1. Identifying Cancer Recurrence within Patient Care Pathways across Linked National Clinical Datasets

2. Summary of Research (abstract)

Information on cancer recurrence is not available in routine national clinical data, preventing important cancer research from being carried out in population-based studies.[1,2] Providing this outcome in routine data would allow researchers and analysts to estimate the risk of recurrence for different groups of patients, to provide much needed evidence on the best combinations of treatments, and to evaluate the care and outcomes of patients whose cancer has recurred.

With the increasing availability of linked national clinical datasets, a very detailed picture of patient care can be constructed, from diagnosis and treatment, through surveillance, to later investigations and treatments. Cancer recurrence will signal in the data as a burst of healthcare activity, including imaging, blood tests, outpatient appointments, treatments and possibly A&E attendances.

The research aims to develop and validate methods to identify, or *phenotype*, bowel cancer recurrence after curative treatment in linked national clinical datasets, and assessing how well the methods extend to breast and prostate cancer recurrence. The research will include six work packages (WPs), first linking and synthesising the information across datasets (WP1), next developing one clinical rule-based (WP2), one statistical modelling and two machine learning (ML; supervised and unsupervised) indicators (WP3) for bowel cancer recurrence, validating the indicators and recommending the optimal approach(WP4). The clinical use of the indicators for bowel cancer will then be demonstrated (WP5) before applying the optimal approach to prostate and breast cancer recurrence (WP6).

WP1: We will construct the diagnostic, care and outcomes pathways of cancer patients across national Cancer Registration data, Cancer Waiting Times data, administrative data (Hospital Episode Statistics admitted, outpatient and A&E data), imaging data (Data Imaging Dataset), chemotherapy (Systemic Anti-Cancer Therapy dataset), radiotherapy (National Radiotherapy Dataset) and mortality data (ONS mortality).[2] The synthesised data will be split into a development dataset and a validation dataset.

WP2: Clinical rule-based indicators will be developed to diagnose when recurrence of bowel cancer occurs, based on recorded information over time. We will use an iterative approach: starting with a forward searching step using pre-defined sets of codes; enhanced by a backward-searching step to identify additional codes missed by the forward searching step; and applying a final review of the additional codes by the clinical panel.[3,4] Treatment for recurrence is likely to be different for very frail or comorbid patients, and this will be taken into account when defining sets of codes.

WP3: Unsupervised ML methods such as the K-means algorithm and hierarchical clustering, will be used to identify clusters of patients with distinct patterns of types and timings of healthcare activity and each cluster will be classified as indicating bowel cancer recurrence or not.[39,40] Statistical regression models and supervised ML methods based on decision trees, such as random forest or boosted tree approaches, will be used to identify patterns of diagnoses and healthcare activity that occur in the care pathway that accurately predict that

the cancer has recurred, using any signal from the data across the entire care pathway.[5,6] Information such as patient comorbidity and frailty will be included in the modelling to take into account the effect of patient heterogeneity on the treatment of cancer recurrence.

WP4: A 5-step approach will be used to validate the bowel cancer indicators within the validation dataset: assessing agreement between indicators; clinical adjudication for a subset of patients; assessing the stability of the indicators over calendar time and across healthcare providers; assessing the clinical plausibility of estimated relationships between each of the cancer recurrence indicators and known correlates of recurrence, such as cancer stage, surgical margins, and long-term cancer survival; assessing the sensitivity of analyses comparing cancer recurrence and recurrence-free survival between groups to the choice of recurrence indicator. Recommendations will be made on the optimal approach. [7,18]

WP5: The value of the indicators will be demonstrated by comparing recurrence-free survival between treatments for bowel cancer.

WP6: The optimal indicator will be applied to breast and prostate cancer recurrence. Clinical experts will advise how to adapt the indicator to the specific cancer site. The indicator will be validated for these cancer sites.

3. Background and Rationale

Background
Curative treatment pathways for cancer are increasingly complex and rapidly evolving, with growing gaps in evidence on the most efficacious treatment combinations. As a result, there is wide variation between care providers in treatments offered for many cancers without understanding the long-term impact on outcomes or cost-effectiveness.[8-10] Cancer recurrence is an endpoint that is captured sooner than death and is strongly correlated with long term survival outcomes.[11] It can also be used to define populations of patients with recurrent disease so that the effectiveness of treatment options later in the patient pathway can be assessed.

Research using national clinical data is highly effective in identifying cancer patients and survival outcomes but not in capturing recurrence. Population-based studies are therefore unable to use cancer recurrence as an outcome in order to provide estimates of the risk of recurrence for different groups of patients, or to evaluate the care and outcomes of patients whose cancer has recurred. Since 2014, the National Cancer Registration and Analysis Service (NCRAS) has requested recurrence data from hospitals but it is very incomplete because the infrastructure is not in place to search patient notes for tests or treatments that suggest recurrence.[1]

With the increasing availability of linked national clinical datasets, we can construct a more detailed picture of patient care than ever before, from diagnosis and treatment, through surveillance, to later investigations and treatments. Cancer recurrence will signal in the data as a change in frequency and type of hospital visits, procedures and treatments.

Literature review
A literature review in PubMed and Science Direct using the search terms ("algorithm" OR "phenotyp*" OR "indicator" OR "identif*") AND ("cancer" OR "malignan*" OR "neoplasm" OR

"carcinoma") AND ("recur*" OR "relapse" OR "progress*") found a systematic review that was published in October 2020 which included 31 studies to identify cancer recurrence in routine data.[12] Although the systematic review was primarily of indicators for breast cancer recurrence, the authors also identified studies to develop indicators of recurrence for other cancers. Our search identified 4 studies published since the systematic review and a further 3 studies that were missed in the systematic review. In total therefore, 38 articles have been found on identifying cancer recurrence in routine data. Large datasets are needed to develop accurate indicators, particularly if they are to make full use of rich information, including combinations of healthcare activity and patterns of visits. The majority of the recurrence indicators developed so far (25 out of 38) were developed in datasets of fewer than 1000 patients. As highlighted by the authors of the systematic review, very few of the studies took patient comorbidity or frailty into account, which could reduce the accuracy of the methods because frail patients are less likely to be treated for their recurrence.

The vast majority of the rule-based approaches used simple code sets, with the majority relying on the presence of any of a set of diagnosis and procedure codes in isolation, and none making use of combinations of codes or exploiting changes in the patterns of hospital attendances. For example, a patient will have regular but infrequent outpatient appointments for several years and then, if their cancer returns, may have for example, an A&E attendance followed by imaging, recurring outpatient appointments and regular chemotherapy doses. Such changes in frequency and types of healthcare activity, compared to the background activity expected for routine surveillance, would allow a more sensitive measure of recurrence than simply identifying specific codes for metastases or procedures for recurrence.

Even using simple code sets, the majority of the studies so far have identified recurrence with reasonable accuracy, identifying at least 80% of recurrences with false positive rates of less than 10%. Linked national clinical datasets provide an increasingly rich source of information about patients over their entire cancer care pathway. For example with details of inpatient, outpatient and A&E attendances, surgical procedures, doses and regimens of chemotherapy, and fractionations of radiotherapy. To fully exploit the richness of the data, a more systematic approach is needed to develop rules about the timing and combinations of codes.[13] Such systematic phenotyping approaches are well developed in other clinical areas.[4,14]

There is also a need to investigate to what extent statistical modelling and machine learning (ML) methods can improve on rule-based algorithms.[15] ML is a field of statistics and computer science that aims to detect patterns in large, heterogeneous, and longitudinal data and which can be highly flexible in modelling complex relationships, including non-linear relationships and interactions between a large number of correlated variables.[16] Very large studies are needed to develop ML algorithms that avoid identifying spurious relationships in the data which would not be replicated in external datasets.[17,18] The largest study to detect cancer recurrence using ML methods was developed for breast cancer and included only 1900 patients with 400 recurrences. It did not make use of patterns of information on timings of hospital attendances, neither did it take into account heterogeneity in the fitness of patients, which could limit its generalisability.[19] Phenotyping in routine data using ML methods has been used with success for a wide range of conditions and outcomes, ranging from asthma to obesity.[13,20].

Pilot work

We carried out pilot work for this application on data for patients who had a curative major resection for non-metastatic bowel cancer. Using only administrative inpatient data we identified in the 9 months to 5 years after major resection: diagnosis codes for metastatic or secondary cancer of lymph nodes; and procedure codes for resection of metastatic cancer or surgery for colorectal cancer recurrence. Despite not making use of the full range of data sources which will be used in the research project, such as cancer registry, chemotherapy, radiotherapy, imaging or outpatient data, and not drawing on combinations of codes or patterns of hospital visits, the results showed the potential of our proposed research.

Specifically, there was evidence of an association between stage at diagnosis, rates of recurrence and long term outcomes. Recurrence was identified in 18% of patients. Of patients who died of cancer in the 3 to 5 years after surgery, 83% were identified as having a recurrence, compared to 10% in those alive 5 years after surgery. Recurrence was identified in 6%, 12% and 31% of patients with stage I, II and III cancer respectively. The median time to recurrence was estimated to be 15.9 months, compared to a median time to death from cancer of 34.3 months.

These provisional results suggest that it will be feasible to identify cancer recurrence in much richer linked national clinical data which incorporates rich data along the full diagnostic and treatment pathway, including timings of patient interactions with hospital services. Work is needed:

1. to develop a systematic approach making full use of the information across linked clinical datasets (imaging, chemotherapy regimens, radiotherapy fractionations, outpatient clinic attendances etc)
2. to apply sophisticated methods making full use of the timing and combinations of hospital visits, tests, diagnoses and treatments
3. to allow the indicators to be validated.

Anticipated impact of the research

The methods developed in the research will allow cancer recurrence to be routinely identified in cancer registries and national cancer audits. Knowledge of the date of relapse will open up a large number of research opportunities. It will enable research into patterns of initial care which may be associated with recurrence, allowing studies with shorter follow-up than those using mortality as their key outcome. Making cancer recurrence indicators routinely available will enable improved performance monitoring of healthcare providers, stimulating local quality improvement. It will also mean that routine health data can be used to ascertain recurrence as an outcome in pragmatic cancer clinical trials, thereby decreasing the burden of patient follow up, increasing efficiency and reducing costs.

Research to date has, in large part, been limited to assessing outcomes from the point of initial diagnosis onwards. Patterns of care and NHS resource utilisation at the time of any recurrence may significantly impact on patient outcome e.g. chance of cure after salvage surgery, or life-expectancy on palliative chemotherapy. The methods developed will also be used in the future to understand whether changes or variation in practices of care are translating into differences in outcomes. In addition, a proportion of patients may choose not to receive or may not be deemed fit for further surgical or oncological treatment but still require the input of primary care and palliative care services. These areas have not previously been a subject for intensive research due to uncertainties in accurately defining this patient population, but they reflect key areas of clinical enquiry.

3a. Evidence explaining why this research is needed now

The research aims to develop methods to identify if and when cancer has recurred following curative treatment for bowel cancer using linked national clinical datasets, and to assess how well the methods extend to breast and prostate cancer. All three cancers have heterogeneous care pathways which makes the methods applicable to other cancers. Together these cancers make up 40% of cancer diagnoses, amounting to over 120,000 new diagnoses per year in England alone.[21] These cancers tend to progress slowly or moderately slowly and there is need for earlier endpoints than survival.[23,24] The prognosis after curative treatment is relatively good for these cancers and the receipt of optimum care has potential to affect large numbers of patients.

Treatment pathways for cancer are increasingly complex and continue to evolve rapidly with the aim of improving patient outcomes such as overall survival and local tumour control. However, the rapid evolution in observed practices of care are not necessarily supported by robust evidence. As a result, there is wide variation between care providers in treatments offered for many cancers without understanding the long term impact on outcomes [8-10] or the costs of delivering care.

In addition to survival, the incidence of recurrence (either locally or distant metastasis) and the duration of recurrence-free survival are crucial endpoints for assessing the effectiveness of care.[11,25-27]. These endpoints are captured sooner than survival and as well as being important outcomes in themselves are also correlated with long term survival outcomes and can be used to address the gaps in our understanding of the effectiveness and value of evolving practices of care. Examples of where evidence is lacking on the best curative treatment options are numerous but include: external beam radiotherapy to both pelvic nodes and prostate compared to prostate alone; the benefit of adjuvant chemotherapy for rectal cancer; and sequential versus concurrent chemotherapy in operable HER2-positive breast cancers.

The methods developed in this research will enable future work to provide evidence on the best combinations of curative treatments for recurrence-free survival in three ways:

1. in observational studies of “real-world” populations of patients under the conditions of everyday clinical practice [28-30]
2. in observational studies for comparing treatments that are unlikely to be assessed in randomised clinical trials
3. by using routine healthcare data as an efficient, cost-effective way to provide longer-term outcomes for randomised clinical trials.[31]

Routine clinical data is currently being considered by bodies such as the MHRA as part of routine submission for cancer drug approvals to provide evidence on the outcomes of drugs in the real world.[32] Further, cancer recurrence is a key element in cost-effectiveness models of cancer treatments, and providing this information will improve the allocation of healthcare resources.[33,34]

More accurate national and regional information on patients whose cancer recurs will guide resource planning and provide improved service evaluation and feedback for care providers. National clinical audits provide hospitals with a suite of process and outcome indicators from across the patient pathway, but long-term outcomes are limited. Providing cancer recurrence rates to hospitals would be a step change for cancer audits, strengthening their ability to

stimulate local quality improvement and broadening the indicators used for quality assurance.[35,36]

Men and women affected by cancer in the patient groups advising national clinical audits have highlighted concerns with the adequacy of treatment after recurrence. Whilst primary curative treatments are largely standardised according to clinical guidelines provided by professional bodies and national organisations (e.g. NICE), variation is increasingly seen in relapsed disease, which could have significant implications for both survival and quality of life. For example, a patient who relapses after bowel cancer with two or three metastatic deposits may receive surgery, radiotherapy or systemic therapy. Understanding variation in treatments and both long term outcomes and cost implications (through resource usage) can help to identify gaps in access to care, support standardisation and improve quality of treatment delivery. This would only be feasible through an accurate estimation of the time point of relapse.

In addition, follow up of patients with recurrence can help to identify how patterns of care vary across and within cancer alliances. This is increasingly important when considering how services should be centralised to ensure patients are able to equitably access the relevant expertise and to inform referral pathways within cancer alliances.

Methods and algorithms will be freely disseminated so that cancer recurrence can be routinely identified in cancer registries and national cancer audits for epidemiological, clinical audit or policy purposes. In addition to usual academic outputs, we will engage with our patient partners, professional clinical bodies, cancer charities and NCRAS to ensure wide uptake of methods and results. And the methods developed will feed directly into the three national clinical cancer audits that the research team deliver.

4. Aims and objectives

The research aims to develop and validate methods to phenotype cancer recurrence after curative treatment for bowel cancer, in linked national clinical datasets. The project will:

1. Construct care and outcomes pathways of cancer patients across datasets, from diagnosis and treatment to subsequent investigations and treatments for recurrence.
2. Develop four indicators of the presence and timing of cancer recurrence for bowel cancer, one using clinical rule-based methods, one using statistical modelling and two using ML methods.
3. Validate the four indicators, including using clinical adjudication for a subset of patients and recommend the optimal indicator.
4. Demonstrate the clinical use of the indicators for bowel cancer.
5. Assess how well the optimal indicator extends to breast and prostate cancer.

5. Research Plan / Methods

The research is achievable because:

1. the data can be accessed immediately
2. stakeholder and PPI connections are built into the project (in its design, through members of the Study Steering Committee and by having a PPI co-applicant)

3. the team has health services research / data science / statistical and ML expertise (LSHTM / RCS) and clinical expertise (three clinical collaborators)
4. our pilot work provides evidence that the methods are feasible.

Research Team

The team combines methodological, clinical and PPI partners to ensure robust analyses with clinical validity aligned with patient priorities. The research will be a close collaboration between statisticians, health services researchers, ML experts, surgical and oncology clinicians from across the three cancer sites, and a PPI co-applicant with a wealth of experience as both a PPI representative and a co-applicant in bowel disease research. Team members are highly experienced in using data from the full patient pathway, through running national clinical audits related to the three cancers.

A PPI focused Study Steering Committee (SSC) will meet twice per year to guide the design and delivery of the project, representing key NHS, data provider and clinician stakeholders and including a PPI and a charity representative for each cancer site. The committee will be key in overseeing the planning and delivery of outputs of the project. We have three confirmed PPI representatives in addition to the PPI co-applicant, representatives from Bowel Cancer UK, Breast Cancer Now, Prostate Cancer UK, NHS England and NCRAS. Clinicians from across the patient pathway have confirmed their membership. The research team includes a breast cancer surgeon, a medical oncologist for bowel cancer and a clinical oncologist for prostate cancer. The SSC will also include a urologist, a bowel cancer surgeon, a liver surgeon, a lung surgeon and a clinical nurse specialist.

Study Design and Setting

The research is a cohort study using national routinely collected healthcare data provided by NCRAS. Pseudonymised data will be stored on the secure data environment of the London School of Hygiene and Tropical Medicine. The indicators developed will be for recurrence of disease requiring secondary care. For patients who have curative treatment of a primary cancer the distinction between cancer recurrence and progression is less important than the shift from routine periodic surveillance for a cancer which is understood to be cured to an intense period of investigations, hospital visits and treatments for a cancer that has returned. The vast majority of cancer recurrence following curative treatment will involve, as a minimum, outpatient attendance, and will therefore be detectable in the data. As a sensitivity analysis, the use of data on community-dispensed prescriptions will be explored to examine whether this identifies further patients with recurrence not entering secondary care.

Methods

The development work to phenotype cancer recurrence will start with bowel cancer in the project's first year, because patients having curative treatment have a moderate rate of recurrence and time from recurrence to death, and because there are well-defined care pathways for curative treatment, surveillance and recurrence of bowel cancer (Table 1).[22,23] Once the indicators of bowel cancer recurrence have been developed and validated, they will be used to compare the efficacy of different bowel cancer treatment pathways for which there are currently gaps in knowledge. Guided by the results from the work on bowel cancer, we will extend the research to the two other cancer sites in years two and three. The research will include six work packages.

Table 1: Typical recurrence and treatment pathways of the three cancers to be included

Cancer	Rate of recurrence after curative treatment	Time from recurrence to cancer death	Curative treatment modalities	Surveillance after curative treatment	Recurrence treatment
Bowel	Moderate	Moderate	Surgery ± neoadjuvant CRT ± adjuvant SACT	2 CT scans & 2 blood tests per year for 3 years, scope at 1 year.	Surgery, ablation, RT and/or SACT, supportive care.
Breast	Low	Long	Surgery +/- SACT +/- RT +/-HT	Annual mammography	Surgery, RT, SACT, HT, targeted therapy
Prostate	Moderate	Long	HT alone, HT + RT, Brachy, surgery, HT+RT+brachy	PSA monitoring 3-4 monthly year 1 then 6 monthly	RT, HT +/- SACT, stereotactic body RT, cryotherapy, brachy, HIFU, watchful wait

CRT = Chemoradiotherapy *SACT = systemic anticancer therapy* *RT = radiotherapy*
HT = hormone therapy *Brachy = brachytherapy*

WP1. Construct the care and outcome pathways of patients across national datasets [Pre-start up to month 3]

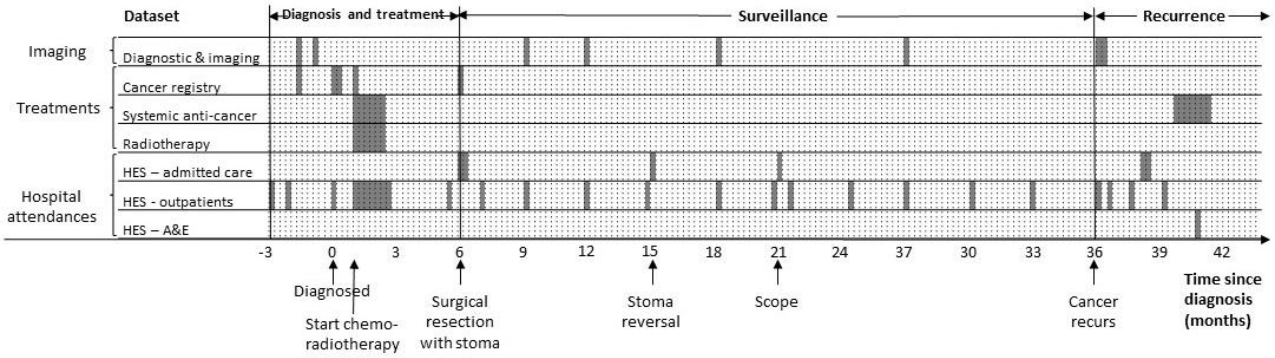
Approvals and one data access request across all three cancer sites to NCRAS for the datasets in Table 2, linked at patient level, will start as soon as the application is successful. We have discussed with NCRAS in detail the data requirements for this project and they have confirmed that the timescales are realistic.

Table 2: Datasets to be obtained from the National Cancer Registration and Analysis Service (NCRAS)

Cancer Registration (CR) Data	Demographics, diagnosis, tumour characteristics, care and treatments received, recurrence (very incomplete)
Hospital Episode Statistics (HES) - Admitted Patient Care (HES-APC) - Outpatient Care (HES-OP) - A&E (HES-A&E)	Administrative hospital database Includes: procedures, diagnoses, date of admissions & discharge clinic specialties, appointment dates dates and times, reasons for attendance
Systemic Anti-Cancer Therapy Dataset (SACT)	Dates, drugs, doses, planned treatment, height, weight
National Radiotherapy Dataset (RTDS)	Dates, treatment types, doses
Cancer Waiting Times (CWT)	Dates for cancer waiting times standards, reason for referral - including recurrence (very incomplete)
Diagnostic and Imaging Dataset (DID)	Type of test, body site, test waiting times
Community prescriptions data	Dates, drug details, quantity and dose
ONS mortality	Dates, causes and places of death

Patients having curative treatment undergo an initial period of intense treatment followed by infrequent but regular surveillance. The care and outcome pathway schematic in Figure 1 shows the healthcare activity at 3 phases (diagnosis and treatment, surveillance, recurrence) for a typical rectal cancer patient undergoing curative chemo-radiotherapy and surgical resection followed by a period of predictable healthcare activity. This patient's cancer returns at 3 years and we see an intense period of healthcare activity with tests, outpatient visits, surgery for their metastases and new regimens of chemotherapy. If the patient dies from their recurrence, mortality data will provide a date and cause of death for the patient.

Figure 1: Care and outcome pathway for a typical rectal cancer patient undergoing curative treatment whose cancer recurs at 3 years



Some information, such as metastatic cancer, type and date of surgery, radiotherapy and chemotherapy, is collected in multiple datasets. This information may be conflicting, missing, or defined differently. Validity checks will be carried out within and between datasets, and the association between known correlates used to rank the reliability of data items across data sources. A hierarchy of data sources for each data item will be used to resolve conflicts.

Errors in the data linkage can potentially affect the representativeness of the cohort. We will build on experience obtained from our current NIHR-funded research on methods for linking multiple clinical datasets to assess the linkage quality between datasets, over time, by hospital, and by patient characteristics.[37] The results of this ongoing research have demonstrated that overall linkage quality is high. However, where necessary we will restrict the cohort to time-periods or hospitals with high-quality linkage to reduce the potential for linkage bias.

The indicators of recurrence for each cancer site will be developed in a ***development dataset*** containing a random subset of trusts covering 60% of patients and validated in a ***validation dataset*** containing the remaining trusts covering 40% of patients. Included in the analyses will be patients treated January 2014 to March 2015 and followed to March 2020 (or later depending on the most recent data available), to ensure a minimum of 5 years' follow-up for all patients. If necessary, we will avoid the period of disrupted cancer services during the COVID-19 pandemic. Table 3 gives approximate sample sizes for the development and validation datasets for each cancer site.

Table 3: Approximate sample sizes for the development and validation datasets.

Cancer site	Inclusion	Size development data	Size validation data
Bowel	Elective major resection, no metastases	19,000	13,000
Breast	Surgery, no metastases	46,000	30,000
Prostate	Radical prostatectomy/ radical radiotherapy, no metastases	41,000	16,000

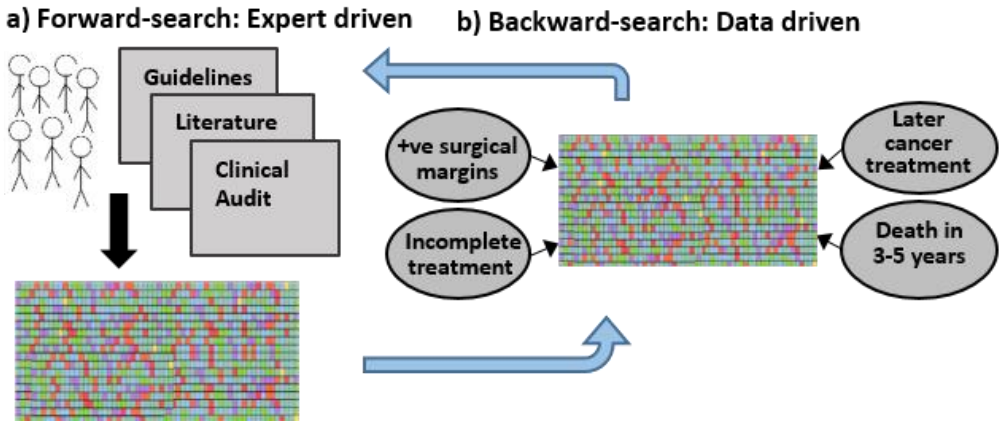
WP2.Develop an indicator to phenotype the presence and timing of recurrence of bowel cancer using clinical rule-based methods [Months 2 to 10]

Curative treatment for bowel cancer consists of local excision or major resection of the tumour, with adjuvant chemotherapy for a sub-set of patients, and preoperative radiotherapy

for most rectal cancer patients. NICE Guidelines recommend a minimum of a surveillance colonoscopy at one year, two CT scans in the first 3 years, and serum carcinoembryonic antigen tests at least every 6 months for the first 3 years.[22] Recurrence of bowel cancer will trigger a burst of healthcare activity, including imaging, blood tests, outpatient appointments, and possibly A&E attendances. Depending on the type and extent of the recurrence, treatment options include liver, thoracic or peritoneal surgery, recurrent colorectal cancer surgery, radiofrequency ablation, stenting (including endoscopic stenting), chemotherapy, radiotherapy, and hyperthermic intraperitoneal chemotherapy.

Within the ***development dataset*** we will use an iterative approach: starting with a forward searching step using pre-defined sets of codes developed by an expert clinical panel; enhanced by a backward-searching step to identify additional codes; and applying a final review of these additional codes by the clinical panel (Figure 2). The research team has successfully used this iterative approach for phenotyping complications and skeletal-related events of prostate cancer treatment.[3,4]

Figure 2: Clinical rule-based methods for indicator development



a. **THE FORWARD SEARCHING STEP** is expert driven. A clinical panel will use guidelines, audit, research and clinical experience to generate a resource detailing the possible combinations of ***timings*** and ***types*** of diagnostic codes and healthcare activity for each of the three phases in the care pathway: diagnosis and treatment; surveillance; and cancer recurrence. Within each phase in the pathway, the types of healthcare activity will be classified into 8 domains:

Domain		**Types of diagnosis / treatment / activity**
1.	Specific diagnosis codes	Diagnosis types, cancer site
2.	Imaging	Test modality, cancer site
3.	SACT	Regimens, curative intent, cancer site
4.	Radiotherapy	Dose, fractionation, curative intent, cancer site
5.	Surgical / endoscopic / radiological therapy	Therapy type, surgical / radiological approach
6.	Hospital admissions	Duration, mode of admission, specialty, diagnoses
7.	Outpatient attendances	Specialty
8.	A&E attendances	Investigations, treatments, diagnoses

The resource will be translated into codes and timings and used to distinguish cancer recurrence from the other two phases. When defining the forward searching algorithm the

clinical panel will take into account the heterogeneity of patients, such as differences in patient frailty and comorbidities, recognising that the aggressiveness of treatment will vary between groups of patients. They will also take into account geographical variation in care pathways, such as different approaches to surveillance and treatment combinations.

b. THE BACKWARD SEARCHING STEP is data driven. It identifies additional common coding patterns across the 8 domains in patients who are highly likely to have had a recurrence, for example, curative patients going on to have a cancer-related treatment, patients whose resected lymph nodes contain malignant cells, patients with positive surgical margins, and patients who die of their cancer in the 3+ years after initial treatment. The backward searching step picks up unpredictable idiosyncrasies of coding practices not included in the forward searching step. The additional coding patterns (reflecting combinations of **timings** and **types** of diagnoses and healthcare activity) identified in the backward searching step will be reviewed by the clinical panel. If they are considered to strengthen the discrimination between patients very likely and unlikely to have a recurrence, they will be included in the definition of the recurrence indicator, again taking into account patient and treatment heterogeneity (Figure 2).

As a sensitivity analysis, the use of data on community-dispensed prescriptions will be explored to examine whether this identifies further patients with recurrence not entering secondary care. For the majority of patients having curative treatment for primary cancer, any recurrence will be treated in secondary care, with outpatient attendances and imaging as a minimum. First-line hormone therapy is a treatment option for breast and prostate cancer patients with recurrence, and this may only be picked up through prescribing data.

Finally, the date of the earliest code included in the coding pattern defining the recurrence indicator will define the date of recurrence.

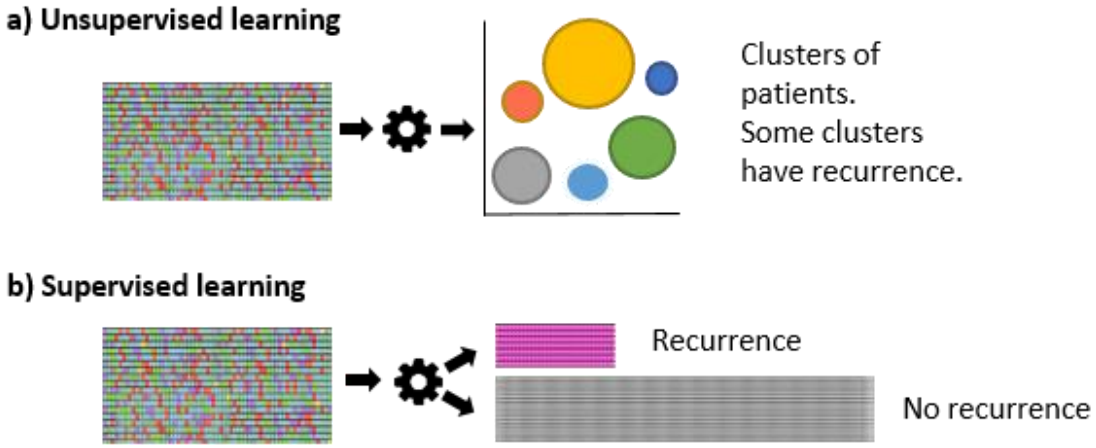
WP3. Develop indicator to phenotype the presence and timing of recurrence of bowel cancer using two machine learning methods [Months 6 to 14]

In the previous work package, a challenge is that the care received by patients, and how it is recorded in routine data, is likely to vary. Some patients will fulfil some criteria but not others such that it may be unclear how to categorise them – clinical knowledge will be used to develop rules for these ambiguous combinations of diagnoses/ healthcare activity across heterogeneous groups of patients in the rule-based algorithms.

An alternative approach which will be used in this work package is to “learn” definitions of cancer recurrence from the data based on statistical regression models and ML. ML is a field of statistics and computer science that aims to detect patterns in large, heterogeneous, and longitudinal data.[38] The algorithms used can be highly flexible in modelling complex relationships, including non-linear relationships and interactions between a large number of correlated variables.[16]

ML approaches can be classified into unsupervised and supervised learning.[13] We will use both, as each has potential advantages for phenotyping cancer recurrence (Figure 3). For both approaches, information such as patient comorbidity and frailty will be included in the modelling to take the effect of patient heterogeneity on the treatment of cancer recurrence into account.

Figure 3: Machine learning approaches to indicator development



UNSUPERVISED LEARNING will identify clusters of patients with distinct patterns of types and timings of healthcare activity across the 8 domains in WP2.[39,40] We will define a set of variables corresponding to each of these domains and their types and timings of diagnoses, treatments and healthcare activities, such as a chemotherapy regimen and dose in the 18th month after curative treatment, an unplanned hospital admission for colon cancer in the 21st month etc. An algorithm can then be applied to these variables to identify the patient clusters in the dataset. One cluster may be patients with no record of further treatment beyond their primary curative treatment. Another could be patients who develop metastatic disease and have palliative chemotherapy but no further surgery.

Methods such as the K-means algorithm and hierarchical clustering will categorise patients into distinct, non-overlapping clusters in the ***development dataset***. Guided by clinical knowledge, each cluster will be classified as indicating recurrence or not, and into sub-types of recurrence. Within recurrence clusters the earliest code linked to a significant diagnosis, event or treatment for recurrence will define the date of recurrence.

In **SUPERVISED LEARNING**, a specific outcome variable, which is considered to be the true label, is predicted from other variables. The aim is to identify patterns of diagnoses and healthcare activity that occur in the care pathway that accurately predict that the cancer has recurred, using any signal from the data across the entire care pathway.[5,6]. The same set of variables described above will be used, corresponding to the types and timings of each of the 8 domains of healthcare activity. Although cancer recurrence data are unavailable (which is the motivation for this study), we will use a proxy measure for recurrence. We can assume that in patients who undergo curative treatment and who then die from cancer 3 or more years later, their cancer has recurred in the interim. Cancer-specific death after 3 years will be a “delayed” proxy for cancer recurrence.

Prognostic models, such as those estimated via multivariable logistic regression, are a conventional example of supervised learning. These statistical regression models will be the first models developed for the supervised learning approach.[7] ML methods can model complex relationships in very rich datasets more flexibly, modelling many features across multiple domains of healthcare and over time. Conventional statistical models and ML methods based on decision trees, such as random forest or boosted tree approaches, amongst others will be used to predict each patient’s probability of recurrence in the ***development dataset***.[41,42] A threshold will be chosen above which recurrence is considered to have occurred. For patients with recurrence identified, the earliest code linked

to a significant diagnosis, event or treatment for recurrence will define the date of recurrence.

Potential disadvantages of ML methods are that they: can lack efficiency unless datasets are very large; can be less transparent; may fit spurious interactions; may be less precise about the date of recurrence; and may be less accessible to analysts without ML expertise. [43] However, they have potential to phenotype cancer recurrence (and its sub-types) more accurately than clinical rule-based methods because they can model complex relationships between many more variables.[5,44,45]

WP4. Validation of the indicators [Months 12 to 18]
Standard approaches have been developed for validating prognostic models and data linkage quality.[7,37] Adapting and extending these approaches to this setting, we will follow 5 steps to validate the four indicators ***within the validation dataset***:

1. Report the agreement between the 4 indicators (clinical rule-based, unsupervised ML, prognostic modelling and supervised ML) on recurrence status and date of recurrence, for the cohort as a whole and separately by key patient and tumour characteristics such as stage, age, comorbidity and frailty.
2. For a sub-set of patients, use clinical adjudication to assess whether or not each patient's cancer has recurred. A sample will be taken, stratified by the number of indicators identifying recurrence and whether or not the patient died from cancer in the 3+ years after treatment (Table 4). Patients will be selected from the validation dataset to provide independence from development of the algorithms. Dependent on the level of agreement between indicators, approximately 50 patients will be sampled per stratum. The exception is stratum 4 (patients did not die of cancer in 3+ years and recurrence identified in no indicators) for which we expect very few recurrence events, and this stratum will include 100 patients. Around 350 patients will be included in total, depending on the level of agreement between the indicators

Table 4: Patients selected from the validation dataset for clinical adjudication

		Cancer recurrence identified in		
		0 indicators	1-2 indicators	3-4 indicators
Died of cancer in 3+ years after treatment	Yes	**Stratum 1** n=50	**Stratum 2** n=50	**Stratum 3** n=50
	No	**Stratum 4** n=100	**Stratum 5** n=50	**Stratum 6** n=50

At least 10 clinical experts will act as adjudicators. They will be blinded to the classifications of the 4 indicators and will be provided with fully anonymised detailed clinical information over time derived from the linked datasets (all hospital visits, diagnoses, procedures, imaging, chemotherapy drugs and doses, radiotherapy fractionations and schedules, (incomplete) cancer recurrence indicators from the Cancer Registry and Cancer Waiting Times data, and dates and causes of death). Each clinical adjudicator will review the records of approximately 100 patients, resulting in 3 reviews per patient. The cancer will be considered to have recurred if classified as recurrence by at least 2 out of 3 adjudicators. Sensitivity and specificity will be reported for each indicator and comparisons will be made between the sensitivity and specificity of the indicators using McNemar's test. The between-adjudicator agreement will also be reported.

3. Assess the stability of the four indicators over calendar time and across healthcare providers. Large shifts in the estimated rate of recurrence will highlight potential deficiencies in the indicators.
4. Estimate the relationship between each of the cancer recurrence indicators and known correlates of recurrence, such as cancer stage, surgical margins (for the primary resection), and long-term survival in all patients from initial treatment, and evaluate the plausibility of the results, compared to published findings.
5. Assess the sensitivity of analyses comparing cancer recurrence and recurrence-free survival between groups to the choice of recurrence indicator. These comparisons will include prognostic factors, treatments, and healthcare providers.

Recommendations will be made on the optimal approach, weighing up any increased accuracy from the model-based and ML algorithms against clinical face validity and simplicity of application.

WP5: Demonstrate the clinical use of the optimal indicator for bowel cancer [Months 19 to 27]

The value of the optimal indicator will be demonstrated using one example from each of the following three distinct purposes of the cancer recurrence indicators:

1. For comparing the effectiveness of different treatments using recurrence-free survival as the outcome (specific question to be determined by the research team under guidance from the study steering committee, taking into account the importance to clinicians and patients, the potential biases and errors, and statistical power)
2. Assessing the value of cancer recurrence after curative treatment as a performance indicator for benchmarking healthcare providers
3. Describing sociodemographic and geographic disparities in care pathways and outcomes of patients whose cancer recurs

Examples of questions addressing current gaps in knowledge for which cancer recurrence information is needed are:

1. Is total neo-adjuvant treatment (chemo-radiotherapy + surgery + chemotherapy) more effective at preventing recurrence than chemo-radiotherapy + surgery?
2. Are there differences in recurrence rates between recipients of robotic and open surgery?
3. Is chemo-radiotherapy + surveillance less effective at preventing recurrence than chemo-radiotherapy + surgery in patients who achieved a complete radiological response to chemo-radiotherapy?
4. Is dual agent chemo-radiotherapy (5FU + oxaliplatin) + surgery more effective than single agent chemo-radiotherapy (Capecitabine) + surgery?

Statistical power will be one of the factors considered when selecting the treatment comparison. There will be a total of approximately 32,000 curative bowel cancer patients available with at least 5 years of follow-up. As an example of one possible comparison, approximately 3,500 of these will be rectal cancer patients undergoing neoadjuvant chemo-radiotherapy. In a comparison of approximately 1,100 patients undergoing chemo-radiotherapy + surgery + chemotherapy versus 2,400 undergoing chemo-radiotherapy + surgery, there would be 85% power to detect a difference in recurrence of 14% versus 18%.

These comparisons will use “state of the art” statistical methods to minimise bias due to the observational nature of the datasets created. These will include comprehensive confounder adjustment using multivariable modelling and propensity score weighting methods. Propensity scores are used as weights to account for selection assignment differences between treatment and comparison groups. Very rich information on patient, tumour and treatment characteristics is available from the linked data for estimating the propensity score, including demographics, deprivation, ethnicity, comorbidities, functional status, frailty, tumour site and staging, surgical procedure, urgency and approach, and other treatments.

Amongst patients having curative treatment data completeness is high for most data items. For example, 88% of bowel cancer patients undergoing major resection have complete cancer stage, 96% have complete ASA grade, and 95% can be linked to HES to obtain information on comorbidities. Having multiple sources of data items reduces the amount of missing data. Multiple imputation will be used for items that are missing across all data sources. Many ML methods, such as random forest, “impute” the missing data automatically, from the assumed functional form.

WP6. Assess how well the optimal indicator extends to breast and prostate cancer [Months 28 to 34]

The optimal indicator for bowel cancer recurrence (rule-based or ML but not both) will be extended to breast and prostate cancer. The rule-based indicator will comprise a sub-algorithm for each of the eight domains of healthcare activity at each of the three phases in the pathway (diagnosis and treatment, surveillance, recurrence). Should this be the optimal indicator, clinical experts will advise how to adapt the relevant sub-algorithms to the specific cancer site. For example, guidelines differ on surveillance for each cancer, and different regimens of chemotherapy are used for recurrence. The same coding principles will be used to identify patterns of care across the cancers, and across all of the sub-algorithms much of the coding will be the same as for bowel cancer.

ML algorithms do not rely on clinical information about the specific cancer type. They can be applied in other cancers without any prior adjustment, given that they will adapt themselves to different recurrence-related events and treatments. Should a ML algorithm be the optimal indicator we will learn from any methodological and convergence issues encountered in work package 3.

The validation process in work package 4 will modified so that it does not make use of agreement between indicators. It will be applied for each cancer using an adjudication dataset and at least 10 clinical adjudicators for each cancer site, who will be blinded to the classifications of the recurrence indicator. The validity of the indicator may differ between cancer sites because of their different care pathways, rates of recurrence and time from recurrence to death. This validation approach will also demonstrate how the indicators can be made applicable to other cancers beyond the three included in this research.

6. Dissemination, Outputs and anticipated Impact

Dissemination and outputs

The results will be relevant to all NHS stakeholders: patients, public, cancer charities, practitioners, health-service managers, academics, and health policy experts. The planning and delivery of outputs will be informed by the PPI-focused Steering Committee and will include:

- Publishing algorithms and methods in full, ensuring they are reproducible, in peer reviewed articles, including those targeting general clinical audiences, cancer specialists and methodologists
- Development of public facing outputs including patient summaries and liaising with cancer charities through the PPI lead to publicise the findings of the project with their members/supporters.
- Publishing the algorithms on the HDRUK Innovation Gateway (https://www.healthdatagateway.org/) which is a go-to repository of methods and algorithms for data science
- Advising NCRAS, the Healthcare Quality Improvement Partnership, national cancer audit providers, and the Welsh Cancer Network, to enable cancer recurrence to become a standard data item for national cancer audits and cancer registries.
- Using e-communication portals (e.g. http://ecancer.org/) to describe the project and its results.
- Providing a research report for the NIHR HS&DR programme detailing research methods, findings and conclusions of all WPs, including recommendations for practice and an extensive summary for patients and the wider public.
- Publicising through stakeholders on the Study Steering Committee, including representatives from NCRAS, clinical professional bodies, national clinical audits, NIHR ARC North Thames and cancer charities.
- Launching the publications in parallel with presentations during relevant conferences and events accompanied by press releases, website updates and social media, such as Facebook and twitter accounts of clinical audits, professional bodies, data providers and charities.
- Transferring the methods for direct use in the three national clinical cancer audits that the research team deliver and using them as exemplars to promote the methods to other cancer audits and other researchers using national clinical cancer data.

Impact:
The research will accelerate improvements in cancer services by directly or indirectly enabling or informing:
1. Prediction of prognosis to better inform patients
2. Evidence on the optimum modalities of care for patients with cancer
3. Identifying patient sub-groups who will benefit from specific treatments
4. Conduct of clinical trials (by identifying patients who experienced a recurrence for inclusion in trials as well as by providing recurrence as an outcome measure)
5. Allocation of healthcare resources by providing outcomes for public health research
6. Service evaluation and feedback to cancer care providers for quality improvement
7. Discovery of aetiological factors and novel cancer treatment targets
8. Improving cost-effectiveness evaluations for cancer.

The results of this research will enable these activities, initially for three common cancers but with the potential to be extended across other cancer sites. The resulting algorithms will give more representative estimates of cancer recurrence or recurrence-free survival for patients according to disease stage and physical fitness and other specific patient characteristics, treatments, and characteristics of providers of cancer services. An important contribution is that it will enhance clinical trials that typically have limited applicability in ‘real-world’ settings due to the limited ability to have long-term follow-up.

The methods and developed algorithms will be disseminated with full transparency so that

cancer recurrence can be incorporated as an indicator within cancer registries and national cancer audits for epidemiological, clinical audit, clinical trial or policy purposes.

We also expect NHS England and regional commissioners of cancer services to be aided in various ways through the availability of better information on cancer recurrence facilitated by our algorithms. In particular, cancer recurrence is a key element in models evaluating the cost-effectiveness of cancer treatments. More accurate national and regional information on the number of cancer patients who experience a recurrence will guide resource planning. The cancer recurrence indicators can be used to compare outcomes between treatments and providers, with subsequent implications for performance management and quality improvement.

7. Project / research timetable

As soon as the application is successful, approvals will be applied for and data requested from NCRAS to ensure that the data are available from the start of the project.

Months	**Delivery**
Pre-start	Recruitment of Research Fellow. Data approvals and requests.
0 to 3	WP1. Construct the care and outcome pathways of patients across national datasets.
2 to 10	WP2. Develop an indicator to phenotype the presence and timing of recurrence of bowel cancer using clinical rule-based methods.
6 to 14	WP3. Develop indicator to phenotype the presence and timing of recurrence of bowel cancer using statistical and ML methods.
12 to 18	WP4. Validation of the bowel cancer indicators. Write publications from WP2 and WP3 and development of patient outputs.
19 to 27	WP5. Demonstrate the clinical use of the indicators for bowel cancer. Write publications from WP5 and development of patient outputs.
28 to 34	WP6. Evaluate the cancer recurrence indicator developed for bowel cancer in breast and prostate cancer.
32 to 36	Dissemination activities. Final report for NIHR.

8. Project management

Dr Kate Walker (20% FTE) will, as principal investigator, take overall responsibility for leadership of the study, supported by Prof Jan Van der Meulen (10% FTE). She will lead the research team’s monthly meetings, held to discuss all relevant methodological, practical and logistical issues. Other co-applicants and collaborators will be involved when necessary and appropriate. Dr Julie Nossiter will be responsible for the project management (10% FTE).

Administrative support (15% FTE) will be available to help with arranging meetings, dealing with day to day queries and budget management.

The study steering committee will oversee the implementation of the study and comprise of all co-applicants, stakeholders and the research fellow. It will monitor the progress of completion of tasks against the project's timeline and consider remedial action if needed. The group will also discuss the implications of findings, and decide how they should be disseminated. The group's meetings will be face-to-face with an option of video-conferencing facilities for those who require it.

The work for this project will be carried out by a full-time research fellow who will be supervised on a daily basis by Dr Kate Walker, supported by Prof Jan Van der Meulen and the rest of the research team. The research team will communicate on a regular basis with the study steering committee members to seek their input on all key issues related to research design, method development, data analysis and interpretation and reporting. The research team will meet monthly to discuss all relevant methodological, practical and logistical issues,

9. Ethics / Regulatory Approvals

National linked electronic health datasets will be requested from Public Health England. These anonymised datasets will only include the following patient information: age (in years), sex, ethnicity and the LSOA reflecting their area of residence. The datasets will be housed in a secure data environment at LSHTM. Governance procedures are already in place to use and store patient level datasets. Given that the proposed research will only involve the use of these anonymised datasets, NHS REC approval will not be sought in accordance with their guidelines. Approval from the LSHTM Observational/Interventions Research Ethics Committee will be sought.

10. Project / research expertise

Kate Walker is PI and, together with Jan Van der Meulen and the rest of the research team, will oversee the successful delivery of the project, in particular as line manager to the research fellow employed to carry out the research. She is a senior statistician specialising in complex methodological issues in health services research, linkage of multiple national clinical datasets, developing risk-adjustment models and clinical indicators, and is lead methodologist for the National Bowel Cancer Audit and a senior methodological advisor to several national clinical audits.

Jan Van der Meulen (co-PI) is lead methodologist for the National Prostate Cancer Audit, senior methodologist for the National Bowel Cancer Audit and lead methodologist for the National Maternity and Perinatal Audit. He has decades of experience in research that focuses on the study of determinants of variation in processes and outcomes of surgical care using routinely collected electronic health data. He will provide senior methodological and project oversight.

Ajay Aggarwal is a Consultant Clinical Oncologist specialising in the delivery of systemic and radiation therapies for the management of prostate cancer at all stages. He holds an NIHR Advanced Fellowship studying integrated care systems for specialist cancer treatments using routinely collected healthcare data. He will provide clinical expertise on the care pathways for prostate cancer, as well as other cancers, from an oncology perspective.

Michael Braun is Clinical Co-lead for the National Bowel Cancer Audit. He has a wealth of experience of interpreting and understanding the linked data for bowel cancer patients, and was a member of the NICE Committee for updating the latest colorectal cancer guidelines.

He will provide clinical expertise on the care pathways for bowel cancer, as well as other cancers, from an oncology perspective.

Kieran Horgan is Clinical Co-lead for the National Audit of Breast Cancer in Older Patients (NABCOP) and representative of the Association of Breast Surgery. He brings to the project his in-depth knowledge of linked data for breast cancer patients and will provide clinical expertise on the care pathways for breast cancer, as well as other cancers, from a surgical perspective.

Karla Diaz Ordaz is a senior statistician who will bring to the project her expertise on ML approaches using high-dimensional electronic health records. She holds a Wellcome Trust-Royal Society Sir Henry Dale Fellowship and is co-lead in a collaborative research project on developing statistical ML methods based at the Alan Turing Institute.

Linda Sharples is Professor of Medical Statistics with expertise in applying rigorous statistical analysis in observational and experimental studies. She will contribute senior statistical expertise, in particular on modelling routinely collected healthcare data, missing data, and incorporating changes in health status over time to provide a comprehensive picture of the way in which diseases and conditions develop.

Thomas Cowling currently holds a MRC Skills Development Fellowship on using linked national clinical datasets to develop prediction algorithms using conventional statistical methods and ML methods. He has in-depth knowledge and methodological expertise of linked data for prostate and bowel cancer patients and will bring statistical and ML expertise to the project.

David Cromwell is lead methodologist for the National Audit of Breast Cancer in Older Patients and the National Oesophago-Gastric Cancer Audit. A quantitative health services researcher with experience of using linked datasets to evaluate patterns of surgery and patient outcomes, he will bring to the project his expertise in health services research using routine clinical data, also building on his MRC-funded methodology research on developing methods to assess the quality of clinical datasets.

Julie Nossiter is a senior project manager who has been Audit Lead for National Prostate Cancer Audit (NPCA) since its inception in 2013 responsible for coordinating activities across different organisations and ensuring the timely delivery of the audit outputs. Working closely with Prostate Cancer UK and Tackle prostate cancer, she set-up a standalone NPCA PPI Forum to ensure that the voice of patients and carers is heard and valued. Her research focuses primarily on evaluating the performance and quality of prostate care services in England and Wales using routine clinical data. She brings to the project a wealth of project management experience including establishing and maintaining robust risk and issue management procedures based upon PRINCE 2 principles, collaborative working with the National Cancer Registration and Analysis Service, as well as information governance expertise.

PPI Lead

The PPI Lead is the co-applicant Professor Robert Arnott (of Green Templeton College, Oxford), who is Chair of the Patient Liaison Group of the ACPGBI and the Patient and Carer Panel of the NHS National Bowel Cancer Audit and a member of the GI Cancer Project Board. He was for sixteen years a trustee of the Bowel Disease Research Foundation and was a founder of Bowel Research UK. He brings to the project years of experience with preparing and activating the PPI for several research projects funded by the NIHR, the Leverhulme Trust and other funding bodies.

He is costed according to INVOLVE guidance and will contribute one day per month for the duration of the project. He has already advised on setting the PPI strategy, and the membership of the PPI focused study steering committee. He will lead on refining the PPI strategy as the project progresses.

As an active member of the research team he will oversee all of the PPI plans, related activities and outcomes of the project. He will ensure that people with living experience (both patients and their carers) are involved in each stage of the project and he will lead in explaining how they can be involved and what they can expect when they do. He will lead (in the interests of patients), the production of patient information sheets and other patient / public facing outputs of the project, as well as writing the PPI sections for project reports. He will keep records of all PPI activity throughout the project and use these to evaluate and report on the PPI impact.

11. Success criteria and barriers to proposed work

We expect to produce a minimum of six high impact publications on the development and validation of the cancer recurrence indicators and demonstrations of their clinical use. Findings of each will be presented at national and international meetings and conferences..

The Study Steering Committee and PPI Advisory Board will ensure outputs are relevant, patient focused and have the potential to translate to every day practice and policy.

The dissemination strategy ensures that benefits accrued from this work are made available to a wide group of stakeholders and we anticipate that, as a results of this research, cancer recurrence will become a standard outcome for national cancer audits, cancer registries, analysts and epidemiologists using national cancer datasets.

In the event of a delay obtaining data from PHE, we will be able to access data from the three national cancer audits from the start of the project. The audits are based between LSHTM and the RCS.

There are likely to be issues with the quality of the data. However, the research team is highly experienced in working with the national datasets required for this project and have developed peer-reviewed, methodological approaches to handling missing data and assessing data quality and linkage quality.[37,46-47]

There is very little risk of a lack of stakeholder engagement as we already have confirmed membership of the Study Steering Committee across all relevant stakeholders.