Creation of the first national linked colorectal cancer dataset in Scotland

International Journal of Population Data Science

Journal Website: www.ijpds.org

Creation of the first national linked colorectal cancer dataset in Scotland: prospects for future research and a reflection on lessons learned

Catherine R Hanna[1], Elizabeth Lemmon[2], Holly Ennis[3], Robert J Jones[1], Joy Hay[4], Roger Halliday[5], Steve Clark[6], Eva Morris[7], and Peter Hall[8]

Submission History

Submitted:	21/01/2021
Accepted:	19/02/2021
Published:	31/03/2021

[1]CRUK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, 1042 Great Western Road, Glasgow, G12 OYN
[2]Edinburgh Health Economics, University of Edinburgh,NINE BioQuarter 9 Little France Road Edinburgh EH16 4UX
[3]Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, NINE, 9 Little France Road, Edinburgh BioQuarter, Edinburgh EH16 4UX
[4]Electronic Data Research and Innovation Service (eDRIS) Public Health Scotland, NINE BioQuarter 9 Little France Road Edinburgh EH16 4UX
[5]University of Glasgow and Chief Statistician, Scottish Government, St Andrew's house, Regent Road, Edinburgh, EH1 3DG
[6]Patient Public Group Member, Bowel Cancer Intelligence (BCI) UK, University of Leeds, LIDA, Worsely Building, Leeds, LS2 9JT
[7]Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Big Data Institute, Nuffield Department of Population Health, University of Oxford, Old Road Campus OX3 7LF
[8]Edinburgh Cancer Research Centre and Edinburgh Health Economics, University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh EH4 2XR

Abstract

Introduction
Current understanding of cancer patients, their treatment pathways and outcomes relies mainly on information from clinical trials and prospective research studies representing a selected sub-set of the patient population. Whole-population analysis is necessary if we are to assess the true impact of new interventions or policy in a real-world setting. Accurate measurement of geographic variation in healthcare use and outcomes also relies on population-level data. Routine access to such data offers efficiency in research resource allocation and a basis for policy that addresses inequalities in care provision.

Objective
Acknowledging these benefits, the objective of this project was to create a population level dataset in Scotland of patients with a diagnosis of colorectal cancer (CRC).

Methods
This paper describes the process of creating a novel, national dataset in Scotland.

Results
In total, thirty two separate healthcare administrative datasets have been linked to provide a comprehensive resource to investigate the management pathways and outcomes for patients with CRC in Scotland, as well as the costs of providing CRC treatment. This is the first time that chemotherapy prescribing and national audit datasets have been linked with the Scottish Cancer Registry on a national scale.

Conclusions
We describe how the acquired dataset can be used as a research resource and reflect on the data access challenges relating to its creation. Lessons learned from this process and the policy implications for future studies using administrative cancer data are highlighted.

Keywords
administrative data; cancer; colon; rectum; Scotland; data linkage; healthcare; costs

*Corresponding Author:
Email Address: catherine.hanna@glasgow.ac.uk (Catherine R Hanna)

https://doi.org/10.23889/ijpds.v6i1.1654

Introduction

Administrative healthcare data – that is data collected as part of a person's routine healthcare - can provide information on screening, surveillance, co-morbidities, diagnosis, treatments and patient outcomes, as well as providing information on the real-world cost of healthcare. Linking datasets together provides more information than one dataset alone.

Cancer is a good example of where there is significant potential to generate public benefit with administrative data, specifically because of its high prevalence and disease burden. Countries such as Denmark, Sweden and the Netherlands, have a long history of using national registries for cancer research [1–3]. Similarly, in the United States, the government funded Surveillance, Epidemiology and End Results cancer database program [4] is linked to Medicare records and this data has been used widely to research patterns of cancer care, the cost of that care and patient outcomes. These examples scratch the surface of the enormous potential offered by going deeper into linked health records.

In the United Kingdom (UK), administrative healthcare datasets are generally held separately within the devolved nations. In Scotland, a cancer registry collects information on cancer specific variables such as date of diagnosis, type and stage of cancer and an indication of the treatments delivered. Although the registry offers a wealth of information, it does not include detailed information on the names or doses of systemic treatments delivered. It is therefore impossible to accurately identify variation in management approaches across the country, to understand the relative successes, or to estimate the costs associated with treatment. Scottish health service costs are publicised annually by sector, for example as hospital, community or family healthcare costs, but with no indication of how much is spent on cancer services specifically [5]. To build up a full picture of what the cancer trajectory looks like in Scotland, registry data must be linked to other administrative healthcare datasets.

There are efforts in Scotland to streamline and improve the use of cancer specific administrative datasets. For example, the Cancer Medicines Outcomes Project (CMOP) was commissioned by the Scottish government in response to the 2016 Beating Cancer: Ambition and Action report [6]. The overarching aim of this programme is to maximise the use of electronic records to understand outcomes for patients treated with cancer medicines in Scotland. One of the key objectives of CMOP is to test the scalability of linking cancer medicines datasets at a national level [7]. A separate major programme of work, the Scottish Cancer Registry and Intelligence Service (SCRIS) [8], was established in 2017 with the aim of creating a national Cancer Intelligence Platform with national reporting of cancer outcomes and treatments available to approved users via a dashboard. In late 2020, chemotherapy prescribing data (ChemoCare) covering 100% of the population was added to this platform. Due to data privacy concerns, this system is not intended to grant access to researchers to analyse individual patient level data, rather its primary function is for use by service providers in their delivery of cancer services in Scotland.

Recognising the gap that currently exists for researchers to access and analyse national chemotherapy and other healthcare administrative data, and considering colorectal cancer (CRC) specifically, we have focused on acquiring and linking administrative datasets in Scotland, including granular treatment records. Worldwide, colon and rectal cancer are the fourth and eighth most common cancers and causes of cancer deaths respectively, with over 1.8 million cases diagnosed each year [9]. It is estimated that at least €13 billion is spent on CRC care in the European Union alone each year [10]. The aim of this project was to create a linked dataset to allow mapping of the CRC landscape in Scotland and to identify differences in treatment pathways and the outcomes associated with different treatment approaches. This would enable practices with the best outcomes to be replicated and any variation in treatments to be addressed. In addition, the ambition was to use the dataset to provide a description of the real-world interaction of cancer patients with the Scottish healthcare system and to calculate the cost of healthcare resource use required for CRC diagnosis and management on a national scale.

This project is a part of a larger initiative, the Cancer Research UK (CRUK) funded COloRECTal Repository (CORECT-R), which is a UK wide programme aiming to quantify the characteristics of, and any variation in, CRC and its management [11]. One of the major work-streams for CORECT-R is to create a single colorectal research data system to enable analysis of data collected across all four UK nations. Due to the distinct data governance regulations, organisations and databases in devolved nations, the process of accessing and linking cancer datasets is being performed in parallel in each nation. This project represents the Scottish arm of that work-stream and has a particular focus on using the Scottish data to investigate the health economics of CRC care.

The aim of this paper is to document the process of creating this disease specific cancer dataset for research in Scotland. Specifically, we:

1. Outline the datasets available for this purpose in Scotland

2. Describe the current data linkage infrastructure in Scotland and provide an overview of the processes and timelines involved in successfully accessing and linking data

3. Describe the potential for use of these data in future for patient and public benefit and summarise the key challenges we encountered.

Methods

Navigating the administrative data linkage infrastructure in Scotland

In Scotland, since the 1970s, on first registration to the healthcare service every patient is assigned a unique Community Health Index (CHI) number. A CHI number consists of a patient's date of birth plus four additional digits. CHI numbers are beneficial in terms of data linkage because they serve as a unique identifier across a range of healthcare datasets. This is one reason why Scotland is particularly suited to linking its health records for research purposes.

Datasets relevant to this project

The datasets that are included in the CORECT-R Scotland specification are outlined in Table 1. Several datasets are held centrally and governed by Public Health Scotland (PHS, previously the Information Services Division (ISD) in National Services Scotland (NSS)), whereas others are held by external data controllers. Systemic anti-cancer therapy prescriptions and Cancer Quality Performance Indicator (QPI) data in particular, are held at local or regional levels, with no central repository. An arbitrary time-point (2006) for the earliest Cancer Registry records accessed was chosen to allow at least 10 years of follow up after the first patient cancer diagnosis. Datasets providing information on deaths or cancer diagnosis were also requested from 2006. Inpatient and outpatient resource use datasets were requested from 1997 to allow analysis of resource use prior to a cancer diagnosis and to allow calculation of the Charlson co-morbidity index using data from five years prior to diagnosis. Any dates of access beginning after 2006 were dictated by the years of data available for those specific datasets.

Additional information on co-morbidity, provided as the Charlson co-morbidity index [12], socio-economic deprivation, provided as Scottish Index of Multiple Deprivation (SIMD) and the Carstairs and Morris Index [13, 14], as well as rurality of patient's residence, is provided by PHS as derived variables included within a number of the above named datasets. Patient level information cost system (PLICS) variables are currently available for individual episodes of care within a number of SMR datasets but there are currently no PLICS variables linked with the cancer registry and little work has been undertaken to assign costs to cancer care specifically in Scotland.

Key stages in accessing and linking datasets

There were four main stages (Figure 1) in accessing and linking datasets on a national level. In what follows, we describe the processes undertaken for the datasets that have already been successfully linked for this project.

Stage 1

The first stage in accessing data was to define the study requirements in order to apply to the Public Benefit and Privacy Panel (PBPP) for Health and Social Care in Scotland [15]. PBPP have responsibility for weighing up the benefits to the public from granting access to healthcare data against the risk that the sharing of the data poses to an individual's privacy. All applications to PBPP go to a Tier 1 panel for proportionate review. Some applications will be referred on for further review by a Tier 2 Committee. There are lay representatives who assess submissions at both Tiers. The application for this project was reviewed at Tier 2. A separate application was required to access SICSAG data and this was submitted to the SICSAG Steering Committee.

Stage 2

The second stage of the project was acquisition of datasets for transfer into the National Safe Haven (NSH). The NSH is a research platform operated by Edinburgh Parallel Computing Centre on behalf of PHS. The NSH provides a secure analytical environment where data controllers can allow administrative data to be used for research purposes when it is not practical to obtain individual patient consent, whilst protecting patient privacy and identity. eDRIS were the principal department of PHS responsible for overseeing data transfer. During this second stage, clarification of the content of the datasets required a PBPP amendment, which also included a clearer explanation of the cohort generation and indexing process.

Datasets held by PHS (Table 1) did not require transfer because they are already held centrally within NSS. Figure 2 shows the data transfer process that occurred to transfer data externally from PHS to eDRIS, and internal data within PHS. A trusted third-party indexing team (CHI Indexing and Linkage Service (CHILIS)) facilitated transfer for the cohort generation and indexing of datasets. This third party and extra step was required to ensure the privacy of patient information. Specifically, this meant that no identifiable data was sent directly from data controllers external to PHS to the eDRIS team. Instead, patient identifiers were replaced with a unique patient identifier and the data was subsequently considered pseudonymised because the link between unique identifiers and CHI numbers was held by a trusted third party (CHILIS). In addition, under General Data Protection Regulation (GDPR), health data is considered sensitive category personal data and therefore cannot be considered fully anonymised. The cohort of patients included in the final dataset (Figure 2 "Master Cohort List") was defined using a combination of Cancer Registry and chemotherapy prescribing data.

Stage 3

Each dataset that was to be released to the research team for analysis was checked by eDRIS to confirm it matched the approved specification. This checking followed a series of steps and was repeated as required if any issues were identified. After the main analyst completed the extraction, a second analyst checked the coding used. The research co-ordinator then checked the data files and confirmed that each approved researcher had submitted the required documentation to grant access to the data (detailed below). Once the data files were ready for release and the researchers met their required obligations, the co-ordinator requested final checks and authorisation of release from a senior manager who had permission to fulfil that role.

Deterministic linkage of pseudonymised datasets was performed by the eDRIS team within their NSH using individual unique identifiers. Essentially each of the unique indexed identifiers supplied in Step 4 of Figure 1 was replaced with the master index in Step 7 so each patient had the same unique identifier across all datasets. Final checks on the linkage step was also done by eDRIS prior to release into the designated project area within the NSH.

Stage 4

After linkage was performed, the pseudonymised dataset was transferred to the researcher-facing NSH. Access to data within the NSH was limited to the project team named on the most recently approved PBPP application. Prior to accessing the data, each named person demonstrated up to

Table 1: Datasets included in the CORECT-R Scotland specification

Dataset	Data controller	Description	Years requested for the purposes of this project
NRS Deaths	NRS	This dataset is collected by National Records Scotland (NRS), which is a Scottish government institution. It contains information on date, cause and place of death for all deaths registered in Scotland since 1974. PHS is granted access to extracts from this dataset for research/linkage purposes.	2006–2018
Outpatients (SMR00)	PHS	This dataset contains patient level episode data on outpatient appointments across all specialities (except A & E and Genito-urinary medicine). Data collection began in Scotland in the 1990s. Data collection within 6 weeks of outpatient attendance.	1997–2018
Inpatients and day cases (SMR01)	PHS	SMR01 comprises patient level episode data on hospital inpatient and day case discharges from acute specialities in Scotland. Data is available in computerised format from 1968.	1997–2018
Mental health (SMR04)	PHS	SMR04 contains data for patients receiving care in Mental Health facilities (inpatient and day cases).	2006–2018
Cancer Registry (SMR06)	PHS	SMR06 is also known as the Scottish Cancer Registry and was established 1954. This dataset collects information relevant to the diagnosis and management of malignant neoplasms, as well as carcinoma in situ and some benign tumours. Data is collected annually. CORECT-R Scotland has requested information on patients with a diagnosis of CRC only. This dataset contains CRC staging information based on both TNM and Duke's staging classifications.	2006–2018
SACT (ChemoCare WoSCAN)	WoS Cancer Network	Regional chemotherapy prescribing dataset. ChemoCare is the electronic data system that captures prescription of systemic anticancer agents and supportive medications in hospitals in Scotland. This does not include all hospital prescriptions, and would not, for example, include medications written on a drug prescription chart for patients during an inpatient stay. It would provide information on chemotherapy and supportive medications received by a patient whether they received that treatment as an inpatient or an outpatient.	2006–2018 but reliable data from 2012 onwards
SACT (ChemoCare SCAN)	SCAN Cancer Network	Regional chemotherapy prescribing dataset (as for WoSCAN).	2012–2018
SACT (ChemoCare Grampian)	Grampian Cancer Network	Regional chemotherapy prescribing dataset (as for WoSCAN).	2006–2018 but reliable data from 2012 onwards
SACT (ChemoCare Tayside)	Tayside Cancer Network	Regional chemotherapy prescribing dataset (as for WoSCAN).	2006–2018 but reliable data from 2012 onwards
SACT (ChemoCare Highlands)	Highlands Cancer Network	Regional chemotherapy prescribing dataset (as for WoSCAN).	2006–2018 but reliable data from 2012 onwards
Cancer Audit (QPI WoS)	NHS Greater Glasgow and Clyde	National prospective audit dataset collected and stored regionally on an annual basis (April each year). NHS boards are required to report their activity against QPIs as part of a mandatory national cancer quality programme. Healthcare Improvement Scotland is responsible for the external quality assurance of cancer services against tumour specific QPIs. This dataset contains CRC staging information based on both TNM and Duke's staging classifications. It also includes surgical operation codes and anaesthetic data such as the ASA score.	2013–2018

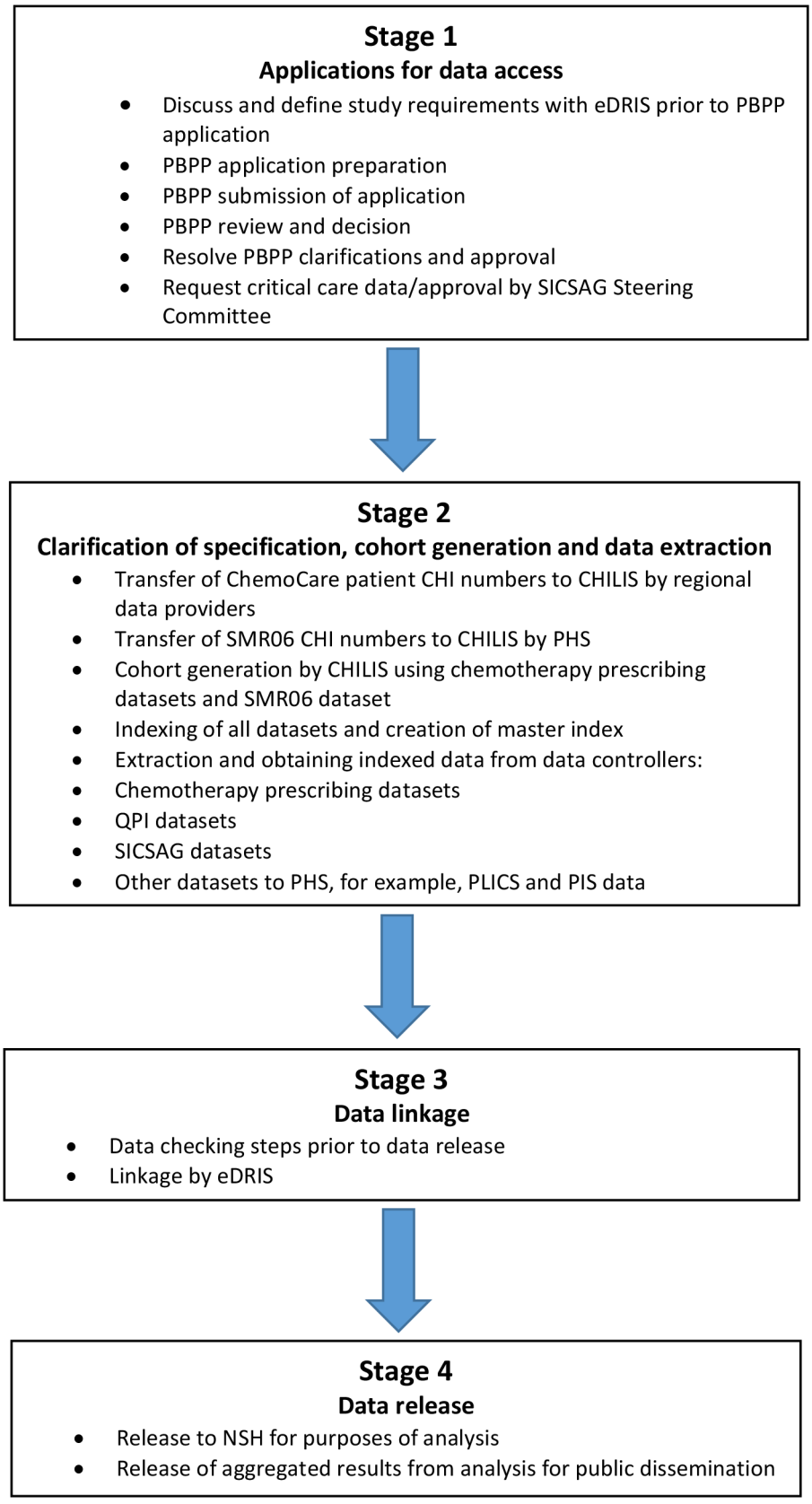
Continued

Table 1: Continued

Dataset	Data controller	Description	Years requested for the purposes of this project
Cancer Audit (QPI SCAN)	NHS Lothian	National prospective audit dataset (as for WoS).	2013–2018
Cancer Audit (QPI NoSCAN)	NHS Tayside, NHS Grampian and NHS Highland	National prospective audit dataset (as for WoS).	2013–2018
Costs (PLICS)	PHS	PLICS is the patient level information costing system and contains cost variables derived from SMR00, SMR01 and SMR04. This top down costing system was developed to allow hospital costs to be attributed to patient activity in a detailed way to reflect key cost drivers such as length of stay. PLICS apportions hospital site and speciality specific direct costs to individual patient records on admission, per day, for theatre time and specific high cost items. PLICS is not yet available for SMR06.	Financial year 2014/15–2017 (SMR01 2012 start)
SICSAG	Scottish Intensive Care Society Audit Group	SICSAG is the Scottish Intensive Care Society Audit Group dataset and contains both episode level and daily information provided for each patient.	2006–2018
Radiotherapy	NA	Detailed radiotherapy data is currently not available on a national basis in Scotland and instead is held loco-regionally by radiotherapy centres. Key information on radiation treatment delivered (for example if radiotherapy was delivered and date of treatment) is currently available within the Scottish Cancer Registry (SMR06). However, granular radiotherapy data (for example dose, technique and modality) is currently held by individual hospital institutions which deliver radiotherapy. A process is in development to make radiotherapy data available nationally – this involves Scottish radiotherapy centres sending data extracts to Public Health England, who curate the data to a common standard prior to returning to Public Health Scotland.	Not currently available
Prescribing Information System	PHS	The Prescribing Information System (PIS) is a data source for all prescribing of medicines (and their costs) that are prescribed and dispensed in the community in Scotland. This includes medications prescribed in hospital but dispensed in the community but not those dispensed in hospital. Information for this dataset is supplied by the Practitioner and Counter Fraud Services Division.	2010–2018
Accident and Emergency	PHS	This dataset was originally established in 2007 to monitor compliance of each NHS board with the maximum four hour waiting time target. Departments may submit individual episode level data (detailed information on each patient attendance) or aggregate level data (often smaller minor injury units). Sites that submit episode level data account for 94% of national A and E attendances.	2011–2018
GP Out of Hours	PHS	A Scottish government commissioned (2014) dataset to improve understanding of activity, demand and capacity at a national level for primary care out of hours services.	2014–2018
Scottish Ambulance Service (SAS)	PHS	The SAS dataset contains individual level records of all patient contact with the service.	2011–2018
NHS 24	PHS	The NHS 24 dataset contains individual level records of all patient contact with the service.	2011–2018

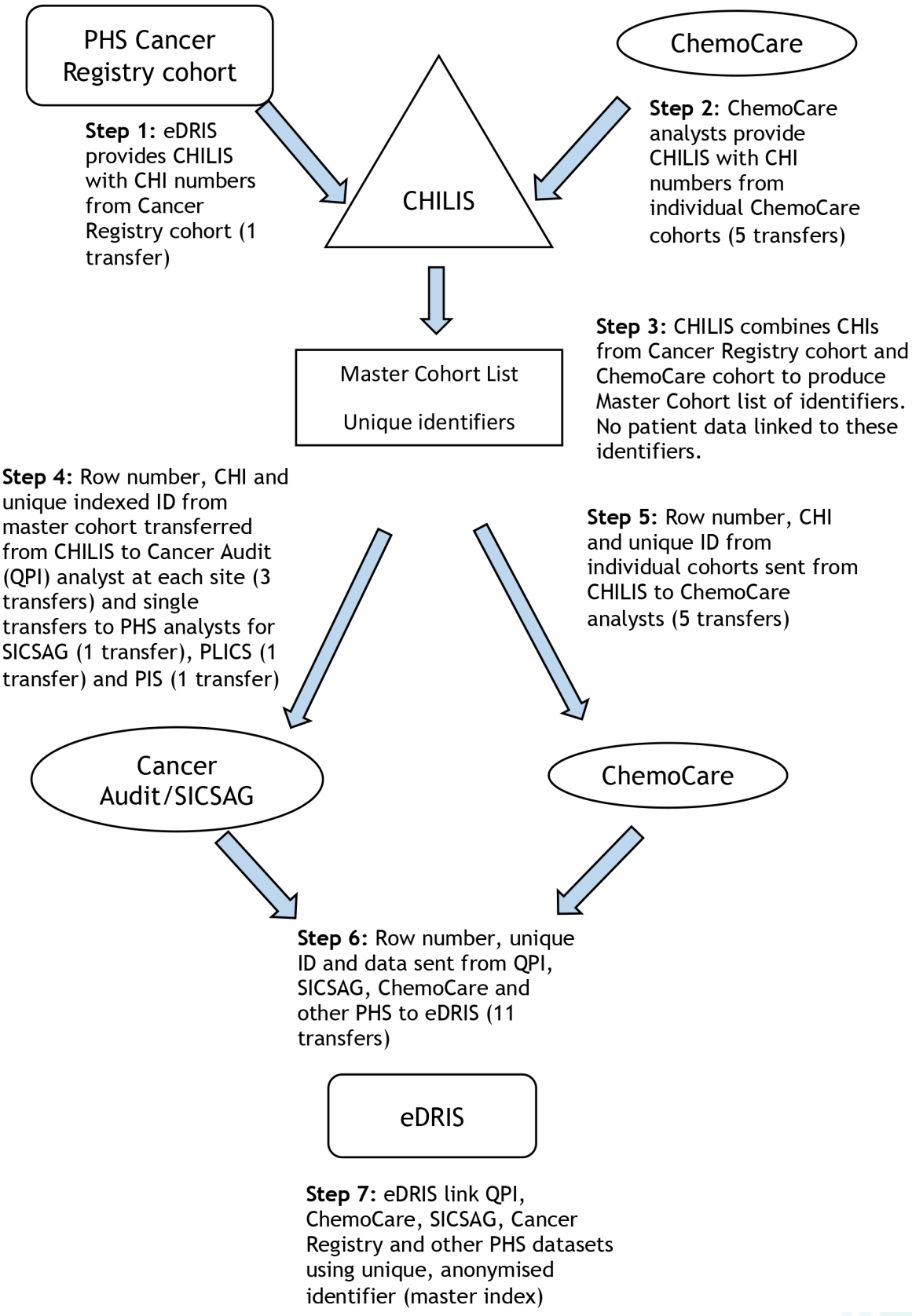
Abbreviations: NRS, National Records Scotland; SMR, Scottish Morbidity Records; A&E, Accident and Emergency; PHS, Public Health Scotland; QPI, Quality Performance Indicator; WoSCAN, West of Scotland Cancer Network; SCAN, South East Scotland Cancer Network; NoSCAN, North of Scotland Cancer Network; PLICS, Patient Level Costing System; NHS, National Health Service; GP, General Practitioner; SICSAG, Scottish Intensive Care Society Audit Group; PIS, Prescribing Information System; SACT, Systemic Anticancer Therapy; ASA, American Society of Anaesthesiologists; TNM, Tumour, Nodes, Metastases.

Figure 1: Flowchart outlining the main stages to access and link data



Abbreviations: NSH, National Safe Haven; eDRIS, electronic Data Research and Innovation Service; PHS, Public Health Scotland; QPI, Quality Performance Indicator; SICSAG, Scottish Intensive Care Society Audit Group; CHILIS, CHI Indexing and Linkage Service; SMR, Scottish Morbidity Record; PBPP, Public Benefit and Privacy Panel for Health and Social Care.

Figure 2: Cohort definition and transfer of datasets to Public Health Scotland



Abbreviations: PHS, Public Health Scotland; QPI, Quality Performance Indicator; CHILIS, CHI Indexing and Linkage Service; eDRIS, electronic Data Research and Innovation Service; CHI, Community Heath Index; SICSAG, Scottish Intensive Care Society Audit Group; PLICS, patient level information costing system.

date, approved information governance training and completed an eDRIS User Agreement, which stipulated the framework for data access and was signed by the named researcher's organisational lead. A requirement for working with this linked data in future will be that all outputs (for the purposes of internal discussion within the research team and/or for public dissemination of research findings) must undergo a disclosure-controlled release. For this to occur, it will require two eDRIS employees to check the outputs, one of whom has special authorisation to approve data for release from the NSH.

Results

Datasets accessed and linked

At the time of writing, 'release one' and 'release two' of the data are currently available within the NSH for use by named project researchers. Interim datasets for release one were available in April 2020, prior to confirmation of the linkage step, and fully available at the end of July 2020. Release two was available in December 2020. These data releases contain information on 47,157 patients diagnosed with CRC in Scotland and links thirty-two separate datasets and a demographics file. This information represents the majority of the administrative data approved for this project by PBPP (see Table 1 for full list) and access to the remaining datasets for unscheduled care data is pending. Figure 3 provides an overview of the data sources and the number of patients included in each of the datasets obtained thus far. The final master cohort contains information on all patients aged 18+ who had a CRC diagnosis between January 2006 and April 2018 in Scotland.

Overall timings

First contact with an eDRIS co-ordinator was made in January 2018 and the PBPP application was submitted in April 2018. PBPP approval was granted in October 2018 after Tier 2 review by a full panel of PBPP committee members. Therefore, Stage 1 took approximately 9 months in total. A substantial amendment was necessary and this was approved in February 2020.

Stage 2 (data acquisition) took approximately 1.5 years. Figure 4 provides a more detailed time-line for this part of the process. It involved initiating and continuing a dialogue and discussion with the relevant data controllers/data providers in the NHS Boards and other analytical teams in PHS. The ChemoCare and QPI datasets required the use of a secure transfer platform to transfer the data into PHS. To facilitate this, both ChemoCare and QPI data providers were given separate login details and passwords to access this platform when they transferred data to CHILIS versus eDRIS. In total, 28 successful secure transfers were performed (see Figure 2) to transfer data from external data providers to PHS.

Organisation of datasets and linkage by eDRIS took approximately ten months. Finally, stage four was transfer of the data to the NSH to be analysed by the researchers. In total, from first contact with eDRIS to access release one of the data by the researchers was around 2.6 years and from PBPP approval to data access was approximately 1.9 years. Access to release two of the data was made available five months after release one.

Estimated direct costs and resource use incurred during this process

An estimation of the costs and resource use required to achieve data access so far for the purposes of this project are outlined in Figure 5. Many of the individuals were involved over the majority of the 2.6-year period. From the perspective of the institution facilitating data access, the main study was considered a bespoke large project for costing purposes. A closely aligned PhD project, submitted for PBPP in parallel with the main project, was charged at a small project rate.

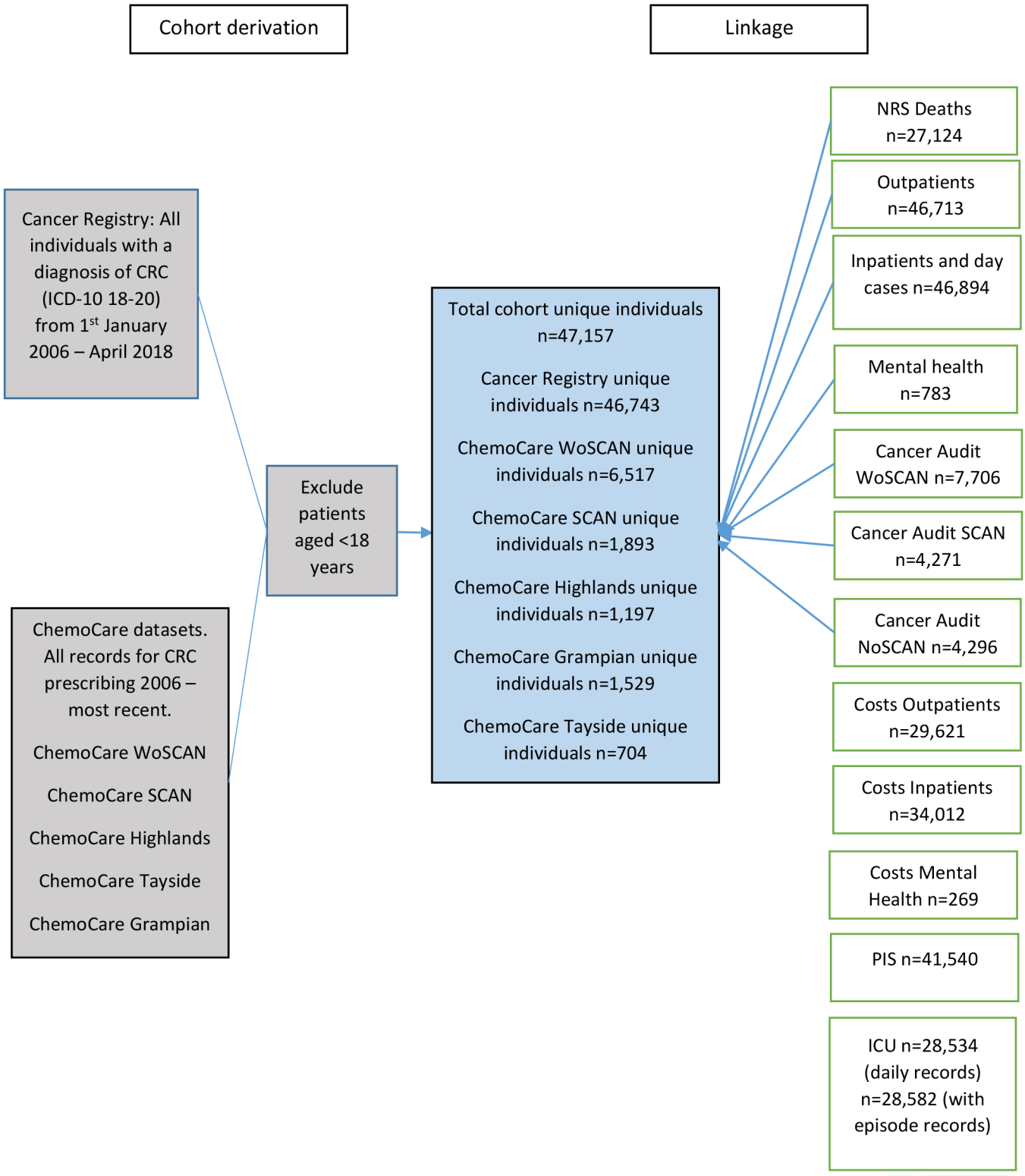
Discussion

Population-level administrative data sets offer significant potential to add to our understanding of how patients interact with health care services and provide a detailed picture of the care they receive. Data on a whole nation allows the characterisation of variation in disease management and clinical pathways, potentially identifying opportunities for more efficient resource allocation and improved care. Access to this type of data however, has been a significant challenge for researchers.

We have demonstrated, using CRC in Scotland as an exemplar, that the creation of a national linked administrative dataset, spanning more than ten years and collated from multiple data providers, is now possible, albeit with substantial efforts and extensive collaboration between researchers and the central team co-ordinating and performing data transfers and linkage. In particular, we have linked regionally held individual patient chemotherapy prescribing data (including co-prescribed medications) and quality performance data for cancer services to other nationally held health care datasets for all CRC patients in Scotland (2006–2018). This is a novel accomplishment. QPI data will be particularly beneficial in ensuring the robustness of any analysis given that it is prospectively collected according to a national proforma and all data entry undergoes validation. Accessing and linking chemotherapy prescribing data has been uniquely challenging because it is held in hospitals within digital prescribing systems that are not linked to the community prescribing platform that currently exists. Indeed, even in Scandinavian countries with a strong history of successful data linkage, accessing systemic anti-cancer treatment data on a national basis and linking it to the central cancer registry has been identified as a challenge [16, 17].

By describing the process of creating such a dataset in Scotland, we hope that we can facilitate data access for future researchers. We also want to allow comparison between our strategy and other important programmes of work (such as C-MOP and SCRIS in Scotland), to improve transparency and promote collaboration on a Scotland, UK or wider basis within the discipline of real-world data analysis for public and patient benefit.

Figure 3: Datasets included in release one and two with the number of patients contained in each dataset

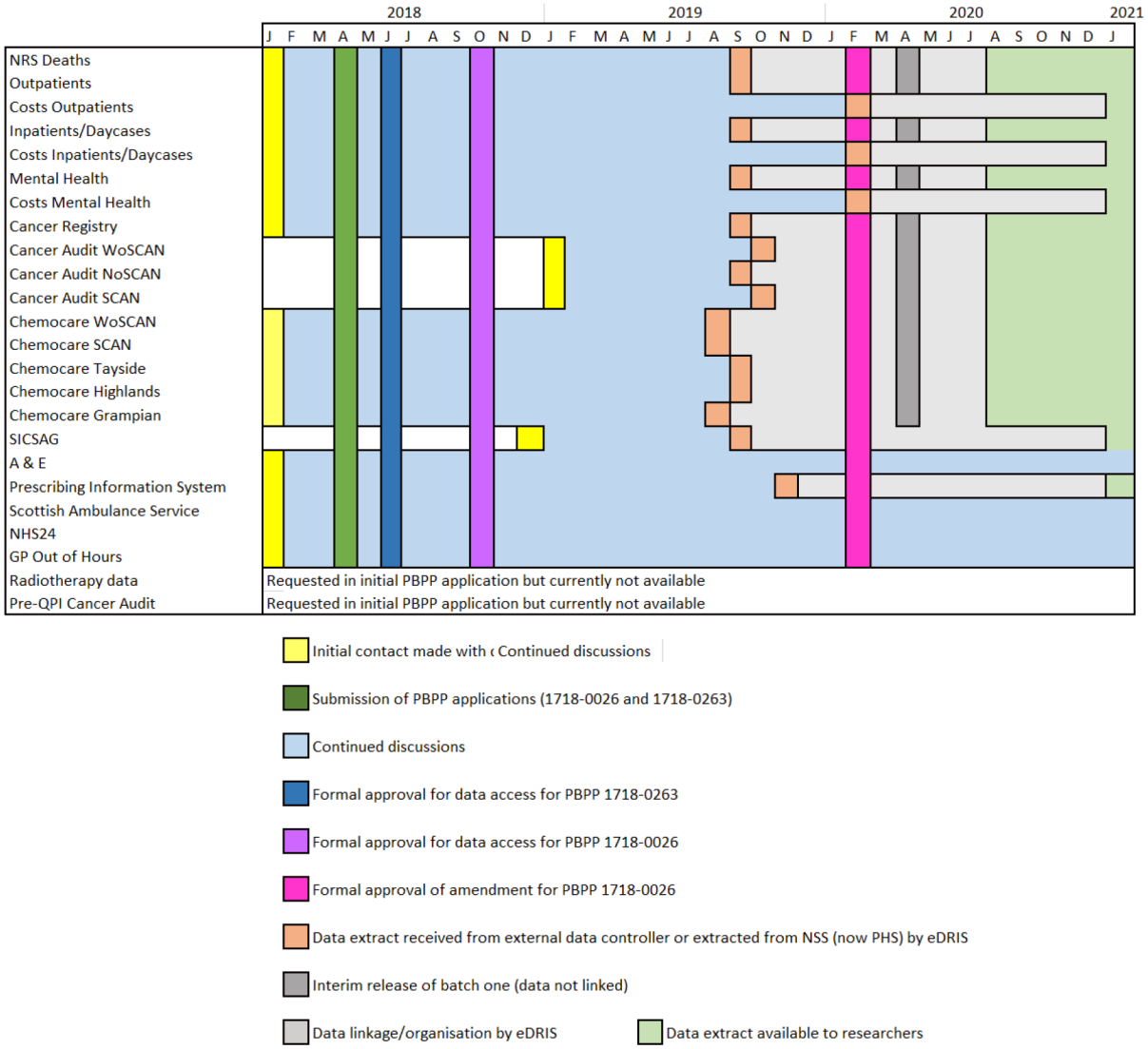


Abbreviations: ICD, International Classification of Disease; CRC, Colorectal cancer; WoSCAN, West of Scotland Cancer Network; SCAN, South East Scotland Cancer Network; NoSCAN, North of Scotland Cancer Network; SMR, Scottish Morbidity Record; QPI, Quality Performance Indicator; NRS, National Registry Scotland; SICSAG, Scottish Intensive Care Society Audit Group; PIS, Prescribing Information System; PLICS, patient level information costing system.

Future research opportunities using this dataset

For the first time, this dataset will allow a description of the real-world patient characteristics, treatment pathways, outcomes and cost of treating CRC in Scotland. Previously this has not been possible due to the lack of detailed prescribing information included in the Scottish Cancer Registry. Initial use of the data will be to describe the cohort of patients diagnosed with CRC in Scotland and the treatments they receive, with

Figure 4: Timeline for transfer of datasets to PHS. As of January 2020, 32 individual data files were available



There was also a demographics file which contained all patients in the master cohort, which was provided to the research team with release one of the data. PIS datasets were provided as nine separate data files, one for each year (2010–2019). ChemoCare Grampian and Highlands data were provided each as three separate files. ChemoCare Grampian provided an additional file with information regarding body surface area, height and weight. SICSAG information consisted of two files (episodes and daily information).
Abbreviations: WoSCAN, West of Scotland Cancer Network; SCAN, South East Scotland Cancer Network; NoSCAN, North of Scotland Cancer Network; QPI, Quality Performance Indicator; NRS, National Registry Scotland; PIS, Prescribing Information System; A&E, Accident and Emergency; GP, General Practice; SICSAG, Scottish Intensive Care Society Audit Group.

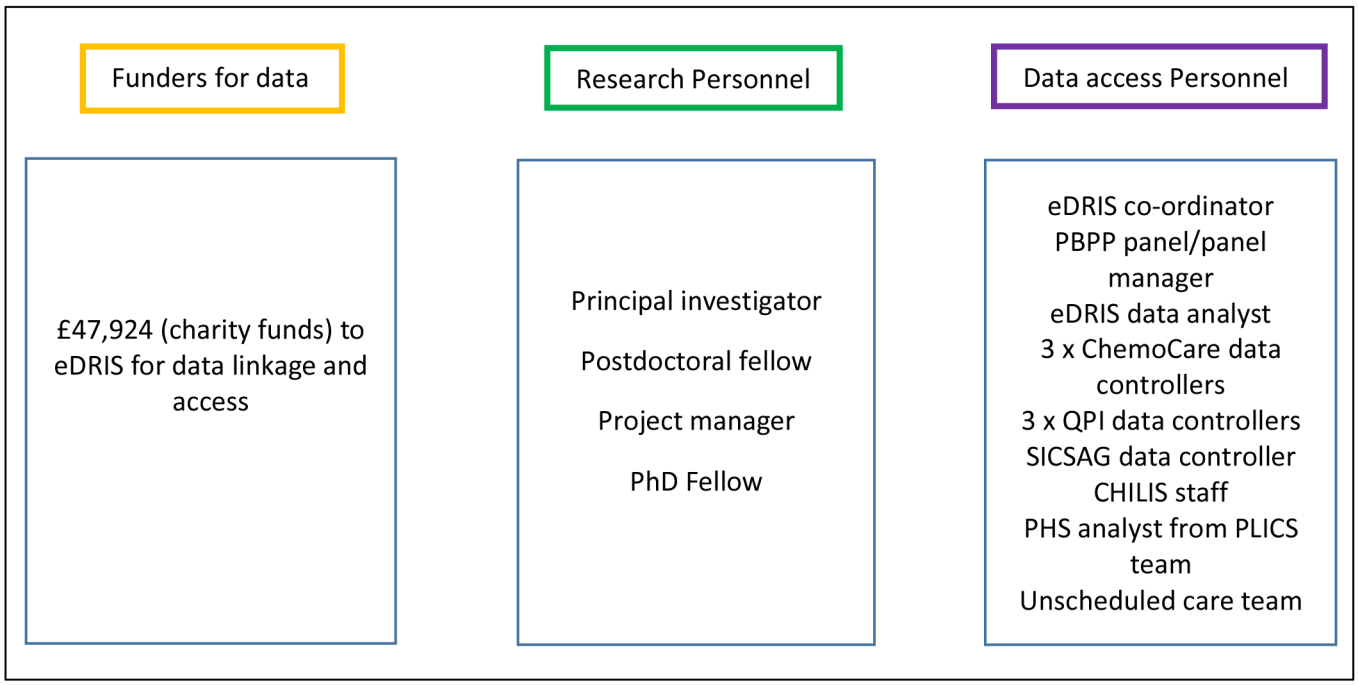
a specific focus on variation and the identification of outliers. This will be followed by a specific focus on post-operative treatment with the aim of understanding if management heterogeneity exists in Scotland, as previously demonstrated in England [16], and what this means for healthcare service costs. A subset of this dataset will be used to investigate the impact of clinical trial findings in the real world setting (PBPP application number 1718-0263) and the budget impact of implementing trial findings on a national scale.

A research repository as a future model for rapid data availability

The current model for access to administrative health care data in Scotland is one of 'link and destroy' for individual research projects [17]. In this way, multiple researchers might apply for the same or similar data to be linked, but this is all done in silos. Although it addresses many of the privacy concerns from data controllers, it is not an efficient use of data providers' resources or researchers' time. This type of model may pose a threat to scientific progress, research integrity and transparency, in the sense that it is difficult to reproduce results that have stemmed from a bespoke linked dataset which has subsequently been destroyed.

Recent developments within the Scottish Government recognise the potential benefits of preserving the linkage between datasets and storing data indefinitely for use by multiple research projects, whilst maintaining appropriate information governance protocols. This avoids duplication of data linkage and cleaning required for the initial

Figure 5: Direct costs and resource use



Abbreviations: eDRIS: electronic Data Research and Innovation Service, PhD: Doctor of Philosophy, QPI: Quality Performance Indicators, SICSAG: Scottish Intensive Care Society, CHILIS: CHI Indexing and Linkage Service, PHS: Public Health Scotland, PLICS: Patient Level Information Costing System, PBPP: Public Benefit and Privacy Panel for Health and Social Care.

linkage and preparation. Specifically, Research Data Scotland was launched in 2019 as part of the Programme for Government [21]. This is a not for profit organisation which aims to improve the economic, social and environmental wellbeing of Scottish residents by enabling access to linked data, not limited to healthcare datasets, for research in the public good. Further developments within this sphere in Scotland are also developing locally [18]. Research data repositories have the advantage of enhancing research integrity and transparency, since the reproducibility of research is maintained. Indeed, the UK Colorectal Cancer Intelligence Hub programme, of which this project is part, has developed the CORECT-R for English data, and other non-cancer examples exist [19].

We have seen during the Covid-19 pandemic how critical it is that researchers can access data in a timely manner to address urgent, real-world research questions. Looking to the future, preserving the linked CRC datasets we have established for this project should be considered essential to support research and inform clinical practice. At present, resources necessary to link datasets are potentially prohibitive for future researchers seeking to carry out research projects using individual patient level chemotherapy data on a national scale in Scotland. It would therefore be in both patient and public interest to preserve the dataset as a repository to which future researchers could apply. As a further step, linking this repository with the CORECT-R in England would increase the strength of any conclusions drawn from the Scottish data by allowing analysis of costs and outcomes relevant to a much larger group of patients and by making cross-country comparisons. Attempts to link these English and Scottish datasets are being pro-actively pursued.

Challenges

Alongside the successes and potential use of this information, we have experienced a number of unforeseen challenges that arose in the process of achieving access to this novel dataset. We believe that if similar issues are repeated in future, this may lead to important research with potential public benefit not being undertaken. Whilst this harm is difficult to quantify, others have documented concerns over the non-use of health data in research [20]. There is a significant patient voice in favour of making patient records available for research which is now overshadowing concerns over privacy that have headlined in the mainstream media [21]. Our intention is not to undermine any of the information governance processes that currently exist, but to address barriers to safe researcher access that we believe are disproportionate.

Data specification

No national data dictionary existed to describe the information that is held in each ChemoCare system. This made dialogue with ChemoCare analysts to plan the project specification more difficult, ultimately leading to the requirement for a substantial PBPP amendment to reflect the actual data received versus the variables originally requested and approved. We have partly addressed this problem by disseminating a list of the core variables held in all ChemoCare systems across Scotland [22] as part of a larger data dictionary describing all variables in release one of the data.

Table 2: Recommendations for creating a linked healthcare dataset

	Recommendation*
Data specification	Data dictionaries for datasets being linked are a requirement to record in advance, which data variables will be accessed and linked.
Capacity	Investment is required to ensure sufficient staff capacity at regional sites so that resource is not being diverted from service provision for the purpose of research and development without proper recognition of this effort. This is an urgent requirement specifically for ChemoCare sites in Scotland.
Data transfer	All parties involved (for example, data controller or analyst transferring data, the institution accepting the data and a third party) in a data transfer from a regional to central site need to prioritise communicating effectively within the same time window (often 2-3 days) regarding a data transfer if the data transfer is to be successful. A secure data transfer platform is required. It should be straightforward to use by central and regional data analysts, with ready access to information technology support if any technical issues arise (such as resetting passwords).
Data linkage	Easier data linkage would be possible if the data being linked was held by a central data controller. Data held by different regional data controllers makes data transfer and linkage more difficult. This was demonstrated in our work for ChemoCare and QPI datasets.
Data Access	A secure research environment to store and analyse data is required to meet data governance and privacy requirements. The Scottish National Safe Haven is one example of this type of research environment. Others exist and some are industry-led, for example, AIMES Management Services Ltd. Preserving linked data, such as the datasets described in this project, as a repository should be a priority. This will facilitate data access for future researchers and reduce wastage of resources.
Resources	Training of staff at regional sites is required to ensure they have the skills required for efficient extraction, analysis and transfer of large datasets. These staff also need access to proper information technology infrastructure that can deal with large datasets. This is particularly urgent for ChemoCare sites in Scotland. Staff capacity at central sites needs to be sufficiently robust so that there is no slowing of data transfer and linkage set-up due to external pressures such as annual leave/sickness/other projects. There should be continuity in the staff managing data transfer and linkage. A co-ordinating team whose role is to oversee and organise information governance approvals, data transfer and linkage helps to streamline the process.

*Regional datasets = the same information for different locations within the same country are held by individual data controllers at a regional level, for example ChemoCare datasets. Regional sites = the organisations holding regional datasets. Central datasets = datasets which are stored and maintained at a national level, for example SMR datasets in Scotland. Central sites = the organisations holding central datasets.

Capacity

The length of time required to obtain ChemoCare datasets in particular was also partly attributable to a lack of capacity for staff within regional cancer networks to engage with the process. The responsibility for physically downloading reports from the ChemoCare system was often performed by an individual whose major responsibility was service provision. In addition, one ChemoCare site had specific difficulties with the software required to store large datasets. In future, ChemoCare and/or QPI data should be stored centrally, as is the case currently for SMR datasets in Scotland and as occurs currently with the Systemic Anti-Cancer Therapy database in England [23]. This would significantly enhance the ease of any linkage of chemotherapy datasets in future. We suggest that investment should be made to help realise this in Scotland.

Data transfer

Each transfer of data from data providers external to PHS required careful communication between the sender and the recipient because data deposited in the secure transfer environments was automatically deleted if not picked up within 72 hours. Launch of a new secure file transfer system coincided with the data transfer and indexing process (stage 2) and at times, it was necessary to utilise a second, separate data transfer platform because of problems with the new system. In addition, the fact that the project's master cohort was defined by both SMR06 as well as five ChemoCare cohorts, meant that transfer of ChemoCare data required additional transfers for each site compared to if SMR06 alone was being used.

Data linkage

Once the datasets had all been transferred to eDRIS, data linkage took longer than anticipated because the first attempt at linkage experienced technical problems. The time between the first and second linkage was approximately four months. This unexpected difficulty was partly due to the number of datasets being linked and the impact of the Covid-19 pandemic when resources reprioritisation was required within PHS.

Data access

The dataset we describe in this paper represents most of the full dataset that was approved, although unscheduled care data is still not available. It had been anticipated these datasets

would be included in release two, and the delay is partly a result of competing resources due to the pandemic.

Resources

A substantial portion of the time-line stipulated by the funder for this project was dedicated to data access. The timeliness of data access has previously been documented as a barrier in several other UK projects [24, 25] and raises a broader issue around the ability of early career researchers to use nationally linked cancer datasets that include chemotherapy data in Scotland within the current landscape. We have outlined that the cost for data access correlates with the number of datasets external to PHS being linked, which also makes it infeasible for an early career researcher to use data that relies on bespoke project-specific linkage.

Regulatory changes and external forces

Several unforeseen regulatory changes and events occurred during the project that caused delays. For example, at the time of submission, the legal requirements to demonstrate that accessing data complied with the European Union GDPR changed and required alteration of the submission documents. Moreover, changes to the key institutions co-ordinating data transfer and access took place during the project. For example, the NSS Information Services Division changed to Public Health Scotland. Finally, the onset of the global Covid-19 pandemic affected working environments and led to an acute increase in workload for PHS from approximately March 2020 as they justifiably prioritised research classified as part of the health service response to the pandemic. Although these issues are specific to this project, it is strongly recommended that the occurrence of similarly unforeseen events are anticipated with any similar projects in future and flexibility in timelines is incorporated to account for this possibility.

Reflecting on the successes and lessons learned from this process, Table 2 outlines our recommendations for conducting this process for other tumour types in future. In particular, these suggestions are aimed at reducing many of the delays we encountered and streamlining this process.

Conclusion

The siloed nature of modern healthcare is clearly detrimental for sharing learnings and improving care. This is evidenced in numerous ways throughout the NHS and is not conducive to optimising patient care. We have outlined our experience over several years to access national level chemotherapy prescribing data for a single tumour type, linked with several other datasets in Scotland, and deposited in an anonymised format specifically for research use. We strongly believe that transparent communication around the efforts that are ongoing to improve access to administrative data for the purposes of research are essential if we are to learn from previous failures and successes. We hope that describing this process will benefit researchers, the institutions helping to provide and co-ordinate linkage of administrative cancer information, the data controllers and ultimately patients who will gain from discoveries made using this data. We feel that setting up this linked dataset has been a valuable investment given the huge potential public and patient benefit from using real world cancer data to improve patient outcomes and service delivery. We welcome any efforts by national policy makers to streamline this process going forward and we hope this project can serve as the basis for future work to build a better landscape for administrative cancer data linkage, storage and access in Scotland.

Acknowledgements

The authors would like to acknowledge the support of colleagues within Public Health Scotland in the eDRIS Team, CHI Indexing and Linkage Service Team, and PBPP, for their involvement in obtaining approvals, indexing, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven.

Conflicts of interest

The authors have no conflicts of interest to declare.

Ethics approval and consent to participate

This project was approved under the favourable ethics opinion of the East of Scotland NHS Research Ethics Service for the secondary analysis of PHS data within the NSH for UK based researchers and also had ethical approval via the broader CORECT-R initiative [26]. CORECT-R Ethical Approval Reference: South West Central Bristol Research Ethics Committee 18/SW/0134

PBPP project number (main application): 1718-0026

PBPP project number (PhD project): 1718-0263

Consent for publication

No identifiable individual person data included.

Funding

This project is funded by Cancer Research UK (CRUK) Ref: C23434/A23706 "Creating a UK Colorectal Cancer Intelligence Hub".

The PhD project is funded by the Beatson Cancer Charity.

CRH holds a Clinical Trials Fellowship Grant from CRUK and the University of Glasgow (Grant ID: C61974/A2429)

Authors' contributions

CRH was responsible for the concept of this manuscript, writing the manuscript, editing, proofreading and final preparation of the manuscript and approved the final version of the manuscript.

EL is the postdoctoral research fellow working on this project and was responsible for the concept of this manuscript,

writing the manuscript, editing and proofreading of the manuscript and approved the final version of the manuscript.

HE is the project manager for the CORECT-R Scottish work-stream and was responsible for writing, editing and proofreading of the manuscript and approved the final version of the manuscript.

RJJ is the PhD supervisor for CRH and was responsible for editing and proofreading of the manuscript and approved the final version of the manuscript.

JH is the eDRIS co-ordinator for the PBPP projects relevant to this study and was responsible for writing, editing and proofreading of the manuscript and approved the final version of the manuscript.

RH is the Chief Statistician for the Scottish Government and Chairs the Transition Board of Research Data Scotland. RH was responsible for editing and proofreading of the manuscript and approved the final version of the manuscript.

SC is the patient representative within the research group and was responsible for editing and proofreading of the manuscript and approved the final version of the manuscript.

EM is the principal investigator for CORECT-R and was responsible for writing, editing, proofreading of the manuscript and approved the final version of the manuscript.

PH is the principal investigator for the Scottish work-stream of CORECT-R and was responsible for the concept of this manuscript, writing the manuscript, editing and proofreading of the manuscript and approved the final version of the manuscript.

References

1. Lundgren C, Lindman H, Rolander B, Ekholm M. Good adherence to adjuvant endocrine therapy in early breast cancer – a population-based study based on the Swedish Prescribed Drug Register. Acta Oncologica. 2018;57(7):935–40. https://doi.org/10.1080/0284186X.2018.1442932

2. Gjerstorff ML. The Danish Cancer Registry. Scand J Public Health. 2011;39(7 Suppl):42–5. https://doi.org/10.1177/1403494810393562

3. Liu L, Neven A, Giusti F, Maraldo MV, Meijnders P, Aurer I, et al. Using both clinical research and population-based cancer registry in long-term research- a case study using EORTC trials and the Dutch national cancer registry (IKNL). Journal of Cancer Policy. 2020;24:100226. https://doi.org/https://doi.org/10.1016/j.jcpo.2020.100226

4. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Medical care. 2002;40(8 Suppl):Iv-3-18. https://doi.org/10.1097/01.mlr.0000020942.47004.03

5. UK CR. Where next for cancer services in Scotland? An evaluation of priorities to improve cancer outcomes.; 2017. URL: https://www.cancerresearchuk.org/sites/default/files/where_next_for_cancer_services_in_scotland_june_2017_-_full_report.pdf

6. Scottish Government. Beating Cancer: Ambition and Action (2016) update: achievements, new action and testing change. 2016. URL: https://www.gov.scot/publications/beating-cancer-ambition-action-2016-update-achievements-new-action-testing-change/

7. Cancer Medicines Outcome Programme Team. Cancer Medicines Outcomes Programme (CMOP) Phase 1 Report (October 2016 – March 2020). 2020 August 2020.

8. Public Health Scotland. Scottish Cancer Registry 2020 [Available from: https://www.isdscotland.org/Health-Topics/Cancer/Scottish-Cancer-Registry/

9. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394–424. https://doi.org/10.3322/caac.21492

10. Luengo-Fernandez R, Leal J, Gray A, Sullivan R. Economic burden of cancer across the European Union: a population-based cost analysis. The Lancet Oncology. 2013;14(12):1165–74. https://doi.org/10.1016/s1470-2045(13)70442-x

11. Oxford Uo. UK Colorectal Cancer Intelligence Hub CORECT-R 2020 [Available from: https://www.ndph.ox.ac.uk/corectr/corect-r

12. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases. 1987;40(5):373–83. https://doi.org/10.1016/0021-9681(87)90171-8

13. Scotland PH. GPD Support Deprivation The Carstairs and Morris Index 2020 [Available from: https://www.isdscotland.org/Products-and-Services/GPD-Support/Deprivation/Carstairs/

14. Carstairs V, Morris R. Deprivation: explaining differences in mortality between Scotland and England and Wales. British Medical Journal. 1989;299(6704):886. https://doi.org/10.1136/bmj.299.6704.886

15. PUBLIC BENEFIT AND PRIVACY PANEL FOR HEALTH AND SOCIAL CARE-HSC-PBPP 2021 [Available from: https://www.informationgovernance.scot.nhs.uk/pbpphsc/

16. Taylor JC, Swinson D, Seligmann JF, Birch RJ, Dewdney A, Brown V, et al. Addressing the variation in adjuvant chemotherapy treatment for colorectal cancer: Can a regional intervention promote national change? International journal of cancer. 2020. https://doi.org/10.1002/ijc.33261

17. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc. 2017;4(2). https://doi.org/10.1177/2053951717745678

18. Edinburgh Uo. DataLoch 2020 [Available from: https://www.ed.ac.uk/usher/data-driven-innovation/dataloch

19. Smith DA, Wang T, Freeman O, Crichton C, Salih H, Matthews PC, et al. National Institute for Health Research Health Informatics Collaborative: development of a pipeline to collate electronic clinical data for viral hepatitis research. BMJ Health & Care Informatics. 2020;27(3):e100145. https://doi.org/10.1136/bmjhci-2020-100145

20. Jones KH, Laurie G, Stevens L, Dobbs C, Ford DV, Lea N. The other side of the coin: Harm due to the non-use of health-related data. International journal of medical informatics. 2017;97:43–51. https://doi.org/10.1016/j.ijmedinf.2016.09.010

21. Use MY data 2020 [Available from: http://www.usemydata.org/

22. Lemmon E. CORECT-R Data Dictionary 2020 [Available from: https://blogs.ed.ac.uk/ectu_ehe/wp-content/uploads/sites/769/2020/12/Data_Dictionary.pdf

23. Public Health England. SACT Systemic Anti-cancer Therapy Chemotherapy Dataset 2011 [Available from: http://www.chemodataset.nhs.uk/home

24. Lugg-Widger FV, Angel L, Cannings-John R, Hood K, Hughes K, Moody G, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. International Journal of Population Data Science. 2018;3(3). https://doi.org/10.23889/ijpds.v3i3.432

25. Dattani N, Hardelid P, Davey J, Gilbert R. Accessing electronic administrative health data for research takes time. Archives of disease in childhood. 2013;98(5):391–2. https://doi.org/10.1136/archdischild-2013-303730

26. NHS Health Research Authority. Establishing a UK Colorectal Cancer Intelligence Hub 2018 [Available from: https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/establishing-a-uk-colorectal-cancer-intelligence-hub/

Abbreviations

A&E:	Accident and Emergency
CHI:	Community Health Index
CHILIS:	CHI Indexing and Linkage Service
CMOP:	Cancer Medicines Outcomes Project
CORECT-R:	Colorectal Repository
CRC:	Colorectal Cancer
CRUK:	Cancer Research UK
eDRIS:	Electronic Data Research and Innovation Service
GDPR:	General Data Protection Regulation
GP:	General Practitioner
ISD:	Information Services Division
NHS:	National Health Service
NoSCAN:	North of Scotland Cancer Network
NRS:	National Records Scotland
NSH:	National Safe Haven
NSS:	National Services Scotland
PBPP:	Public Benefit and Privacy Panel for Health and Social Care
PhD:	Doctor of Philosophy
PHS:	Public Health Scotland
PIS:	Prescribing Information System
PLICS:	Patient Level Costing System
QPI:	Quality Performance Indicator
SACT:	Systemic Anti-Cancer Therapy
SCAN:	South East Scotland Cancer Network
SCRIS:	Scottish Cancer Registry and Intelligence Service
SICSAG:	Scottish Intensive Care Society Audit Group
SIMD:	Scottish Index of Multiple Deprivation
SMR:	Scottish Morbidity Records
UK:	United Kingdom
USA:	United States
WoSCAN:	West of Scotland Cancer Network