BMJ Health & Care Informatics

Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study

Andres Tamm,[1,2] Helen JS Jones,[1,3] William Perry ,[1,3] Des Campbell,[4,5] Rachel Carten,[4,6] Jim Davies,[1,7] Algirdas Galdikas,[8,9] Louise English,[10] Alex Garbett,[11,12] Ben Glampson,[8,9] Steve Harris,[1,7] Khurum Khan,[10,13] Stephanie Little,[1,3] Lee Malcomson,[11,12] Sheila Matharu,[4,5] Erik Mayer,[9,14] Luca Mercuri,[8,9] Eva JA Morris,[1,2] Rebecca Muirhead,[1,3] Ruth Norris,[11] Catherine O'Hara,[11,12] Dimitri Papadimitriou,[8,9] Niels Peek ,[11,15] Andrew Renehan,[11,12] Gail Roadknight ,[1,3] Naureen Starling,[4,5] Marion Teare,[4,5] Rachel Turner,[4,5] Kinga A Várnai,[1,3] Harpreet Wasan,[8,16] Kerrie Woods,[1,3] Chris Cunningham[1,3]

To cite: Tamm A, Jones HJS, Perry W, *et al*. Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study. *BMJ Health Care Inform* 2022;**29**:e100535. doi:10.1136/bmjhci-2021-100535

AT, HJJ and WP are joint first authors.

Received 27 December 2021
Accepted 25 May 2022

For numbered affiliations see end of article.

Correspondence to
Chris Cunningham;
Chris.Cunningham@ouh.nhs.uk

ABSTRACT

Objective Colorectal cancer is a common cause of death and morbidity. A significant amount of data are routinely collected during patient treatment, but they are not generally available for research. The National Institute for Health Research Health Informatics Collaborative in the UK is developing infrastructure to enable routinely collected data to be used for collaborative, cross-centre research. This paper presents an overview of the process for collating colorectal cancer data and explores the potential of using this data source.

Methods Clinical data were collected from three pilot Trusts, standardised and collated. Not all data were collected in a readily extractable format for research. Natural language processing (NLP) was used to extract relevant information from pseudonymised imaging and histopathology reports. Combining data from many sources allowed reconstruction of longitudinal histories for each patient that could be presented graphically.

Results Three pilot Trusts submitted data, covering 12 903 patients with a diagnosis of colorectal cancer since 2012, with NLP implemented for 4150 patients. Timelines showing individual patient longitudinal history can be grouped into common treatment patterns, visually presenting clusters and outliers for analysis. Difficulties and gaps in data sources have been identified and addressed.

Discussion Algorithms for analysing routinely collected data from a wide range of sites and sources have been developed and refined to provide a rich data set that will be used to better understand the natural history, treatment variation and optimal management of colorectal cancer.

Conclusion The data set has great potential to facilitate research into colorectal cancer.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Colorectal cancer is a major source of mortality and morbidity worldwide and further research is needed to improve outcomes.

WHAT THIS STUDY ADDS

⇒ This study outlines the potential of a multicentre colorectal cancer data set from routinely collected National Health Service data.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Research using such a data set will inform clinical practice and aid governing bodies in the development of colorectal cancer care pathways to reduce disparities and improve overall patient outcomes.

INTRODUCTION

Globally, 1.93 million people were diagnosed with colorectal cancer in 2020.[1] Further, some 9.4% of cancer mortality was attributed to colorectal malignancy.[1] In the UK, it is one of the most common cancers with approximately 42 000 new cases registered each year.[2]

Current global epidemiological estimates for colorectal cancer are provided by the WHO's Global Cancer Observatory[3] and by the Institute of Health Metrics and Evaluation's Global Burden of Disease Estimates.[4] Both use complex statistical modelling to overcome data limitations to produce estimates of fatal and non-fatal outcomes. Various smaller national databases exist, including those that specifically explore colorectal cancer.[5]

Such databases provide an opportunity to better understand the burden of colorectal cancer and outcomes, alongside improving treatment guidelines, however they are

limited by both a lack of automated input of routinely captured clinical data and their adaptability and applicability for research. These limitations have been acknowledged via initiatives such as the UK Colorectal Cancer Intelligence Hub[6] (which promotes the generation of colorectal cancer intelligence by compiling and using administrative data in the COloRECTal cancer data Repository) but higher resolution and more timely information remains in demand.

This need could be met via automated collation of routinely collected high-resolution clinical data from hospital systems. This would provide further opportunity to alleviate administrative burden and allow for expansive data sets that capture a large volume and expanding number of touchpoints for every patient and every healthcare interaction. The challenge of making such data available for research led to the development of the National Institute for Health Research (NIHR) Health Informatics Collaborative (HIC).[7]

The NIHR HIC is a partnership of 29 National Health Service (NHS) Trusts and health boards, including the 20 hosting NIHR Biomedical Research Centres (BRCs). The NIHR HIC network aims to facilitate development of clinical informatics infrastructure to enable the reuse and sharing of routinely collected NHS clinical information to better inform research, patients and NHS staff. The utility of this programme in addressing viral hepatitis has already been demonstrated.[8]

The Colorectal Cancer theme of the NIHR HIC was established to develop and produce a descriptive analysis of colorectal cancer in the UK and address contemporary research questions. Specifically, the theme aims to develop an automatically collated high-resolution data set, validate national colorectal cancer patient data, create a longitudinal patient record of treatment for colorectal cancer patients, improve national reporting, and provide data and research outcomes to improve the delivery of colorectal cancer care across the UK.

This study aimed to collate routinely collected colorectal cancer data across three pilot sites. Further, it aimed to document both the process of doing so and the wider potential of the HIC platform for colorectal cancer research.

METHODOLOGY

All member Trusts of the NIHR HIC were invited to partake in the colorectal cancer theme, led out of Oxford University Hospitals (OUH) NHS Foundation Trust (FT) in collaboration with the NIHR Oxford BRC's Clinical Informatics and Big Data theme. Of those Trusts which joined the Collaborative, Imperial College Healthcare NHS Trust (ICHT), The Royal Marsden NHS FT (RMT) and OUH NHS FT submitted data as part of this pilot study.

Patient population

All patients with International Classification of Diseases Version-10 (ICD-10) diagnosis codes C18, C19 and C20 from 1 January 2012 through 28 February 2021 were eligible for inclusion.

Defining data capture

Data points for capture were specified by a group of experts from across the NIHR HIC Colorectal Cancer theme using a modified-Delphi framework. This group was comprised of colorectal surgeons and oncologists from institutions partaking in the wider NIHR HIC colorectal cancer theme: ICHT, The RMT, OUH NHS FT, Guys and St Thomas' NHS FT, Leeds Teaching Hospitals NHS Trust, The Christie NHS FT, University College London Hospitals NHS FT and University Hospitals Birmingham NHS FT. The group met virtually on a bi-weekly basis during construction of the data points. The National Bowel Cancer Audit (NBOCA) data set,[5] the Commissioning Data Sets[9] and the National Cancer Registration and Analysis Service data sets including Cancer Outcomes and Services Data Set,[10] Systemic Anti-Cancer Therapy Data Set[11] and National Radiotherapy Data Set[12] were used as a reference. The data points proposed by the group were then tested against a series of hypothetical research questions to ensure data captured could drive descriptive research in colorectal cancer before they were finalised. The model was designed so that it could be expanded without compromising the integrity of any contemporaneous data. The NHS Spine[13] was interrogated on a regular basis to update mortality data.

Data collation

Data were initially collated at each Trust using an internal and secure data warehouse in an identifiable form. Each Trust reviewed their regional data to ensure accuracy of data capture. Lead clinicians were responsible for ensuring accuracy of longitudinal data representation, with any discrepancies addressed and integrated into a quality improvement cycle. It was then processed to remove all directly identifying patient information from the records prior to transfer. Data were then transmitted via the NHS Health and Social Care Network (HSCN) using a LabKey[14] portal to the NIHR HIC Colorectal Cancer research database, where patients were assigned a unique pseudonymised study identifier for subsequent analysis.

The NIHR HIC Colorectal Cancer research database was built using Microsoft MySQL Server[15] and hosted by OUH NHS FT. The anonymous data were processed and stored in accordance with the NIHR HIC Data Sharing Framework. Code to extract data at each site and all transformations applied thereafter were stored securely, allowing the entire database to be recreated with minimal effort if required. Once collated, data points were parsed through logic and linkage validation.

Natural language processing

The OUH team developed rule-based algorithms to extract cancer staging and recurrence from local free-text imaging and pathology reports using natural language

processing (NLP). Data extraction included tumour, node, metastases (TNM) classification, extramural venous invasion, circumferential resection margin (CRM) involvement, distance to the CRM, Kikuchi and Haggitt subcategories of T stage, and the presence of recurrence and metastasis. Each algorithm was designed to look for target words in the context of other keywords, or for variable sequences of TNM categories. A lightweight app was also created in Shiny,[16] an R package[17] run on Rstudio,[18] to facilitate the labelling of reports. The output of NLP was cross-referenced with the free text to ensure accuracy. These algorithms were shared with the ICHT team to allow implementation prior to data collation and transfer to the NIHR HIC Colorectal Cancer research database.

Analysis

Baseline characteristics

Baseline characteristics were reported as median and IQR (shown as 25th and 75th percentiles), and number and percentage. Age at diagnosis was derived using the date of the first ICD-10 C18–C20 diagnosis code. Average body mass index (BMI) was computed for each patient after excluding erroneous values less than 10 or greater than 100. Neoadjuvant treatment was defined as chemotherapy and/or radiotherapy without surgery or preceding surgery by up to 180 days. Adjuvant treatment was defined as chemotherapy or radiation (eg, postlocal excision) within 180 days of surgery. Surgery consisted of local excision (Office of Population Censuses and Surveys Classification of Intervention and Procedures (OPCS-4) codes starting with H402, H412 and H34) or radical resection (OPCS-4 codes starting with H04–H11, H29, H33, X14). Length of follow-up was computed as the number of years from the first colorectal cancer diagnosis code to last contact date or date of last check against NHS Spine, whichever was later. Analysis was undertaken in Python V.3.8.5 using the pyodbc (V.4.0.32)[19] and pandas (V.1.1.3)[20 21] libraries.

Recurrence and T stage

A rule-based algorithm was used to extract T stage for each patient for whom relevant clinical reports were available. To summarise staging in the patient cohort, the highest T stage was selected: for patients who had local excision or radical resection, the highest histopathological staging up to 6 weeks after surgery was used; for patients who only had chemotherapy and/or radiotherapy, the highest staging given in imaging reports up to 6 weeks before therapy was used; in all other cases, the highest staging given at any point in time was used (with a preference for pathological staging). For patients with colon cancer (C18) who had undergone radical resection, presurgical and postsurgical T stages given closest to the time of surgery were visualised using a Sankey diagram created with plotly (V.5.1.0).[22]

A separate algorithm was used to extract references to recurrence and metastasis from relevant endoscopy, imaging and pathology reports. Additional instances of metastasis were extracted using ICD-10 diagnosis codes (starting with C76–C80). Metastases occurring up to 6 weeks before or after the first known colorectal cancer diagnosis code were classified as part of the primary presentation.

Longitudinal plotting

Longitudinal pathway plots were created using Matplotlib (V.3.3.2)[23 24] to visually represent individual patient pathways with colon and rectal cancer. The sequence of events to define groups of patients were predesignated by authors (AT, HJJ, WP, CC) as outlined in figure 1. All longitudinal plots presented in this paper are hypothetical. They do not depict any real patient but rather provide a representation of the plotting achieved in order to preserve anonymity.

RESULTS

A total of 12 903 unique patients who had a diagnosis of colorectal cancer between 1 January 2012 and 28 February 2021 were submitted to the NIHR HIC Colorectal Cancer research database across the three pilot sites. An overview of baseline demographics and outcomes is provided in table 1. In total, the database contained 32 tables and 336 data fields. The number of records captured per selected data item is outlined in table 2.

Data captured included all surgical procedures, courses of chemotherapy and radiotherapy, endoscopy and imaging events, blood test results and clinical diagnoses. NLP was used for 4150 patients. T stage was identifiable in 2444 (58.9%) of these patients when applied to endoscopy, imaging and histopathology reports. Some 1931 (46.5%) of these patients were identified to have recurrence or metastases of which 1119 (27.0%) were found at the time of diagnosis. T stage was more readily identified in those who had undergone surgery (94.7%, table 3).

A Sankey plot based on NLP of imaging and histopathology reports is provided in figure 2. The left side of the plot shows the pretreatment T stage for 204 colon cancers determined by NLP of CT and MRI reports. The right side shows the T stage based on NLP of the histopathology report issued close to the time of surgery.

Patient events were represented on longitudinal pathway plots. Figure 1 shows hypothetical pathway plots for patients with four common pathways. A representation of 10 patient pathways is shown for each group.

Two individual timelines are expanded in figures 3 and 4 to demonstrate the level of detail that could theoretically be obtained through this process. Figure 3 shows a single timeline for a hypothetical patient who had rectal cancer managed by neoadjuvant treatment then surgery. Figure 4 shows a timeline for a hypothetical patient with rectal cancer managed by local excision.

DISCUSSION

Nine years of colorectal cancer data were successfully collected across three pilot sites and collated in a

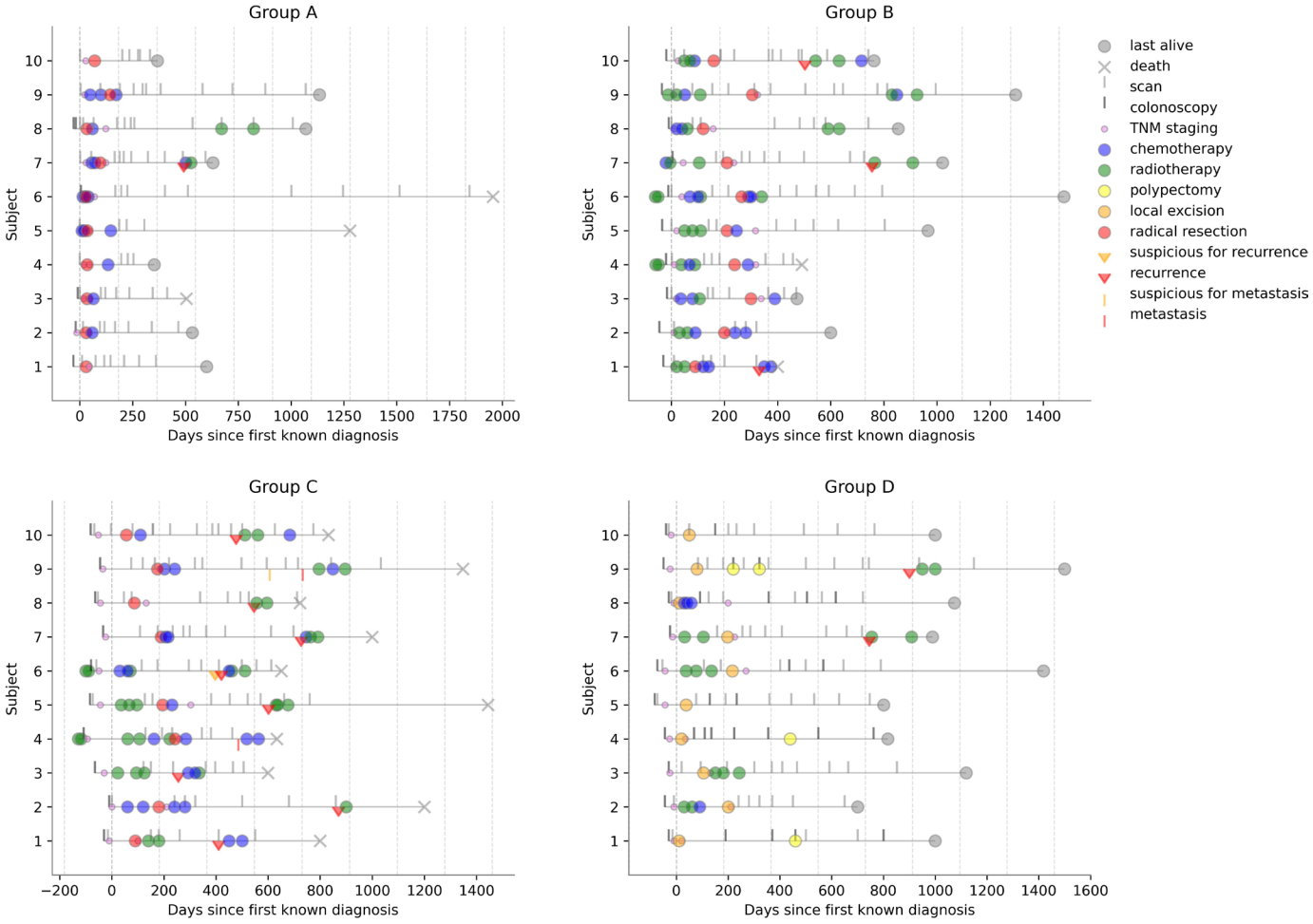

Figure 1 Hypothetical patient timelines that show specific treatment and surveillance patterns. Group A: Timelines of patients with colon cancer that follow the pattern 'diagnosis, scan, surgery, scan'. Group B: Patients with rectal cancer with 'diagnosis, scan, chemoradiotherapy, radical resection, chemo(radio)therapy, scan'. Group C: Patiens with colorectal cancer with 'diagnosis, treatment, scan, recurrence, treatment, death'. Group D: Patients with rectal cancer with local excision. Timelines for 10 patients were created to illustrate each group. TNM, tumour, node, metastases.

centralised research database as part of the NIHR HIC. This process demonstrated that it is possible to create an automated data-rich longitudinal research-focused database from routinely collected health data. In doing so, this paper highlights the potential of this database in future colorectal cancer research. To our knowledge, this is the first such database of its type.

NLP of histopathology, imaging and endoscopy reports has the potential to create a depth of data beyond coding, synoptic reporting and manually entered data. Several algorithms were successfully created to extract key data points and successfully implement them for different source data. This paper only presents one Trust's NLP results, however the algorithms have been shared and implemented across the other pilot sites. This ability to share complex validated algorithms has the potential to take the database beyond common epidemiological or research parameters, especially when using an open source philosophy. In this context, open source facilitates sharing, collaboration, personalisation and rapid advancement of algorithm development and thus data processing.

The pathway plots developed are valuable in identifying groupings of patients with similar pathways to aid future analysis. Group A (patients with a diagnosis of colon cancer, with a pretreatment staging scan, followed by surgical resection of the cancer) and Group B (diagnosis of rectal cancer who had a pretreatment staging scan, neoadjuvant treatment followed by surgical resection then further treatment, either adjuvant or for recurrence) provide apt examples: Group A plots provided a visual indication of the proportion of the group who had adjuvant treatment, the completeness of the follow-up regime and the incidence of disease recurrence, while Group B plots provided insight into temporal variability in adjuvant treatment.

These particular groups were selected to illustrate the potential of this form of representation of the data, rather than address particular research questions. The plots also clarified issues with the data that needed to be addressed. For example, several pathways recorded

Table 1 Demographic baseline of patients captured with colorectal cancer in the NIHR HIC colorectal cancer research database form three pilot sites

Characteristic	Value
Number of participants	12 903 (100%)
Cancer site	
C18—colon	4920 (38.1%)
C19—rectosigmoid	1171 (9.1%)
C20—rectum	2699 (20.9%)
Not known yet	4997 (38.7%)
Age at diagnosis	
Age, years, median (IQR)	68.4 (58.1–77.3)
Not known yet	4997 (38.7%)
Sex	
Male	7223 (56.0%)
Female	5504 (42.7%)
Not known	176 (1.4%)
Ethnicity	
White	9222 (71.5%)
Not stated	1575 (12.2%)
Other ethnic groups	786 (6.1%)
Asian or Asian British	554 (4.3%)
Black or Black British	396 (3.1%)
Mixed	77 (0.6%)
Not known	293 (2.3%)
Smoking status	
Current or ex-smoker	2912 (22.6%)
Non-smoker	3793 (29.4%)
Not known	6964 (54.0%)
Body mass index	
Median (IQR)	25.8 (22.9–29.2)
Not known	3220 (25.0%)
Treatment	
Neoadjuvant	3708 (28.7%)
Adjuvant	962 (7.5%)
Local excision	291 (2.3%)
Radical resection	3715 (28.8%)
Not known yet	5749 (44.6%)
Mortality and follow-up	
Number of deaths	5090 (39.4%)
Years from diagnosis to mortality, median (IQR)	0.9 (0.3–2.0)
Years from diagnosis to last follow-up, median (IQR)	2.4 (0.8–4.6)

HIC, Health Informatics Collaborative; NIHR, National Institute for Health Research.

treatment or even recurrence well before the initial diagnosis. Certain groups were able to be identified within the data set, such as those primarily managed at a peripheral hospital before referral to a specialist tertiary unit, that need further attention and processing before the data are used for research analysis.

Table 2 Number of records per selected data items in the NIHR HIC colorectal cancer research database

Field	Number of records
Laboratory tests	5071 605
Inpatient episodes	13 737
Diagnosis codes	443 762
Procedure codes	154 363
Radiotherapy	7726
Chemotherapy	17 452
Histology reports	15 311
Relevant histology reports	9226
Imaging reports	96 330
Endoscopy reports	13 737
Relevant endoscopy reports	11 352

Relevant histology reports contain references to colon or rectum; relevant imaging reports correspond to certain investigations, such as MRI of pelvis and rectum; and relevant endoscopy reports represent colonoscopies and flexible sigmoidoscopies.
HIC, Health Informatics Collaborative; NIHR, National Institute for Health Research.

Although not specifically explored in this pilot study, such data capture has the potential to explore variation in practice. For example, there is significant variation in the use of neoadjuvant treatment across the UK, especially

Table 3 T stage, recurrence and metastatic disease identified in the NIHR HIC colorectal cancer research database through NLP of imaging, endoscopy and/or histopathology reports for all patients who had surgical excision at one of the pilot sites

Characteristic	Radical resection or local excision
Number of participants	2124 (100%)
Maximum T stage	
0	31 (1.5%)
is (in situ)	0 (0%)
1	195 (9.2%)
2	369 (17.4%)
3	954 (44.9%)
4	460 (21.7%)
X	2 (0.1%)
Not known	113 (5.3%)
Recurrence or metastasis	
Recurrence or metastasis detected	769 (36.2%)
Metastasis present around time of diagnosis	286 (13.5%)
Not detected	1355 (63.8%)

HIC, Health Informatics Collaborative; NIHR, National Institute for Health Research; NLP, natural language processing.

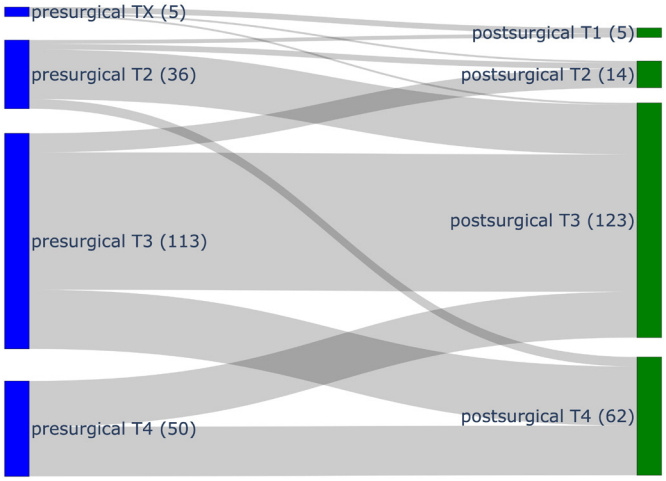

Figure 2 Presurgery and postsurgery T staging for patients with colon cancer (C18) who had a major resection, determined by natural language processing (NLP) of imaging reports (presurgery) and histopathology reports (postsurgery). Number of patients is given in brackets.

for higher rectal tumours.[25] The breadth of this database has the potential to identify variance in greater detail, and provide insight into outcomes across various patient cohorts.

The processes developed for this pilot study will be applied to generate a much larger database as other centres contribute data and the time period is extended. Further, the data set can be expanded and adapted to match the requirements of research questions. This does not replace other data collection programmes such as the NBOCA,[5] which plays an important role in quality of colorectal cancer care above all else. The attributes of this data set, however, provide a unique research opportunity to investigate novel strategies in the management of colorectal cancer.

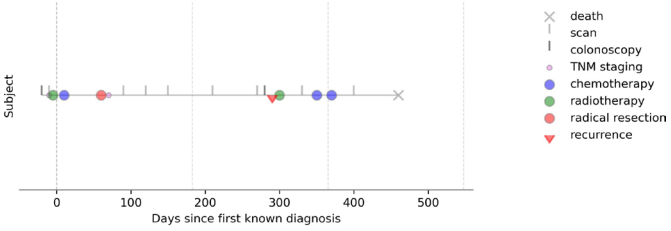

Figure 3 Longitudinal pathway plot of a hypothetical patient with rectal cancer treated with neoadjuvant therapy then radical resection. After a colonoscopy and around the time of diagnosis the patient had neoadjuvant radiotherapy and chemotherapy as identified by the green and blue circles. They then proceeded to surgery, after which TNM staging was available (small pink circles). The next time point for this patient (light grey line) shows a scan done as part of the follow-up regime, with several further thereafter. Nearly 300 days since diagnosis a scan and colonoscopy led to the diagnosis of recurrence and further radiotherapy and chemotherapy. The final 'X' signifies death, although it does not show whether death was related to the cancer or not. TNM, tumour, node, metastases.

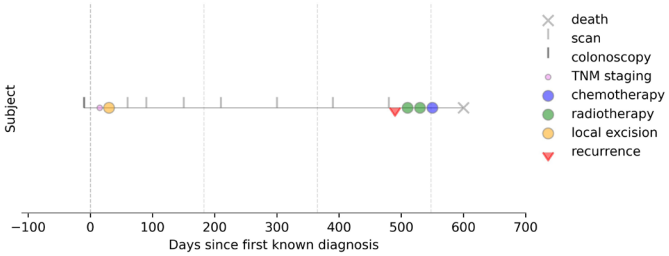

Figure 4 Longitudinal pathway plot of a hypothetical patient with rectal cancer who underwent local excision. Rectal cancer was picked up on colonoscopy as indicated by the dark grey line, and treated by local excision as indicated by the orange circle. After a disease-free surveillance period of approximately 18 months, the patient had recurrence as shown by the first red arrow. This was followed by radiotherapy and chemotherapy prior to death. TNM, tumour, node, metastases.

Alongside the research potential illustrated by this study, the process also highlighted challenges in such data extraction. Trusts were readily able to obtain inpatient data points, however outpatient data were more difficult to capture which explains some of the variables still missing in the results (table 1). This highlights the importance of greater collaboration across inpatient and outpatient facilities while demanding a greater focus on this aspect of data extraction in the longer term. Further, several therapy points, including neoadjuvant, surgical and adjuvant therapy were missing when treatments were provided at facilities outside the central Trust, reflecting the centralisation of certain services at a regional level in the NHS. The database will ultimately need to be expanded to include more centres across the UK to maximise the research potential.

Although NLP was successful in capturing more complex data components, it currently has a low capture rate. For example, T staging has not yet been identified in some 41% of patients for whom NLP was undertaken. However, when analysis was restricted to patients who had surgical resection recorded at the site, T stage was obtained for 95% of patients. The algorithms have only been applied to imaging and histopathology reports so far, and require these reports to specifically mention T stage. It is expected that data capture will increase as the algorithm is improved, and as it is applied across a wider range of data sources, for example, including multidisciplinary meeting reports and operative notes.

The database is only as accurate as the data inputted. While it is possible to build in simple validation checks to exclude or correct nonsensical values, for example in age or BMI, more complex issues such as errors in reporting that result in misclassification will not be detected at the 'big data' level. Such errors may be detected in smaller scale research projects where original data are scrutinised, but at the larger scale, the assumption is that the incidence of such errors will be relatively small and not significantly impact the overall results. This is however a

limitation of the database and a focus for optimisation as the database continues to be developed.

In summary, automated collation of routinely collected clinical data does not only promise to alleviate administrative burden but allows for expansive data sets that capture a theoretically unlimited and expanding number of touchpoints for every patient. Ultimately, research using catalogued, comparable, comprehensive and longitudinal patient data will inform clinical practice and aid governing bodies in the development of colorectal cancer care pathways to reduce disparities and improve overall patient outcomes.

Author affiliations
[1]NIHR Oxford Biomedical Research Centre, Oxford, UK
[2]Big Data Institute and the Nuffield Department of Population Health, University of Oxford, Oxford, UK
[3]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[4]Royal Marsden NHS Foundation Trust, London, UK
[5]NIHR Biomedical Research Centre at The Royal Marsden and The Institute of Cancer Research (ICR), London, UK
[6]Croydon University Hospital, Croydon, UK
[7]Big Data Institute, University of Oxford, Oxford, UK
[8]NIHR Imperial Biomedical Research Centre, London, UK
[9]Imperial College Healthcare NHS Trust, London, UK
[10]NIHR University College London Hospitals Biomedical Research Centre, London, UK
[11]NIHR Manchester Biomedical Research Centre, Manchester, UK
[12]The Christie NHS Foundation Trust, Manchester, UK
[13]University College London Hospitals NHS Foundation Trust, London, UK
[14]Department of Surgery & Cancer, Imperial College London, London, UK
[15]Division of Informatics, Imaging & Data Sciences, The University of Manchester, Manchester, UK
[16]iCare & Imperial College Healthcare NHS Trust, London, UK

Acknowledgements This work uses data provided by patients and collected by the NHS as part of their care and support. The authors thank the UK Colorectal Cancer Intelligence Hub programme's Bowel Cancer Intelligence UK Patient-Public Group for their support and feedback on this project. This project is conducted using NIHR HIC data resources and supported by NIHR Biomedical Research Centres (BRCs) at Imperial, Marsden, Oxford and Manchester. The authors thank all staff including clinicians, projects managers, governance and contracts teams, informaticians, and data managers at Imperial College Healthcare NHS Trust, The Royal Marsden NHS Foundation Trust, Oxford University Hospitals NHS Foundation Trust, Guys and St Thomas' NHS Foundation Trust, Leeds Teaching Hospitals NHS Trust, The Christie NHS Foundation Trust, University College London Hospitals NHS Foundation Trust and University Hospitals Birmingham NHS Foundation Trust.

Contributors AT, HJ and WP contributed equally as joint first authors. All authors made significant contributions to the conception and design of the work. CC lead the collaborative with the assistance of WP, JD, HJ, GR, SL, KV and KW. CC, WP, HJ, EM, RM and AR defined the data set. AT, DC, RC, JD, AG, LE, AG, BG, SH, KK, SL, LMa, SM, LMe, RN, EJAM, CO, DP, NP, GR, NS, MT, RT, KV, HW and KW made substantial contributions to the acquisition of data. AT, HJ, SH made substantial contributions in analysis of the data. CC is the guarantor.

Funding AT is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The protocol for the collection and management of the data for the NIHR HIC Colorectal Cancer research database has been reviewed and approved by the East Midlands - Derby Research Ethics Committee (REF Number: 21/EM/0028).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
William Perry http://orcid.org/0000-0001-8718-8306
Niels Peek http://orcid.org/0000-0002-6393-9969
Gail Roadknight http://orcid.org/0000-0002-1158-0181

REFERENCES

1 Global Cancer Observatory. Global Cancer Observatory Colorectal Factsheet, 2020. Available: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf [Accessed Sep 2021].
2 Cancer Research UK. Bowel cancer incidence statistics. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence [Accessed Sep 2021].
3 World Health Organization. Global Health Observatory. Geneva: World Health Organization, 2020. Available: https://www.who.int/data/gho/
4 Institute for Health Metrics and Evaluation (IHME). GBD. Seattle, WA: IHME, University of Washington. Available: http://www.healthdata.org [Accessed Jan 2020].
5 NHS Digital. National bowel cancer audit. Available: https://digital.nhs.uk/data-and-information/clinical-audits-and-registries/national-bowel-cancer-audit [Accessed Aug 2021].
6 Downing A, Hall P, Birch R, *et al*. Data resource profile: the colorectal cancer data Repository (CORECT-R). *Int J Epidemiol* 2021;50:1418–1418k.
7 National Institute of Health Research. Health informatics collaborative, 2020. Available: https://hic.nihr.ac.uk/
8 Smith DA, Wang T, Freeman O, *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 Inform* 2020;27:e100145.
9 NHS Digital. Commissioning data sets. Available: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/commissioning-data-sets [Accessed Aug 2021].
10 National Cancer Registration and Analysis Service (NCRAS) datasets. Cancer outcome and services data set (COSD), 2021. Available: http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd
11 National Cancer Registration and Analysis Service (NCRAS) datasets. Systemic anti-cancer therapy dataset (SACT), 2021. Available: http://www.ncin.org.uk/collecting_and_using_data/data_collection/chemotherapy
12 National Cancer Registration and Analysis Service (NCRAS) datasets. National radiotherapy dataset (RTDS), 2021. Available: http://www.ncin.org.uk/collecting_and_using_data/rtds
13 NHS Digital. Spine. Available: https://digital.nhs.uk/services/spine [Accessed Nov 2021].
14 Nelson EK, Piehler B, Eckels J, *et al*. LabKey server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics* 2011;12:71.
15 Microsoft Corporation. Microsoft SQL server 2016, 2016. Available: https://www.microsoft.com/en-us/sql-server/sql-server-2016 [Accessed Mar 2022].
16 Chang W, Cheng J, Allaire J. Shiny: web application framework for R, 2021. Available: https://CRAN.R-project.org/package=shiny [Accessed Oct 2021].
17 R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing, 2021. Available: https://www.R-project.org/
18 RStudio Team. RStudio: integrated development environment for R. RStudio, PBC, 2021. Available: http://www.rstudio.com
19 pyodbc Development Team. pyodbc 4.0.32(v4.0.32), 2021. Available: https://github.com/mkleehammer/pyodbc/ [Accessed Oct 2021].
20 McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th python in science conference*. , 2010: 445, 51–6.
21 The pandas development team. (2020) pandas-dev/pandas: pandas 1.1.3 (v1.1.3). Zenodo, 2021. Available: https://doi.org/10.5281/zenodo.4067057
22 Plotly Technologies Inc. Collaborative data science. Montréal, Qc, 2015. Available: https://plot.ly [Accessed Oct 2021].

23 Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;9:90–5.
24 Caswell TA, Droettboom M, Lee A. 2021 matplotlib/matplotlib: REL: v3.3.2 (v3.3.2). Zenodo, 2020. Available: https://doi.org/10.5281/zenodo.4030140
25 Morris EJA, Finan PJ, Spencer K, *et al*. Wide variation in the use of radiotherapy in the management of surgically treated rectal cancer across the English National health service. *Clin Oncol* 2016;28:522–31.