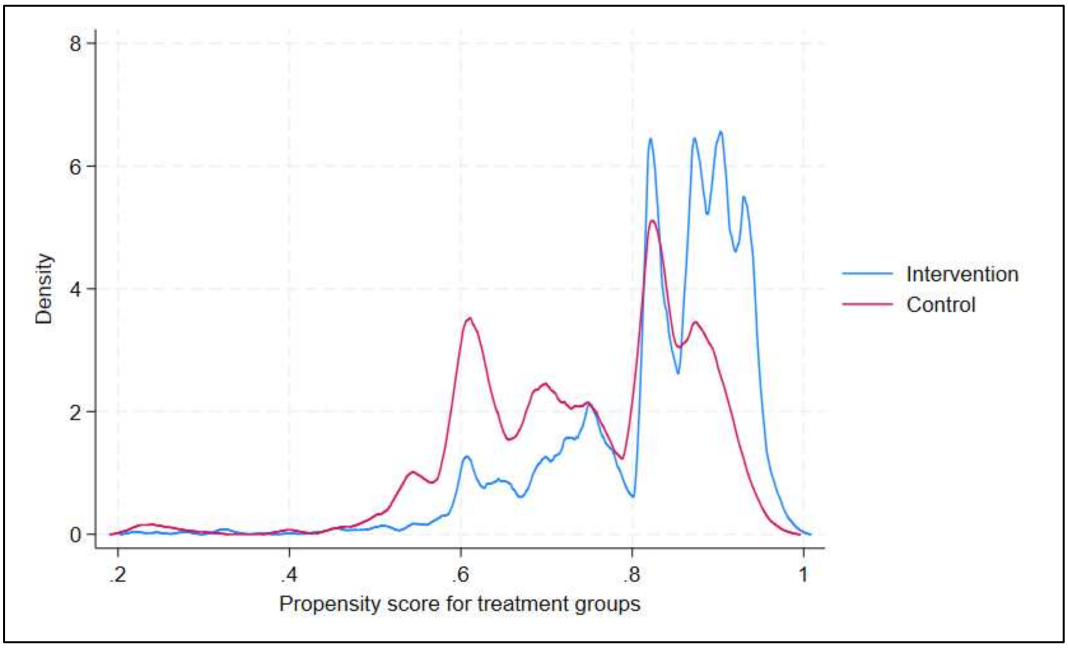


Figure 2: Example propensity score overlap graph

Results comparisons

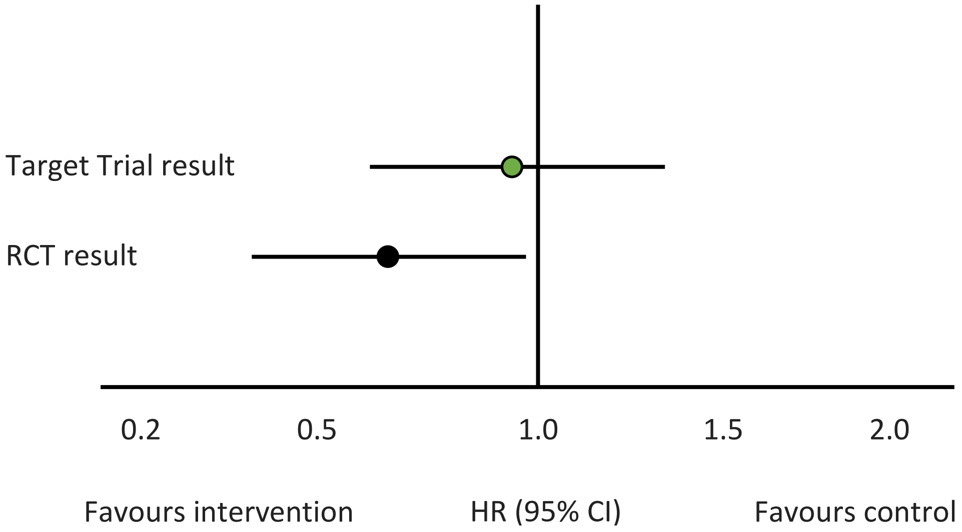


Figure 3: Example Overall survival estimates from RCT and TTs estimates

9.4. Example summary tables

Example data completeness summary report

Table 9-7: Data completeness summary

Variable	Complete n (%)			Missing n (%)	Total number of patients n (%)
	Treatment group 1	Treatment group 2	Not eligible		
Tumour stage	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Performance score	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)

Variable	Complete n (%)			Missing n (%)	Total number of patients n (%)
	Treatment group 1	Treatment group 2	Not eligible		
Number of tumour lesions	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Height	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Weight	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Sex	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Ethnicity	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Found a match in HES inpatient data	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Found a match in HES A & E data	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Found a match in HES outpatients	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
…..	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
…..	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)

Example demographics of patients screened

Table 9-8: Eligible and not eligible patient demographics summary

Characteristics	Statistics	Eligible (n=)	Not eligible (n=)	Total (n=)
Sex	N	XX	XX	XX
	Male, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Female, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Age	N	XX	XX	XX
	Mean (SD)	XX	XX	XX
	Median (Min, Max)	XX (XX,XX)	XX (XX,XX)	XX (XX,XX)
Ethnicity	N	XX	XX	XX
	White, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Black/African/Caribbean/Black British, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Asian/Asian British, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Mixed, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Other, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
…..				
…..				

Example baseline summary statistics report

Table 9-9: Baseline summary statistics

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
Sex	N	XX	XX	XX
	Male, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Female, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Age	N	XX	XX	XX
	Mean (SD)	XX	XX	XX
	Median (Min, Max)	XX (XX,XX)	XX (XX,XX)	XX (XX,XX)
Ethnicity	N	XX	XX	XX
	White, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Black/African/Caribbean/Black British, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Asian/Asian British, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Mixed, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Other, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Cancer stage	N	XX	XX	XX
	Stage II, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Stage III, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Stage IV, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Time to treatment (in days)	N	XX	XX	XX
	Mean (SD)	XX	XX	XX
	Median (Min, Max)	XX (XX,XX)	XX (XX,XX)	XX (XX,XX)
…..				
…..				

Example Baseline treatment compliance report

Table 9-10: Treatment compliance and treatment changes summary

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
Total number of regimens	N	XX	XX	XX
	One, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Two, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Number of cycles	N	XX	XX	XX
	Mean (SD)	XX	XX	XX
	Median (Min, Max)	XX (XX,XX)	XX (XX,XX)	XX (XX,XX)
Second-line treatments	N	XX	XX	XX
	Drug A, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Drug B, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Drug C, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Drug D, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	Other, n (%)	XX (XX%)	XX (XX%)	XX (XX%)

Example follow-up summary report

Table 9-11: Follow-up duration summary

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
Follow-up duration	N	XX	XX	XX
	Mean (SD)	XX	XX	XX
	Median (Min, Max)	XX (XX,XX)	XX (XX,XX)	XX (XX,XX)

Example post-baseline summary report

Table 9-12: Concurrent or adjuvant treatment summary

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
Surgical treatment received	N	XX	XX	XX
	Yes, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	No, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Radiotherapy received	N	XX	XX	XX
	Yes, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	No, n (%)	XX (XX%)	XX (XX%)	XX (XX%)

Table 9-13: Post-baseline adverse events summary

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
Hospital admission	N	XX	XX	XX
	Yes, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	No, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
Number of hospital admissions	N	XX	XX	XX
	1, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	2, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	3, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	4, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
	5 or more, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
A & E attendance	N	XX	XX	XX
	Yes, n (%)	XX (XX%)	XX (XX%)	XX (XX%)

Characteristics	Statistics	Treatment group 1 (n=)	Treatment group 2 (n=)	Total (n=)
	No, n (%)	XX (XX%)	XX (XX%)	XX (XX%)
…..				