University of Sheffield | Sheffield Centre For Health & Related Research

HEALTH ECONOMICS & DECISION SCIENCE

Discussion Paper Series

HEDS Discussion Paper 24.05

Title: A comparison of the effectiveness of different treatment regimens for pancreatic cancer using English cancer registry data

Author: Nick Latimer

Corresponding author: Nick Latimer, Professor of Health Economics, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK. Tel: +44 (0) 114 222 0821 , Email: n.latimer@shef.ac.uk

A comparison of the effectiveness of different treatment regimens for pancreatic cancer using English cancer registry data

Study Protocol

11th December 2025

Project Initiation	
Project Title	A comparison of the effectiveness of different treatment regimens for pancreatic cancer using English cancer registry data
Project Objective	This project aims to investigate whether or not English cancer registry data is sufficient for reliably comparing the effectiveness of different cancer treatments given in the NHS.
Principal Investigator	Dr Nicholas Latimer, University of Sheffield
Project team members	Professor Jim Chilcott (Health economist/modeller, University of Sheffield) Professor Paul Tappenden (Health economist/modeller, University of Sheffield) Dr Peter Hall (Medical oncologist/Health economist, University of Edinburgh) Professor Jonathan Wadsley (Clinical oncologist, Sheffield Teaching Hospital) Professor Uwe Siebert (Health decision scientist/epidemiologist, UMIT TIROL – University for Health Sciences and Technology, Austria) Dr Rebecca Smittenaar (Principal Clinical Scientist, GRAIL, Inc. Formerly Analytical Lead, Public Health England) Dr Ellie Murray (Epidemiologist, Boston University)
PI's institution	University of Sheffield
Project Funder	Yorkshire Cancer Research
Study protocol version	1.5

Project Amendment	
Study protocol amendment history	v1.1. 4th February 2020: amended title to "A comparison of the effectiveness of different treatment regimens for pancreatic cancer using English cancer registry data", from "A comparison of the effectiveness of different treatment regimens for adjuvant and advanced/metastatic pancreatic cancer using English cancer registry data"
	v1.2. 20th October 2020: - Amended data request summary table in response to discussions with ODR staff. This includes more detail on specific variables required and Charlson scores (pg. 20-23). - Clarified request for geographic data (pg.19, pg.21).
	v1.3. 4th November 2020: removed request for duplicated variables (i.e. the same variable requested from multiple datasets), taking the advice of the Public Health England analyst on which dataset to most appropriately request the data from.
	V1.4. 30th October 2024:

	- The aim of this update was to provide more detail on the analysis plan, prior to conducting analyses, such that key analytical details could be pre-specified and placed in the public domain. - The study was considerably delayed due to delays in accessing the data and limited researcher capacity. Since the original data application was made, the process for applying for data has changed, as have some of the National Cancer Registration and Analysis Service (NCRAS) datasets and variables. Therefore, parts of the “Data Requirements” section of this protocol are outdated – however, they are representative of the situation when the data application was made. - Version 1.4 includes updates to study team details, an abstract, minor updates to the “Background” section, and major updates to the “Analysis Plan” section. The “Data Requirements”, “Project Administration and Governance”, “Ethical Approval” and “Timelines and Dissemination” sections are left unchanged and are in some places out-dated, but are retained in this version of the document for completeness.
	V1.5. 11th December 2025: correction to formula used for agreement criteria (criterion 3)

Contents

Abstract

Large amounts of data are collected on cancer patients in the NHS, held by the National Cancer Registration and Analysis Service (NCRAS). Data are collected on patient and tumour characteristics, treatments, and can be linked to hospital episode statistics. Usually new cancer treatments are investigated in randomised controlled trials (RCTs), which are widely considered to represent the gold standard approach for comparing interventions. However, sometimes it is not possible to run an RCT, due to feasibility or ethical issues. In addition, RCTs often have strict and restrictive eligibility criteria. Whilst RCTs might tell us about the comparative effectiveness of treatments in highly selected trial populations, they are less useful for investigating comparative effectiveness in more general populations. Therefore, it is important to investigate the use of NCRAS data as a resource for estimating the comparative effectiveness of cancer treatments in the "real world", that is, under routine conditions. This project aims to investigate whether or not English cancer registry data is sufficient for deriving valid causal estimates of the comparative effectiveness of different cancer treatments given in the NHS.

This document provides a protocol for carrying out four Target Trial Emulations (TTE), using NCRAS data. Each TTE seeks to emulate as closely as possible an existing RCT investigating treatments for pancreatic cancer. We describe each TTE in detail, and specify agreement criteria that will be used to evaluate the success of each emulation.
This study will provide valuable evidence on whether it is possible to derive robust and valid causal estimates of comparative effectiveness of cancer treatments given in the NHS. If we are able to successfully emulate existing RCTs, our study will provide evidence that obtaining such estimates is possible, and will provide the basis for designing analyses that seek to answer questions not addressed by RCTs. If we are not able to successfully emulate existing RCTs, our study will seek to identify key weaknesses in the registry datasets, with the intention of determining how these datasets could be improved. Therefore, our study has the potential to provide valuable insights for healthcare decision-makers, clinicians, and patients.

Background

Large amounts of data are collected on cancer patients in the NHS, held by the National Cancer Registration and Analysis Service (NCRAS). The Systemic Anti-Cancer Therapy (SACT) database, part of the NCRAS dataset, purports to be the world's first comprehensive database, collecting information on systemic-anti cancer therapies on a national scale. The dataset collects information at patient and tumour level and is designed to be linked to other data sources (such as hospital episode statistics (HES) and radiotherapy datasets) to provide a complete picture of the cancer patient pathway. In fact, NCRAS has been commissioned by NHS England (NHSE) to provide data and analysis for the evaluation of drugs that are in the Cancer Drugs Fund (CDF), with the aim of using the data to resolve uncertainties around the effectiveness and cost-effectiveness of cancer treatments placed in the CDF. Despite this, as yet, no attempts have been made to assess whether the data held by NCRAS is sufficient for reliably comparing the effectiveness of different cancer treatments.

Usually new cancer treatments are compared in randomised controlled trials (RCTs). RCTs are usually considered to represent the gold standard approach for comparing interventions, with the purpose of random assignment being to avoid selection bias in the assignment of treatment options (i.e. "confounding by indication"); that is, ensuring that characteristics of patients that may influence the outcome are randomly distributed between groups, so that any difference in outcome can be explained only by the treatment.[1]

However, sometimes it is not practical to run an RCT. For example, consider the case of the Cancer Drugs Fund (CDF). The National Institute for Health and Care Excellence (NICE) assesses the effectiveness and cost-effectiveness of new treatments. For cancer treatments, where clinical uncertainty means that NICE is unsure whether the new treatment is cost-effective – but there is a plausible case that it might be – NICE is able to recommend that the treatment is only available in the CDF, rather than in routine commissioning. Often an updated data cut from an existing RCT will represent the main source of evidence used to resolve NICE's uncertainties, but, sometimes the pivotal RCT is not ongoing and comparative effectiveness uncertainties remain. In this situation, it is highly unlikely to be feasible (or, possibly, ethical) to recruit patients into an RCT to address NICE's uncertainties and instead there may be a need to use NCRAS data.

In addition, RCTs have strict and usually restrictive eligibility criteria. Frequently they are not representative of the general population. Hence, whilst RCTs might tell us about the comparative effectiveness of treatments in highly selected trial populations, they are less useful for investigating comparative effectiveness in more general cancer populations.

For these reasons, it is important to investigate the use of NCRAS data as a resource for estimating the comparative effectiveness of cancer treatments in the "real world", that is, under routine conditions. This project aims to investigate whether or not English cancer registry data is sufficient for deriving valid causal estimates of the comparative effectiveness of different cancer treatments given in the NHS. A case study comparing adjuvant and metastatic pancreatic cancer treatments will be used. Analysis will be undertaken using causal inference methods in a Target Trial framework.[2] Results will be compared to those found in recently published RCTs to assess the reliability of the analyses.

Analysing observational data

Estimating comparative effectiveness using observational data is known to be prone to important biases – those present due to the absence of randomisation. For example, confounding by indication at baseline is an important issue, because the treatment that

patients receive may be strongly influenced by their prognostic characteristics, creating a selection bias. In addition, time-dependent confounding can also be an important issue, if treatments change over time and simultaneously affect the confounding variable. Hence, it is necessary to use advanced causal inference analytical methods such as g-methods in an attempt generate comparability between treatment groups ("exchangeability") to avoid bias.[3]

Traditional statistical methods such as multivariate regression analysis or propensity score matching fail in the presence of time-dependent confounding – they cannot control adequately for time-dependent confounding variables.[2] Marginal structural models, which incorporate inverse probability weighting, a 'g-method' developed by Robins and colleagues,[4,5] are able to adjust appropriately for baseline and time-dependent confounding. G-methods require the assumption of "no unmeasured confounding" – any patient characteristics that influence the treatment choice and the outcome of interest must be measured and included in the analysis. Hence, data collection is critical and we will test the adequacy of the data included in the NCRAS datasets for conducting causal analyses in our study. In addition, it is critical that there is some overlap in patient characteristics with respect to patients receiving different treatments; that is, patients with similar prognostic characteristics should have received different treatments (this is known as the "positivity" assumption). Thus, we will need data not only on treatment received, cancer type and stage, and relevant outcomes (such as survival times), but also on any potentially prognostic information measured at baseline such as age, sex, diagnosis date, or excision margin, and on prognostic information measured over time – for example, biomarker values, tumour size, clinical signs/symptoms, performance status, and hospital episodes. If data on these characteristics is inadequately captured in NCRAS data, our treatment effect estimates may be subject to bias.

Target Trial Framework

Hernan and Robins recently introduced their "Target Trial" framework for conducting comparative effectiveness analyses using observational data.[2] The framework is based on the rationale that if, for any reason, an RCT cannot be run, observational data analysis should be designed so as to emulate the RCT that would have been run had it been possible. A key aim of the framework is to protect against time-related biases (e.g., immortal time bias) that be particularly problematic in analyses of observational data. The framework outlines seven key components to the research design:

- Eligibility criteria
- Treatment strategies
- Assignment procedures
- Follow-up period
- Outcome
- Causal contrasts of interest
- Analysis plan

The Target Trial framework is currently being used in the United States to assess whether US cancer registry datasets are suitable for estimating the comparative effectiveness of different cancer treatments, primarily as part of the RCT DUPLICATE project.[6-11]. These ongoing studies are attempting to emulate existing RCTs (sometimes referred to as "benchmark studies") using registry data, as a way of testing whether comparative effectiveness analyses based on the registry data are reliable. This means that (1) the analysis is restricted, as far as possible, to the population included in the relevant RCT, and (2) the analysis is conducted using a similar design to that used in the RCT. If the registry-based analysis provides similar results to that observed in the RCT, we may have

confidence that we can infer estimates of causal treatment effects from the registry analysis. This then provides confidence that we could use the registry data to address questions that were not answered by the RCT – for example, estimating effectiveness in patients who do not meet the strict eligibility criteria typically used in RCTs.

In this study our primary aim is to determine whether English cancer registry datasets are suitable for estimating the comparative effectiveness of different cancer treatments. We will investigate this by emulating existing RCTs using the NCRAS data, following the benchmarking approach used in RCT DUPLICATE.[9-11] However, we also recognise that analyses of registry data allows questions to be investigated that have not been addressed by RCTs. Therefore, for each emulated RCT, we will undertake further analyses that are not constrained by the eligibility criteria and follow-up times used in the existing RCT, broadening the populations included and using unrestricted follow-up times, allowing estimation of treatment effects that are applicable to broader populations.

Analysis Plan

We have identified pancreatic cancer as a suitable disease area for undertaking Target Trial analyses using NCRAS data. In the following section, we justify this choice, and provide background information on pancreatic cancer and treatment options in England. We then specify four Target Trial emulation analyses that we will undertake. Finally we specify the NCRAS data required to perform these analyses.

Pancreatic Cancer

In 2016, approximately 10,000 patients were diagnosed with pancreatic cancer in the United Kingdom,[12] and often pancreatic cancer is diagnosed at an advanced stage.[13] The prognosis is poor even for patients diagnosed at an early stage of pancreatic cancer, where surgical resection is possible, with 5-year survival rates estimated at between 7% and 25%.[14] Survival rates are extremely poor for patients with metastatic disease, with median survival of between 2 and 6 months if untreated.[13]

A NICE Guideline on the diagnosis and management of pancreatic cancer, published in 2018, recommends that gemcitabine plus capecitabine should be offered as adjuvant treatment for patients who have had sufficient time to recover after pancreatic cancer resection.[15] Gemcitabine monotherapy should be considered for patients who are not well enough to tolerate combination chemotherapy. FOLFIRINOX, a combination regimen consisting of oxaliplatin, inrinotecan, leucovorin and fluorouracil, is not mentioned in the NICE guideline, but is beginning to be offered as adjuvant treatment in the NHS, due to trial results published in December 2018.[16]

For metastatic pancreatic cancer, the NICE guideline recommends that FOLFIRINOX should be offered to patients with an Eastern Cooperative Oncology Group (ECOG) performance status of 0-1.[15] Gemcitabine combination therapy should be considered for patients not well enough to tolerate FOLFIRINOX, with the first combination option being gemcitabine plus capecitabine.[13,15] For patients for whom FOLFIRINOX and gemcitabine plus capecitabine are unsuitable gemcitabine plus nab-paclitaxel is an option.[13] Gemcitabine monotherapy should be offered to patients not well enough to tolerate combination chemotherapy.[15]

These guidelines seem to present a clear hierarchy of treatments for adjuvant and metastatic pancreatic cancer, and seem to suggest that there might be little overlap in prognostic characteristics of patients receiving different treatments. However, the NICE technology appraisal of gemcitabine plus nab-paclitaxel notes that some patients for whom FOLFIRINOX is otherwise suitable choose not to have this treatment because of its considerable toxicity.[13] Further, it is noted that the current treatment options have a number of limitations, including serious adverse effects – in particular, the most effective treatment option (FOLFIRINOX) is associated with the most significant adverse events, whereas the least effective (gemcitabine monotherapy) is associated with the least significant adverse events.[13] In addition, it is unfortunately the case that prognosis remains poor even with the most effective treatment. Therefore, it is likely that due to patient choice, there will be overlap in prognostic characteristics between patients who receive FOLFIRINOX and patients who receive gemcitabine for metastatic pancreatic cancer. Similarly, because gemcitabine combination therapies have lower effectiveness and toxicity than FOLFIRINOX, and higher effectiveness and toxicity than gemcitabine monotherapy, it is likely that there is some overlap in prognostic characteristics between patients who receive FOLFIRINOX, gemcitabine combination therapies, or gemcitabine monotherapy. The NICE technology appraisal guidance for gemcitabine plus nab-paclitaxel states that there is evidence of use of gemcitabine doublet chemotherapy for pancreatic cancer in the NHS.[13]

Similar is likely to be true for adjuvant treatment for pancreatic cancer, where gemcitabine plus capecitabine is more effective than gemcitabine monotherapy, but where toxicity is lower for the monotherapy option and prognosis is relatively poor with both treatment options.

Hence, it is likely that there is variation in treatments received for adjuvant and metastatic pancreatic cancer in the NHS, with an overlap in characteristics of patients receiving different treatments. This echoes clinical expert opinion from Professor Jonathan Wadsley, who states that for both adjuvant and metastatic pancreatic cancer there is substantial overlap between patients receiving different treatments. For adjuvant treatment, Professor Wadsley believes that due to the additional side effects and limited increase in effectiveness associated with combination treatment, some patients choose gemcitabine monotherapy instead of gemcitabine plus capecitabine, and in fact some patients choose no treatment at all. For metastatic disease, Professor Wadsley believes that treatment with gemcitabine monotherapy remains common, with patients choosing it instead of the highly toxic FOLFIRINOX regimen, whilst some patients receive gemcitabine combination therapy.

To be able to infer causal estimates for the comparative effectiveness of different treatment options in registry data there needs to be some overlap in prognostic characteristics between patients receiving the different treatments ("positivity"). Based on statements made by clinical and patient experts in NICE technology appraisal documents and information from a practicing clinician who treats patients with pancreatic cancer, we are confident that such positivity/overlap exists for the treatment of both adjuvant and metastatic pancreatic cancer in the NHS.

Target Trial Analyses

We have identified four pancreatic cancer trials that we will try to replicate using NCRAS data, using Hernan and Robins' Target Trial [2] framework.[17-20].

For each Target Trial, multiple sets of analyses will be completed. Analysis Set 1 will be undertaken whereby the population analysed will match that included in the RCT being emulated as closely as possible, based on the eligibility criteria of the RCT. These analyses will be compared to the RCT results, allowing us to determine whether or not it has been possible to successfully emulate the RCT. Analysis Set 2 will consider a broader population, not restricted to criteria around characteristics such as age and performance status specified by the RCT. For example, in Target Trial 1, the ESPAC-4 RCT included strict eligibility criteria (shown in the Table below). In Analysis Set 1 we will attempt to replicate the trial population as closely as possible using these eligibility criteria. In Analysis Set 2 we will include all patients aged 18 or older who received adjuvant treatment for pancreatic cancer with gemcitabine monotherapy or gemcitabine plus capecitabine, irrespective of other eligibility criteria (such as treatment within 12 weeks of having curative surgery, performance status, or history of cancer/treatment). Analysis Set 1 will allow us to compare results to the emulated benchmark RCT, whereas Analysis Set 2 will allow us to estimate the effectiveness of treatment in a more general real-world population.

In addition, other analysis sets (denoted Analysis Set 3+) may be developed for each Target Trial depending on the characteristics of the data provided. For example, if missing data means that one or more eligibility criteria results in a drastic reduction in patient numbers, analyses will be run with and without including those eligibility criteria. Similarly, if several eligibility criteria are problematic to emulate, analyses will be run using those eligibility criteria considered by clinical experts to be most important. This will allow us to identify key issues associated with variables included in (and excluded from) the NCRAS datasets. For each eligibility criteria we will report our emulation approach, and any assumptions or issues

associated with this (for example, whether proxy variables were required, and whether missing data was an issue). We will therefore be transparent around the extent to which emulation was possible for each Target Trial.

Within each Analysis Set, a number of analyses will be run. It is anticipated that analyses will require adjustment for baseline and time-dependent confounding variables. Available variables and data will be presented to clinical experts and variables used to adjust for baseline confounding will be selected based upon discussion using directed acyclic graphs as a decision aid. It is anticipated that scenario and sensitivity analyses will be carried out using "complete" models (that include all variables considered to be potential confounders), and "reduced" models (that include variables considered to be the most important confounders). Potential residual confounding due to missing data or missing variables will be discussed and reported. In addition, when emulating existing RCTs as closely as possible, it is important to use minimum and maximum follow-up times that match those used in the RCTs. However, this would mean excluding longer-term data that may be available in the NCRAS data. Therefore, within each Analysis Set we will run analyses with minimum and maximum follow-up times matching those in the target trial, but also with no restriction on follow-up times. Finally, when weighting methods are used to adjust for confounding, it is possible to use stabilised or unstabilised weights – we will conduct analyses using both techniques.[4,5]

The Analysis Sets we will include are described in Table 1, below.

Table 1. Analysis sets to be included in Target Trial analyses

Analysis Group	Analysis characteristics	Weighting technique
Analysis Set 1: Emulating the existing benchmark RCT as closely as possible	With "complete" adjustment models	With stabilised weights
		With unstabilised weights
	With "reduced" adjustment models	With stabilised weights
		With unstabilised weights
	With minimum and maximum follow-up times matching those in the target RCT	With stabilised weights
		With unstabilised weights
	With no restriction on minimum and maximum follow-up times	With stabilised weights
		With unstabilised weights
Analysis Set 2: Estimating comparative effectiveness of the treatments investigated in the RCT in a broader population (to be defined more specifically for each Target Trial)	With "complete" adjustment models	With stabilised weights
		With unstabilised weights
	With "reduced" adjustment models	With stabilised weights
		With unstabilised weights
	With minimum and maximum follow-up times matching those in the target RCT	With stabilised weights
		With unstabilised weights
	With no restriction on minimum and maximum follow-up times	With stabilised weights
		With unstabilised weights
Analysis Set 3+: Emulating the existing RCT partially, where specific problems are identified with the emulation – for example when one or more eligibility criteria are problematic to emulate. The specifics of this Analysis Set will be	With "complete" adjustment models	With stabilised weights
		With unstabilised weights
	With "reduced" adjustment models	With stabilised weights
		With unstabilised weights
	With minimum and maximum follow-up times matching those in the target RCT	With stabilised weights
		With unstabilised weights
	With no restriction on minimum and maximum follow-up times	With stabilised weights
		With unstabilised weights

determined when data have been received – but before any analyses are undertaken		

Evaluating Emulation Success

The success of our Target Trial emulations will be based on comparisons of the results of analyses contained within Analysis Set 1 with results published for each of the benchmark RCTs. If problems with emulating specific eligibility criteria mean that analyses contained within Analysis Set 1 are unreliable or highly uncertain, benchmark comparisons may also be made using analyses contained within Analysis Set 3+.

Four assessment criteria will be used to assess alignment between the results of the benchmark RCTs and their emulated counterparts. In each of the existing benchmark RCTs that we will seek to emulate, the primary endpoint was overall survival, and therefore all our assessments of alignment will be based on overall survival estimates. Criteria 1-3 are based on the criteria used in the RCT DUPLICATE project,[9-11] and involve an assessment of relative treatment effects – that is, the hazard ratio (HR) for overall survival. We have added Criterion 4 in order to examine absolute outcomes, as we wish to investigate whether our emulated trials result in similar estimates of relative treatment effects *and* absolute survival outcomes. We believe this is important because there is a possibility that emulated trials could produce similar estimates of relative effects, whilst absolute outcomes could differ considerably, indicating sub-optimal emulation. This approach is similar to a recently published Target Trial benchmarking study published by Chang *et al.*[21]

Criterion 1: Regulatory Agreement. This assesses whether the emulated Target Trial using NCRAS data replicates the benchmark RCT's results with respect to the direction and statistical significance of the HR for overall survival.

Criterion 2: Estimate Agreement. This assesses whether the point estimate of the HR for overall survival estimated by the emulated Target Trial using NCRAS data falls within the 95% confidence interval (CI) of the benchmark RCT.

Criterion 3. Standardised Differences. This criterion assesses whether there is a statistically significant difference in the HR for overall survival estimated by the emulated Target Trial and the benchmark RCT, based on standardised differences, calculated as:

$$Z = \frac{\hat{\theta}_{NCRAS} - \hat{\theta}_{RCT}}{\sqrt{\hat{\sigma}^2{}_{NCRAS} + \hat{\sigma}^2{}_{RCT}}}$$

where $\hat{\theta}$ are treatment effect estimates from the NCRAS and benchmark RCT analyses, and $\hat{\sigma}^2$ the associated variances. The null hypothesis of no difference between the treatment effects will be rejected if $|Z| > 1.96$.[9]

Criterion 4. Absolute Survival Curve Agreement. This assesses whether the point estimates (over time) of the Kaplan-Meier survival curve estimated by the emulated Target Trial fall within the 95% CI of the Kaplan-Meier curve for the benchmark RCT. To assess this, we will reconstruct patient-level survival data for each of the benchmark RCTs using published Kaplan-Meier curves and Guyot et al.'s digitisation method,[22] allowing us to re-create the Kaplan-Meier curves from each study with the addition of confidence intervals (since CIs for Kaplan-Meier curves were not included in any of the study publications).

Target Trial Components

Details of the Target Trial components, under the headings used by Hernan and Robins, are presented for each of the four Target Trials in the following four tables.

Target Trial 1. Comparing gemcitabine monotherapy with gemcitabine plus capecitabine in patients with adjuvant pancreatic cancer

Trial	ESPAC-4. Comparing gemcitabine monotherapy with gemcitabine plus capecitabine in patients with adjuvant pancreatic cancer [17]	Target Trial 1. Emulation of ESPAC-4 using NCRAS data
Eligibility criteria	Patients aged 18 or older who had undergone complete macroscopic resection for ductal adenocarcinoma of the pancreas (R0 or R1 resection) with histological confirmation and with no evidence of malignant ascites, live or peritoneal metastatis, or spread to other distant abdominal, or extra-abdominal organs. A clear CT scan of the chest, abdomen, and pelvis was required within 3 months before randomisation. No restriction was placed on randomisation on the basis of postoperative carbohydrate antigen 19-9 (CA19-9) concentrations. Other specific inclusion criteria were full recovery from surgery, randomised within 12 weeks of surgery, a WHO performance score of two or less, creatinine clearance of at least 50 mL/min, and a life	Analysis Set 1: Target Trial eligibility criteria: to match ESPAC-4 as far as possible. Patients aged 18 or older who had undergone complete macroscopic resection for ductal adenocarcinoma of the pancreas (R0 or R1 resection) and had TNM stage I, II, or III disease. ECOG performance status of 2 or less (ECOG is the same measure as the WHO performance score), and who started either of the treatments studied in ESPAC-4. Patients who had previously had chemotherapy and with pancreatic lymphoma, macroscopically remaining tumours (R2 resection), or TNM stage IV disease to be excluded. No previous or concurrent malignancy, except basal cell carcinoma of skin, carcinoma in situ of cervix. Permitted tumour locations will be based on ICD-10 codes. Presence (and therefore absence) of metastases will be based on recorded stage of disease. Completion of R0 or R1 resection will be based on data on excision margins, the OPCS-4 classification of interventions and procedures received, and recorded surgical interventions (which are classified in cancer registry datasets as “curative”, “non curative” or “type unknown”). Previous cancers and treatments will be identified from the cancer registry and SACT datasets. For criteria related to comorbidities, Charlson scores will be used where relevant. It is expected that it will not be possible to emulate all criteria completely – for each criteria the approach used for emulation will be recorded and reported. Clinical expert assistance will be used when proxy variables are required. Minimum follow-up in ESPAC-4 was 18 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated

	expectancy of more than 3 months. Patients who were pregnant, or who had previously had chemotherapy and with pancreatic lymphoma, macroscopically remaining tumours (R2 resection), or TNM stage IV disease were excluded. No previous or concurrent malignancy diagnoses (except curatively-treated basal cell carcinoma of skin, carcinoma in situ of cervix).	treatment 18 months or longer before the cut-off date of the NCRAS data currently available. Analysis Set 2: Patients aged 18 or older who receive adjuvant treatment with gemcitabine monotherapy or gemcitabine plus capecitabine for pancreatic cancer.
Treatment strategies	Patients were eligible to be randomised if curative surgery had been received within the last 12 weeks, with treatment then starting within 2 weeks of randomisation. Randomisation was to receive gemcitabine or gemcitabine plus capecitabine. Gemcitabine was delivered as a 1000 mg/m² intravenous infusion administered once a week for three of every 4 weeks (one cycle) for six cycles (24 weeks). Capecitabine was administered orally for 21 days followed by 7 days' rest (one cycle) for six cycles (24 weeks) at a daily dose of 1660 mg/m².	Treatment to have begun within 12 weeks of curative surgery. Treatment strategies are initiation of gemcitabine monotherapy, or initiation of gemcitabine plus capecitabine. Patients who meet the eligibility criteria set out above but did not initiate gemcitabine or gemcitabine plus capecitabine are not relevant for the analysis and are excluded. Time zero will be the time of initiation of gemcitabine monotherapy or gemcitabine plus capecitabine, with the restriction that that time-point must fall within 14 weeks of their curative surgery (matching the 12+2 weeks used in the trial as the period in which treatment could be initiated). In ESPAC-4 there could be a 2-week lag between randomisation and treatment initiation. This represents an aspect of the trial that cannot be perfectly emulated, which could cause differences in analytical results. We cannot emulate this 2 week "grace period" because we will not have an intention-to-treat (ITT) date. Therefore, we must use the time of treatment initiation as time zero. This has two implications: a) All patients in our emulated analysis initiated one of the target trial treatments. In ESPAC-4, 1 out of 366 patients randomised to gemcitabine and 6 out of 364 patients randomised to gemcitabine + capecitabine did not receive study treatment; b) Survival analysis (e.g. Kaplan-Meier curves and hazard ratio estimates) in ESPAC-4 included time up to 2 weeks before treatment initiation, whereas

		in our emulated analyses these analyses will begin at the time of treatment initiation. The potential impacts of these emulation imperfections will be discussed in analysis reports.
Assignment procedures	Eligible patients were randomly assigned (1:1) to receive gemcitabine or gemcitabine plus capecitabine within 12 weeks of surgery. Randomisation was based on a minimisation routine with a random element of 20%. Resection margin (negative or positive) and country were used as stratification factors. Participants and study investigators were not masked to treatment allocation.	To emulate the random assignment of strategies at baseline, we need to adjust for all confounding factors required to ensure comparability (exchangeability) of the groups defined by initiation of the treatment strategies. This will be performed using covariate adjustment using all potentially prognostic variables available at the time of treatment initiation. In ESPAC-4, univariate survival analyses showed that smoking, preoperative, and postoperative CA19-9 concentrations, preoperative C-reactive protein concentrations, resection margin status, tumour grade, lymph nodes status, maximum tumour size, tumour stage, venous resection, and local invasion were all associated with survival, whilst a multivariable model identified resection margin status, postoperative CA19-9 concentrations, tumour grade, lymph node status, and maximum tumour size as significant independent factors of overall survival. These variables will not all be available in the NCRAS datasets. Available variables and data will be presented to clinical experts and variables used to adjust for baseline confounding will be selected based upon discussion using directed acyclic graphs as a decision aid. It is anticipated that scenario and sensitivity analyses will be carried out using “complete” models (that include all variables considered to be potential confounders), and “reduced” models (that include variables considered to be the most important confounders). Potential residual confounding due to missing data or missing variables will be discussed and reported. Participants and investigators were not blinded in ESPAC-4, and therefore for the trial emulation it is not a problem that we cannot emulate blinding.
Follow-up period	Randomisation was carried out between Nov 10, 2008, and Sept 11, 2014, with data cut-off on March 9, 2016. Patients alive and still in follow-up at 5 years were censored at that point.	Minimum follow-up in ESPAC-4 was 18 months. The maximum possible follow-up was 88 months, with published Kaplan-Meier curves ending at 80 months. Therefore, for our main analysis, patients are only included in our trial emulation if they initiated treatment 18 months or longer before the cut-off date of the NCRAS data available, and patients remaining alive at 80 months will be censored.

		Supplementary analyses will be included that do not place restrictions on minimum or maximum follow-up times.
Outcome	The primary outcome in ESPAC-4 was overall survival, measured as the time from randomisation until death from any cause.	Overall survival, measured as the time from treatment initiation until death from any cause (subject to the minimum and maximum follow-up restrictions referred to in the "Follow-up period" section of this table).
Causal contrasts of interest	The primary effect measure used was the overall survival hazard ratio (HR) between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 12 months and 24 months were presented for both treatment arms. Analyses were undertaken on an ITT basis, i.e. the comparative effect of being assigned to the treatment strategies at baseline, irrespective of any protocol deviations with the exception of patients who withdrew consent between randomisation and the start of therapy. A per-protocol treatment effect was also estimated but results are not reported in the trial publication.	The emulated primary effect measure will be the overall survival HR between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 12 months and 24 months will also be presented for both treatment arms, for each of the analyses included in "Analysis plan". Analyses will represent an analogue of the ITT effect – i.e. the comparative effect will be estimated according to treatment strategy initiated irrespective of whether these strategies continued to be followed after initiation. An analogue of a per-protocol effect will also be estimated, to represent the effect according to if patients followed treatment pathways that are representative of those followed in ESPAC-4. It is possible that treatment pathways followed in the cancer registry dataset will deviate from the treatment pathways received in ESPAC-4, if patients in the registry data switch onto treatments that were not available or were not commonly used during the conduct of ESPAC-4. ESPAC-4 publications report some information on post-study treatments received, and these will be compared to subsequent treatments received by patients identified in the NCRAS data. Clinical expert opinion will be sought to determine which treatment switches represent deviations from the treatment pathways received in ESPAC-4. Hence, the purpose of our per-protocol analysis is to develop an analysis that more closely emulates the primary ITT analysis used in ESPAC-4, if the treatment pathways present in the cancer registry dataset do not adequately resemble those followed in ESPAC-4. We will also examine the extent to which treatment received in the NCRAS dataset reflect the treatment received in ESPAC-4 – for example, with respect to duration of treatment. As previously noted, the *intention* to treat cannot be perfectly emulated, and the time zero used in

		our emulation does not perfectly match the time zero used in ESPAC-4 (because there could be up to a 2-week lag between randomisation and initiation of study treatment). Therefore, our ITT analogue has imperfections. However, given the relatively short 2-week "grace period" used in ESPAC-4, and given that 99.7% of patients assigned to gemcitabine, and 98.6% of patients assigned to gemcitabine + capecitabine, received their study treatment, we expect the impact of these imperfections to be minor.
Analysis plan	All efficacy analyses were done in the ITT population retaining all patients in their initially randomised groups irrespective of any protocol deviations with the exception of patients who withdrew consent between randomisation and the start of therapy. A per-protocol analysis was also conducted but results are not reported in the trial publication. A Cox proportional hazards model was used to estimate the overall survival HR, with country and resection margin as stratification factors. Confidence intervals were presented. A log-rank test (stratified by country and resection margin) was used to test for a statistically significant difference in survival. Kaplan-Meier survival curves, median survival, and 12- and 24-month survival proportions were presented for each treatment arm. Confidence intervals	Analysis sets will be undertaken as detailed in Table 1 (Analysis sets to be included in Target Trial analyses). Analysis Set 1 will emulate the target trial as closely as possible. Analysis Set 2 will consider a broader population, encompassing patients aged 18 or older who receive adjuvant treatment for pancreatic cancer. Other analysis sets (denoted Analysis Set 3+) will be developed depending on the data available. For example, if missing data means that one or more eligibility criteria results in a drastic reduction in patient numbers, analyses will be run with and without including those eligibility criteria. Similarly, if several eligibility criteria are problematic to emulate, analyses will be run using those eligibility criteria considered by clinical experts to be most important. For each Analysis Set a number of analyses will be run: - With "complete" adjustment models (see "Assignment procedures, above) - With "reduced" adjustment models (see "Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights. For Analysis Set 1 (and for Analysis Set 3+, if this analysis set is required due to problems emulating one or more eligibility criteria), analyses will be undertaken using the ITT and per-protocol analogues described in "Causal contrasts of interest".

	were reported for median survival, and 12- and 24-month survival proportions.	The ITT analysis analogue will estimate the comparative effect according to the treatment strategy initiated, irrespective of whether these strategies continued to be followed after initiation. The per-protocol analysis analogue will estimate the comparative effect adjusting for any treatment switches that occur in the NCRAS data that are not representative of treatment pathways received by patients in ESPAC-4. Both the ITT and per-protocol analyses included in Analysis Set 1 (and Analysis Set 3+, if required) will be subject to the minimum and maximum follow-up restrictions referred to in the “Follow-up period” section of this table. For the ITT-based analysis, inverse probability weighting will be used to adjust for baseline confounders. For the per-protocol analysis, patients who deviate from the defined treatment strategies will be censored at that time-point and therefore adjustment for baseline and post-baseline confounding is necessary. Inverse probability weighting using time-varying weights will be used for this purpose. For the analogue of the ITT analysis and for the per-protocol analysis it is possible that selection bias could be present due to informative loss to follow up. If this is apparent, inverse probability of censoring weighting using time varying weights will be used. These weights will be combined with the weights used to adjust for baseline confounding in the ITT-based analysis, and with the time-dependent weights used to address treatment deviations in the per-protocol analysis. For each analysis, Cox models that incorporate inverse probability weights to adjust for baseline (and where relevant, time-dependent) confounding will be used to estimate overall survival HRs and the log-rank test will be used to test for differences in survival. Where it is necessary to attempt to control for time-dependent confounding, marginal structural Cox models will be used. The HRs will be compared to the HRs for overall survival estimated in ESPAC-4. These HR estimates will be used to assess emulation agreement, using agreement criteria 1-3 described in the “Evaluating Emulation Success” section of this protocol. Agreement criterion 4 will be assessed by comparing the Kaplan-Meier survival curves

		presented in the ESPAC-4 publication (digitised and with confidence intervals added, as described in the “Evaluating Emulation Success” section of this report) to weighted Kaplan-Meier curves constructed for each analysis and analysis set previously described. The ESPAC-4 publication also reported median overall survival (with confidence intervals) and survival proportions at 12- and 24-months. We will report these statistics for our emulated analyses to allow further assessment of agreement between the results of our emulation and those reported for ESPAC-4. However, as previously stated, it is the overall survival HR that will be used to formally assess agreement criteria 1-3, because overall survival was the primary endpoint in ESPAC-4 and the study was designed based on this HR effect measure. Stratification factors of country and resection margin were used in the Cox model used to estimate the HR for overall survival in ESPAC-4. Country is not relevant for our emulated trial. Resection margin will be included as a stratification factor in our analyses if data on these margins are available. Analysis Set 2 is purposely not comparable to ESPAC-4, as it will include a broader population. As such, for this analysis we will not draw formal comparisons to ESPAC-4 results, and per-protocol analogues designed to be consistent with treatment pathways received in ESPAC-4 are not necessary. Therefore, for Analysis Set 2, only the ITT analogue analysis will be undertaken. However, as for Analysis Sets 1 and 3+ (if required), the overall survival HR, median survival, Kaplan-Meier survival curves, and survival proportions at 12- and 24-months will be reported. Also, a range of sensitivity and scenario analyses will be reported, as previously described: - With “complete” adjustment models (see “Assignment procedures, above) - With “reduced” adjustment models (see “Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights.

Notes: CT: Computed tomography; WHO: World Health Organisation; TNM: Tumour, nodes, metastasis, Classification of Malignant Tumours; ECOG: Eastern Cooperative Oncology Group; ICD: International Classification of Diseases; OPCS: Office

of Population Censuses and Surveys; SACT: Systemic Anti-Cancer Therapy; ITT: Intention-to-treat; NCRAS: National Cancer Registration and Analysis Service; HR: Hazard ratio

Target Trial 2. Comparing FOLFIRINOX to gemcitabine in patients with metastatic pancreatic cancer

Trial	ACCORD. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer [18]	Target Trial 2. Emulation of ACCORD using NCRAS data
Eligibility criteria	Patients were eligible to be included in the study if they were 18 years of age or older and had histologically and cytologically confirmed, measurable metastatic pancreatic adenocarcinoma that had not previously been treated with chemotherapy. Other inclusion criteria were an Eastern Cooperative Oncology Group (ECOG) performance status score of 0 or 1 and adequate bone marrow (granulocyte count, ≥1500 per cubic millimeter; and platelet count, ≥100,000 per cubic millimeter), liver function (bilirubin ≤1.5 times the upper limit of the normal range), and renal function. Exclusion criteria were an age of 76 years or older, endocrine or acinar pancreatic carcinoma, previous radiotherapy for anterior abdominal measurable lesions, previous chemotherapy, cerebral metastases, a history of another major cancer (i.e. except cancer in situ of the cervix, skin	Analysis Set 1: Target Trial eligibility criteria: to match ACCORD as far as possible. Patients aged 18-75 with metastatic pancreatic adenocarcinoma (TNM stage IV) that had not been treated previously with chemotherapy, and who started either of the treatments studied in ACCORD. ECOG performance status of 0 or 1. Patients will be excluded if they have endocrine or acinar pancreatic carcinoma, previous radiotherapy for measurable lesions, cerebral metastases, a history of another major cancer (i.e. except cancer in situ of the cervix, skin cancer). Permitted tumour locations will be based on ICD-10 codes. Presence of metastases will be based on recorded stage of disease. Previous cancers and treatments will be identified from the cancer registry and SACT datasets. For criteria related to comorbidities, Charlson scores will be used where relevant. It is expected that it will not be possible to emulate all criteria completely – for each criteria the approach used for emulation will be recorded and reported. Clinical expert assistance will be used when proxy variables are required. Minimum follow-up in ACCORD was 6 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 6 months or longer before the cut-off date of the NCRAS data currently available. Analysis Set 2: Patients aged 18 or older who received treatment with FOLFIRINOX or gemcitabine for metastatic pancreatic cancer.

	cancer), pregnant or breast feeding women, active infection, chronic diarrhea, a clinically significant history of cardiac disease, and pregnancy or breast-feeding.	
Treatment strategies	Patients were assigned to receive FOLFIRINOX or gemcitabine. Gemcitabine, at a dose of 1000 mg per square meter of body-surface area, was delivered by 30-minute intravenous infusion weekly for 7 weeks, followed by a 1-week rest, then weekly for 3 weeks in subsequent 4-week courses. FOLFIRINOX consisted of oxaliplatin at a dose of 85 mg per square meter, given as a 2-hour intravenous infusion, immediately followed by leucovorin at a dose of 400 mg per square meter, given as a 2-hour intravenous infusion, with the addition, after 30 minutes, of irinotecan at a dose of 180 mg per square meter, given as a 90-minute intravenous infusion through a Y-connector. This treatment was immediately followed by fluorouracil at a dose of 400 mg per square meter, administered by intravenous bolus, followed by a continuous intravenous infusion	Treatment strategies are initiation of FOLFIRINOX, or initiation of gemcitabine monotherapy. Patients who meet the eligibility criteria set out above but did not initiate FOLFIRINOX or gemcitabine are excluded from the analysis. Time zero will be the time of initiation of FOLFIRINOX or gemcitabine, with the restriction that that time-point must fall at a point at which eligibility criteria are satisfied. In ACCORD there could be a 1-week lag between randomisation and treatment initiation. This represents an aspect of the trial that cannot be perfectly emulated, which could cause differences in analytical results. We cannot emulate this 1 week “grace period” because we will not have an intention-to-treat (ITT) date. Therefore, we must use the time of treatment initiation as time zero. This has two implications: a) All patients in our emulated analysis initiated one of the target trial treatments. In ACCORD, 4 out of 171 patients randomised to FOLFIRINOX and 2 out of 171 patients randomised to gemcitabine did not receive study treatment; b) Survival analysis (e.g. Kaplan-Meier curves and hazard ratio estimates) in ACCORD included time up to 1 week before treatment initiation, whereas in our emulated analyses these analyses will begin at the time of treatment initiation. The potential impacts of these emulation imperfections will be discussed in analysis reports.

	of 2400 mg per square meter over a 46-hour period every 2 weeks. Treatment was to be initiated within 1 week of enrolment.	
Assignment procedures	Patients were randomly assigned to receive FOLFIRINOX or gemcitabine within 1 week after enrollment. Randomisation was performed centrally in a 1:1 ratio with stratification according to center, performance status (0 vs. 1), and primary tumour localisation (the head vs. the body or tail of the pancreas).	To emulate the random assignment of strategies at baseline, we need to adjust for all confounding factors required to ensure comparability (exchangeability) of the groups defined by initiation of the treatment strategies. This will be performed using inverse probability weighting using all potentially prognostic variables available at the time of treatment initiation. In ACCORD, randomisation was stratified according to ECOG performance status and primary tumour location. In addition, synchronous metastases, a low baseline albumin level, hepatic metastases, and an age of more than 65 years were identified as independent adverse prognostic factors for overall survival. These variables – or potential proxies for them – will be considered for inclusion in our analysis. Not all relevant variables will not all be available in the NCRAS datasets. Available variables and data will be presented to clinical experts and variables used to adjust for baseline confounding will be selected based upon discussion using directed acyclic graphs as a decision aid. It is anticipated that scenario and sensitivity analyses will be carried out using “complete” models (that include all variables considered to be potential confounders), and “reduced” models (that include variables considered to be the most important confounders). Potential residual confounding due to missing data or missing variables will be discussed and reported.
Follow-up period	Randomisation was carried out between December 2005, and October 2009, with data cut-off on April 16 2010. Patients were followed until death or were censored at April 16 2010 if alive at that point.	Minimum follow-up in ACCORD was 6 months. The maximum possible follow-up was 52 months, with published Kaplan-Meier curves ending at 48 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 6 months or longer before the cut-off date of the NCRAS data available, and patients remaining alive at 48 months will be censored. Supplementary analyses will be included that do not place restrictions on minimum or maximum follow-up times.
Outcome	The primary outcome in ACCORD was	Overall survival, measured as the time from treatment initiation until death from any cause

	overall survival, measured as the time from randomisation until death from any cause	(subject to the minimum and maximum follow-up restrictions referred to in the “Follow-up period” section of this table).
Causal contrasts of interest	The primary effect measure used was the overall survival hazard ratio (HR) between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 6, 12 and 18 months were presented for both treatment arms. Analyses were undertaken on an ITT basis, i.e. the comparative effect of being assigned to the treatment strategies at baseline, irrespective of any protocol deviations.	The emulated primary effect measure will be the overall survival HR between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 6, 12, and 18 months will also be presented for both treatment arms, for each of the analyses included in “Analysis plan”. Analyses will represent an analogue of the ITT effect – i.e. the comparative effect will be estimated according to treatment strategy initiated irrespective of whether these strategies continued to be followed after initiation. An analogue of a per-protocol effect will also be estimated, to represent the effect according to if patients followed treatment pathways that are representative of those followed in ACCORD. It is possible that treatment pathways followed in the cancer registry dataset will deviate from the treatment pathways received in ACCORD, if patients in the registry data switch onto treatments that were not available or were not commonly used during the conduct of ACCORD. ACCORD publications report some information on post-study treatments received, and these will be compared to subsequent treatments received by patients identified in the NCRAS data. Clinical expert opinion will be sought to determine which treatment switches represent deviations from the treatment pathways received in ACCORD. Hence, the purpose of our per-protocol analysis is to develop an analysis that more closely emulates the primary ITT analysis used in ACCORD, if the treatment pathways present in the cancer registry dataset do not adequately resemble those followed in ACCORD. We will also examine the extent to which treatment received in the NCRAS dataset reflect the treatment received in ACCORD – for example, with respect to duration of treatment. As previously noted, the *intention* to treat cannot be perfectly emulated, and the time zero used in our emulation does not perfectly match the time zero used in ACCORD (because there could be up to a 1-week lag between randomisation and initiation of study treatment). Therefore, our ITT analogue has imperfections. However, given the

		relatively short 1-week “grace period” used in ACCORD, and given that 97.7% of patients assigned to FOLFIRINOX, and 98.8% of patients assigned to gemcitabine, received their study treatment, we expect the impact of these imperfections to be minor.
Analysis plan	All efficacy analyses were done in the ITT population retaining all patients in their initially randomised groups irrespective of any protocol deviations. A Cox proportional hazards model was used to estimate the overall survival HR, with center, performance status (0 vs. 1), and primary tumour localisation (the head vs. the body or tail of the pancreas) as stratification factors. Confidence intervals were presented. A log-rank test (stratified by the above factors) was used to test for a statistically significant difference in survival. Kaplan-Meier survival curves, median survival, and 6-,12- and 18-month survival proportions were presented for each treatment arm. Confidence intervals were reported for median survival, but not for 6-, 12- and 18-month survival proportions.	Analysis sets will be undertaken as detailed in Table 1 (Analysis sets to be included in Target Trial analyses). Analysis Set 1 will emulate the target trial as closely as possible. Analysis Set 2 will consider a broader population, encompassing patients aged 18 or older who receive treatment for metastatic pancreatic cancer. Other analysis sets (denoted Analysis Set 3+) will be developed depending on the data available. For example, if missing data means that one or more eligibility criteria results in a drastic reduction in patient numbers, analyses will be run with and without including those eligibility criteria. Similarly, if several eligibility criteria are problematic to emulate, analyses will be run using those eligibility criteria considered by clinical experts to be most important. For each Analysis Set a number of analyses will be run: - With “complete” adjustment models (see “Assignment procedures, above) - With “reduced” adjustment models (see “Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights. For Analysis Set 1 (and for Analysis Set 3+, if this analysis set is required due to problems emulating one or more eligibility criteria), analyses will be undertaken using the ITT and per-protocol analogues described in “Causal contrasts of interest”. The ITT analysis analogue will estimate the comparative effect according to the treatment strategy initiated, irrespective of whether these strategies continued to be followed after initiation.

		The per-protocol analysis analogue will estimate the comparative effect adjusting for any treatment switches that occur in the NCRAS data that are not representative of treatment pathways received by patients in ACCORD. Both the ITT and per-protocol analyses included in Analysis Set 1 (and Analysis Set 3+, if required) will be subject to the minimum and maximum follow-up restrictions referred to in the “Follow-up period” section of this table. For the ITT-based analysis, inverse probability weighting will be used to adjust for baseline confounders. For the per-protocol analysis, patients who deviate from the defined treatment strategies will be censored at that time-point and therefore adjustment for baseline and post-baseline confounding is necessary. Inverse probability weighting using time-varying weights will be used for this purpose. For the analogue of the ITT analysis and for the per-protocol analysis it is possible that selection bias could be present due to informative loss to follow up. If this is apparent, inverse probability of censoring weighting using time varying weights will be used. These weights will be combined with the weights used to adjust for baseline confounding in the ITT-based analysis, and with the time-dependent weights used to address treatment deviations in the per-protocol analysis. For each analysis, Cox models that incorporate inverse probability weights to adjust for baseline (and where relevant, time-dependent) confounding will be used to estimate overall survival HRs and the log-rank test will be used to test for differences in survival. Where it is necessary to attempt to control for time-dependent confounding, marginal structural Cox models will be used. The HRs will be compared to the HRs for overall survival estimated in ACCORD. These HR estimates will be used to assess emulation agreement, using agreement criteria 1-3 described in the “Evaluating Emulation Success” section of this protocol. Agreement criterion 4 will be assessed by comparing the Kaplan-Meier survival curves presented in the ACCORD publication (digitised and with confidence intervals added, as described in the “Evaluating Emulation Success” section of this report) to weighted Kaplan-Meier curves constructed for each analysis and analysis set

		previously described. The ACCORD publication also reported median overall survival (with confidence intervals) and survival proportions at 6-, 12- and 18-months. We will report these statistics for our emulated analyses to allow further assessment of agreement between the results of our emulation and those reported for ACCORD. However, as previously stated, it is the overall survival HR that will be used to formally assess agreement criteria 1-3, as the primary relative effect measure used in ACCORD. Stratification factors of center, performance status, and primary tumour localisation were used in the Cox model used to estimate the HR for overall survival in ACCORD. These variables will be included as stratification factors in our analyses if data are available. Analysis Set 2 is purposely not comparable to ACCORD, as it will include a broader population. As such, for this analysis we will not draw formal comparisons to ACCORD results, and per-protocol analogues designed to be consistent with treatment pathways received in ACCORD are not necessary. Therefore, for Analysis Set 2, only the ITT analogue analysis will be undertaken. However, as for Analysis Sets 1 and 3+ (if required), the overall survival HR, median survival, Kaplan-Meier survival curves, and survival proportions at 6, 12, and 24 months will be reported. Also, a range of sensitivity and scenario analyses will be reported, as previously described: - With “complete” adjustment models (see “Assignment procedures, above) - With “reduced” adjustment models (see “Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights.

Notes: CT: Computed tomography; TNM: Tumour, nodes, metastasis, Classification of Malignant Tumours; ECOG: Eastern Cooperative Oncology Group; ICD: International Classification of Diseases; SACT: Systemic Anti-Cancer Therapy; ITT: Intention-to-treat; NCRAS: National Cancer Registration and Analysis Service; HR: Hazard ratio

Target Trial 3. Comparing gemcitabine to gemcitabine plus capecitabine in patients with metastatic pancreatic cancer

Trial	CRUK-GEM-CAP. Gemcitabine versus gemcitabine plus	Target Trial 3. Emulation of CRUK-GEM-CAP using NCRAS data

	capecitabine for metastatic pancreatic cancer [19]	
Eligibility criteria	Patients were eligible if they had histologically or cytologically proven ductal adenocarcinoma or undifferentiated carcinoma of the pancreas, presence of locally advanced or metastatic disease precluding curative surgical resection, macroscopic residual disease following resection confirmed by positive histology in postresection tissue biopsies from the tumor bed (R2 resection), or unidimensionally measurable disease as assessed by computed tomography. Other eligibility criteria included no previous chemotherapy, radiotherapy, or other investigation drug treatment for either (neo)adjuvant or advanced disease settings; World Health Organization performance status (PS) of 0, 1, or 2; adequate bone marrow, liver, and renal functions; no significant cardiac history; and no known malabsorption.	Analysis Set 1: Target Trial eligibility criteria: to match CRUK-GEM-CAP as far as possible. Patients with locally advanced or metastatic pancreatic adenocarcinoma (TNM stage III or IV) who did not subsequently have surgical resection or who had previously had an R2 resection, and who started either of the treatments studied in CRUK-GEM-CAP. No previous chemotherapy, radiotherapy, or other investigation drug treatment for either (neo)adjuvant or advanced disease settings; ECOG performance status (PS) of 0, 1, or 2. Permitted tumour locations will be based on ICD-10 codes. Presence of metastases will be based on recorded stage of disease. Previous cancers and treatments will be identified from the cancer registry and SACT datasets. For criteria related to comorbidities, Charlson scores will be used where relevant. It is expected that it will not be possible to emulate all criteria completely – for each criteria the approach used for emulation will be recorded and reported. Clinical expert assistance will be used when proxy variables are required. Minimum follow-up in CRUK-GEM-CAP was 26 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 26 months or longer before the cut-off date of the NCRAS data available. Analysis Set 2: Patients aged 18 or older with locally advanced or metastatic pancreatic cancer who receive treatment with gemcitabine or gemcitabine plus capecitabine.
Treatment strategies	Patients were assigned to receive gemcitabine plus capecitabine or gemcitabine. Patients randomly allocated to gemcitabine alone received gemcitabine	Treatment strategies are initiation of gemcitabine plus capecitabine, or initiation of gemcitabine monotherapy. Patients who meet the eligibility criteria set out above but did not initiate gemcitabine plus capecitabine or gemcitabine monotherapy are excluded from the analysis.

	intravenously at 1,000 mg/m2 over 30 minutes weekly 7 followed by 1 week rest, then weekly 3 every 4 weeks. Patients randomly allocated to the gemcitabine plus capecitabine arm received gemcitabine intravenously at 1,000 mg/m2 weekly 3 every 4 weeks. Capecitabine was administered orally at 1,660 mg/m2 /d (830 mg/m2 twice daily)for 3 weeks followed by 1 week's rest. All treatment was given until disease progression or intolerable toxicity.	Time zero will be the time of initiation of gemcitabine plus capecitabine, or gemcitabine monotherapy, with the restriction that that time-point must fall at a point at which eligibility criteria are satisfied. The published CRUK-GEM-CAP documents do not state if there was an allowable "grace period" between randomisation and treatment initiation. Therefore, we cannot speculate as to whether this represents an aspect of the trial that cannot be perfectly emulated, which could cause differences in analytical results. Such grace periods cannot be emulated because we will not have an intention-to-treat (ITT) date and instead must use the time of treatment initiation as time zero. The published CRUK-GEM-CAP documents also do not state what number of patients initiated their assigned study treatment. It is stated that 247/266 assigned to gemcitabine monotherapy, and 251/267 assigned to gemcitabine + capecitabine received at least one cycle of treatment, but this is not the same as simply initiating treatment. If any patients did not initiate treatment at all, this would have two implications: a) All patients in our emulated analysis initiated one of the target trial treatments; b) Survival analysis (e.g. Kaplan-Meier curves and hazard ratio estimates) in CRUK-GEM-CAP included time from randomisation, which may have occurred some period before treatment initiation, whereas in our emulated analyses these analyses will begin at the time of treatment initiation. The potential impacts of these emulation imperfections will be discussed in analysis reports
Assignment procedures	Patients were randomly assigned to each treatment arm on a 1:1 basis according to a computer-generated variable-size blocked randomisation method. Randomisation was stratified by performance status (0, 1 versus 2) and extent of disease (locally advanced stage III/IVA versus metastatic stage IVB).	To emulate the random assignment of strategies at baseline, we need to adjust for all confounding factors required to ensure comparability (exchangeability) of the groups defined by initiation of the treatment strategies. This will be performed using inverse probability weighting using all potentially prognostic variables available at the time of treatment initiation. In CRUK-GEM-CAP, randomisation was stratified according to performance status and disease stage. These variables will be considered for inclusion in our analysis. Not all relevant variables will not all be available in the NCRAS datasets. Available variables and data will be presented to clinical experts and variables used to adjust for baseline confounding will be

		selected based upon discussion using directed acyclic graphs as a decision aid. It is anticipated that scenario and sensitivity analyses will be carried out using “complete” models (that include all variables considered to be potential confounders), and “reduced” models (that include variables considered to be the most important confounders). Potential residual confounding due to missing data or missing variables will be discussed and reported.
Follow-up period	Randomisation was carried out between May 2002, and January 2005, with data cut-off on March 31 2007. Patients were followed until death or were censored at March 31 2007 if alive at that point.	Minimum follow-up in CRUK-GEM-CAP was 26 months. The maximum possible follow-up was 59 months, with published Kaplan-Meier curves ending at 27 months (99% of patients had died). Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 26 months or longer before the cut-off date of the NCRAS data available, and patients remaining alive at 27 months will be censored. Supplementary analyses will be included that do not place restrictions on minimum or maximum follow-up times.
Outcome	The primary outcome in in CRUK-GEM-CAP was overall survival, measured as the time from randomisation until death from any cause.	Overall survival, measured as the time from treatment initiation until death from any cause (subject to the minimum and maximum follow-up restrictions referred to in the “Follow-up period” section of this table).
Causal contrasts of interest	The primary effect measure used was the difference in 1-year survival rates. The overall survival hazard ratio (HR) between treatment arms, Kaplan-Meier survival curves, and median survival were also presented. Analyses were undertaken on an ITT basis, i.e. the comparative effect of being assigned to the treatment strategies at baseline, irrespective of any protocol deviations	The emulated primary effect measure will be the difference in 1-year survival rates. The overall survival HR between treatment arms, Kaplan-Meier survival curves, and median survival will also be presented, for each of the analyses included in “Analysis plan”. Analyses will represent an analogue of the ITT effect – i.e. the comparative effect will be estimated according to treatment strategy initiated irrespective of whether these strategies continued to be followed after initiation. An analogue of a per-protocol effect will also be estimated, to represent the effect according to if patients followed treatment pathways that are representative of those followed in CRUK-GEM-CAP. It is possible that treatment pathways followed in the cancer registry dataset will deviate from the treatment pathways received in CRUK-GEM-CAP, if patients in the registry data switch onto treatments that were not available or were not

		commonly used during the conduct of CRUK-GEM-CAP. Unfortunately, CRUK-GEM-CAP publications do not report information on post-study treatments received. Therefore, we will have to rely on clinical expert opinion to determine which treatment switches are likely to represent deviations from the treatment pathways received in CRUK-GEM-CAP. Hence, the purpose of our per-protocol analysis is to develop an analysis that more closely emulates the primary ITT analysis used in CRUK-GEM-CAP, if the treatment pathways present in the cancer registry dataset do not adequately resemble those likely to have been followed in CRUK-GEM-CAP. We will also examine the extent to which treatment received in the NCRAS dataset reflect the treatment received in CRUK-GEM-CAP – for example, with respect to duration of treatment. As previously noted, the *intention* to treat cannot be perfectly emulated, and the time zero used in our emulation may not perfectly match the time zero used in CRUK-GEM-CAP (though it was not reported whether there was a lag between randomisation and initiation of study treatment in CRUK-GEM-CAP). Therefore, our ITT analogue may have imperfections.
Analysis plan	All efficacy analyses were done in the ITT population retaining all patients in their initially randomised groups irrespective of any protocol deviations. A Cox proportional hazards model was used to estimate the overall survival HR. Results for the HR and log-rank test (testing for a statistically significant difference in survival) were presented both with and without stratification factors included in the regression (performance status [0, 1 versus 2] and extent of disease [locally advanced	Analysis sets will be undertaken as detailed in Table 1 (Analysis sets to be included in Target Trial analyses). Analysis Set 1 will emulate the target trial as closely as possible. Analysis Set 2 will consider a broader population, encompassing patients aged 18 or older who receive treatment for locally advanced or metastatic pancreatic cancer. Other analysis sets (denoted Analysis Set 3+) will be developed depending on the data available. For example, if missing data means that one or more eligibility criteria results in a drastic reduction in patient numbers, analyses will be run with and without including those eligibility criteria. Similarly, if several eligibility criteria are problematic to emulate, analyses will be run using those eligibility criteria considered by clinical experts to be most important. For each Analysis Set a number of analyses will be run:

	stage III/IVA versus metastatic stage IVB]). Confidence intervals were presented for the HR, 1-year survival rates, and median survival. Kaplan-Meier survival curves were presented for each treatment arm.	- With “complete” adjustment models (see “Assignment procedures, above) - With “reduced” adjustment models (see “Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights. For Analysis Set 1 (and for Analysis Set 3+, if this analysis set is required due to problems emulating one or more eligibility criteria), analyses will be undertaken using the ITT and per-protocol analogues described in “Causal contrasts of interest”. The ITT analysis analogue will estimate the comparative effect according to the treatment strategy initiated, irrespective of whether these strategies continued to be followed after initiation. The per-protocol analysis analogue will estimate the comparative effect adjusting for any treatment switches that occur in the NCRAS data that are not representative of treatment pathways received by patients in CRUK-GEM-CAP. Both the ITT and per-protocol analyses included in Analysis Set 1 (and Analysis Set 3+, if required) will be subject to the minimum and maximum follow-up restrictions referred to in the “Follow-up period” section of this table. For the ITT-based analysis, inverse probability weighting will be used to adjust for baseline confounders. For the per-protocol analysis, patients who deviate from the defined treatment strategies will be censored at that time-point and therefore adjustment for baseline and post-baseline confounding is necessary. Inverse probability weighting using time-varying weights will be used for this purpose. For the analogue of the ITT analysis and for the per-protocol analysis it is possible that selection bias could be present due to informative loss to follow up. If this is apparent, inverse probability of censoring weighting using time varying weights will be used. These weights will be combined with the

		weights used to adjust for baseline confounding in the ITT-based analysis, and with the time-dependent weights used to address treatment deviations in the per-protocol analysis. For each analysis, Cox models that incorporate inverse probability weights to adjust for baseline (and where relevant, time-dependent) confounding will be used to estimate overall survival HRs and the log-rank test will be used to test for differences in survival. Where it is necessary to attempt to control for time-dependent confounding, marginal structural Cox models will be used. The HRs will be compared to the HRs for overall survival estimated in CRUK-GEM-CAP. For our emulation of CRUK-GEM-CAP, we will use these HR estimates and estimates of 1-year survival rates to assess emulation agreement, using agreement criteria 1-3 described in the "Evaluating Emulation Success" section of this protocol. HRs will be used to be consistent with our other Target Trials, but 1-year survival rates will also be used as these were used as the primary means to design the CRUK-GEM-CAP study. Agreement criterion 4 will be assessed by comparing the Kaplan-Meier survival curves presented in the CRUK-GEM-CAP publication (digitised and with confidence intervals added, as described in the "Evaluating Emulation Success" section of this report) to weighted Kaplan-Meier curves constructed for each analysis and analysis set previously described. The CRUK-GEM-CAP publication also reported median overall survival (with confidence intervals). We will report this for our emulated analyses to allow further assessment of agreement between the results of our emulation and those reported for CRUK-GEM-CAP. In the CRUK-GEM-CAP study, HRs were calculated both with and without including stratification factors of performance status and extent of disease in the Cox model. We will emulate both these analyses, with the caveat that stratification factors will only be included if suitable data are available. Analysis Set 2 is purposely not comparable to CRUK-GEM-CAP, as it will include a broader population. As such, for this analysis we will not draw formal comparisons to CRUK-GEM-CAP results, and per-protocol analogues designed to be consistent with treatment pathways received in CRUK-GEM-CAP are not necessary. Therefore, for Analysis Set 2, only the ITT analogue analysis will be undertaken. However, as for Analysis Sets

		1 and 3+ (if required), the overall survival HR, median survival, Kaplan-Meier survival curves, and survival proportions at 1 year will be reported. Also, a range of sensitivity and scenario analyses will be reported, as previously described: - With "complete" adjustment models (see "Assignment procedures, above) - With "reduced" adjustment models (see "Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights.

Note: Note, there are two RCTs of gem vs gem+cap for advanced/metastatic pancreatic cancer [19,23]. As a slightly more recent, slightly bigger, UK based, and more inclusive RCT, we have chosen to attempt to emulate the Cunningham et al [19] trial.
TNM: Tumour, nodes, metastasis, Classification of Malignant Tumours; ECOG: Eastern Cooperative Oncology Group; PS: Performance status; ICD: International Classification of Diseases; SACT: Systemic Anti-Cancer Therapy; ITT: Intention-to-treat; NCRAS: National Cancer Registration and Analysis Service; HR: Hazard ratio

Target Trial 4. Comparing gemcitabine to gemcitabine plus nab-paclitaxel in patients with metastatic pancreatic cancer

Trial	MPACT. Gemcitabine versus gemcitabine plus nab-paclitaxel for metastatic pancreatic cancer [20]	Target Trial 4. Emulation of MPACT using NCRAS data
Eligibility criteria	Eligible patients were ≥18 years of age with a Karnofsky performance status (KPS) score of 70 or higher and histologically or cytologically confirmed metastatic adenocarcinoma of the pancreas. Disease was required to be measurable by RECIST version 1.0. Additional eligibility criteria included adequate hepatic, hematologic, and renal function (including a bilirubin level ≤ the upper limit of the normal range, an absolute neutrophil count ≥ 1.5×109 /L, and a hemoglobin	Analysis Set 1: Target Trial eligibility criteria: to match MPACT as far as possible. Patients with metastatic pancreatic adenocarcinoma (TNM stage IV), and who started either of the treatments studied in MPACT. No previous chemotherapy, radiotherapy or surgery for metastatic disease. Exclude patients with islet cell neoplasms or locally advanced adenocarcinoma, and patients who had received cytotoxic doses of any systemic chemotherapy, including gemcitabine, in the adjuvant setting. Treatment with fluorouracil or gemcitabine as a radiation sensitizer in the adjuvant setting allowed if given at least six months prior to random assignment. ECOG score must be 0, 1 or 2, to be approximately equivalent to a Karnofsky performance status of 70 or higher. Metastatic disease to have been diagnosed within 6 weeks before treatment initiation. No brain metastases. No history of malignancy in previous 5 years, except basal cell carcinoma of skin, carcinoma in situ of cervix. Permitted tumour locations will be based on ICD-10 codes. Presence of metastases will be based on recorded stage of disease. Previous cancers

	level ≥ 9g/dL). Treatment with fluorouracil or gemcitabine as a radiation sensitizer in the adjuvant setting was allowed if given at least six months prior to random assignment. Previous chemotherapy, radiotherapy or surgery for metastatic disease was an exclusion criterion for this study. Patients with islet cell neoplasms or locally advanced adenocarcinoma were also excluded, as were patients who had received cytotoxic doses of any systemic chemotherapy, including gemcitabine, in the adjuvant setting. Metastatic disease had to have been diagnosed within 6 weeks before randomisation. Patients must not have had known brain metastases, unless previously treated and well-controlled for at least 3 months. Patients were excluded if they had a history of malignancy in the last 5 years but patients with prior history of in situ cancer or basal or squamous cell skin cancer were eligible. Patients with other malignancies were eligible if they were cured by surgery alone or surgery plus radiotherapy and	and treatments will be identified from the cancer registry and SACT datasets. For criteria related to comorbidities, Charlson scores will be used where relevant. It is expected that it will not be possible to emulate all criteria completely – for each criteria the approach used for emulation will be recorded and reported. Clinical expert assistance will be used when proxy variables are required. Minimum follow-up in MPACT was 6 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 6 months or longer before the cut-off date of the NCRAS data available. Analysis Set 2: Patients aged 18 or older who receive treatment with gemcitabine monotherapy or gemcitabine plus nab-paclitaxel for metastatic pancreatic cancer.

	have been continuously disease-free for at least 5 years.	
Treatment strategies	Patients were assigned to receive gemcitabine plus nab-paclitaxel or gemcitabine. Patients randomly allocated to gemcitabine plus nab-paclitaxel received a 30-to40–minute intravenous infusion of nab-paclitaxel at a dose of 125 mg per square meter, followed by an infusion of gemcitabine according to the gemcitabine label at a dose of 1000 mg per square meter, on days 1, 8, 15, 29, 36, and 43. Patients assigned to gemcitabine alone received a dose of 1000 mg per square meter weekly for 7 of 8 weeks (cycle 1). In subsequent cycles, all patients were administered treatment on days 1, 8, and 15 every 4 weeks. Treatment continued until disease progression or until there was an unacceptable level of adverse events. Per protocol, crossover was not allowed at any time after randomisation.	Treatment strategies are initiation of gemcitabine plus nab-paclitaxel, or initiation of gemcitabine monotherapy. Patients who meet the eligibility criteria set out above but did not initiate gemcitabine plus nab-paclitaxel or gemcitabine are excluded from the analysis. Time zero will be the time of initiation of gemcitabine monotherapy or gemcitabine plus nab-paclitaxel, with the restriction that that time-point must at a point at which eligibility criteria are satisfied. In MPACT there could be a 3-day lag between randomisation and treatment initiation. This represents an aspect of the trial that cannot be perfectly emulated, which could cause differences in analytical results. We cannot emulate this 3 day “grace period” because we will not have an intention-to-treat (ITT) date. Therefore, we must use the time of treatment initiation as time zero. This has two implications: a) All patients in our emulated analysis initiated one of the target trial treatments. In MPACT, 11 out of 431 patients randomised to gemcitabine + nab-paclitaxel, and 27 out of 403 patients randomised to gemcitabine monotherapy did not receive study treatment; b) Survival analysis (e.g. Kaplan-Meier curves and hazard ratio estimates) in MPACT included time up to 3 days before treatment initiation, whereas in our emulated analyses these analyses will begin at the time of treatment initiation. The potential impacts of these emulation imperfections will be discussed in analysis reports.
Assignment procedures	Patients were randomly assigned to each treatment arm on a 1:1 basis. Patients were stratified according to performance status, presence or absence	To emulate the random assignment of strategies at baseline, we need to adjust for all confounding factors required to ensure comparability (exchangeability) of the groups defined by initiation of the treatment strategies. This will be performed using inverse probability weighting using all potentially prognostic variables available at the time of treatment initiation.

	of liver metastases, and geographic region.	In MPACT, randomisation was stratified according to performance status and presence or absence of liver metastases. These variables – or potential proxies for them – will be considered for inclusion in our analysis. Not all relevant variables will not all be available in the NCRAS datasets. Available variables and data will be presented to clinical experts and variables used to adjust for baseline confounding will be selected based upon discussion using directed acyclic graphs as a decision aid. It is anticipated that scenario and sensitivity analyses will be carried out using "complete" models (that include all variables considered to be potential confounders), and "reduced" models (that include variables considered to be the most important confounders). Potential residual confounding due to missing data or missing variables will be discussed and reported.
Follow-up period	Randomisation was carried out between May 2009, and April 2012, with data cut-off on September 17 2012. Patients were followed until death or were censored at September 17 2012 if alive at that point.	Minimum follow-up in MPACT was 6 months. The maximum possible follow-up was 41 months, with published Kaplan-Meier curves ending at 38 months. Therefore, for our main analysis, patients are only to be included in our trial emulation if they initiated treatment 6 months or longer before the cut-off date of the NCRAS data available, and patients remaining alive at 38 months will be censored. Supplementary analyses will be included that do not place restrictions on minimum or maximum follow-up times.
Outcome	The primary outcome in MPACT was overall survival, measured as the time from randomisation until death from any cause.	Overall survival, measured as the time from treatment initiation until death from any cause (subject to the minimum and maximum follow-up restrictions referred to in the "Follow-up period" section of this table).
Causal contrasts of interest	The primary effect measure used was the overall survival hazard ratio (HR) between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 6, 12, 18 and 24 months were presented for both treatment arms.	The emulated primary effect measure will be the overall survival HR between treatment arms. Kaplan-Meier survival curves, median survival, and survival proportions at 6, 12, 18 and 24 months will also be presented for both treatment arms, for each of the analyses included in "Analysis plan". Analyses will represent an analogue of the ITT effect – i.e. the comparative effect will be estimated according to treatment strategy initiated irrespective of whether these strategies continued to be followed after initiation. An analogue of a per-protocol effect will also be estimated, to represent the effect according to if

	Analyses were undertaken on an ITT basis, i.e. the comparative effect of being assigned to the treatment strategies at baseline, irrespective of any protocol deviations	patients followed treatment pathways that are representative of those followed in MPACT. It is possible that treatment pathways followed in the cancer registry dataset will deviate from the treatment pathways received in MPACT, if patients in the registry data switch onto treatments that were not available or were not commonly used during the conduct of MPACT. MPACT publications report some information on post-study treatments received, and these will be compared to subsequent treatments received by patients identified in the NCRAS data. Clinical expert opinion will be sought to determine which treatment switches represent deviations from the treatment pathways received in MPACT. Hence, the purpose of our per-protocol analysis is to develop an analysis that more closely emulates the primary ITT analysis used in MPACT, if the treatment pathways present in the cancer registry dataset do not adequately resemble those followed in MPACT. We will also examine the extent to which treatment received in the NCRAS dataset reflect the treatment received in MPACT – for example, with respect to duration of treatment. As previously noted, the *intention* to treat cannot be perfectly emulated, and the time zero used in our emulation does not perfectly match the time zero used in MPACT (because there could be up to a 3-day lag between randomisation and initiation of study treatment). Therefore, our ITT analogue has imperfections. However, given the short 3-day "grace period" used in MPACT, and given that 97.4% of patients assigned to gemcitabine + nab-paclitaxel, and 93.7% of patients assigned to gemcitabine alone, received their study treatment, we expect the impact of these imperfections to be minor.
Analysis plan	All efficacy analyses were done in the ITT population retaining all patients in their initially randomised groups irrespective of any protocol deviations. A Cox proportional hazards model was used to estimate the overall survival HR, with performance	Analysis sets will be undertaken as detailed in Table 1 (Analysis sets to be included in Target Trial analyses). Analysis Set 1 will emulate the target trial as closely as possible. Analysis Set 2 will consider a broader population, encompassing patients aged 18 or older who receive adjuvant treatment for pancreatic cancer. Other analysis sets (denoted Analysis Set 3+) will be developed depending on the data available. For example, if missing data means that one or more

	status, presence or absence of liver metastases, and geographic region as stratification factors. Confidence intervals were presented. A log-rank test (stratified by the above factors) was used to test for a statistically significant difference in survival. Kaplan-Meier survival curves, median survival, and 6-, 12-, 18- and 24-month survival proportions were presented for each treatment arm. Confidence intervals were reported for all measures.	eligibility criteria results in a drastic reduction in patient numbers, analyses will be run with and without including those eligibility criteria. Similarly, if several eligibility criteria are problematic to emulate, analyses will be run using those eligibility criteria considered by clinical experts to be most important. For each Analysis Set a number of analyses will be run: - With "complete" adjustment models (see "Assignment procedures, above) - With "reduced" adjustment models (see "Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights. For Analysis Set 1 (and for Analysis Set 3+, if this analysis set is required due to problems emulating one or more eligibility criteria), analyses will be undertaken using the ITT and per-protocol analogues described in "Causal contrasts of interest". The ITT analysis analogue will estimate the comparative effect according to the treatment strategy initiated, irrespective of whether these strategies continued to be followed after initiation. The per-protocol analysis analogue will estimate the comparative effect adjusting for any treatment switches that occur in the NCRAS data that are not representative of treatment pathways received by patients in MPACT. Both the ITT and per-protocol analyses included in Analysis Set 1 (and Analysis Set 3+, if required) will be subject to the minimum and maximum follow-up restrictions referred to in the "Follow-up period" section of this table. For the ITT-based analysis, inverse probability weighting will be used to adjust for baseline confounders. For the per-protocol analysis, patients who deviate from the defined treatment strategies will be censored at that time-point and therefore adjustment for baseline and post-baseline

		confounding is necessary. Inverse probability weighting using time-varying weights will be used for this purpose. For the analogue of the ITT analysis and for the per-protocol analysis it is possible that selection bias could be present due to informative loss to follow up. If this is apparent, inverse probability of censoring weighting using time varying weights will be used. These weights will be combined with the weights used to adjust for baseline confounding in the ITT-based analysis, and with the time-dependent weights used to address treatment deviations in the per-protocol analysis. For each analysis, Cox models that incorporate inverse probability weights to adjust for baseline (and where relevant, time-dependent) confounding will be used to estimate overall survival HRs and the log-rank test will be used to test for differences in survival. Where it is necessary to attempt to control for time-dependent confounding, marginal structural Cox models will be used. The HRs will be compared to the HRs for overall survival estimated in MPACT. These HR estimates will be used to assess emulation agreement, using agreement criteria 1-3 described in the "Evaluating Emulation Success" section of this protocol. Agreement criterion 4 will be assessed by comparing the Kaplan-Meier survival curves presented in the MPACT publication (digitised and with confidence intervals added, as described in the "Evaluating Emulation Success" section of this report) to weighted Kaplan-Meier curves constructed for each analysis and analysis set previously described. The MPACT publication also reported median overall survival (with confidence intervals) and survival proportions at 6-, 12-, 18-, and 24-months. We will report these statistics for our emulated analyses to allow further assessment of agreement between the results of our emulation and those reported for MPACT. However, as previously stated, it is the overall survival HR that will be used to formally assess agreement criteria 1-3, as the primary relative effect measure used in MPACT. Stratification factors of performance status, presence or absence of liver metastases, and geographic region were used in the Cox model used to estimate the HR for overall survival in MPACT. These variables will be included as stratification factors in our analyses if data are available.

		Analysis Set 2 is purposely not comparable to MPACT, as it will include a broader population. As such, for this analysis we will not draw formal comparisons to MPACT results, and per-protocol analogues designed to be consistent with treatment pathways received in MPACT are not necessary. Therefore, for Analysis Set 2, only the ITT analogue analysis will be undertaken. However, as for Analysis Sets 1 and 3+ (if required), the overall survival HR, median survival, Kaplan-Meier survival curves, and survival proportions at 6, 12, 18, and 24 months will be reported. Also, a range of sensitivity and scenario analyses will be reported, as previously described: - With "complete" adjustment models (see "Assignment procedures, above) - With "reduced" adjustment models (see "Assignment procedures, above) - With minimum and maximum follow-up times matching those in the target trial - With no restriction on minimum and maximum follow-up times - With stabilised and unstabilised weights used for inverse probability weights.

Notes: KPS: Karnofsky Performance Status; TNM: Tumour, nodes, metastasis, Classification of Malignant Tumours; ECOG: Eastern Cooperative Oncology Group; ICD: International Classification of Diseases; SACT: Systemic Anti-Cancer Therapy; ITT: Intention-to-treat; NCRAS: National Cancer Registration and Analysis Service; HR: Hazard ratio

Geographical Descriptive Statistics

The focus of our study is on estimating comparative effectiveness using the Target Trial framework. However, we plan to supplement this analysis with descriptive information about the treatments received in different areas of England. Hence, we also request access to geographic data. This is unlikely to be used in our estimation of comparative effectiveness (though instrumental variables analyses may be considered if treatment received is highly associated with organisation codes), but may be interesting if we are able to reliably estimate comparative effectiveness and if treatments received differ substantially by geographical area. If we find that very few patients (less than 5) received a specific treatment regimen in a geographical area any related publication would suppress this information in order to avoid potential identification of patients.

Data Requirements

Request Summary

Summary of request - Please provide a summary of the data being requested, outlining which of the available datasets are being requested
We will require linked data from the following datasets - Cancer registration (patient table) - Cancer registration (tumour table) - Cancer registration (treatment table) - SACT dataset - Radiotherapy dataset - HES admitted care - HES outpatient - HES accident and emergency - Route to diagnosis To allow us to complete our Target Trial analysis, we need detailed information on patients who meet the inclusion/exclusion criteria of our Target Trials. Hence, we need detailed information on patient characteristics, tumours and treatments for patients with pancreatic cancer (ICD: C25x), and this is reflected by the variables we are requesting access to in the tables below. However, importantly, **we only need data for patients who received some kind of systemic anti-cancer therapy for their pancreatic cancer**. Patients who were diagnosed with pancreatic cancer but did not receive systemic anti-cancer therapy can be excluded from the data extract. In addition, we need a selection of other derived variables so that we can identify which patients meet the inclusion/exclusion criteria for the different Target Trials that we plan to run. For the following variables, we do not need detailed information on the tumours, treatments and malignancies that these variables refer to and to avoid requesting excessive amounts of data we are instead requesting derived variables: - Previous treatment with systemic anti-cancer therapy (yes/no) - Previous or concurrent malignancy (i.e. ICD C00-C43, C45-C96, D00-D05, D07-49) (except basal cell carcinoma of skin (C44), carcinoma in situ of cervix (D06)): yes/no, what the ICD code was and date of diagnosis. For this, it may be easiest to extract the data as follows: Include two columns for each previous malignancy that a patient has had, one column for the ICD code of that malignancy, and one column for date of diagnosis [diagnosisdatebest] of that malignancy. Information on C44 and D06 malignancies would be included here. Some patients will have had several previous malignancies and some will have had none (and so these columns will be empty). Then we will be able to derive whether or not patients meet the eligibility criteria of the different Target Trials and will derive the "yes" "no" variable myself. - Previous malignancy in 5 years prior to diagnosis of metastatic disease (i.e. ICD C00-C43, C45-C96, D00-D05, D07-49) (except basal cell carcinoma of skin (C44), carcinoma in situ of cervix (D06)): yes/no, and what the ICD code was and date of diagnosis. The easiest approach for extracting this data might be as described in the previous bullet, except limited to the 5 years prior to treatment for pancreatic cancer (since we acknowledge that date of metastatic disease diagnosis is not available)

- Radiotherapy previous to metastatic pancreatic cancer diagnosis (ever): yes/no
- Previous treatment with fluorouracil or gemcitabine as a radiation less than 6 months prior to diagnosis of metastatic disease: yes/no. (This will likely require a rule such as: has gemcitabine/flourouracil been used concurrently with radiotherapy less than 6 months prior to diagnosis of metastatic disease)
- Development of another cancer after their pancreatic cancer diagnosis: yes/no and date of new diagnosis

Whilst patients with more than one tumour may be excluded from our Target Trial emulation analyses, we will also conduct a second (more “real world”) analysis for each of the Target Trials, which will include all patients with the relevant pancreatic cancer diagnosis, irrespective of other patient characteristics. Hence, we need information on all patients with pancreatic cancer who received some kind of systemic anti-cancer treatment irrespective of the number of tumours, but also need information on the number of tumours (and the other factors listed above) to allow us to identify who should be included in the trial emulation analysis and who should be included in the more "real world" analysis.

Cancer Sites/Morphologies – Please provide the cancer sites and/or morphologies required for the request separated by commas and the coding system used. If combinations of site/ morphology are required please separate site and morphology with hyphens. If all codes within a tumour site grouping are required an ‘x’ may be used to suffix the 3 character grouping. (For example: C18x, C19x, C20x, C44 – 80903, C44 – 81703, C56.1, C56.2)

C25x

Geography or treatment provider criteria – Please provide us with the geography for the data provided, if data are required for all of England please state this. If data are required for particular geographies/provider please state the geography level, the required geographies and how these geographies should be applied to the data. (For example: CCGs 07X, 08V, 08B defined by patient treatment within trust located in one of these CCGs)

Data are required for all of England. We are also requesting geographic data. This is unlikely to be used in our estimation of comparative effectiveness (though we may consider instrumental variable approaches, if treatments received are highly associated with organisation codes), but we plan to supplement our analysis with descriptive information about the treatments received in different areas of England. If very few patients (less than 5) received a specific treatment regimen in an area any related publication would suppress this information in order to avoid potential identification of patients.

Time period criteria - individual years or a range of years. Time period should also describe which dataset time period applies to, e.g. all patients with a diagnosis date between 2000-2010 or patients with any inpatient HES activity in the trusts defined above in 2015. Please also indicate clearly if the date is diagnostic date, treatment date, event date or a combination.

Patients diagnosed from April 2012 until 6 months prior to final data cut-off available. Follow-up data is requested for all patients up to the latest data cut available.

We also believe that it will be important to attempt to construct co-morbidity weights to account for different prognoses in patients. We plan to base this on four factors: (i) A

Charlson score (based upon information on prior inpatient diagnoses over a 6 year period); (ii) Per-patient total inpatient length of stay over a 6 year period; (iii) Number of inpatient admissions over a 6 year period; (iv) Total number of outpatient appointments over a 6 year period. Hence, **for patients who received some kind of systemic anti-cancer therapy for their pancreatic cancer** we would like history of cancer information and hospital inpatient and outpatient data for the 6 years prior to their diagnosis of adjuvant/metastatic pancreatic cancer. For patients with adjuvant pancreatic cancer, we need this data for the 6 years prior to the incidence of pancreatic cancer. For patients with metastatic pancreatic cancer, we need this data for the 6 years prior to the date of metastasis. We understand data on date of metastases is not available, so instead we would like this data for the 6 years prior to receipt of SACT treatment for their pancreatic cancer.

For (i) I understand that it is possible for ODR to provide Charlson scores - these are not in the data dictionary but if these are available as specified below, we would request these scores, which would avoid the need for us to be provided with the data detailed in the table below. We understand that Charlson scores are available with a lookback period of 27 to 3 months before diagnosis, or 78 to 6 months before diagnosis. We request the 78 to 6 months lookback data. We also understand that data are available either on a total Charlson score (out of 17), or data can be provided on 16 of the 17 categories separately (excluding HIV). We request data on the 16 categories separately.

If Charlson scores are not available, the data may most usefully be in the form of a series of “yes/no” variables for the ICD codes included in Charlson calculations. These are given in the table below. Note that to avoid including the incidence cancer in the comorbidity calculation records of pancreatic cancer or secondary cancer occurring within 6 months of the incidence date should be excluded (i.e. ICD10 codes C25* or C77 to C80).

The Charlson approach was used by Gray et al. (2019),[24] with more detail provided by the authors on a wiki page.[25]

Condition	ICD09	ICD10
Acute Myocardial Infarction	410, 412	I21, I22, I252
Congestive Heart Failure	428,4254,4255,4257,4258,4259, 39891, 40201, 40211, 40291, 40401, 40403, 40411, 40413, 40491, 40493	I43,I50,I099,I110,I130,I132,I255, I420,I425, I426,I427,I428,I429,P290
Peripheral Vascular Disease	440,441,0930, 4373, 4431, 4432, 4438, 4439, 4471, 5571, 5579	I70, I71, I731, I738, I739, I771, I790, I792, K551, K558, K559, V434, Z958, Z959
Cerebral Vascular Disease/ Accident	430, 431, 432, 433, 434, 435, 436, 437, 438, 36234	G45, G46, I60, I61, I62, I63, I64, I65, I66, I67, I68, I69, H340
Dementia	290, 2941, 3312,	F00, F01, F02, F03, F051, G30, G311
Chronic Pulmonary Disease	490, 491, 492, 493, 494, 495 ,496, 500, 501, 502, 503, 504, 505 ,4168, 4169, 5064, 5081, 5088	J40, J41, J42, J43, J44, J45, J46, J47, J60, J61, J62, J63, J64, J65, J66, J67,I278,I279,J684,J701,J703
Connective Tissue/ Rheumatologic Disease	4465, 7100, 7101, 7102, 7103, 7104, 7140, 7141, 7142, 7148, 725	M05, M06, M32, M33, M34, M315, M351, M353, M360
Peptic ulcer	531, 352, 533, 534,	K25, K26, K27, K28
Diabetes without complications	2500, 2501, 2502, 2503, 2508, 2509	E100, E101, E106, E108, E109, E110, E111, E116, E118, E119, E120, E121, E126, E128, E129, E130, E131, E136, E138, E139, E140, E141, E146, E148, E149
Diabetes with complications	2504, 2505, 2506, 2507	E102, E103, E104, E105, E107, E112, E113, E114, E115, E117, E122, E123, E124, E125, E127, E132, E133, E134,

		E135, E137, E142, E143, E144, E145, E147
Hemiplegia, Paraplegia	3341, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3449, 342, 343	G81, G82, G041, G114, G801, G802, G830, G831, G832, G833, G834, G839
Renal Disease	582, 585, 586, V56, 5830, 5831, 5832, 5836, 5837, 5880, C420, V451, 40301, 40311, 40391, 40402, 40403, 40412, 40413, 40492, 40493	N18, N19, N052, N053, N054, N055, N056, N057, N250, I120, I131, N032, N033, N034, N035, N036, N037, Z490, Z491, Z492, Z940, Z992
Cancer - Any	140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 170, 171, 172, 174, 175, 176, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 200, 201, 202, 203, 204, 205, 206, 207, 208	C00, C01, C02, C03, C04, C05, C06, C07, C08, C09, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19, C20, C21, C22, C23, C24,C25, C26, C30, C31, C32, C33, C34, C37, C38, C39, C40, C41, C43, C45, C46, C47, C48, C49, C50, C51, C52, C53, C54, C55, C56, C57, C58, C60, C61, C62, C63, C64, C65, C66, C67, C68, C69, C70 , C71, C72,C73, C74, C75, C76, C81, C82, C83, C84, C85, C88, C90, C91, C92, C93, C94, C95, C96, C97
Cancer - Metastatic carcinoma	C77, C78, C79, C80	C77, C78, C79, C80
Liver disease – mild	07022, 07023, 07032, 07033, 07044, 07054, 0706, 0709, 5733, 5734, 5738, 5739, 570, 571	B18, K73, K74, K700, K701, K702, K703, K709, K717, K713, K714, K715, K760, K762, K763, K764, K768, K769, V427, Z944
Liver disease – moderate/severe	4560, 4561, 4562, 5722, 5723, 5724, 5728	K704, K711, K721, K729, K765, K766, K767, I850, I859, I864, I982
HIV/Aids	042, 043, 044	B20, B21, B22, B24

For (ii), (iii) and (iv) derived variables for total length of stay, total number of inpatient admissions and total number of outpatient appointments in the same 6 year periods as outlined above would be sufficient.

The variables available and required from each dataset are presented in detail below, using the table formatting provided in the NCRAS Data Dictionary at the time the application for data was made.

Note, we acknowledge that in some cases the same variable is requested from many fields. We are not sure which is the best dataset to source these variables from, so we are happy to leave this to the analyst who extracts the data. In some cases we acknowledge it is possible to derive one variable from another already requested. The ODR may decide to only provide the original variable in such cases. However, in some cases it might be preferable to have both, allowing easy alternation between variables in different analyses - in case one turns out to be more useful than another. For example, for HES diagnosis codes there are both 3 digit and 4 digit codes available: it might be that there is no additional valuable information in the 4 digit code, making it reasonable to use the 3 digit in analyses, or we may find that the 4 digit codes are useful. We would prefer to be provided with both variables, but if the ODR prefers, we are happy to derive the 3 digit variable from the 4 digit variable.

Cancer Registration (patient table)

Data item	Field name	Description of field content	Request field (mark required variables with x)	Justification - detail why the field is necessary for your analysis
Pseudonymised patient ID	PATIENTID	Project specific ID for each person	x	To allow linking of patient data
NHS number	**NHSNUMBER**	**Valid NHS Number or blank.**		
Alias check flag - patient	ALIASFLAG	0,1 (Indicates that this patient record has been deduplicated with another patient and the tumour(s) moved to that other patientid)		
Date of Birth	**BIRTHDATEBEST**	**ddmmyyyy**		
Month of birth	MONTH_DOB	mm	x	Age is likely to be an important prognostic variable in our analyses – exact date of birth not required, but month and year useful
Year of birth	YEAR_DOB	yyyy	x	Age is likely to be an important prognostic variable in our analyses – exact date of birth not required, but month and year useful
Date of Birth check flag - patient	BIRTHDATEFLAG	0,1,2,3 (Set to 0 if the date was fully specified, 1 if the month and year of diagnosis are known, but the day was not specified, 2 if the year is fully known, but the month and day are not specified, and 3 if the date was less specific than any of these)		
Sex	SEX	0=Not known, 1=Male, 2=Female, 9=Not specified	x	Sex may be an important prognostic variable
Ethnicity	ETHNICITY	A = (White) British, B =(White) Irish, C = Any other White background, D = White and Black Caribbean, E = White and Black African, F = White and Asian, G = Any other mixed background, H = Indian, J = Pakistani, K = Bangladeshi, L = Any other Asian background, M = Caribbean , N = African, P = Any other Black background, R = Chinese, S = Any other ethnic group, Z = Not stated, X = Not Known	x	Ethnicity may be an important prognostic variable

Ethnic group	ETHNICITYNAME	(White) British, (White) Irish, Any other White, background, White and Black Caribbean, White and Black African, White and Asian, Any other mixed background, Indian, Pakistani, Bangladeshi, Any other Asian background, Caribbean , African, Any other Black background, Chinese, Any other ethnic group, Not stated, Not Known		
Broad ethnic group	Option to group ethnicities (e.g. white/ non-white/ unknown)	Derived as per applicant requirements		
Vital status of the patient	VITALSTATUS	A =Alive, D =Dead, X =Exit posting	x	Essential for estimating comparative effectiveness of treatments
Date of death of the patient	DEATHDATEBEST	ddmmyyyy	x	Essential for estimating comparative effectiveness of treatments. Actual date is required rather than an interval (e.g. time from diagnosis to death) because the staging of different events over time will be important
Month of death of the patient	MONTH_DOD	MM		
Year of death of the patient	YEAR_DOD	YYYY		
Days from another event to date to death	Option to provide number of days from another event to death (e.g. days from diagnosis to death)	Derived as per applicant requirements		
Date of death imputed flag	DEATHDATEFLAG	0,1,2,3 (Set to 0 if the date was fully specified, 1 if the month and year of diagnosis are known, but the day was not specified, 2 if the year is fully known, but the month and day are not specified, and 3 if the date was less specific than any of these)	x	Useful information for interpreting date of death data
Embarkation flag	EMBARKATION	Y or blank	x	Useful for censoring in the dataset
Date of embarkation	EMBARKATIONDATE	ddmmyyyy	x	Useful for censoring in the dataset
Month of embarkation	Month of embarkation	mm		
Year of embarkation	Year of embarkation	yyyy		
Days from another event to embarkation	Option to provide number of days from another event to embarkation (e.g. days from diagnosis to embarkation	Derived as per applicant requirements		
As provided with death notification	DEATHCAUSECODE_1A	Text – no validation		
As provided with death notification	DEATHCAUSECODE_1B	Text – no validation		

As provided with death notification	DEATHCAUSECODE_1C	Text – no validation		
As provided with death notification	DEATHCAUSECODE_2	Text – no validation		
As provided with death notification	DEATHCAUSECODE_UNDERLYING	Text – no validation		
Code of the location (type) where the patient died, e.g. patients home, hospice etc.	DEATHLOCATIONCODE	1, 2, 3, 4, 5, 6, X, blank		
Description of the location (type) where the patient died, e.g. patients home, hospice etc.	DEATHLOCATIONDESC	CARE HOME, HOSPICE NOS, HOSPITAL, NHS HOSPICE / SPECIALIST PALLIATIVE CARE UNIT, NURSING HOME, OTHER, PRIVATE HOME, UNKNOWN, VOLUNTARY HOSPICE / SPECIALIST PALLIATIVE CARE UNIT, blank		
Code of institution at which death takes place	SITECODEOFDEATH	Valid institution code		
Pseudonymised code of institution at which death takes place	SITECODEOFDEATH (pseudonymised)			
Indicates whether a post-mortem took place	POSTMORTEM	8, 9, N, Y, blank		
Count of every tumour assigned to this PatientID.	TUMOURCOUNT	Number	x	Useful to allow analysis of co-morbidities/multiple cancers
Count of every tumour assigned to this PatientID in range C00-97 excl C44	BIGTUMOURCOUNT	Number	x	Useful to allow analysis of co-morbidities/multiple cancers

Cancer registration (tumour table)

Data item	**Field name**	**Description of field content**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**
Pseudonymised tumour ID	TUMOURID	Project specific ID for each tumour	x	To allow analyses specific to tumours for each patient
Pseudonymised patient ID	PATIENTID	Project specific ID for each person	x	To allow linking between datasets
NHS Number	**NHSNUMBER**	**Valid NHS Number or blank.**		
Date of Birth	**BIRTHDATEBEST**	**ddmmyyyy**		
Month of birth	MONTH_DOB	MM		
Year of birth	YEAR_DOB	YYYY		
Age at diagnosis	AGE	Number or blank	x	Age at diagnosis may be an important prognostic factor
Age at diagnosis in 5 year age bands (0-4 etc.)	FIVEYEARAGEBAND	0 - 4 YRS | 5 - 9 YRS | 10 - 14 YRS | 15 - 19 YRS | 20 - 24 YRS | 25 - 29 YRS | 30 - 34 YRS | 35 - 39 YRS | 40 - 44 YRS | 45 - 49 YRS | 50 - 54 YRS | 55 - 59 YRS | 60 - 64 YRS | 65 - 69 YRS | 70 –		

		74 YRS | 75 - 79 YRS | 80 - 84 YRS | Blank)		
Sex	SEX	0=Not known, 1=Male, 2=Female, 9=Not specified.		
Postcode at Diagnosis	**POSTCODE**	**Postcode-7 format.**		
Outward postcode	POSTCODE_OUTWARD	The area and district component of the Postcode		
Broader geographic area/ IMD quintile	Option to provide geography as deprivation score or aggregate to larger geographic areas such as MSOA or county.	Derived as per applicant requirements	x	Geographic area (county) requested for descriptive statistics of treatment received
Ethnicity	ETHNICITY	A = (White) British, B =(White) Irish, C = Any other White background, D = White and Black Caribbean, E = White and Black African, F = White and Asian, G = Any other mixed background, H = Indian, J = Pakistani, K = Bangladeshi, L = Any other Asian background, M = Caribbean , N = African, P = Any other Black background, R = Chinese, S = Any other ethnic group, Z = Not stated, X = Not Known		
Broad ethnic group	Option to group ethnicities (e.g. white/ non-white/ unknown)	Derived as per applicant requirements		
Earliest date when the diagnosis may have taken place	DIAGNOSISDATE1	ddmmyyyy	x	Age at diagnosis may be an important prognostic factor
Latest date when the diagnosis may have taken place	DIAGNOSISDATE2	ddmmyyyy	x	Age at diagnosis may be an important prognostic factor
Diagnosis date	DIAGNOSISDATEBEST	ddmmyyyy	x	Age at diagnosis may be an important prognostic factor. In addition, this is needed in order to calculate timelines of events (e.g. if/when surgery, chemotherapy, radiotherapy occurred in relation to each other and in relation to diagnosis)
Month of diagnosis	DIAGNOSISMONTH	mm		
Year of diagnosis	DIAGNOSISYEAR	yyyy		
Days from another event to date to diagnosis	Option to provide number of days from another event to diagnosis (e.g. days from birth to diagnosis)	Derived as per applicant requirements		

Date of diagnosis imputed flag	DIAGNOSISDATEFLAG	A flag set to inform if any part of the diagnosis date has been imputed	x	Useful information to inform interpretation of diagnosis date variables
Financial year of diagnosis	FINANCIALYEAR	yyyy		
Basis of diagnosis of the tumour	BASISOFDIAGNOSIS	Non-microscopic: 0 = Death certificate 1 = Clinical: Diagnosis made before death without (2-7) 2 = Clinical investigation: Includes all diagnostic techniques without a tissue diagnosis 4 = Specific tumour markers: Includes biochemical and/or immunological markers which are site specific Microscopic: 5 = Cytology: Examination of cells whether from a primary or secondary site, including fluids aspirated using endoscopes or needles. Also including microscopic examination of peripheral blood films and trephine bone marrow aspirates6 = Histology of a metastases: Includes autopsy specimens 7 = Histology of a primary tumour: Includes all cutting and bone marrow biopsies. Also includes autopsy specimens of a primary tumour 9 = Unknown, e.g. PAS or HISS record only	x	Informs selection of patients to be included in analyses according to Target Trial eligibility criteria
Diagnosis death certificate only	DCO	Y = Yes, N = No		
Site of neoplasm (4-character ICD-10-O2 code)	SITE_ICD10_O2	Valid 4 digit ICD-10 codes in the range C00-D48 plus D76, E85, O01, Q85 or blank	x	Site of pancreatic cancer may be an important prognostic factor
Site of neoplasm (3-character ICD-10-O2 code)	SITE_ICD10_O2_3CHAR	Valid 3 digit ICD-10 codes in the range C00-D48 plus D76, E85, O01, Q85 or blank	x	To confirm pancreatic cancer and location within pancreas: ICD C25.x
Site of the cancer	SITE_CODED	Site of the cancer, in the coding system that the tumour was originally coded in.	x	Site of pancreatic cancer may be an important prognostic factor
Description of the code in SITE_CODED	SITE_CODED_DESC	Text description of the code in SITE_CODED	x	Site of pancreatic cancer may be an important prognostic factor
3 digit version of SITE_CODED	SITE_CODED_3CHAR	Three digit version of site_coded	x	Site of pancreatic cancer may be an important prognostic factor
The coding system used to register the tumour	CODING_SYSTEM	1 = ICD-8, 2 = ICD-9, 3 = ICD-10/O-2, 4 = ICD-10/O-3, 5 = ICD-O-3, 6 = ICD-7, 7 = ICD-8pre1971, 8 = ICD-O-2, 9 = ICD-O, 10 = ICD-O-3 (2011), 11 = ICD-10rev4/O-2, 12 = MOTNAC, 14 = SNOMED/O(TCR), 15 = SNOMED/O-1, 16 = SNOMED/O-2, 17 = SNOMED/O-3	x	Useful for interpretation of site variables
Description of coding system used in registration	CODING_SYSTEM_DESC	TBC	x	Useful for interpretation of site variables
Morphology	MORPH_CODED	TBC	x	Morphology may be an important

				prognostic factor, and important for identifying eligibility for Target Trial analyses
Morphology of the cancer, in the ICD-10-O2 system	MORPH_ICD10_O2	Number 8000-9990 or blank	x	Morphology may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Behaviour of the cancer, in the ICD-10-O2 system	BEHAVIOUR_ICD10_O2	0, 1,2,3,5,6,9,XXX,XXXX, blank	x	Behaviour may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Numeric behaviour code	BEHAVIOUR_CODED	0 = Benign, 1 = In situ, 2 = Malignant, 3 =Malignant, metastatic / secondary site, 5 = Malignant, uncertain whether primary or metastatic, 6 = Micro-invasive, 9 = Uncertain	x	Behaviour may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Description of behaviour code	BEHAVIOUR_CODED_DESC	Description of behaviour code	x	Histology may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Histology code	HISTOLOGY_CODED	Histology code	x	Histology may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Description of histology code	HISTOLOGY_CODED_DESC	Text – no validation	x	Histology may be an important prognostic factor, and important for identifying eligibility for Target Trial analyses
Grade of tumour	GRADE	GX = Grade of differentiation is not appropriate or cannot be assessed G1 = Well differentiated G2 = Moderately differentiated G3 = Poorly differentiated G4 = Undifferentiated / anaplastic	x	Grade may be an important prognostic factor
Size of the largest dimension of the tumour, in mm	TUMOURSIZE	Number or blank	x	Tumour size may be an important

				prognostic factor
Number of nodes excised	Nodes_excised_new	Number or blank	x	Number of nodes may be an important prognostic factor
Number of nodes involved	nodes_involved_new	Number or blank	x	Number of nodes may be an important prognostic factor
Laterality	LATERALITY	L = Left, R = Right, M = Midline, B = Bilateral, 8 = Not applicable, 9 = Not Known		
Multifocal	MULTIFOCAL	N= No, Y = Yes, 8 = Not applicable, 9 = Not known		
Oestrogen receptor status of the tumour	ER_STATUS	N = negative, P = positive, X = not performed		
Oestrogen receptor score of the tumour.	ER_SCORE	ER Allred score (range 0, 2-8)		
Progesterone receptor status of the tumour	PR_STATUS	N = negative, P = positive, X = not performed		
Progesterone receptor score of the tumour	PR_SCORE	ER Allred score (range 0, 2-8)		
HER2 status of the tumour	HER2_STATUS	N = negative, P = positive, X = not performed		
Nottingham Prognostic Index Score	NPI	Number (two decimal places) or blank		
Dukes' stage	DUKES	A = Dukes' A: Tumour confined to wall of bowel, nodes negative B = Dukes' B: Tumour penetrates through the muscularis propria to involve extramural tissues, nodes negative C1 = Dukes' C1: Metastases confined to regional lymph nodes (node/s positive but apical node negative) C2 = Dukes' C2: Metastases present in nodes at mesenteric artery ligature (apical node positive) D = Dukes D: Metastatic spread outside the operative field 99 = Not Known		
FIGO stage	FIGO	0, 1, 1a, 1a1, 1a2, 1b, 1b1, 1b2, 1c, 1c1, 1c2, 1c3, 2, 2a, 2a1, 2a2, 2b, 2c, 3, 3a, 3b, 3c, 3c1, 3c2, 4, 4a, 4b, I, IA, IA1, IA2, IB, IB1, IB2, IC, II, IIA, IIA2, IIB, IIC, III, IIIA, IIIB, IIIC, IIIC1, IIIC2, IV, IVA, IVB, blank		
Clark's stage	CLARKS	1, 2, 3, 4, 5, blank		
Breslow thickness of tumour	BRESLOW	Number or range, x, or blank		
Gleason primary pattern	GLEASON_PRIMARY	1-5, 8 = not applicable		
Gleason secondary pattern	GLEASON_SECONDARY	1-5, 8 = not applicable		
Gleason tertiary pattern	GLEASON_TERTIARY	1-5, 8 = not applicable		
Combined Gleason primary and secondary scores	GLEASON_COMBINED	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, blank		
T stage (pre-treatment)	T_IMG	UICC code	x	TNM may be an important prognostic variable
N stage (pre-treatment)	N_IMG	UICC code	x	TNM may be an important prognostic variable

M stage (pre-treatment)	M_IMG	0 = no distant metastasis 1, 1a, 1b, 1c, 1e = distant metastasis X = unknown	x	TNM may be an important prognostic variable
Stage at diagnosis derived from imaging	STAGE_IMG	Text	x	Stage at diagnosis may be an important prognostic variable
System used to record imaging stage at diagnosis	STAGE_IMG_SYSTEM	5 = 5th, 6 = 6th, 7 = 7th, 20 = UICC 5, 21 = UICC 6, 22 = UICC 7, 23 = AJCC 7, 24 =Unknown	x	Stage at diagnosis may be an important prognostic variable
T stage (pathology)	T_PATH	UICC code	x	TNM may be an important prognostic variable
N stage (pathology)	N_PATH	UICC code	x	TNM may be an important prognostic variable
M stage (pathology)	M_PATH	0, 1, 1a, 1b, 1c, 1e, 2, 3, 4, 9, X, blank	x	TNM may be an important prognostic variable
Pathological stage at diagnosis	STAGE_PATH	0, 0A, 0IS, 1, 1A, 1A1, 1A2, 1B, 1B1, 1B2, 1C, 1E, 2, 2A, 2B, 2C, 2E, 3, 3A, 3B, 3C, 3E, 4, 4A, 4B, 4C, 5, 6, ?, U, X, blank	x	Pathological stage may be an important prognostic variable
System used to record pathological stage at diagnosis	STAGE_PATH_SYSTEM	5, 6, 7, 20, 21, 22, 23,24, blank	x	Pathological stage may be an important prognostic variable
Pathological stage at diagnosis (pre-treatment)	STAGE_PATH_PRETREATED	Y = Yes, N = No	x	Pathological stage may be an important prognostic variable
T stage flagged by the registry as the ‘best’ T stage	T_BEST	UICC code	x	Best TNM may be an important prognostic variable
N stage flagged by the registry as the ‘best’ N stage	N_BEST	UICC code	x	Best TNM may be an important prognostic variable
M stage flagged by the registry as the ‘best’ M stage	M_BEST	UICC code	x	Best TNM may be an important prognostic variable
Best ‘registry’ stage at diagnosis of the tumour	STAGE_BEST	0, 0A, 0IS =Stage 0 1, 1A, 1A1, 1A2, 1B, 1B1, 1B2, 1C, 1E = Stage 1 2, 2A, 2A1, 2A2, 2B, 2C, 2E, 2S = Stage 2 3, 3A, 3B, 3C, 3E, 3S = Stage 3 4, 4A, 4B, 4C, 4S = Stage 4 6 = not stageable ? = insufficient information U = unstageable, X = not staged	x	Best registry stage may be an important prognostic variable
System used to record best registry stage at diagnosis	STAGE_BEST_SYSTEM	5 = 5th, 6 = 6th, 7 = 7th, 20 = UICC 5, 21 = UICC 6, 22 = UICC 7, 23 = AJCC 7, 24 =Unknown	x	Best registry stage may be an important prognostic variable
Code for the place where the diagnosis episode took place	DIAGNOSISPROVIDER_CODE	Valid provider code		

Pseudonymised diagnosis provider code	DIAGNOSISPROVIDER_CODE (pseudonymised)	To be derived on request		
Description of DIAGNOSISPROVIDER_CODE	DIAGNOSISPROVIDER_NAME	Text - no validation		
Code for the Trust at diagnosis	DIAGNOSISTRUST_CODE	Valid trust code		
Pseudonymised diagnosis trust code	DIAGNOSISTRUST_CODE (pseudonymised)	To be derived on request		
Name of the trust at diagnosis	DIAGNOSISTRUST_NAME	Text - no validation		
Tumour registration status	STATUSOFREGISTRATION	F= registration is final; P= provisional		
Excision margin	EXCISIONMARGIN	01 = Excision margins are clear (distance from margin not stated) 02 = Excision margins are clear (tumour >5mm from the margin) 03 = Excision margins are clear (tumour >1mm but less than or equal to 5mm from the margin 04 = Tumour is less than or equal to 1mm from excision margin, but does not reach margin 05 = Tumour reaches excision margin 06 = Uncertain 07 = Margin not involved =>1mm 08 = Margin not involved <1mm 09 = Margin not involved 1-5mm 98 = Not applicable 99 = Not Known	x	Essential for selection of patients for inclusion in adjuvant pancreatic cancer Target Trial, and likely to be an important prognostic factor
Screen detected cancer	SCREENDETECTED	N = No, Y = Yes, 8 = Not applicable, 9 = Not known		
Screening status of the tumour	SCREENINGSTATUSCOSD_CODE	TBC		
Description of SCREENINGSTATUSCOSD_CODE	SCREENINGSTATUSCOSD_NAME	Text - no validation		
Full detailed screening status of the tumour	SCREENINGSTATUSFULL_CODE	TBC		
Description of SCREENINGSTATUSFULL_CODE	SCREENINGSTATUSFULL_NAME	Text - no validation		
Date of first recorded event in treatment table	DATE_FIRST_EVENT	ddmmyyyy	x	Essential for analysis of treatment received
Month of first recorded event in treatment table	Month of first recorded event in treatment table	mm		
Year of first recorded event in treatment table	Year of first recorded event in treatment table	yyyy		
Days from another event to first recorded event	Option to provide number of days from another event to the first recorded event in the treatment table (e.g. days from diagnosis to first treatment event)	Derived as per applicant requirements		
Trust code of first recorded event in treatment table	TRUSTCODE_FIRST_EVENT	Valid trust code		
Pseudonymised trust code of first event	TRUSTCODE_FIRST_EVENT (Pseudonymised)	Derived as per applicant requirements		
Name of trust for first recorded event in treatment table	TRUSTNAME_FIRST_EVENT	Text - no validation		
Date of first recorded surgery in treatment table	DATE_FIRST_SURGERY	ddmmyyyy	x	Useful for including in analysis of

				treatment post-surgery
Month of first recorded surgery in treatment table	Month of first recorded surgery in treatment table	mm		
Year of first recorded surgery in treatment table	Year of first recorded surgery in treatment table	yyyy		
Days from another event to first recorded surgery in treatment table	Option to provide number of days from another event to the first recorded surgery (e.g. days from diagnosis to first recorded surgery)	Derived as per applicant requirements		
Trust code of first recorded surgery in treatment table	TRUSTCODE_FIRST_SURGERY	Valid trust code		
Pseudonymised trust code of first recorded surgery	TRUSTCODE_FIRST_SURGERY (pseudonymised)	Derived as per applicant requirements		
Name of trust for first recorded surgery in treatment table	TRUSTNAME_FIRST_SURGERY	Text - no validation		
2011 Lower Super Output Area	LSOA11_CODE	ONS code format: X00000000, blank		
2001 Lower Super Output Area	LSOA01_CODE	ONS code format: X00000000, blank		
2011 Middle Super Output Area	MSOA11_CODE	ONS code format: X00000000, blank		
2001 Middle Super Output Area	MSOA01_CODE	ONS code format: X00000000, blank		
Clinical Commissioning Group code (at diagnosis)	CCG_CODE	Code format: 00X, blank		
Name of the Clinical Commissioning Group	CCG_NAME	Text - no validation		
Primary Care Trust code the patient was resident in when the tumour was diagnosed	PCT_CODE	3 digit PCT code, blank		
Name of the Primary Care Trust the patient was resident in when the tumour was diagnosed	PCT_NAME	Text - no validation		
Local Authority Unitary Authority code the patient was resident in when the tumour was diagnosed	LAUA_CODE	00XX UA code		
Name of the Local Authority Unitary Authority the patient was resident in when the tumour was diagnosed	LAUA_NAME	Text - no validation		
Upper tier Local Authority code the patient was resident in when the tumour was diagnosed	UTLA_CODE	00XX UA code, or number, or blank		
Name of the upper tier Local Authority the patient was	UTLA_NAME	Text – no validation		

resident in when the tumour was diagnosed				
Strategic Clinical Network code the patient was resident in when the tumour was diagnosed	SCN_CODE	N44, N50, N51, N52, N53, N54, N55, N56, N57, N58, N59, N60, N61, N95, N96, Z99, blank		
Name of the Strategic Clinical Network the patient was resident in when the tumour was diagnosed	SCN_NAME	Text – no validation		
Cancer network code the patient was resident in when the tumour was diagnosed	CNET_CODE	N01, N02, N03, N06, N07, N08, N11, N12, N20, N21, N22, N23, N24, N25, N26, N27, N28, N29, N30, N31, N32, N33, N34, N35, N36, N37, N38, N39, N95, N96, Z99, blank		
Name of the cancer network the patient was resident in when the tumour was diagnosed	CNET_NAME	Text – no validation		
County code the patient was resident in when the tumour was diagnosed	COUNTY_CODE	11, 12, 16, 17, 18, 19, 21, 22, 23, 24, 26, 29, 30, 31, 32, 33, 34, 36, 37, 38, 40, 41, 42, 43, 44, 45, 47, blank	x	For descriptive statistics on treatment by location
Name of the county the patient was resident in when the tumour was diagnosed	COUNTY_NAME	Text – no validation	x	For descriptive statistics on treatment by location
Government office region code the patient was resident in when the tumour was diagnosed	GOR_CODE	A, B, D, E, F, G, H, J, K, blank		
Name of the government office region the patient was resident in when the tumour was diagnosed	GOR_NAME	East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, Yorkshire and The Humber		
Cancer registry catchment area code the patient was resident in when the tumour was diagnosed	CREG_CODE	Y0201, Y0301, Y0401, Y0801, Y0901, Y1001, Y1101, Y1201, Y1701, Z9999		
Name of the cancer registry catchment area the patient was resident in when the tumour was diagnosed	CREG_NAME	Eastern Cancer Registration & Information Centre, North West Cancer Intelligence Service, Northern & Yorkshire Cancer Registry & Information Service, Oxford Cancer Intelligence Unit, South West Cancer Intelligence Service, Thames Cancer Registry, Trent Cancer Registry, Welsh Cancer Intelligence & Surveillance Unit, West Midlands Cancer Intelligence Unit		
Country code the patient was resident in when the tumour was diagnosed	CTRY_CODE	11, 12, 16, 17, 18, 19, 21, 22, 23, 24, 26, 29, 30, 31, 32, 33, 34, 36, 37, 38, 40, 41, 42, 43, 44, 45, 47, blank		
Name of the country the patient was resident in	CTRY_NAME	Text - no validation		

when the tumour was diagnosed				
Cancer registry code which finalised the case and was responsible for sending it to ONS if it was an in-region case	CENTRE	0101, 0201, 0202, 0301, 0302, 0401, 0402, 0403, 0404, 0500, 0600, 0801, 0802, 0901, 1001, 1002, 1201, 1301, 1401, 1501, 1702, NBTR, blank,		
Name of the cancer registry which finalised the case and was responsible for sending it to ONS if it was an in-region case	CENTRENAME	ECRIC BEDFORD, ECRIC CAMBRIDGE, ECRIC IPSWICH, ECRIC NORWICH, FHSA, MERSEY MERSEYSIDE AND CHESHIRE CANCER REGISTRY,		

Cancer registration (treatment table)

Data item	**Field name**	**Description of field content**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**
Pseudonymised event ID	EVENTID	Project specific ID for each event	x	To allow analysis to take into account events
Pseudonymised tumour ID	TUMOURID	Project specific ID for each tumour	x	To allow analyses specific to tumours for each patient
Pseudonymised patient ID	PATIENTID	Project specific ID for each person	x	To allow linking between datasets
Age at diagnosis	AGE	Number or blank		
Age at diagnosis in 5 year age bands (0-4 etc.)	FIVEYEARAGEBAND	0 - 4 YRS | 5 - 9 YRS | 10 - 14 YRS | 15 - 19 YRS | 20 - 24 YRS | 25 - 29 YRS | 30 - 34 YRS | 35 - 39 YRS | 40 - 44 YRS | 45 - 49 YRS | 50 - 54 YRS | 55 - 59 YRS | 60 - 64 YRS | 65 - 69 YRS | 70 – 74 YRS | 75 - 79 YRS | 80 - 84 YRS | Blank		
Age at diagnosis in x year age bands	Option to provide age in broad categories (e.g. =<45, 46-55, 56-65, >65)	Derived as per applicant requirements		
Sex	SEX	0=Not known, 1=Male, 2=Female, 9=Not specified		
Diagnosis date	DIAGNOSISDATEBEST	ddmmyyyy		
Month of diagnosis	DIAGNOSISMONTH	mm		
Year of diagnosis	DIAGNOSISYEAR	yyyy		
Days from another event to date to diagnosis	Option to provide number of days from another event to diagnosis (e.g. days from birth to diagnosis)	Derived as per applicant requirements		
Number of tumours affected by this event	NUMBER_OF_TUMOURS	Number	x	To allow interpretation of event data

Type of event code	EVENTCODE	01a = Surgery – curative, 01b = Surgery - not curative, 01z = Surgery etc. - type unknown, 02 = Cytotoxic Chemotherapy, 03 = Hormone Therapy, 05 = RT – Teletherapy, 06 =RT – Brachytherapy, 15 = Immunotherapy, 97 = Other Treatment, 99 = Treatment unknown, CTX = CT – Other, IM = Imaging, RTX = RT - Other/NK	x	To allow analysis of event data
Description of the event	EVENTDESC	Text – no validation	x	To allow analysis of event data
Date the event took place	EVENTDATE	ddmmyyyy	x	To allow analysis of event data
Month the event took place	Month of the year the event took place	MM		
Year the event took place	EVENTYEAR	YYYY		
Days from another event to this event	Option to provide number of days from another recorded event to this event (e.g. days from diagnosis to event)	Derived as per applicant requirements		
Treatment provider (organisation code)	PROVIDERCODE	Valid institution code		
Pseudonymised treatment provider code	PROVIDERCODE (pseudonymised)	Derived as per applicant requirements		
Name of the organisation where the event took place	PROVIDERDESC	Text – no validation		
Code of the NHS Trust where the event took place	TRUST_CODE	Valid Trust code		
Pseudonymised NHS Trust code where the event took place	TRUST_CODE (pseudonymised)	Derived as per applicant requirements		
Name of the NHS Trust where the event took place	TRUST_NAME	Text – no validation		
Consultant code	PRACTITIONERCODE	Valid consultant or GP code		
Consultant code (pseudonymised by default)	PRACTITIONERCODE (pseudonymised)	To be derived for the applicant		
Consultant name	PRACTITIONERDESC	Text – no validation		
Cancer registry catchment area code the patient was resident in when the tumour was diagnosed	CREG_CODE	Y0201, Y0301, Y0401, Y0801, Y0901, Y1001, Y1101, Y1201, Y1701, Z9999		
Treatment within 6 months of diagnosis - check flag	WITHIN_SIX_MONTHS_FLAG	0 = No, 1 = Yes	x	Speed of treatment may be an important prognostic factor. Other data requested should allow us to calculate this ourselves, but this variable would provide a useful check

Treatment six months from date of diagnosis - check flag	SIX_MONTHS_AFTER_FLAG	0 = No, 1 = Yes	x	Speed of treatment may be an important prognostic factor. Other data requested should allow us to calculate this ourselves, but this variable would provide a useful check
Operations, procedures and interventions (OPCS-4)	OPCS4_CODE	Valid OPCS4 code	x	Information on procedure/intervention may be an important prognostic factor
Name of the operations, procedures and interventions	OPCS4_NAME	Text - no validation	x	Information on procedure/intervention may be an important prognostic factor
Radiotherapy code	RADIOCODE	1 = 1 + 2, 2 = 1 + 4, 3 = Brachytherapy, 4 = External beam, 5 = Intracavitary or interstitial, 8 = Other, B = Radioactive isotopes, X = Unknown / inapplicable	x	Information on any radiotherapy received may be an important prognostic factor
Radiotherapy description	RADIODESC	Text - no validation	x	Information on any radiotherapy received may be an important prognostic factor
Imaging code – internal coding system	IMAGINGCODE	Text - no validation	x	Information on imaging may be an important prognostic factor
Description of imaging	IMAGINGDESC	Text - no validation	x	Information on imaging may be an important prognostic factor
Site on body where imaging occurred	IMAGINGSITE	Text - no validation	x	Information on imaging may be an important prognostic factor
List of all systemic anti-cancer therapy drugs	CHEMO_ALL_DRUGS	Text - no validation	x	Important for analysis of treatments
Name or acronym of known drug combinations derived from CHEMO_ALL_DRUGS (e.g. R-CHOP or FEC-T)	CHEMO_DRUG_GROUP	Text - no validation	x	Important for analysis of treatments

Size in millimetres of the diameter of a lesion (histology)	LESIONSIZE	Number or blank	x	Lesion size may be an important prognostic factor

SACT dataset

Data item	Field name	Request field (mark required variables with x)	Justification - detail why the field is necessary for your analysis
Demographics and consultant			
Pseudonymised patient ID	PATIENTID	x	To allow linking of data
Pseudonymised tumour ID	TUMOURID	x	To allow linking of data
NHS number	**NHS_Number**		
NHS number status indicator code	NHS_Number_Status		
Date of birth	**Date_Of_Birth**		
Month of birth	MONTH_DOB		
Year of birth	YEAR_DOB		
Gender code (current)	Gender_Current		
Ethnicity	Ethnicity	x	Ethnicity may be an important prognostic factor
Broad ethnic group	Option to group ethnicities (e.g. white/ non-white/ unknown)		
Postcode	**Postcode**		
Broader geographic area/ IMD quintile	Option to provide geography as deprivation score or aggregate to larger geographic areas such as MSOA or county.	x	Deprivation score and geographic area (county) may be important prognostic factors
General medical practice code (patient registration)	GP_Practice_Code		
Consultant code (initiated SACT)	**Consultant_GMC_Code_Clean**		
Consultant code (pseudonymised)	Consultant_GMC_Code (pseudonymised)		
Care professional main speciality code (start SACT)	Consultant_Speciality_Code	x	Carer specialty could influence treatment given
Organisation code	Organisation_Code_of_Provider	x	Geographic area requested for descriptive statistics of treatment received
Organisation code (pseudonymised)	Organisation_Code_of_Provider (pseudonymised)		
Clinical status			
Primary diagnosis (on SACT initiation)	Primary_Diagnosis	x	Important for selection of patients in analysis
Morphology (ICD-O on SACT initiation)	Morphology_clean	x	Important for selection of patients in analysis
Pre- treatment (final) TNM stage	Stage_at_Start	x	Important for selection of

			patients in analysis
Programme and regimen			
SACT programme number	Programme_Number	x	Line of treatment is useful for summarising treatment history and determining eligibility for the Target Trial analyses. We recognise that this variable may be poorly completed, but in itself this is important to investigate so we request the data
Anti-cancer regimen number	Regimen_Number	x	Important for analysis of different treatments
Drug treatment intent	Intent_of_Treatment	x	May be an important prognostic factor
Regimen analysis grouping	Analysis_Group	x	Important for analysis of different treatments
Regimen grouping (benchmark reports)	Benchmark_Group	x	Important for analysis of different treatments
Patient's height (metres (m))	Height_At_Start_of_Regimen	x	Height and weight combined may represent a prognostic factor
Patient's weight (kilograms (kg))	Weight_At_Start_of_Regimen	x	Height and weight combined may represent a prognostic factor
Performance Status (Adult)	Performance_Status_at_Start_of_Regimen_Clean	x	Performance status is likely to represent an important prognostic factor
Performance Status (Young Person)	Performance_Status_at_Start_of_Regimen_Clean		
Co-morbidity adjustment indicator	Comorbidity_Adjustment	x	Whether comorbidity affected the clinicians decision making is important information as this could represent a confounding factor
Decision to treat date (Drug regimen)	Date_Decision_To_Treat	x	Speed of treatment may represent an important

			prognostic factor
Month of decision to treat (Drug regimen)	Month of decision to treat		
Year of decision to treat (Drug regimen)	Year of decision to treat		
Days from another event to decision to treat date	Option to provide number of days from another event to the date of the decision to treat (e.g. days from diagnosis to date of decision to treat)		
Start date (Drug regimen)	Start_Date_of_Regimen	x	Start date of treatment essential for analysis of treatment effectiveness
Month of start date for drug regimen	Month of start date of drug regimen		
Year of start date for drug regimen	Year of start date of drug regimen		
Days from another event to drug regimen start date	Option to provide number of days from another event to the start date of the drug regimen (e.g. days from date of decision to treat to start date of regimen)		
Clinical trial indicator	Clinical_Trial	x	Whether or not the person is in a clinical trial could be an important prognostic factor
Chemo-radiation indicator	Chemo_Radiation	x	Whether chemo-radiation is received could be an important prognostic factor
Number of planned systemic anti-cancer therapy cycles	Number_of_Cycles_Planned	x	Planned treatment is useful to compare to treatment actually received
Cycle			
Cycle identifier	Cycle_Number	x	Data over time is essential for comparative effectiveness analysis
Start date (Cycle)	Start_Date_of_Cycle	x	Data over time is essential for comparative effectiveness analysis
Month of start date of cycle	Month of start date of cycle		
Year of start date of cycle	Year of start date of cycle		
Days from another event to start date of cycle	Option to provide number of days from another event to the start date of the cycle (e.g. days from diagnosis to start date of cycle)		
Patient's Weight (Kilograms (kg))	Weight_At_Start_Of_Cycle	x	Weight over time could be an important prognostic factor
Performance Status (Adult)	Performance_Status_At_Start_Of_Cycle_Clean	x	Performance status is likely to represent an important prognostic factor

Performance Status (Young Person)	Performance_Status_At_Start_Of_Cycle_Clean		
Primary procedure (OPCS)	OPCS_Procurement_Code	x	Information on procedure could represent important prognostic information
Drug details			
Drug analysis grouping	Drug_Group	x	Details on treatment important for analysing effectiveness of treatment options
Actual dose	Actual_Dose_Per_Administration	x	Details on treatment important for analysing effectiveness of treatment options
SACT drug route of administration	Administration_Route	x	Details on treatment important for analysing effectiveness of treatment options
SACT administration date	Administration_Date	x	Details on treatment important for analysing effectiveness of treatment options
Organisation code (provider)	Organisation_Code_of_Drug_Provider	x	Geographic area requested for descriptive statistics of treatment received
Pseudonymised organisation code (provider)	Organisation_Code_of_Drug_Provider (pseudonymised)		
Primary procedure (OPCS)	OPCS_Delivery_Code	x	Information on procedure could represent important prognostic information
Outcome			
Start date (Final therapy)	Date_of_Final_Treatment	x	Data over time is essential for comparative effectiveness analysis
Month of final therapy	Month of final therapy		
Year of final therapy	Year of final therapy		
Days from another event to start date of final therapy	Option to provide number of days from another event to the start date of the final therapy (e.g. days from diagnosis to start date of final therapy)		
Regimen modification indicator (dose reduction)	Regimen_Modification_Dose_Reduction	x	Data over time on treatment changes is essential for comparative effectiveness analysis

Regimen modification indicator (time delay)	Regimen_Modification_Time_Delay	x	Data over time on treatment changes is essential for comparative effectiveness analysis
Regimen modification indicator (days reduced)	Regimen_Modification_Stopped_Early	x	Data over time on treatment changes is essential for comparative effectiveness analysis
Planned treatment change reason	Regimen_Outcome_Summary	x	Data over time on treatment changes is essential for comparative effectiveness analysis

Radiotherapy dataset

Data item	Description	Field name	Request field (mark required variables with x)	Justification - detail why the field is necessary for your analysis
PATIENT ID (Pseudonymised)	Project specific patient ID	PATIENTID	x	To allow linking of data
RADIOTHERAPY EPISODE IDENTIFIER (Pseudonymised)	Any identifier that is unique for each radiotherapy episode.	RADIOTHERAPYEPISODEID	x	To identify radiotherapy episodes
APPOINTMENT DATE	Date when PATIENT is to be seen by or be in contact with one or more CARE PROFESSIONALS.	APPTDATE	x	Information on radiotherapy received may represent important prognostic information
Month of appointment	Derived from APPOINTMENT DATE field	Month of appointment		
Year of appointment	Derived from APPOINTMENT DATE field	Year of appointment		
Days from another event to appointment date	Derived from APPOINTMENT DATE field	Option to provide number of days from another event to the appointment date (e.g. days from diagnosis to appointment date)		
DECISION TO TREAT DATE (RADIOTHERAPY TREATMENT EPISODE)	The date on which it was decided that the PATIENT required a specific Planned Cancer Treatment. This is the date that the consultation between the PATIENT and the clinician took place and a Planned Cancer Treatment was agreed.	DECISIONTOTREATDATE	x	Speed of treatment may be an important prognostic factor
Month of decision to treat date	Derived from DECISION TO TREAT DATE field	Month of decision to treat date		

Year of decision to treat date	Derived from DECISION TO TREAT DATE field	Year of decision to treat date		
Days from another event to decision to treat date	Derived from DECISION TO TREAT DATE field	Option to provide number of days from another event to the decision to treat date (e.g. days from diagnosis to date of decision to treat)		
EARLIEST CLINICALLY APPROPRIATE DATE	This is the first date that the patient would have been available to start radiotherapy.	EARLIESTCLINAPPROPRIATEDATE		
Month of earliest clinically appropriate date	Derived from EARLIEST CLINICALLY APPROPRIATE DATE field	Month of earliest clinically appropriate date		
Year of earliest clinically appropriate date	Derived from EARLIEST CLINICALLY APPROPRIATE DATE field	Year of earliest clinically appropriate date		
Days from another event to decision to earliest clinically appropriate date	Derived from EARLIEST CLINICALLY APPROPRIATE DATE field	Option to provide number of days from another event to the earliest clinically appropriate date (e.g. days from diagnosis to earliest clinically appropriate date)		
RADIOTHERAPY PRIORITY	The priority for this course of therapy as classified by the requesting clinician.	RADIOTHERAPYPRIORITY	x	Priority of therapy may provide important prognostic information
TREATMENT START DATE (RADIOTHERAPY TREATMENT EPISODE)	The start of a stay, an episode, period covered by a plan or other time period. This may be used to calculate the length of the period, or to classify by financial year or other time-based criterion.	TREATMENTSTARTDATE	x	Information on radiotherapy received may represent important prognostic information
Month of treatment start date	Derived from TREATMENT START DATE field	Month of treatment start date		
Year of treatment start date	Derived from TREATMENT START DATE field	Year of treatment start date		
Days from another event to treatment start date	Derived from TREATMENT START DATE field	Option to provide number of days from another event to the treatment date (e.g. days from diagnosis to treatment start date)		
RADIOTHERAPY DIAGNOSIS (ICD)	This is the PATIENT DIAGNOSIS for: • Patients with cancer, the primary tumour diagnosis code or • non-cancer diagnoses, the main condition being treated during the episode of radiotherapy Note: The definition of this field is different from that of the Primary Diagnosis in CDS.	RADIOTHERAPYDIAGNOSISICD	x	When linking data, useful corroborative information
RADIOTHERAPY INTENT	The intent of the delivered beam radiation.	RADIOTHERAPYINTENT	x	Intent of treatment may provide important

				prognostic information
PRESCRIPTION IDENTIFIER (Pseudonymised)	Any identifier that is unique for each radiotherapy prescription.	PRESCRIPTIONID		
RADIOTHERAPY TREATMENT REGION	The specific area to be treated with radiotherapy.	RTTREATMENTREGION		
ANATOMICAL TREATMENT SITE (RADIOTHERAPY)	The part of the body to which the RADIOTHERAPY ACTUAL DOSE is administered.	RTTREATMENTANATOMICALSITE	x	Site of radiotherapy may provide important prognostic information
NUMBER OF TELETHERAPY FIELDS	The prescribed number of fields of a Teletherapy Treatment Course.	NUMBEROFTELETHERAPYFIELDS		
RADIOTHERAPY PRESCRIBED DOSE	The total prescribed absorbed radiation dose in Grays	RTPRESCRIBEDDOSE		
PRESCRIBED FRACTIONS	The prescribed number of Fractions or hyperfractionation of a Teletherapy Treatment Course	PRESCRIBEDFRACTIONS		
RADIOTHERAPY ACTUAL DOSE	The total actual absorbed radiation dose given in Grays. *This item may be omitted from all but the ultimate fraction for this prescription.*	RTACTUALDOSE		
ACTUAL FRACTIONS	The total number of Fractions or hyperfractionation of a Teletherapy Treatment Course administered. *This item may be omitted from all but the ultimate fraction for this prescription.*	RTACTUALFRACTIONS		
RADIOTHERAPY TREATMENT MODALITY	The type of treatment delivered during a RADIOTHERAPY PRESCRIPTION (Teletherapy or Brachytherapy).	RTTREATMENTMODALITY		
MACHINE IDENTIFIER	A unique code ascribed to the radiotherapy equipment used to treat this exposure. This identifier is made up of: Five character NACS site code (R----) Two character equipment type code (LA/CO/KV/OT) Four digit unique sequence number (issued by RTDS).	MACHINEID		
MACHINE IDENTIFIER (pseudonymised by default)	A pseudonymised code ascribed to the radiotherapy equipment used to treat this exposure. This identifier is made up of:	MACHINEID (pseudonymised)		

	Five character NACS site code (R----) Two character equipment type code (LA/CO/KV/OT) Four digit unique sequence number (issued by RTDS).			
RADIOISOTOPE	The type of radioactive source used to deliver radiotherapy with brachytherapy. To record the isotope in standard scientific notation (e.g.: I123 or Ir192)	RADIOISOTOPE		
RADIOTHERAPY BEAM TYPE	The prescribed type of beam of a Teletherapy Treatment Course.	RADIOTHERAPYBEAMTYPE		
RADIOTHERAPY BEAM ENERGY	Beam energy in MeV/MV/MVp. Record kVp energies as decimals (e.g. 250kV = 0.25MV). Only for multi-modality machines.	RADIOTHERAPYBEAMENERGY		
TIME OF EXPOSURE	Time when the exposure was initiated	TIMEOFEXPOSURE		
ORGANISATION CODE (CODE OF PROVIDER)	This is the ORGANISATION CODE of the ORGANISATION acting as a Health Care Provider.	ORGCODEPROVIDER		
ORGANISATION CODE (CODE OF PROVIDER) - pseudonymised by default	This is a pseudonymised ORGANISATION CODE of the ORGANISATION acting as a Health Care Provider.	ORGCODEPROVIDER (pseudonymised)		
PROCEDURE (OPCS)	Procedure carried out and recorded for CDS or HES purposes.	PRIMARYPROCEDUREOPCS	x	Procedure carried out may provide important prognostic information
PROCEDURE DATE	The date of the occurrence of the CLINICAL INTERVENTION.	PROCEDUREDATE	x	Date of procedure may provide important prognostic information
Month of procedure	Derived from PROCEDURE DATE field	Month of procedure		
Year of procedure	Derived from PROCEDURE DATE field	Year of procedure		
Days from another event to procedure date	Derived from PROCEDURE DATE field	Option to provide number of days from another event to the treatment date (e.g. days from diagnosis to procedure date)		

HES admitted care

Data item	**Field name**	**Notes**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**

Patient				
Pseudonymised patient ID	PATIENTID		x	To allow linking of data
Administrative & legal status of patient	category	Available from 1989/90 to 2001/2002		
Administrative category	admincat	From 2001/2002 onwards		
Age at start of episode	startage		x	Age may represent an important prognostic factor
Date of birth - patient	dob			
Month of birth	Month of birth			
Year of birth	Year of birth			
Ethnic category	ethnos	From 1995/1996 onwards	x	Ethnic category may represent an important prognostic factor
Broad ethnic group	Option to group ethnicities (e.g. white/ non-white/ unknown)			
Postcode district of patient residence	postdist			
Postcode of patient residence	**homeadd**			
Broader geographic area/ IMD quintile	Option to provide geography as deprivation score or aggregate to larger geographic areas such as MSOA or county.			
Sex of patient	Sex			
Admissions				
Date of admission	admidate		x	HES admitted care data may provide important prognostic information, date of admission is important for linking with SACT treatment being received
Month of admission	Month of admission			
Year of admission	Year of admission			
Days from another event to admission	Option to provide number of days from another event to date of admission (e.g. days from diagnosis to admission)			
Date of decision to admit	elecdate		x	Waiting time may provide important prognostic information
Month of decision to admit	Month of decision to admit			
Year of decision to admit	Year of decision to admit			
Days from another event to date of decision to admit	Option to provide number of days from another event to date of decision to admit (e.g. days from decision to admit to admission)			
Method of admission	admimeth		x	Method of admission may provide important prognostic information
Source of admission	admisorc			
First regular day or night admission	firstreg		x	Admission history may provide important prognostic information
Waiting time	elecdur		x	Waiting time may provide important prognostic information
Class of patient	classpat			

Discharges				
Date of discharge	disdate		x	Date of discharge may provide important prognostic information
Month of discharge	Month of discharge			
Year of discharge	Year of discharge			
Days from another event to date of discharge	Option to provide number of days from another event to date of discharge (e.g. days from admission to discharge)			
Destination on discharge	disdest			
Method of discharge	dismeth			
Episodes and Spells				
Bed days within the year	bedyear			
Beginning of spell	spelbgin		x	Spell duration may provide important prognostic information
Date episode ended	epiend		x	Episode duration may provide important prognostic information
Month the episode ended	Month the episode ended			
Year the episode ended	Year the episode ended			
Days from another event to date episode ended	Option to provide number of days from another event to date episode ended (e.g. days from diagnosis to date episode ended)			
Date episode started	epistart		x	Episode duration may provide important prognostic information
Month the episode started	Month the episode started			
Year the episode started	Year the episode started			
Days from another event to date episode started	Option to provide number of days from another event to date episode ended (e.g. days from diagnosis to date episode started)			
Duration of spell	speldur			
End of spell	spelend		x	Spell duration may provide important prognostic information
Episode duration	epidur		x	Episode duration may provide important prognostic information
Episode order	epiorder		x	Episode order may provide important prognostic information
Episode type	epitype		x	Episode type may provide important prognostic information
Hospital provider spell number (pseudonymised by default)	provspno	From 1997/1998 onwards	x	Spell number may provide important prognostic information
Clinical				
All diagnosis codes	diag_4n	4 digit code up to 24 positions	x	HES admitted care data may provide important prognostic information
All diagnosis codes	diag3_3n	3 digit code up to 24 positions	x	HES admitted care data may provide important prognostic information

All operative procedure codes	opertn_nn	These fields reflect all procedures and interventions recorded through OPCS 4	x	HES admitted care data may provide important prognostic information
Date of operations	opdate_nn	These fields reflect all procedures and interventions recorded through OPCS 4	x	HES admitted care data may provide important prognostic information
Month of operations	Month of operations			
Year of operations	Year of operations			
Days from another event to date of operation	Option to provide number of days from another event to date of operation (e.g. days from admission to date of operation)			
Operation status code	operstat	From 1997-1998 onwards	x	HES admitted care data may provide important prognostic information
Intended management	intmanig		x	HES admitted care data may provide important prognostic information
Main specialty	mainspef		x	HES admitted care data may provide important prognostic information
Treatment specialty	tretspef		x	HES admitted care data may provide important prognostic information
Healthcare Resource Groups				
Dominant procedure	domproc	From 2003-2004 onwards		
Healthcare resource group (Applied HRG code from 2006-07 onwards)	hrg_3.5			
NHS-generated HRG code	hrgnhs			
NHS-generated HRG code version number	hrgnhsvn	Available from 2009/10 onwards		
SUS generated core spell HRG	suscorehrg	Available from 2009/10 onwards		
SUS generated HRG	sushrg	Available from 2009/10 onwards		
SUS generated HRG version number	sushrgvers	Available from 2009/10 onwards		
SUS generated spell ID	susspellid			
Organisation				
Commissioner code	purcode	From 1995-1996 onwards		
Commissioner code status	purval			
Commissioner's regional office	purro			
Commissioner's strategic health authority	purstha	From 2000-2001 onwards		
Commissioning serial number	csnum	From 2000-2001 onwards		
Health authority where patients GP was registered	gppracha	Available from 1999-2000 to 2000-2001 onwards		
Primary care group	pcgcode	Historically derived from 1997-1998 to 2001-2002 on same basis as 2002-2003		
Primary care trust of responsibility - historic	pctcode	Available from 2006-2007		
Primary care trust of responsibility - current	pctcode06	Historically derived from 1999-1998 to 2001-2002 on same basis as 2002-2003		

Primary care trust area where patient's GP was registered	gpprpct			
Provider code - 5 character	procode			
Pseudonymised provider code - 5 character	pseudoprocode			
Provider code - 3 character	procode3			
Pseudonymised provider code - 3 character	pseudoprocode3			
Provider code - treatment centre	procodet	Available from 1997/1998		
Site code of treatment	sitetret	Available from 2003-2004 onwards		
Pseudonymised site code of treatment	pseudositetret			
Provider type	protype	From 2000-2001 onwards		
Regional office area where patient's GP was registered	gppracro	Historically derived from 1999-1998 to 2001-2002 on same basis as 2002-2003		
Strategic health authority area where patient's GP was registered	gpprstha			
Broader geographical area where patient's GP was registered	Option to provide a broader geographic area that the patient's GP was registered (e.g. country or country)			
Geographical				
Census output area 2001	oacode	From 2003-2004 onwards		
Census output area 2001 (6 character)	oacode6			
County of residence	rescty			
Local Authority district	resladst			
Local authority district & current electoral ward	resladst_currward			
Electoral ward in 91	ward91			
Government office region of residence	resgor			
Government office region of treatment	gortreat			
Health authority of residence	resha			
Health authority of treatment	hatreat			
Patient's health authority/PCT of residence provide by NHS	pctnhs	Historically derived from 1996-1997 to 2001-2002 on same basis for 2002-2003. Derived from 2006-2007 on same basis as 2002-2003.		
Patient's primary care trust of residence - historic	respct	Available from 2006-2007 onwards		
Patient's primary care trust of residence - current	respct06	Historically derived from 1996-1997 to 2001-2002 on same basis for 2002-2003. Derived from 2006-2007 on same basis as 2002-2003?		
Patients strategic health authority of residence - historic	resstha	Available from 2006-2007 onwards		

Patients strategic health authority of residence - current	resstha06	Historically derived from 1999-1998 to 2001-2002 on same basis as 2002-2003		
Primary care trust area of treatment	pcttreat			
Region of treatment	rotreat			
Regional office of residence	Resro	Historically derived from 1999-1998 to 2001-2002 on same basis as 2002-2003		
Strategic health authority area of treatment	sthatret			
Practioner				
Code of GP practice	gpprac	Available from 1995-1996		
Code of GP practice (Pseudonymised by default)	pseudogpprac			
Consultant code	consult	Available from 1995-1996		
Consultant code (pseudonymised by default)	pseudoconsult			
Code of patient's registered or referring general medical practitioner	reggmp	Available from 1995-1996		
Code of patient's registered or referring general medical practitioner (Pseudonymised by default)	pseudoreggmp			
Person referring patient	referrer			
Referring organisation Code	referorg			
Referring organisation Code (pseudonymised)	referorg			
System Data				
Record Identifier (pseudonymised by default)	epikeyanon			
Datayear	datayear			

HES outpatient

Data Item	**Field Name**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**
Patient			
Pseudonymised patient ID	PATIENTID	x	For linking data
Administrative category	Admincat		
Ethnic category	ethnos	x	Ethic status might represent an important prognostic factor
Broad ethnic group	Option to group ethnicities (e.g. white/ non-white/ unknown)		
Appointments			
Appointment date	apptdate	x	HES outpatient data may provide important

			prognostic information. Appointment dates are important for mapping out patient timelines
Month of appointment	Month of appointment		
Year of appointment	Year of appointment		
Days from another event to appointment date	Option to provide number of days from another event to the appt date (e.g. days from diagnosis to appointment)		
Attendance identifier	attendid	x	HES outpatient data may provide important prognostic information
Attendance type	atentype		
Attended or did not attend	attended	x	HES outpatient data may provide important prognostic information
First attendance	firstatt		
Last DNA or patient cancelled date	DNAdate		
Medical staff type seeing patient	stafftyp		
Outcome of attendance	outcome	x	HES outpatient attendance outcome may provide important prognostic information
Priority type	priority	x	Priority of HES outpatient attendance may provide important prognostic information
Referral request received date	reqdate		
Service type requested	servtype		
Source of referral for outpatients	refsourc		
Days waiting	waiting	x	Waiting times may influence prognosis
Waiting/waiting calculation indicator also known as waiting quality indicator	wait_ind		
Clinical			
All diagnosis codes	diag_nn	x	HES outpatient attendance outcome may provide important prognostic information
Primary diagnosis - 4 character	diag_4	x	HES outpatient attendance outcome may provide important prognostic information
Primary diagnosis - 3 character (derived)	diag3	x	HES outpatient attendance outcome may provide important prognostic information
All operation codes	opertn_nn	x	HES outpatient attendance outcome may

			provide important prognostic information
Main operation	opertn_01	x	HES outpatient attendance outcome may provide important prognostic information
Main operation - 3 character (derived)	opertn3	x	HES outpatient attendance outcome may provide important prognostic information
Operation Status code	operstat	x	HES outpatient attendance outcome may provide important prognostic information
Main Specialty	mainspef	x	HES outpatient attendance outcome may provide important prognostic information
Treatment Specialty	tretspef	x	HES outpatient attendance outcome may provide important prognostic information
Healthcare Resource Groups			
NHS generated HRG code	hrgnhs		
NHS generated HRG code version number	hrgnhsvn		
SUS generated HRG	sushrg		
SUS generated HRG version number	sushrgvers		
Organisations			
Commissioner code	purcode		
Commissioner code (pseudonymised by default)	purcode (pseudonymised)		
Provider code - treatment	procodet		
Pseudonymised provider code	procodet (pseudonymised)		
Provider type	protype		
Geographical			
Patients census output area (2001) (10 character)	oacode01		
Patients census output area (2001) (6 character)	oacode6		
County of residence	rescty		
Government office region of residence	resgor		
Government office region of treatment	gortreat		
Patients electoral ward in 1991	ward91		
Patients Primary Care Trust of residence - current	respct06		
Patients Primary Care Trust of residence - historic	respct		
Patients Strategic Health Authority of Residence - current	resstha06		
Patients Strategic Health Authority of Residence - historic	resstha		
Practioner			
Code of GP practice	gpprac		

Code of GP practice (Pseudonymised)	gpprac (pseudonymised)		
Consultant Code	consult		
Consultant Code (Pseudonymised)	consult (pseudonymised)		
Code of patient's registered or referring general medical practitioner	reggmp		
Code of patient's registered or referring general medical practitioner (Pseudonymised)	reggmp (pseudonymised)		
Person referring patient	referrer		
Referring organisation Code	referorg		
Pseudonymised referring organisation code	referorg (pseudonymised)		
Systems data			
Record Identifier (pseudonymised by default)	attendkeyanon		
Datayear	datayear		

HES accident and emergency

Data item	**Field name**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**
Patient			
Pseudonymised patient ID	PATIENTID	x	To link data
Ethnic category	ethnos	x	Ethnic status may provide important prognostic information
Attendances			
Arrival mode	aearrivalmode		
Attendance category	aeattendcat	x	HES accident and emergency data may provide important prognostic information
Attendance disposal	aeattenddisp	x	HES accident and emergency data may provide important prognostic information
Department type	aedepttype		
Duration to assessment	initdur		
Duration to treatment	tretdur		
Duration to conclusion	concldur		
Duration to departure	depdur		
Incident location type	aeincloctype		
Patient group	aepatgroup		
Source of referral	aerefsource		
Arrival date	arrivaldate	x	HES accident and emergency data may provide important prognostic information
Day of the week of the arrival	Option to provide the day of the week the A&E arrival took place		
Arrival on a: weekday / weekend	Option to provide whether the A&E arrival was on a weekday or at the weekend		

Days from arrival date to another event	Option to provide number of days from arrival to another event (e.g. days from A&E arrival to diagnosis)		
Arrival time	arrivaltime		
Arrival time occurring in the: morning / afternoon / evening	Option to provide the part of the day the patient arrived		
Clinical diagnosis			
A&E diagnosis	diag_n	x	HES accident and emergency data may provide important prognostic information
A&E diagnosis - 2 character	diag2_n	x	HES accident and emergency data may provide important prognostic information
A&E diagnosis - Anatomical area	diaga_n	x	HES accident and emergency data may provide important prognostic information
A&E diagnosis - Anatomical side	diags_n	x	HES accident and emergency data may provide important prognostic information
Clinical Investigation			
A&E investigation	invest_n	x	HES accident and emergency data may provide important prognostic information
Clinical treatment			
A&E treatment	treat_n	x	HES accident and emergency data may provide important prognostic information
A&E treatment - 2 character	treat2_n	x	HES accident and emergency data may provide important prognostic information
Residence			
2001 Census output area	oacode		
2001 Census output area (6 character)	oacode6		
County of residence	rescty		
Current electoral ward	currward		
Current PCT of residence	respct06		
Current SHA of residence	resstha06		
Government Office Region of residence	resgor		
Health authority of residence	resha		
Historic PCT of residence	respct02		
Historic SHA of residence	resstha02		
LA district of residence	resladst		
Region of residence	resro		

Treatment			
Government Office Region of treatment	gortreat		
Health Authority of treatment	hatreat		
PCT of treatment	pcttreat		
Region of treatment	rotreat		
SHA of treatment	sthatret		
HRG data			
Dominant procedure	domproc		
Trust derived HRG value	hrgnhs		
Version no. of trust derived HRG	hrgnhsvn		
SUS generated HRG (available 2009-2010)	sushrg		
SUS generated HRG version number (Available 2009 - 2010)	sushrgvers		
Organisation data			
Provider code 3 - character	procode3		
Provider code 3 - character (pseudonymised)	procode3 (pseudonymised)		
Provider code 5 - character	procode		
Pseudonymised Provider code 5 - character	procode (pseudonymised)		
Provider code - treatment	procodet		
Pseudonymised treatment provider code	procodet (pseudonymised)		
Provider type	protype		
Patient Pathway			
Org code of patient path ID issuer	orgpppid		
RTT period start	rttperstart		
RTT period status	rttperstat		
RTTP period end	rttperend		
Duration of wait (referral to treatment period)	waitdays		
Practitioner Data			
GP practice code	gpprac		
GP practice code (pseuondymised by default)	gpprac (pseudonymised)		
System Data			
Record Identifier (pseudonymised by default)	aekeyanon		
Datayear	datayear		

Route to Diagnosis

Data item	**Field name**	**Request field (mark required variables with x)**	**Justification - detail why the field is necessary for your analysis**
Tumour level pseudo ID (for linkage)	TUMOURID	x	For linking data
Route to diagnosis code (the code assigned to a route for the purpose of the algorithm)	ROUTE_CODE	x	Route to diagnosis may provide useful prognostic information
Finalised route to diagnosis (the published route with all datasets types accounted for)	FINAL_ROUTE	x	Route to diagnosis may provide useful prognostic information

Project Administration and Governance

Dr Nicholas Latimer will undertake all analyses. Professor James Chilcott and Professor Paul Tappenden are supervising Dr Latimer's Yorkshire Cancer Research Senior Fellowship and will provide advice. Professor Jonathan Wadsley and Dr Peter Hall will provide clinical expert advice. Dr Ellie Murray and Professor Uwe Siebert will provide support relating to causal inference methods. Dr Rebecca Smittenaar will provide support relating to the linked datasets and the analysis plan.

Data Management Plan

A data sharing agreement with ODR will be required. Data will be held at the University of Sheffield and will not be shared with third parties. Data already exists and no new data will be collected for this study. The variables available and required from each existing dataset are presented in the previous section, using the table formatting provided in the NCRAS Data Dictionary. A de-personalised data extract will be performed by Public Health England and provided to Dr Nicholas Latimer at the University of Sheffield. The data are owned by Public Health England. The data will be stored securely on centrally provisioned University of Sheffield virtual servers and research data storage infrastructure as Stata datasets for a period of two years. Access control is by authorised University computer account username and password. Off-site access is facilitated by secure VPN connection authenticated by University username and remote password. By default, two copies of data are kept across two physical plant rooms, with a 28 day snapshot made of data and backed up securely offsite at least daily. This service is maintained by the University's Corporate Information and Computing Services. We will comply with the Data Protection Act and the University's own Information Security and Data Protection Policies as well as the School of Health and Related Research (ScHARR) Information Governance Policy. Because the data will be de-personalised rather than completely anonymous data will not be placed in a repository or made publicly available. On or before the effective date of termination or End Date of the data sharing agreement (expected to be 2 years after data receipt), the data provided will be securely and permanently destroyed or erased such that it cannot be recovered or reconstructed, together with all hard or soft copies of the manipulated or derived data generated from the data. In order to allow the analyses conducted during this study to be re-produced detailed information regarding the exact data extract received and the programming code used to analyse it will be recorded and made publicly available. This would allow an interested party to request the same extract of data from ODR, and to re-produce the analyses.

The data will be analysed in Stata by Dr Nicholas Latimer to estimate the comparative effectiveness of treatments for pancreatic cancer, as described above. All analyses will be documented in Stata .do files.

Dr Nicholas Latimer will be responsible for implementing the data management plan, and ensuring it is reviewed and revised if required. ODR operate a cost recovery framework, and charge for the time taken to provide the data extract. Fees will be paid by Dr Nicholas Latimer's research support fund, provided as part of his Yorkshire Cancer Research Senior Research Fellowship.

Information Governance declarations

Dr Nicholas Latimer is a *bona fide* worker at the University of Sheffield.
Dr Nicholas Latimer has been subject to personnel background checks and his employment contract includes compliance with organisational information governance standards.

Information governance awareness and mandatory training procedures are in place and Dr Nicholas Latimer is appropriately trained.

The data can be entrusted to the organisation, in the knowledge that Dr Nicholas Latimer will conscientiously discharge his obligations, including with regard to confidentiality of the data.

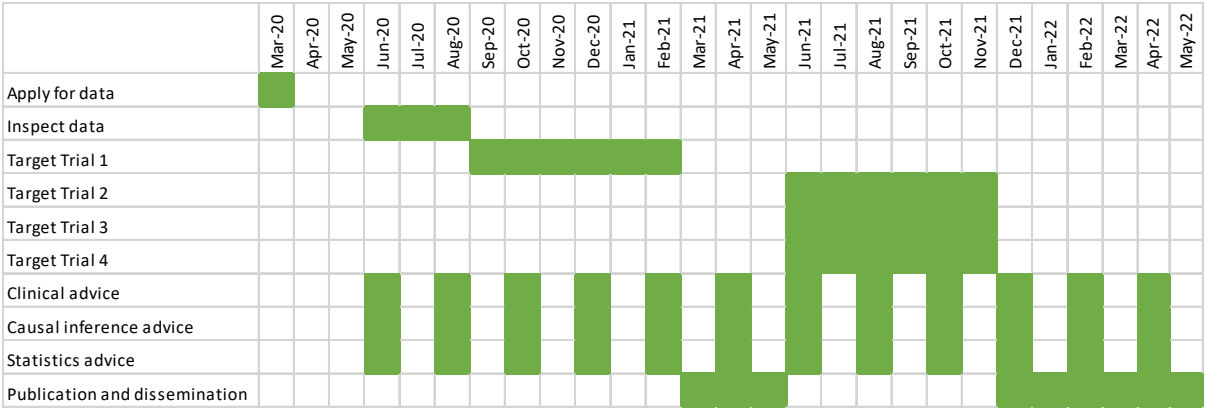
Ethical Approval

We are requesting de-personalised data and therefore have obtained Research Ethics Committee Approval (REC Committee London Bromley, REC reference 20/LO/0057, approved on 19th February 2020).

Timelines and Dissemination

The timelines for the project are shown below. These will be updated when data are obtained. Initially a period of time will be spent familiarising with the data. Then, Target Trial 1 will be completed. This will be done separately from the other Target Trials, because Target Trial 1 investigates adjuvant treatment of pancreatic cancer, whereas Target Trials 2-4 investigate metastatic and locally advanced pancreatic cancer. Upon completion of Target Trial 1 a first round of dissemination will commence, including publications in peer reviewed journals and presentations at national and/or international conferences. Following this, Target Trials 2-4 will be carried out concurrently, which is appropriate because they all involve treatments for metastatic pancreatic cancer. Following completion of these, further dissemination (peer-reviewed journal articles, conference presentations) will be undertaken. Clinical, causal inference, and statistics advice will be sought at regular intervals throughout the project. All study team members will be included in all dissemination activities.

It is possible that the data provided will be of insufficient quality for the Target Trials to be conducted. If this is the case, we will report on the reasons for this, and will comment on the data that would be required in order for appropriate analyses to be undertaken.



References

[1] Roberts C, Torgesson D. Randomisation methods in controlled trials. BMJ1998;317:1301–10.

[2] Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology 2016;183(8):758–764.

[3] Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC (2020).

[4] Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000;56(3):779-788.

[5] Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate theJoint Causal Effect of NonrandomizedTreatments. Journal of the American Statistical Association 2001;96(454):440-448.

[6] Petito LC. Assessing comparative effectiveness of cancer treatments in the SEER-Medicare linked database: A causal approach. Powerpoint presentation, October 25 2018.

[7] Garcia-Albeniz X, Hsu J, Hernan MA. The Value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. European Journal of Epidemiology 2017;32(6):495-500.

[8] Cain LE, Saag MS, Petersen M, May MT, Ingle SM, Logan R et al. Using observational data to emulate a randomised trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. International Journal of Epidemiology 2016;45(6):2038-2049.

[9] Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, Schneeweiss S. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clinical Pharmacol Ther. 2019 Sept 21. doi: 10.1002/cpt.1351.

[10] Franklin JM, Glynn RJ, Suissa S, Schneeweiss.
Emulation Differences vs. Biases When Calibrating Real-World Evidence Findings Against Randomized Controlled Trials. Clinical Pharmacol Ther. 2020 Feb 12. doi: 10.1002/cpt.1793.

[11] Franklin JM, Patorno E, Desai R, Glynn RJ, Martin D, Quinto K, Pawar A, Bessettee LG, Lee H, Garry EM, Gautam N, Schneeweiss S. Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative. Circulation. 2020 Dec 17. doi: 10.1161/CIRCULATIONAHA.120.051718.

[12] Cancer Research UK. Pancreatic cancer incidence by sex and UK country. Available from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer/incidence#heading-Zero (accessed 24/7/19).

[13] National Institute for Health and Care Excellence. Paclitaxel as albumin-bound nanoparticles with gemcitabine for untreated metastatic pancreatic cancer. Technology appraisal guidance. TA476. 6th September 2017. Available from www.nice.org.uk/guidance/ta476 (accessed 24/7/19).

[14] Cancer Research UK. Pancreatic cancer. Available from https://about-cancer.cancerresearchuk.org/about-cancer/pancreatic-cancer/survival (accessed 24/7/19).

[15] National Institute for Health and Care Excellence. Pancreatic cancer in adults: diagnosis and management. NICE guideline. NG85. 7th February 2018. Available from www.nice.org.uk/guidance/ng85 (accessed 24/7/19).

[16] Conroy T, Hammel P, Hebbar M, Abdelghani MB, Wei AC, Raoul JL et al. FOLFIRINOX or Gemcitabine as Adjuvant Therapy for Pancreatic Cancer. New England Journal of Medicine 2018;379:2395-2406.

[17] Neoptolemos JP, Palmer DH, Ghaneh P, Psarelli EE, Valle JW, Halloran CM et al. Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial. Lancet 2017;389:1011-1024.

[18] Conroy T, Desseigne F, Ychou M, Bouche O, Guimbaud R, Becouarn Y et al. FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer. New England Journal of Medicine 2011;364:1817-1825.

[19] Cunningham D, Chau I, Stocken DD, Valle JW, Smith D, Steward W et al. Phase III Randomized Comparison of Gemcitabine Versus Gemcitabine Plus Capecitabine in Patients with Advanced Pancreatic Cancer. Journal of Clinical Oncology 2009;27;33:5513-5518.

[20] Von Hoff DD, Ervin T, Arena FP, Chiorean EG, Infante J, Moore M et al. Increased Survival in Pancreatic Cancer with nab-Paclitaxel plus Gemcitabine. New England Journal of Medicine 2013;369:1691-1703.

[21] Chang J-YA, Chilcott JB, Latimer NR. (2024) Leveraging real-world data to assess treatment sequences in health economic evaluations: a study protocol for emulating target trials using the English Cancer Registry and US Electronic Health Records-Derived Database. Report. SCHARR HEDS Discussion Papers (24.01). Sheffield Centre for Health and Related Research, University of Sheffield.

[22] Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12:9.

[23] Herrmann R, Bodoky G, Ruhstaller T, Glimelius B, Bajetta E, Schuller J et al. Gemcitabine Plus Capecitabine Compared with Gemcitabine Alone in Advanced Pancreatic Cancer: A Randomized, Multicenter, Phase III Trial of the Swiss Group for Clinical Cancer Research and the Central European Cooperative Oncology Group. Journal of Clinical Oncology 2007;25(16):2212-2217.

[24] Gray E, Marti J, Brewster DH, Wyatt JC, Piaget-Rossel R, Hall PS. Real-world evidence was feasible for estimating effectiveness of chemotherapy in breast cancer: a cohort study. Journal of Clinical Epidemiology 2019;109:125-132.

[25] Hall P. Edinburgh Cancer Informatics Wiki / Cancer Types. Breast Cancer. https://www.wiki.ed.ac.uk/display/CAN/Breast+Cancer#BreastCancer-CharlsonIndexofCo-Morbidity Last updated 02 November 2019. Accessed on 27th November 2019.