Original Research Article

Health Informatics Journal

Health Informatics Journal
1–17
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14604582221077055
journals.sagepub.com/home/jhi
SAGE

Opportunities and challenges when using record linkage of routinely collected electronic health care data to evaluate outcomes of systemic anti-cancer treatment in clinical practice

Tanja Mueller
Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK

Jennifer Laskey, Kelly Baillie, Julie Clarke and Christine Crearie
NHS Greater Glasgow & Clyde, Glasgow, UK

Kimberley Kavanagh
Department of Mathematics & Statistics, University of Strathclyde, Glasgow, UK

Janet Graham
Beatson West of Scotland Cancer Centre, NHS Greater Glasgow & Clyde, Glasgow, UK

Institute of Cancer Sciences, University of Glasgow, Glasgow, UK

Kathryn Graham and Ashita Waterson
Beatson West of Scotland Cancer Centre, NHS Greater Glasgow & Clyde, Glasgow, UK

Robert Jones
Institute of Cancer Sciences, University of Glasgow, Glasgow, UK

Amanj Kurdi
Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK

Department of Pharmacology, College of Pharmacy, Hawler Medical University, Erbil, Iraq

Division of Public Health Pharmacy and Management, School of Pharmacy, Sefako Makgatho Health Sciences University, Ga-Rankuwa, South Africa

Corresponding author:
Tanja Mueller, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street,
Glasgow G4 0RE, UK.
Email: tanja.muller@strath.ac.uk

David Morrison
Public Health Scotland, Edinburgh, UK

Marion Bennie
Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK

Public Health Scotland, Edinburgh, UK

Abstract

The efficacy and safety of cancer medicines as reported from randomised clinical trials do not always translate into similar benefits in routine clinical practice; hence, post-marketing studies are a useful addition to the evidence base. With recent advances in digital infrastructure and the advent of electronically available health records, linkage of routinely collected data has emerged as a promising evaluation method for these studies. This paper discusses the opportunities and challenges when applying an electronic record linkage methodology with respect to systemic anti-cancer therapy by showcasing exemplar studies conducted over a three-year period in Scotland, and highlights some of the potential pitfalls spanning the entire breadth and depth of the research process. Our experiences as an interdisciplinary team indicate that there is scope to conduct large cohort studies to generate results from routine clinical practice within a reasonable time frame; however, close collaboration between researchers, data controllers and clinicians is required in order to obtain valid and meaningful results.

Keywords
Cancer, clinical practice, record linkage, treatment outcomes

Introduction

Globally, medicines account for a significant proportion of healthcare interventions and resource use.[1] In the setting of cancer, systemic anti-cancer treatments (SACT) are used at different stages in the treatment pathway – increasingly on a continuous basis until disease progression rather than as a fixed course of treatment. With a general trend towards ageing populations,[2,3] improvements in the early detection of cancer, and the availability of advanced treatment options, many more patients are now living with cancer and are being treated with some form of SACT for a significant proportion of their cancer journey.[4,5]

Evidence to support the safety and efficacy of new cancer medicines is obtained through randomised clinical trials (RCTs), which is then used to support licencing applications. However, results based on RCTs – the gold standard with regards to evaluating the effect of drugs on patients – may not always translate into similar benefits in routine clinical practice, with absolute outcomes such as survival frequently being inferior in a real world setting compared to those reported in the pivotal trials.[6,7] Moreover, side effects from medicines, which can be particularly prevalent with SACT, may occur with increased frequency in populations encountered in routine healthcare settings. There are a number of potential reasons for inferior outcomes being observed in routine clinical practice; most importantly, participants included in RCTs are often not fully representative of patient populations, having been carefully selected based on pre-specified inclusion and exclusion criteria. Patients receiving these medicines in routine settings are much more varied – generally being older and frailer than trial participants, with more co-morbidities.[8–10]

The discrepancy between trial results and treatment outcomes observed in clinical practice poses a risk-benefit challenge for prescribers and patients, especially in the context of SACT which often has modest survival gains and may negatively impact quality of life. Hence, to ensure that cancer treatments are of benefit to patients, post-marketing studies evaluating the clinical effectiveness and safety of these medicines in routine practice – as opposed to the tightly controlled settings of RCTs – are essential. Such observational studies are, by their nature, not able to measure the incremental benefits associated with the medicine, as it is rarely possible to derive an appropriate control population; however, they can help clinicians, patients and policymakers better understand the likely outcomes among patients treated in clinical practice.

With recent advances in digital infrastructure and the advent of electronically available health records, linkage of routinely collected data has emerged as a promising evaluation method for observational studies – mainly because record linkage offers the opportunity to conduct large-scale cohort studies, while simultaneously enabling the inclusion of information from much broader sources of data than was previously feasible.[11] Nevertheless, using administrative health data is not without its challenges; although technical difficulties have largely been overcome, certain issues – mainly regarding privacy and data availability/accessibility – remain topical.[12,13] Furthermore, cancer treatment is very complex, not least due to the different purposes of treatment – ranging from neo-adjuvant intent, that is, treatment given as a first step in order to reduce the size of a tumour prior to the main treatment (such as surgery or radiotherapy), to palliative treatment, aimed at alleviating symptoms in patients with incurable disease.

Key requirements for observational studies assessing outcomes of cancer medicines

Crucial aspects to be considered when conducting studies evaluating SACT treatment outcomes in clinical practice fall into three major categories:

- Data – demographic and clinical information to allow patients, treatments and outcomes to be described;
- Information governance – permissions to process data; and
- Interdisciplinary collaboration – the strategic and operational groups enabling and advising the work.

Data requirements mainly relate to the four areas of patient identification, treatment exposure, patient baseline characteristics, and treatment outcomes. While death records and hospitalisation data are vital to facilitate the assessment of treatment outcomes, complete regimen and indication data for the population are primary pre-requisites for any study assessing SACT treatment – allowing the identification of study eligible patients and the reliable definition of the exposure of interest, thereby decreasing the risk of bias. Summarising patient baseline characteristics is necessary to support the interpretation of study findings, and to enable comparisons to trial populations or other published studies; in addition, several disease-related factors are recognised to be important confounders and/or predictors of treatment effectiveness and/or survival.

Information governance demands that data are processed according to relevant laws and guidance, which includes, for instance, the UK General Data Protection Regulation (GDPR) and the Data Protection Act (2018); while the former remained substantially unchanged from EU GDPR, the latter provides additional rules and details UK Data Protection Authority enforcement powers. Furthermore, access to Scottish data requires affiliation with an approved organisation (e.g.

Universities or the National Health Service (NHS)); non-UK based researchers can partner with an approved organisation, or have analyses carried out on their behalf. Formal ethics committee approval may not be required if evaluation of outcomes is considered to be part of clinical audits – a component of good clinical practice; nevertheless, there is usually, at the very least, a requirement to highlight the potential benefit of a proposed study to patients and/or a healthcare system, as well as provisions relating to data anonymisation, researcher training, and the safety and security of work environments and associated IT infrastructure. Although specifics of implemented procedures to obtain access to data may differ depending on context, applications may be complex and, thus, time-consuming.

Considering the complexity of the topic and the amount and nature of data required to conduct studies focussing on outcomes of cancer medicines in clinical practice, interdisciplinarity of the study team and broad academic and health services support is essential; clinical expertise, especially in assessing patients for SACT and SACT prescribing, is crucial. Furthermore, governmental/health system engagement is recommended to facilitate rapid translation of findings into clinical practice.

Following publication of the 'Cancer Action Plan' and in line with the national Digital Strategy,[14,15] in 2016, the Scottish Government made funding available for the Cancer Medicines Outcomes Programme (CMOP), a collaborative project aimed at evaluating the feasibility of using routinely collected electronic health data to determine outcomes of SACT in clinical practice in Scotland. The programme team comprises academics, quantitative and qualitative researchers, and clinicians – spanning a variety of disciplines, specialities and areas of expertise. The main objective of the programme was to set up a series of incremental studies to test existing linkage capabilities and assess data quality with regards to cancer treatment outcomes in local populations.

The purpose of this paper is to discuss the opportunities and challenges when applying an electronic record linkage methodology with respect to SACT by showcasing a series of studies conducted between 2017 and 2020. Although firmly situated within a Scottish setting, the majority of challenges encountered are likely not exclusive to Scotland.

Methods

Study context

Scotland has a population of approximately 5.4 million, and all residents are covered by the National Health Service (NHS) Scotland – a tax-based system offering healthcare free of charge at the point of delivery, including hospital care and prescription medicines. Initial projects, serving as pilot studies, focused on NHS Greater Glasgow & Clyde, the largest Health Board (regional organisations providing services to their respective population) and covering approximately 1.2 million people (or 20% of the overall Scottish population); subsequent projects were intended to test the scalability of the methodology and aimed to include all patients residing within the West of Scotland Cancer Network, covering almost 50% of the total population of Scotland.[3,16]

Data sources

In Scotland, a range of routinely collected, national-level administrative datasets are available for research purposes, ranging from medicines dispensed in primary care to death records; completeness of these datasets is generally high due to their original purposes (planning, payment and/or population monitoring). In addition, clinical datasets providing radiotherapy treatment information and laboratory test results, curated by individual Health Boards, can be made available. Details are presented in Table 1.

Table 1. Main data sources used for cancer medicines outcomes programme record linkage projects.

Data source	Content	Data controller
National Register of Scotland (NRS)[17]	Registration of life events – death records (date and cause of death, ICD10 codes)	Public Health Scotland
Scottish Cancer Registry (SMR06)[17]	Collection of incident cancer registrations from across the health system; date of diagnosis, ICD10 and ICD-O codes, and details relating to the tumour	Public Health Scotland
Scottish Morbidity Records, outpatient attendances (SMR00)[17]	Episode level data on outpatient clinic attendances, including diagnostic codes and procedures undertaken	Public Health Scotland
Scottish Morbidity Records, inpatient and day case dataset (SMR01)[17]	Episode level data on acute hospital admissions, including diagnostic codes (ICD10) and procedures undertaken (OPCS4 codes)	Public Health Scotland
Prescribing Information System (PIS)[17]	Primary care prescribing, including date of prescription and dispensing, and drug name, dose and quantity[a]	Public Health Scotland
Chemotherapy Electronic Prescribing and Administration System (CEPAS)[18]	Systemic anti-cancer therapy prescribing, including diagnosis, drug, dose, indication, date of administration and cycle number[a]	Health Board
Scottish Care Information (SCI store)[19]	Laboratory test results (biochemistry and haematology); comprising test name, date and value	Health Board
ARIA, a radiotherapy management system	Radiotherapy records, including indication (ICD10 codes), dose/fraction and dates of administration	Health Board

ICD10 – International classification of diseases, 10th edition; ICD-O – International classification of diseases for oncology; OPCS4 – (Office of population censuses and surveys) classification of interventions and procedures.

[a] In-hospital prescribing – other than systemic anti-cancer treatment and supporting medicines administered during cancer treatment sessions – are not covered in these datasets.

Study design

All CMOP studies conducted to-date (as listed in Table 2) have been designed as retrospective cohort studies. For practical reasons and to ensure that study cohorts were composed appropriately, patients were identified based on cancer diagnosis and/or treatment, within a defined time frame, through the Chemotherapy Electronic Prescribing and Administration System (CEPAS), a system used to record the prescription of SACT.[18] Deterministic record linkage with other relevant datasets has been enabled through the availability of a unique patient identifier in Scotland, the Community Health Index (CHI) number; CHI numbers are assigned to every resident (at birth or entry into the health system), and their use is mandatory across Health and Social Care in Scotland,[20] that is, all data sourses are easily linkable. While all studies had a very similar design, analytical methods were adapted to fit specific questions depending on cancer type.

Data management

Permission to use all requested datasets has been approved by the Public Benefit and Privacy Panel for Health and Social Care,[22] and access has been granted through the Glasgow Safe Haven, a

Table 2. Cancer medicines outcomes programme studies completed or in-progress, 2017–2020.

Study topic area	Systemic anti-cancer treatment of interest	Cancer registry diagnosis[a]	Study population	Study time period
Metastatic castration-resistant prostate cancer[21]	abiraterone, enzalutamide	C61	NHS Greater Glasgow & Clyde	Cohort inclusion 02.2012–12.2015, end of follow-up 02.2017
Metastatic colorectal cancer	5-FU/capecitabine, cetuximab, FOLFIRI, FOLFOX, XELOX/CAPOX, cetuximab-FOLFIRI, cetuximab-FOLFOX, aflibercept[b]	C18-20	NHS Greater Glasgow & Clyde	Cohort inclusion 01.2015–12.2016, end of follow-up 02.2018
Advanced (unresectable or metastatic) melanoma	dabrafenib, vemurafenib, trametinib, ipilimumab, nivolumab and/or pembrolizumab	C05.1, C06.1, C12, C21, C30-32, C43, C51-52, C69.3, C69.4, C69.6, C69.9, C77-80; ICD-O 87203, 87206, 87213, 87233, 87303, 87403, 87423, 87433, 87443, 87463, 87703, 87723	West of Scotland[b]	Cohort inclusion 11.2010–12.2017, end of follow-up 03.2018
Newly diagnosed non-ovarian gynaecological cancers: cervical, endometrial, vulval/vaginal	All chemotherapy with neoadjuvant intent	C510, C511, C519, C52X, C530, C539, C541, C549	West of Scotland[b]	Cohort inclusion 01.2012–12.2016, end of follow-up 02.2018
Multiple myeloma	Pomalidomide, carfilzomib, panobinostat	C90	West of Scotland[b]	Cohort inclusion 01.2015–12.2017, end of follow-up 11.2018

As a basis, all studies comprised the following datasets: Chemotherapy Electronic Prescribing and Administration System; the Cancer Registry; Scottish Morbidity Records, Inpatient and Outpatient datasets; the Prescribing Information System; and National Records of Scotland. Locally conducted studies, focussing on metastatic castration-resistant prostate cancer and metastatic colorectal cancer, had in addition access to laboratory data and radiotherapy records. All sources have been linked deterministically using Community Health Index (CHI) numbers. 5-FU – fluorouracil; CMOP – Cancer Medicines Outcomes Programme; FOLFIRI – folinic acid, fluorouracil, irinotecan; FOLFOX – folinic acid, fluorouracil, oxaliplatin; ICD-O – International Classification of Disease for Oncology, 3rd edition; XELOX/CAPOX – capecitabine, oxaliplatin

[a] ICD10 codes unless stated otherwise;

[b] NHS Ayrshire & Arran, NHS Forth Valley, NHS Greater Glasgow & Clyde, NHS Lanarkshire.

secure, closed environment.[23] Data has been pseudonymised, and no identifying information has been made available to researchers; results to be released from the Safe Haven are subject to statistical disclosure procedures.

Findings

Identifying patients

For most studies, study populations were initially identified through CEPAS, where patients are assigned both an indication for treatment (the diagnosis) and a regimen (the treatment), accompanied by a treatment intent (e.g. palliative); the Cancer Registry was used to confirm diagnoses. In a study focussing on patients with metastatic castration-resistant prostate cancer (mCRPC), for example, patients were identified based on the treatment prescribed on CEPAS (abiraterone or enzalutamide) and the attached indication (mCRPC), as well as the patients' diagnoses of prostate cancer as recorded in the Cancer Registry.[21] Similarly, patients with metastatic melanoma were identified within CEPAS through a combination of diagnosis and treatment regimen, and eligibility for study inclusion was confirmed via ICD10 codes as recorded in the Cancer Registry (see also Table 2 for further details).

Identification of patients with some other cancers was, however, more challenging. For instance, the identification of patients newly diagnosed with a range of gynaecological cancers (cervical, endometrial and vaginal/vulval cancer) subject to neo-adjuvant treatment required a considerable amount of clinical input, for two main reasons: first, neo-adjuvant therapy was not easily identifiable in CEPAS for some cancer types, as this sub type was not necessarily included within the prescribing setup (where diagnoses can be selected from a drop-down menu); and second, the complex nature of these cancers impacts the reliability and consistency of the data available, since diagnoses may evolve as more information becomes available over time (e.g. from biopsies or scans). Consequently, identifying patients comprised several steps and required information from a number of datasets in order to exclude study ineligible patients (e.g. those with recurrent disease), including: details of received SACT; timing between diagnosis and initiation of SACT; and previous/concurrent treatment for the cancer of interest. The flowchart in Figure 1 illustrates the complexity of the patient identification process.

Defining exposure

Exposure to cancer medicines is comprehensively collected within CEPAS, including the name of the regimen, the drugs given along with their prescribed dose (i.e. the actual dose authorised to be given to the patient), and dates of administration. While defining exposure can be straightforward – as, for instance, in the case of mCRPC patients were the treatments of interest (abiraterone and enzalutamide) were both single agent oral drugs given at a fixed dose – other therapies may require further interpretation of records to enable reliable categorisation (e.g. due to complex treatment schedules involving multiple drugs with some variability of sequencing and/or dosing of medication).

As an example: patients with metastatic colorectal cancer (mCRC) may be treated with a range of different regimens depending on circumstances, comprising both chemotherapy (most prominently fluorouracil (5-FU) and/or oxaliplatin) and/or targeted treatments (e.g. aflibercept or cetuximab). If applicable and appropriate, doublet or triplet therapy (combining more than one drug) is preferred over monotherapy. The number and combination of drugs given might, however, change over time for various reasons; changes may include 'stepping up' of treatment, that is, adding a drug to the

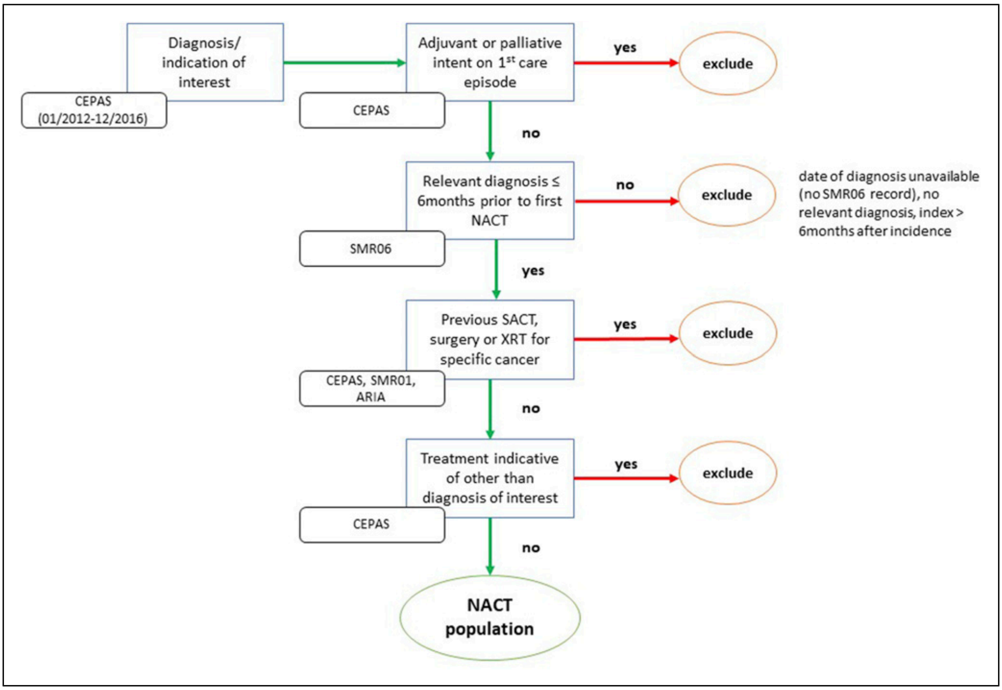
Diagnosis/ indication of interest
CEPAS (01/2012-12/2016)
Adjuvant or palliative intent on 1st care episode
CEPAS
yes
exclude
no
Relevant diagnosis ≤ 6months prior to first NACT
SMR06
no
exclude
date of diagnosis unavailable (no SMR06 record), no relevant diagnosis, index > 6months after incidence
yes
Previous SACT, surgery or XRT for specific cancer
CEPAS, SMR01, ARIA
yes
exclude
no
Treatment indicative of other than diagnosis of interest
CEPAS
yes
exclude
no
NACT population

Figure 1. Generic flowchart detailing identification of a neo-adjuvant chemotherapy patient cohort in gynaecological cancers. ARIA – radiotherapy treatment records dataset; CEPAS – Chemotherapy prescribing and administration system; NACT – neo-adjuvant chemotherapy; SACT – systemic anti-cancer treatment; SMR01 – Scottish morbidity records, inpatient dataset; SMR06 – Scottish cancer registry; XRT – radiotherapy.

previously prescribed drug – this might occur if there are initial concerns with regards to potential side effects and a patient's fitness; or 'stepping down' of treatment, that is, removing a drug from an originally given regimen, usually due to a patient experiencing side effects relating to one of the drugs included in the combination.

Unfortunately, there is some variability in how these cases are handled in CEPAS, and it was not always possible to determine the reason for changes in treatment. Following in-depth discussions with clinicians involved in treating mCRC patients, a set of rules was developed to facilitate the simplification of exposure while acknowledging the prescribing intention; these rules were based on (a) the sequence of drugs prescribed; (b) the number of cycles given; and (c) the timing of different regimens and/or drugs. Figure 2 illustrates two examples.

Describing patient baseline characteristics

In Scotland, data on age and sex are incorporated into a person's CHI number and are included in all datasets. In addition, a number of geographical and socio-demographic information is available from the majority of health-related datasets; these usually include the Health Board of residence and the postcode area; an urban/rural classification[24]; and the Scottish Index of Multiple Deprivation.[25] Information regarding co-morbidities can be obtained from hospital and outpatient clinic records, and data on medicines prescribed and dispensed in primary care may be used as a proxy for long-term conditions that do not require hospital admission; nevertheless, comorbidities might be missed

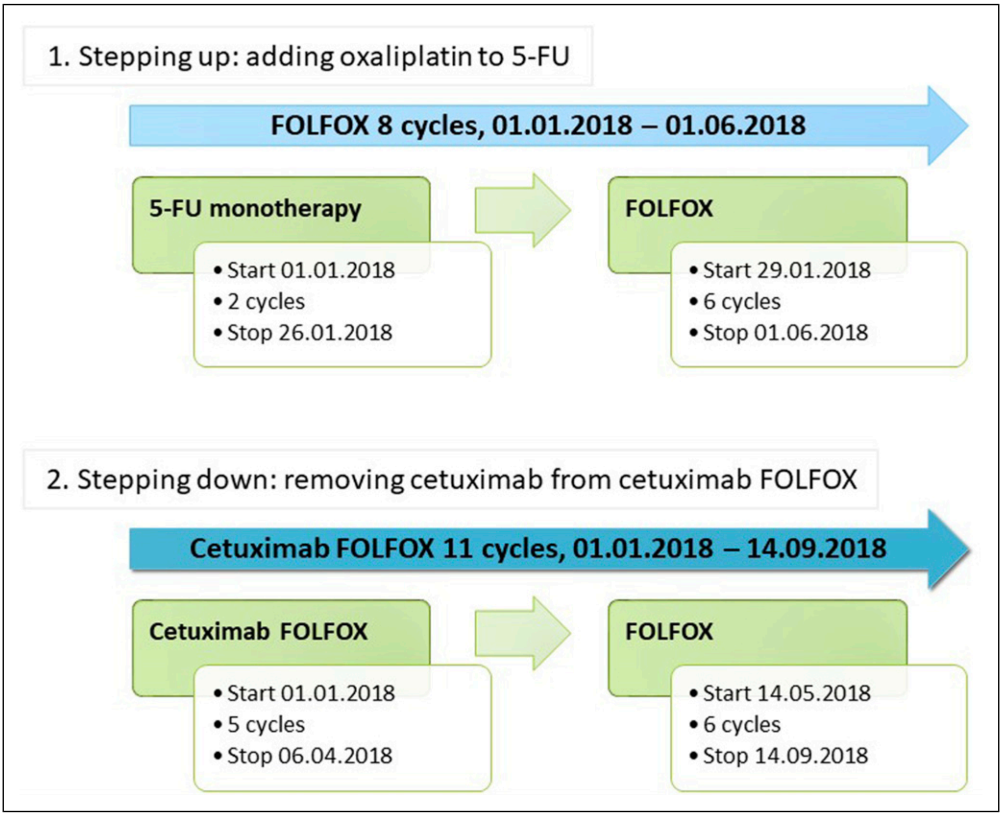

Figure 2. Examples of re-coding exposure due to stepping up or down of treatment. 5-FU – 5-fluoro-uracil; FOLFOX – folinic acid, fluorouracil, oxaliplatin.

if an individual has not been hospitalised and medication cannot easily be used as a proxy for the condition in question. Hospital records also comprise details with regards to procedures undertaken (see also Table 1).

Regrettably, additional information relevant to cancer treatment is not always available in Scotland for record linkage studies for a number of reasons, with potential implications on the accuracy and comparability of results. Data stemming from (molecular) pathology – such as BRAF or RAS status – for instance are recorded locally by the departments conducting these tests, and stored in non-standardised formats (e.g. spreadsheets); while the presence of CHI numbers theoretically enables these datasets to be linked to other data sources, locating, accessing, and processing these files requires considerable resources, both in terms of time and technical expertise. This particularly affects studies aimed at combining patients across Health Boards, as resources and/or processes may differ considerably between regions. The Cancer Registry, in contrast, routinely captures patients from across Scotland; however, completion of all fields is not mandatory. While the Registry offers the opportunity to collect additional useful information related to the disease including pathology/cell type and cancer staging based on the TNM (Tumour-Node-Metastases) system[26] or cancer-type appropriate methods (e.g. FIGO stage in patients with gynaecological cancers),[27] some of this data is frequently incomplete. In addition, records are entered at time of diagnosis and are usually not updated at recurrence – that is, the location of metastases, for example, is only recorded if these were present at the time of diagnosis, but not if patients subsequently develop metastases.

Defining treatment outcomes

The most common outcome measures as reported from both clinical trials and observational studies of cancer medicines are median overall survival (OS) and landmark survival figures (e.g. % alive after 5 years). Death records, comprising both date and cause of death, can easily be linked to other patient records in Scotland; therefore, analyses of these outcomes among cancer patients are well supported. CMOP studies conducted thus far have provided insights into OS among mCRPC patients treated with abiraterone or enzalutamide; patients with metastatic melanoma or mCRC undergoing treatment with chemotherapy, targeted treatment and/or immunotherapies; and patients with gynaecological cancers (cervical, endometrial, and vulval cancer) subject to neo-adjuvant treatment.

Nevertheless, progression-free survival (PFS) is increasingly used to establish early efficacy in advanced cancers in RCTs with small sample sizes and/or short duration.[28,29] PFS for solid tumours is usually based on imaging data; however, in haematological malignancies such as multiple myeloma, it is determined from changes in blood plasma levels of, for example, paraprotein.[30] Laboratory test values are, to a certain extent, available for record linkage in Scotland; however, these are collected and stored locally within individual NHS Health Boards, and the existing infrastructure does not easily support the combination of data from different areas across Scotland for various reasons (including, but not limited to, the use of different lab equipment with diverging measurement standards and variability in the coding of tests). This, in effect, currently limits the capability of conducting studies across regions where laboratory test data is required for analyses. Imaging data – and, in extension, any other data not provided in a standardised format (such as patient-reported outcomes) – is not routinely available for use in record linkage studies. National Language Processing and other Machine Learning techniques enabling analyses of these data are, however, in development, and have in certain areas already proven their usefulness[31,32]; with further progress being made, PFS might be added to the list of treatment outcomes readily available for analyses in the future.

On a related note, side effects of treatment are of considerable interest, especially if these can lead to serious events or may potentially have long-lasting negative effects; for instance, assessing the occurrence of immune-related adverse events – which can range from minor skin irritation to severe pneumonia – is vital in light of the increasing use of immune checkpoint inhibitors in various cancers. Information regarding adverse events can be obtained from hospital records, or using prescribing data as proxy; however, the range of side effects identifiable through these datasets is limited to those that are serious enough to require hospitalisation, or result in the prescribing of specific medication that can be attributed to the side effect in question (e.g. newly initiated levothyroxine due to immune-related adverse events; or anti-nausea medicines prescribed as supporting medicines on CEPAS for traditional chemotherapy). Nevertheless, some instances of CEPAS provide the option to directly record treatment toxicities; with anticipated upgrades to the software in several cancer centres in Scotland, a more comprehensive assessment of cancer treatment side effects may become possible.

Discussion

Observational post-marketing studies substantially add to the evidence base upon which treatment decisions are made; this is true for all clinical areas, but especially in cancer where medicines often have a narrow risk/benefit ratio and promising RCT findings might not entirely translate into benefits in routine practice.[7–9] Although existing disadvantages of observational studies need to be kept in mind – most prominently the potential for confounding and other biases – these studies can offer a useful overview of the effectiveness and safety of medicines in clinical practice.[33,34] Furthermore, they present an invaluable opportunity to understand treatment within the ever-changing clinical environments outwith RCTs; this is of particular importance in light of COVID-19, which has considerably impacted the way SACT is being provided to patients.[35]

CMOP studies conducted to-date have indicated that record linkage of routinely collected data to determine outcomes of treatment with cancer medicines is feasible, albeit currently within certain limits – and keeping in mind the presence of a unique patient identifier in Scottish health records, obviating the necessity to employ complex, probabilistic methods. These studies offer a range of opportunities: informing clinicians of likely treatment outcomes in their local population; enabling patients to make better informed treatment choices, particularly at the end of life; allowing local clinical guidelines to be tailored to support the safe delivery of SACT and reduce the risk of side effects; and supporting the assessment of the relative value of medicines in local populations. The usefulness of record linkage to conduct pharmacoepidemiological studies in the area of cancer is supported by experiences gained in other clinical areas where this method has been used for many years; studies have, for example, used Scottish data to determine the effectiveness and safety of cardiovascular drugs; evaluate the effects of in-utero exposure to antihypertensive drugs on neonatal outcomes; and calculate the cost burden of Clostridium difficile infections, to name just a few.[36–38]

Although linkage of electronic health records has also frequently been used to conduct epidemiological studies in cancer – for example, analysing the association between atopic dermatitis and the risk to develop colorectal cancer[39]; or the impact of previous non-melanoma skin cancer on the occurrence of further primary tumours[40] – there is little evidence to-date with regards to studies focussing specifically on cancer medicine outcomes. Notable exceptions include, for example, studies analysing: outcomes of chemotherapy among patients with ovarian cancer in the Netherlands[41]; factors influencing breast cancer chemotherapy in the United States[42]; and outcomes of immunotherapy in patients with lung cancer, again in the United States.[43] It is, however, worth pointing out that studies conducted in the United States were limited to beneficiaries of specific health insurance schemes, possibly limiting the comparability of findings due to selection bias.

Key challenges

While the description of patient populations and the calculation of median OS are supported by record linkage, potential pitfalls span the entire breadth and depth of the research process – from data acquisition to selecting variables for statistical models. The following issues are not necessarily restricted to cancer studies but deserve particular consideration in this context.

Data

Some data that would be useful for analyses is currently not routinely available for record linkage; or cannot easily be obtained. Most importantly, there are difficulties collating laboratory and molecular pathology data from across Health Boards, potentially affecting the scope of studies. In addition, across Health Boards, different versions of some data sources (particularly CEPAS) are used, making comparability of data and mapping and merging of datasets from different regions challenging.

The non-availability of specific variables might influence the ability to answer certain research questions; information regarding some potential toxicities of medicines is, for instance, not routinely captured across Scotland. Particular attention needs to be paid when attempting to use information that is contained in non-mandatory data fields of a dataset, such as socio-economic data (including, e.g. ethnicity or marital status); this data has most likely a high degree of missing-ness. Unfortunately, the same is true for data pertaining to treatment modifications and/or side effects in CEPAS. Furthermore, there is a considerable lag time (currently approximately 12–18 months) in the collection of Cancer Registry data, meaning that very recent data will likely be unavailable.

Most of these issues are, however, not unique to Scotland. Challenges with incomplete or out-of-date information in Cancer Registry data has, for example, also been reported from the United States,[44,45] whereas details relating to SACT and laboratory test results appear to be incomplete in the Netherlands[46]; previous research conducted in the United States also advised caution when using data based on laboratory tests – not least due to diverging data recording conventions between health service providers.[47] Calls to standardise data collection, coding, and storage, and to implement nation-wide infrastructure with possible re-use of data in mind, have been made for many years.[48–50]

For the majority of datasets regularly used for research purposes, metadata is readily available; it is advisable to consult these prior to making any decisions with respect to study data. Nevertheless, it is highly recommended to collaborate with clinicians familiar with, ideally, both the treatment area of interest and the local systems used to generate and collect data. Many of the issues encountered while conducting the aforementioned studies – spanning from difficulties identifying the correct patients to uncertainties with regards to the appropriate categorisation of treatment exposure – were resolvable by discussing clinical procedures and treatment details with oncologists and other health care professionals working within these clinical settings.

Information governance

In Scotland, the current approval process to access routinely collected electronic data is both complicated and time consuming. Permissions to use data have to be sought from the Public Benefit and Privacy Panel for Health and Social Care,[22] potentially followed by an application to use one of four existing local Safe Havens[51] depending on how data is intended to be accessed. Subsequently, extraction, anonymisation, sense checking, and uploading of data onto the selected Safe Haven to grant researchers access may require input from several parties, depending on the datasets in question. Overall, this process might take several months – this needs to be taken into account at the planning stage of a study.[52]

Interdisciplinary collaboration

CMOP has highlighted the benefit of researchers and clinicians collaborating on ambitious projects aimed at generating results useful for clinical practice by harnessing the wealth of routinely collected data. Study findings have been shared with health care professionals, and clinicians are beginning to use the data to inform discussions regarding treatment decisions with their patients. In addition, initial conversations have taken place with the Scottish Government, intended to potentially impact future decisions related to data collection, recording and access. Activities are already ongoing with regards to improving the Scottish Cancer Registry (level of completeness); upgrading CEPAS to version 6 (with the possibility of unifying data collection across Scotland and adding further information); and – potentially – implementing a Scotland-wide laboratory resource.

Next steps

The Scottish Government has provided additional funding, enabling CMOP to continue its work. Building on the experiences of the first 3 years, the next phase is aimed at solving some of the existing challenges; upscaling select projects to cover Scotland-wide patient populations; and further developing the applied methodology. While emphasising a drive towards quality improvement in Scotland, CMOP will also investigate the utility of data stemming from clinical practice to inform medicines assessment and reimbursement decisions; and explore the integration of patient reported outcomes measures related to cancer medicines into the Health Care System.

Conclusion

Based on the experiences made in Scotland to-date, using electronic record linkage to evaluate the clinical effectiveness and safety of cancer medicines in day-to-day practice is feasible, and may offer scope to conduct large cohort studies to generate results from routine clinical practice within a reasonable time frame. Nevertheless, large-scale collaboration between researchers, data controllers, clinicians, and policy makers is urgently required in order to further improve existing systems and processes.

Acknowledgements

The work of the Cancer Medicines Outcomes Programme is only possible because of the wealth of information collected by the NHS as part of routine clinical practice.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: KB is currently attending the 2019-2021 PM Clinical Leadership in Pharmacy in Scotland programme, which is sponsored by Daiichi Sankyo UK, GlaxoSmithKline, and Napp Pharmaceuticals. RJ has received institutional grants from Astellas and Bayer; and personal fees from Astellas, Janssen, Pfizer, and Bayer, outside the submitted work. JL, KB, JC, CC, JG, KG, AW, DM and MB are employees of the National Health Services Scotland/Public Health Scotland. TM, KK and AK have no competing interests.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Scottish Government. The funder has no role in the study design, data collection, data analyses, or interpretation of data; or in the preparation or review of manuscripts for publication.

Data availability

NHS data is confidential, and is only available upon request subject to approval by a Caldicott Guardian/the Public Benefit and Privacy Panel for Health and Social Care.

Ethical approval

The use of NHS data was approved by the local Caldicott Guardian (NHS Greater Glasgow & Clyde) and/or the Public Benefit and Privacy Panel for Health and Social Care (study numbers 1617-0371/1917-0371; 1819-

0055). In addition, accessing data was approved by the NHS Greater Glasgow & Clyde Safe Haven (study numbers GSH/17/ON/003; GSH/18/ON/012; GSH/19/ON/002); permissions include ethical approval (REC numbers for devolved rights to approve projects: 12/WS/0142; 17/WS/0237). All studies have been conducted in accordance with information governance standards; no identifiable data was available to researchers.

ORCID iD

Tanja Mueller https://orcid.org/0000-0002-0418-4789

References

1. OECD. *Health at a Glance 2019: OECD Indicators*. Paris, France: OECD; 2019. DOI: 10.1787/4dd50c09-en.
2. United Nations. *Ageing*. United Nations: Global Issues. Available at: https://www.un.org/en/sections/issues-depth/ageing/ (accessed 15 October 2020).
3. National Records of Scotland. *Mid-2019 Population Estimates Scotland. Statistics and Data*. Available at: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/mid-2019 (accessed 15 October 2020).
4. Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics. *A Cancer J Clinicians* 2019; 69(5): 363–385. DOI: 10.3322/caac.21565.
5. Forster V. Surviving cancer: How big data is helping patients live longer, healthier lives. Available at: https://www.lshtm.ac.uk/research/research-action/features/surviving-cancer-how-big-data-helping-patients-live-longer (accessed 15 October 2020).
6. Naci H, Davis C, Savović J, et al. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European medicines agency, 2014-16: Cross sectional analysis. *BMJ* 2019: 366: l5221. DOI: 10.1136/bmj.l5221.
7. Di Maio M, Perrone F and Conte P. Real-world evidence in oncology: Opportunities and limitations. *Oncologist* 2020; 25(5): e746–e752. DOI: 10.1634/theoncologist.2019-0647.
8. Jin S, Pazdur R and Sridhara R. Re-evaluating eligibility criteria for oncology clinical trials: Analysis of investigational new drug applications in 2015. *J Clin Oncol* 2017; 35(33): 3745–3752. DOI: 10.1200/JCO.2017.73.4186.
9. Kim ES, Bruinooge SS, Roberts S, et al. Broadening eligibility criteria to make clinical trials more representative: American society of clinical oncology and friends of cancer research joint research statement. *JCO* 2017; 35(33): 3737–3744. DOI: 10.1200/JCO.2017.73.7916.
10. Scher KS and Hurria A. Under-representation of older adults in cancer registration trials: Known problem, little progress. *JCO* 2012; 30(17): 2036–2038. DOI: 10.1200/JCO.2012.41.6727.
11. Pavis S and Morris AD. Unleashing the power of administrative health data: The Scottish model. *Public Health Res Pr* 2015; 25(4): e2541541. DOI: 10.17061/phrp2541541.
12. Künn S. *The Challenges of Linking Survey and Administrative Data*. Bonn: IZA World of Labor, 2015.
13. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. *Big Data Soc* 2017; 4(2): 205395171774567. DOI: 10.1177/2053951717745678.
14. Scottish Government. *Beating Cancer: Ambition and Action*. Scottish Government, 2016. Available at: https://www.gov.scot/publications/beating-cancer-ambition-action/ (accessed 15 October 2020).
15. Scottish Government. *EHealth Strategy 2014-2017*. Scottish Government, 2015Available at: https://www.gov.scot/publications/ehealth-strategy-2014-2017/ (accessed 15 October 2020).
16. NHS Scotland. West of Scotland Cancer Network (WoSCAN). Available at: https://www.woscan.scot.nhs.uk/ (accessed 10 December 2020).

17. Information Services Division. *National Data Catalogue: National Datasets*. ISD Scotland National Data Catalogue. Available at: https://www.ndc.scot.nhs.uk/National-Datasets/index.asp (accessed 15 October 2020).
18. CIS Oncology. *ChemoCare goes live across NHS Scotland*. Available at: https://www.cis-healthcare.com/latest/chemocare-goes-live-across-nhs-scotland/ (accessed 15 October 2020).
19. NHS National Services Scotland. *SCI Store*. Scottish Care Information. Available at: https://www.sci.scot.nhs.uk/products/store/store_main.htm (accessed 15 October 2020).
20. Information Services Division. *Data Dictionary A-Z: CHI number*. ISD Scotland Data Dictionary. Available at: https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128&Title=CHI%20Number (accessed 15 October 2020).
21. Baillie K, Mueller T, Pan J, et al. Use of record linkage to evaluate treatment outcomes and trial eligibility in a real-world metastatic prostate cancer population in Scotland. *Pharmacoepidemiol Drug Saf* 2020; 29(6): 653–663. DOI: 10.1002/pds.4998.
22. Public Health Scotland. *Public Benefit and Privact Panel for Health and Social Care - HSC-PBPP*. Available at: https://www.informationgovernance.scot.nhs.uk/pbpphsc/ (2020, accessed 15 October 2020).
23. NHS Greater Glasgow & Clyde. Glasgow Safe Haven. Available at: https://www.nhsggc.org.uk/about-us/professional-support-sites/safe-haven/ (2020, accessed 15 October 2020).
24. Scottish Government. *Defining Scotland by Rurality*. Scottish Government Urban Rural Classification. Available at: https://www2.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification (2020, accessed 15 October 2020).
25. Scottish Government. *The Scottish Index of Multiple Deprivation*. Statistics. Available at: https://www2.gov.scot/SIMD (2020, accessed 15 October 2020).
26. Union for International Cancer Control. What is TNM? Resources. Available at: https://www.uicc.org/resources/tnm (2020, accessed 18 November 2020).
27. FIGO Committee on Gynecologic Oncology. FIGO staging for carcinoma of the vulva, cervix, and corpus uteri. *Int J Gynecol Obstet* 2014; 125(2): 97–98, DOI: 10.1016/j.ijgo.2014.02.003.
28. Villaruz LC and Socinski MA. The clinical viewpoint: Definitions, limitations of RECIST, practical considerations of measurement. *Clin Cancer Res* 2013; 19(10): 2629–2636. DOI: 10.1158/1078-0432.CCR-12-2935.
29. Zhu J, Yang Y, Tao J, et al. Association of progression-free or event-free survival with overall survival in diffuse large B-cell lymphoma after immunochemotherapy: a systematic review. *Leukemia* 2020; 34(10): 2576–2591. DOI: 10.1038/s41375-020-0963-1.
30. Tate JR. The paraprotein–an enduring biomarker. *Clin Biochem Rev* 2019; 40(1): 5–22.
31. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. *Radiographics* 2016; 36(1): 176–191. DOI: 10.1148/rg.2016150080.
32. McTaggart S, Nangle C, Caldwell J, et al. Use of text-mining methods to improve efficiency in the calculation of drug exposure to support pharmacoepidemiology studies. *Int J Epidemiol* 2018; 47(2): 617–624. DOI: 10.1093/ije/dyx264.
33. Greenfield S and Platt R. Can observational studies approximate RCTs? *Value in Health* 2012; 15(2): 215–216. DOI: 10.1016/j.jval.2012.01.003.
34. Greenfield S. Making real-world evidence more useful for decision making. *Value in Health* 2017; 20(8): 1023–1024. DOI: 10.1016/j.jval.2017.08.3012.
35. Scottish Government. *Recovery and Redesign: An Action Plan for Cancer Services*. Available at: https://www.gov.scot/publications/recovery-redesign-action-plan-cancer-services/ (2020, accessed 12 January 2021).

36. Mueller T, Alvarez-Madrazo S, Robertson C, et al. Comparative safety and effectiveness of direct oral anticoagulants in patients with atrial fibrillation in clinical practice in Scotland. *Br J Clin Pharmacol* 2019; 85(2): 422–431. DOI: 10.1111/bcp.13814.
37. Fitton CA, Fleming M, Steiner MFC, et al. Utero antihypertensive medication exposure and neonatal outcomes. *Hypertension* 2020; 75(3): 628–633. DOI: 10.1161/HYPERTENSIONAHA.119.13802.
38. Robertson C, Pan J, Kavanagh K, et al. Cost burden of clostridioides difficile infection to the health service: A retrospective cohort study in Scotland. *J Hosp Infect* 2020; 106(3): 554–561. DOI: 10.1016/j.jhin.2020.07.019.
39. Chou W-Y, Lai P-Y, Hu J-M, et al. Association between atopic dermatitis and colorectal cancer risk: A nationwide cohort study. *Medicine* 2020; 99(1): e18530. DOI: 10.1097/MD.0000000000018530.
40. Ong ELH, Goldacre R, Hoang U, et al. Subsequent primary malignancies in patients with nonmelanoma skin cancer in England: A national record-linkage study. *Cancer Epidemiol Biomarkers Prev* 2014; 23(3): 490–498. DOI: 10.1158/1055-9965.EPI-13-0902.
41. Houben E, van Haalen HGM, Sparreboom W, et al. Chemotherapy for ovarian cancer in the Netherlands: a population-based study on treatment patterns and outcomes. *Med Oncol* 2017; 34(4): 50. DOI: 10.1007/s12032-017-0901-x.
42. Kurian AW, Lichtensztajn DY, Keegan THM, et al. Patterns and predictors of breast cancer chemotherapy use in kaiser permanente Northern California, 2004-2007. *Breast Cancer Res Treat* 2013; 137(1): 247–260. DOI: 10.1007/s10549-012-2329-5.
43. Khozin S, Carson KR, Zhi J, et al. Real-world outcomes of patients with metastatic non-small cell lung cancer treated with programmed cell death protein 1 inhibitors in the year following U.S. regulatory approval. *The Oncologist* 2019; 24(5): 648–656. DOI: 10.1634/theoncologist.2018-0307.
44. Pezzi CM. Big data and clinical research in oncology: The good, the bad, the challenges, and the opportunities. *Ann Surg Oncol* 2014; 21(5): 1506–1507. DOI: 10.1245/s10434-014-3519-7.
45. Weber SC, Seto T, Olson C, et al. Oncoshare: Lessons learned from building an integrated multi-institutional database for comparative effectiveness research. *AMIA Annu Symp Proc* 2012; 2012: 970–978.
46. van Herk-Sukel MPP, van de Poll-Franse LV, Lemmens VEPP, et al. New opportunities for drug outcomes research in cancer patients: The linkage of the Eindhoven cancer registry and the PHARMO record linkage system. *Eur J Cancer* 2010; 46(2): 395–404. DOI: 10.1016/j.ejca.2009.09.010.
47. Wiitala WL, Vincent BM, Burns JA, et al. Variation in laboratory naming conventions in EHRs within and between hospitals: A nationwide longitudinal study. *Med Care* 2019; 57(4): e22–e27. DOI: 10.1097/MLR.0000000000000996.
48. Clauser SB, Wagner EH, Bowles EJA, et al. Improving modern cancer care through information technology. *Am J Prev Med* 2011; 40(5 Suppl 2): S198–S207. DOI: 10.1016/j.amepre.2011.01.014.
49. Kanas G, Morimoto L, Mowat F, et al. Use of electronic medical records in oncology outcomes research. *Clinicoecon Outcomes Res* 2010; 2: 1–14.
50. Scottish Government. *Medicines Use and Digital Capabilites - Building Capability to Assess Real-World Benefits, Risks and Value of Medicines: Towards a Scottish Medicines Intelligence Unit*. Data Scoping Taskforce. Available at: https://www2.gov.scot/Resource/0054/00540468.pdf (2018, accessed 15 October 2020).
51. Scottish Government. *A Charter for Safe Havens in Scotland*. Scottish Government, 2015. Available at: https://www.gov.scot/binaries/content/documents/govscot/publications/agreement/2015/11/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/documents/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/charter-safe-havens-scotland-handling-unconsented-data-

national-health-service-patient-records-support-research-statistics/govscot%3Adocument/00489000.pdf (accessed 15 October 2020).

52. Hanna C, Lemmon E, Ennis H, et al. Creation of the first national linked colorectal cancer dataset in Scotland: prospects for future research and a reflection on lessons learned: Creation of Scottish colorectal cancer dataset for research purposes. *IJPDS* 2021; 6(1): 1654. DOI: 10.23889/ijpds.v6i1.1654.

Appendix

Abbreviations

CEPAS	Chemotherapy electronic prescribing and administration system
CMOP	Cancer medicines outcomes programme
CHI	Community health index
ICD10	International classification of diseases, 10th edition
mCRPC	Metastatic castration-resistant prostate cancer
mCRC	Metastatic colorectal cancer
NHS	National health service
OS	Overall survival
PFS	Progression-free survival
RCT	Randomised clinical trial
SACT	Systemic anti-cancer treatment