


Designing an electronic medical record alert to identify hospitalised patients with HIV: successes and challenges

Walid El-Nahal ,¹ Thomas Grader-Beck,¹ Kelly Gebo,¹ Elizabeth Holmes,^{1,2} Kayla Herne,¹ Richard Moore,¹ David Thompson,³ Stephen Berry¹

To cite: El-Nahal W, Grader-Beck T, Gebo K, *et al.* Designing an electronic medical record alert to identify hospitalised patients with HIV: successes and challenges. *BMJ Health Care Inform* 2022;**29**:e100521. doi:10.1136/bmjhci-2021-100521

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2021-100521>).

Received 04 December 2021
Accepted 11 May 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

²San Francisco Department of Public Health, San Francisco, California, USA

³Department of Anesthesiology and Critical Care Medicine, John Hopkins University School of Medicine, Baltimore, Maryland, USA

Correspondence to

Dr Walid El-Nahal;
welnaha1@jh.edu

ABSTRACT

Objectives Electronic medical record (EMR) tools can identify specific populations among hospitalised patients, allowing targeted interventions to improve care quality and safety. We created an EMR alert using readily available data elements to identify hospitalised people with HIV (PWH) to facilitate a quality improvement study intended to address two quality/safety concerns (connecting hospitalised PWH to outpatient HIV care and reducing medication errors). Here, we describe the design and implementation of the alert and analyse its accuracy of identifying PWH.

Methods The EMR alert was designed to trigger for at least one of four criteria: (1) an HIV ICD-10-CM code in a problem list, (2) HIV antiretroviral medication(s) on medication lists, (3) an HIV-1 RNA assay ordered or (4) a positive HIV-antibody result. We used manual chart reviews and an EMR database search to determine the sensitivity and positive predictive value (PPV) of the overall alert and its individual criteria.

Results Over a 24-month period, the alert functioned as intended, notifying an intervention team and a data abstraction team about admissions of PWH. Manual review of 1634 hospitalisations identified 18 PWH hospitalisations, all captured by the alert (sensitivity 100%, 95% CI 82.4% to 100.0%). Over the 24 months, the alert triggered for 1191 hospitalisations. Of these, 1004 were PWH hospitalisations, PPV=84.3% (95% CI 82.2% to 86.4%). Using fewer criteria (eg, using only ICD-10-CM codes) identified fewer PWH but increased PPV.

Conclusion An EMR alert effectively identified hospitalised PWH for a quality improvement intervention. Similar alerts might be adapted as tools to facilitate interventions for other chronic diseases.

INTRODUCTION

In the USA, people with HIV (PWH) are hospitalised at a rate 2–3 times the general population.^{1–4} In the past decade, over 90% of these hospitalisations have been for conditions (non-AIDS-defining conditions) not typically associated with HIV.³ Two leading quality and safety concerns among hospitalised PWH are low rates of engagement in outpatient HIV care and high rates of inpatient antiretroviral medication prescription errors.^{5–10} In 2017, we initiated a trial to evaluate the ability of a

WHAT IS ALREADY KNOWN ON THIS TOPIC

Electronic medical record tools may be able to identify special hospitalised patient populations in real time for quality and safety interventions, research or other purposes.

WHAT THIS STUDY ADDS

We incorporated diagnosis (ICD-10-CM) codes, laboratory results and medication lists into an electronic medical record inbox message alert that accurately identified hospitalised persons with HIV for a quality improvement study.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

When designing an alert, balancing its sensitivity and specificity is a function of the criteria used, with ICD-10-CM codes having the highest utility for identifying persons with HIV. Iterative testing of individual criteria is important to improving the accuracy of an electronic medical record alert.

hospital HIV Support Team (HST) to address both issues for patients admitted to our large, academic, urban hospital. The team consisted of a nurse and an HIV-specialist pharmacist who met hospitalised PWH at the bedside.

To facilitate this work, we developed a novel electronic medical record (EMR) alert to identify PWH among all adult hospital admissions. The alert was based on readily available discrete data elements in the EMR and was designed to notify two groups of people by EMR message: (1) our HST ('intervention team') in real time and (2) a data abstraction team to collect data from charts captured by the alert. In this report, we describe the design of our EMR alert, explore challenges in its implementation and analyse the sensitivity and positive predictive value (PPV) of the alert's criteria for identifying admitted PWH. We plan to report results of the HST's effects on HIV care engagement and medication errors in future manuscripts.

METHODS

The trial evaluating the HST's effects was designed as a phased, cluster-randomised ('step-wedge') trial with each of six randomly determined clusters of nine hospital admitting services becoming successively included in the intervention group over contiguous 4-month periods. PWH hospitalised on services not yet included in the intervention group served as controls. We designed the EMR alert to identify all PWH on both intervention and control services. In this manuscript, we describe the EMR alert's function in identifying PWH on all services included in the randomised trial. Further description of the methods and results of the trial itself will be the focus of future manuscripts.

Our hospital uses Epic corporation's Hyperspace software as its EMR in both inpatient and outpatient settings. Our alert, programmed by an Epic physician builder within our health system (TG-B), screened records of adult patients admitted to inpatient status (excluding observation hospitalisations) to determine if they met criteria of PWH. We excluded observation hospitalisations (<48 hours) because these might be too short for the HST to be effective. The alert's output consisted of a message with patient name and medical record number delivered to the EMR's 'In-Basket' system.

The alert was designed to trigger for any one of four criteria, chosen to identify PWH using discreet EMR data elements: (1) an HIV International Classification of Diseases 10th Revision (ICD-10-CM) code (B20, Z21, O98.711–O98.73) in any current or prior outpatient or inpatient problem lists, (2) any antiretroviral therapy (ART) medication(s) which can be used (but are not necessarily specific) for HIV (identified from pharmaceutical subclasses maintained and updated externally by the EMR vendor, online supplemental S1) on the patient's current inpatient or historical outpatient medication lists, (3) an HIV-1 RNA level assay ordered (regardless of result) during the hospitalisation or (4) a positive HIV-antibody (Ab) result during the hospitalisation or at any prior point in time. The third criterion was intended to capture both instances of the clinical team performing virological monitoring of individuals with diagnosed HIV, (in which case the result could be either detectable or undetectable) and instances of diagnosing acute HIV infection during the hospitalisation (in which case the result would be detectable). We did not look at HIV-1 RNA level testing prior to the hospitalisation because we felt this may introduce a high number of false-positives due to prior attempts at diagnosing acute HIV, and because we felt these individuals would be well captured by the fourth criterion.

We initially considered a fifth criterion, a laboratory order for a CD4 cell count during the index admission. In 2 weeks of predeployment testing (40 hospitalisations alerted), this criterion triggered for 3 hospitalisations of HIV-uninfected persons, all of whom had CD4 cell counts ordered to assess immunodeficiency in the setting of cancer chemotherapy. The CD4 criterion did not identify

any PWH who were not identified by one or more other criteria; thus, it was eliminated as an alert trigger for subsequent hospitalisations.

In May 2017, we deployed the alert and began the randomised trial. We reviewed all alert instances in the first month of deployment (approximately 60 charts). We identified a single instance where the alert was activated by the HIV-1 RNA criterion but failed to recognise that the patient also had a positive antibody. In this case, we identified and fixed a coding error that resulted in an unintended upper age limit for the antibody criterion. We then considered our specifications for the HIV alert criteria finalised. The alert build and post go-live support required 74 total hours of physician builder time.

The alert separately notified a data abstraction team tasked with confirming the patient's HIV status (through reviewing chart notes and/or lab results) and the intervention team (HST members), who used the alerts to know which patients to see at the bedside. Rather than manually define individual data abstraction team members and intervention team members, the alert sent messages to separate recipient pools for each of these teams. Individuals could be added and removed from each pool as needed for team member turnover.

To analyse sensitivity, two nurses (EH and KH) conducted manual chart reviews of hospitalisations selected without regard to the EMR alert. A manual review of all adult patients admitted during the 2-year intervention period (approximately 100 000 hospitalisations) was beyond our capacity, so we collected a random sample of charts over a 4-month interval during the midpoint of the intervention period, from admitting services that averaged more than 10 hospitalisations among PWH per year. We aimed to review 1500–2000 charts, approximately 3%–4% of annual hospital volume. The protocol for each review began with reading the admission history and physical note and the most recent progress note looking for HIV (or AIDS) described as an active or historical diagnosis. If there was no indication of HIV (or AIDS) in the clinical notes, the reviewers then examined laboratory, medication and problem list chart sections and finally screened outpatient visits for any visits at the hospital-affiliated HIV clinic.

We also determined the proportion of hospitalisations identified by the full alert that was identified by each alert criterion alone and in two-way combinations. Assuming the full alert would approach 100% sensitivity, these results would then approximate sensitivity estimates for the individual criteria.

For the analysis of PPV, we started with the abstractor team's manual reviews of each alerted hospitalisation indicating whether the patient was, indeed, living with HIV. We then performed a secondary review of all 199 hospitalisations the abstractors initially classified as false-positives (not having HIV infection despite the alert triggering). The secondary review involved detailed examination of current and prior discharge summaries, inpatient progress notes, outpatient office visits, medication lists and a

search of outside records (available through EMR links) including antibody measurements and viral loads. Finally, we performed a retrospective EMR query of the charts that were identified by the alert to analyse which criteria (ICD-10-CM, ART, HIV-1 RNA, antibody or a combination) activated the alert.

Through this post-deployment analysis, we observed that the antibody criterion had never alerted, confounding our expectations. On a targeted review of a sample of 10 charts, 4 of which were known to have a positive antibody result within our EMR, we determined that the alert, as coded, was not successfully capturing antibody tests. Our 2017 audits did not include charts in which the only positive criterion was the antibody, and thus we failed to identify this issue prior to or during the intervention period. The error appears to have originated from failure of a function within the EMR to translate textual results of antibody tests into discrete normal/abnormal data.

To evaluate the potential impact of the antibody criterion failure, we determined from our retrospective EMR query that only 95 of 110 028 (0.09%) adult hospital admissions during the study period (29 unique patients) were associated with a positive HIV antibody and no other criteria. Thus, we considered the percentage of charts missed by the failure of this criterion to be negligible. We limited our analysis of individual criteria to the information available for the three functioning criteria: ICD code, HIV-1 RNA or prescription of any ART. We performed analyses using Stata V.16.1 software (StataCorp)¹¹ with an α value of 0.05 for significance and CI calculations of sensitivity and PPV.

RESULTS

Between May 2017 and May 2019, the EMR alert met criteria for 1191 hospitalisations among 849 unique patients. The majority of identified patients were male (63.0%), and black (72.4%) with a median age of 53.2 (IQR 41.6–60.6) years (table 1). During the 24-month intervention period, 671 (79.0%) patients were hospitalised once, 107 (12.6%) were hospitalised twice, 32 (3.8%) three times and 39 (4.6%) four or more times.

Our random sample to assess sensitivity comprised 1634 hospitalisations (approximately 3% of typical annual hospital volume). Among these hospitalisations, we identified 18 PWH admitted to inpatient status (with a total of 18 hospitalisations over the review period). The three-criteria-based alert identified all 18, yielding a sensitivity of 100% (95% CI 82.4% to 100%).

Using all three criteria, the alert was activated in 1191 instances, of which 1004 were true-positives, PPV=84.3% (95% CI 82.2% to 86.4%) (table 2). Using only two criteria (ICD code and ART) identified 988 (98.4%) of the 1004 true-positives, with a higher PPV of 94.2%. Using only ICD codes identified 947 (94.3%) of all true-positives, and further increased PPV to 99.1%. Results for other individual criteria and combinations are shown in table 2. The HIV-1 RNA criterion had the most

Table 1 Demographic characteristics of patients identified by EMR alert

Patient characteristics	
Total unique patients identified	849 patients (1191 alerts)
Sex at birth	535 (63.0%) male
Median age (IQR), mean	53.2 (41.6–60.6), 50.9 years
Race	
Black	615 (72.4%)
White	176 (20.7%)
Other	58 (6.9%)
Ethnicity	
Hispanic	31 (3.7%)
Non-Hispanic	805 (94.8%)
Unknown	13 (1.5%)
Alerts per person	
1	671 (79.0%)
2	107 (12.6%)
3	32 (3.8%)
4+	39 (4.6%)
EMR, electronic medical record.	

false-positives, 141 out of 1078 (13.1%) alerts, due to this assay being used to evaluate for acute HIV infection and resulting negative. The 59 false-positive (6.1% of 960 total) ART alerts were for instances of individuals taking antiretrovirals for HIV pre-exposure prophylaxis (n=46), HIV post-exposure prophylaxis (n=4), treatment of Hepatitis B (n=3, three medications are approved by the U.S. Food and Drug Administration for both viral infections), or a mixture of indications (n=6) including experimental colorectal cancer treatment¹¹ or research protocols for HIV pre-exposure prophylaxis. The nine false-positive ICD code instances were errors in entering HIV into the problem list, typically for individuals undergoing testing for HIV or receiving pre-exposure prophylaxis. The ICD-10-CM codes in these cases were B20 ‘HIV (HIV) disease’ (n=8) and Z21 ‘Asymptomatic HIV infection’ (n=1). Cases we evaluated in postdeployment analysis (approximately 2 years after the trial concluded) had already been corrected in the EMR by removal of HIV from the problem list.

DISCUSSION

This study has several important findings. First, using a combination of readily available discrete data (ICD-coded problem lists, medication prescriptions and disease-specific laboratory test orders), the EMR alert achieved both a high sensitivity and PPV, correctly identifying most admitted PWH. Second, our results demonstrate the impact of the number and choice of criteria on the balance between PPV and sensitivity of an EMR alert. Finally, our experience offers several practical lessons

Table 2 Rates of true and false-positive detection of HIV by EMR alert and individual criteria

Alert components	Total alerts	True-positive	False-positive	Proportion of the three-criteria alert true-positives	PPV (true-positives/total alerts) (95% CI)
ICD or ART or HIV-1 RNA†	1191	1004	187	100% (reference)**	84.3 (82.2 to 86.4) %
ICD or ART	1049	988	61	98.4 (97.6, 99.2) %	94.2 (92.8 to 95.6) %
ICD or HIV-1 RNA	1137	991	146	98.7 (98.0, 99.4) %	87.2 (85.2 to 89.1) %
HIV-1 RNA or ART	1182	996	186	99.2 (98.7, 99.8) %	84.3 (82.2 to 86.3) %
ICD only alert	956	947	9	94.3 (92.9, 95.8) %	99.1 (98.4 to 99.7) %
ART only alert	960	901	59	89.7 (87.9, 91.6) %	93.9 (92.3 to 95.4) %
HIV-1 RNA only alert	1078	937	141	93.3 (91.8, 94.9) %	86.9 (84.9 to 88.9) %

*The three-criteria alert is used as the standard against which combinations of criteria are measured, and the sensitivity of the three-criteria alert approached 100% (95% CI 82.4% to 100%) based on the manual chart review sample.

†ICD, International Classification of Diseases, 10th Revision.

ART, antiretroviral therapy; EMR, electronic medical record; PPV, positive predictive value.

from successes and limitations of the implementation of our alert.

We demonstrated a high sensitivity and PPV of EMR alerts to accurately determine persons with HIV in order to facilitate a prospective evaluation of a quality improvement intervention. The same alert (potentially with a repaired antibody criterion) could be used to facilitate additional quality improvement interventions for hospitalised PWH. Indeed, there is a growing body of evidence that EMR data can be used to improve diagnosis, linkage and engagement in care for PWH.^{12–15} This analysis adds to that literature by determining sensitivity and PPV of an EMR alert and its individual elements. Because the EMR data elements should be similar if not identical in versions of the same EMR software used at different health systems, this alert might be easily reproduced and used by others for quality interventions among PWH. Finally, we propose that similar alerts might be used for quality and safety interventions in other chronic diseases. We are currently evaluating the performance of an alert using ICD code, medication and laboratory testing data to identify persons with chronic Hepatitis C virus (HCV) infection at the time of hospital admission. Such an alert could be used to target efforts to engage these individuals in outpatient care and to initiate curative HCV therapy. We suspect alerts based on similar discrete data elements might also be applied to non-infectious chronic diseases such as diabetes, inflammatory bowel disease and rheumatological disorders in order to deploy multidisciplinary teams for quality interventions including medication reviews, patient education, targeted case management and/or outpatient linkage to follow-up.

Our findings also highlight the importance of criteria selection. Specifically, as a single criterion, ICD coded problem lists captured the most patients with the highest PPV compared with either HIV-1 RNA or ART. For individuals with previously diagnosed HIV and prior contact within our health system, it is not surprising that ICD

codes entered into problem lists by providers would be highly accurate. Conversely, the CD4 criteria introduced false-positives without adding sensitivity beyond the other criteria, so it was eliminated during early revisions of the alert. The HIV-1 RNA criterion also introduced a notable number of false-positives; however, we considered it important to identifying newly diagnosed PWH who were in-need of initial linkage to care. Our results demonstrate that adding criteria identifies more patients but increases false-positives. The right balance between sensitivity and PPV may vary for different diseases and intended purposes of identifying patients.

Related to creation and implementation of the alert, several design aspects of the alert are notable. The use of EMR medication classes for the ART criteria minimised upkeep, as these classes are updated independently by the software manufacturer, obviating the need to manually update the alert when new medications were brought to market. A disadvantage of their use is the lag between marketing approval for new drugs and updating the classes. However, based on our observation with one recently approved antiretroviral (fostemsavir), we suspect this duration is typically a matter of days or weeks. The use of EMR recipient pools facilitated conducting an interventional trial by having separate data abstractor and intervention team pools and simplified EMR programming when study personnel transitioned.

Our experience with the alert also offered several valuable lessons related to the iterative nature of quality improvement. The original alert included an HIV-antibody criterion, designed to capture newly diagnosed PWH or PWH who otherwise may not ever have engaged in outpatient HIV care such that any provider added an HIV diagnosis to their problem list or prescribed ART. Early iterations of the alert were repeatedly tested, with manual chart reviews of identified patients. Despite this, it was not until later review of a larger sample of charts that we learnt that the HIV-antibody criterion was not

functioning as intended, demonstrating the importance of frequent testing and possibly of targeted exploration of each individual criterion and various criteria in combination. Overall, however, even missing one criterion, the alert in this case remained effective at identifying our target patient population.

The main limitation of the assessment of the sensitivity of the alert was the sample size, as this required a manual chart review. While over 1600 charts were reviewed, amounting to approximately 3% of annual inpatient volume, only 18 patients were PWH that could contribute to a sensitivity calculation. Our study was performed at a large, academic, urban medical centre with a high prevalence of HIV and may not generalise to other care systems.

Conclusion

In summary, EMR alerts have significant potential as tools to identify PWH when hospitalised. The use of such alerts can facilitate the deployment of multidisciplinary inpatient teams for medication review, education, targeted case management and outpatient linkage to follow-up. Our approach using readily available discrete data elements could potentially be applied to other chronic illnesses to facilitate quality and safety interventions. The selection of criteria, however, plays an important role in the functioning of such an alert, and dictates its sensitivity and PPV, which should factor heavily into design. Lastly, as with any quality improvement project, iterative revision and regular monitoring of the intervention itself are clearly important.

Contributors TG-B built the electronic medical record alert. EH and KH reviewed charts. WE-N completed the data analysis, conducted a secondary review of charts, and prepared the initial manuscript draft. SB conceived the project, supervised the findings, and is responsible for the overall content as guarantor of this work.

Funding This work was supported by a grant from the Agency for Healthcare Research and Quality, R01HS024079, PI, SB. WE-N is supported by the National Institutes of Health T32 AI007291.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The Johns Hopkins School of Medicine Institutional Review Board determined the development of the EMR alert (as part of the study of the HST) to be not human subjects/quality improvement research (IRB00100665).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All de-identifiable data are available upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and

responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Walid El-Nahal <http://orcid.org/0000-0001-6834-0405>

REFERENCES

- 1 Lazar R, Kersanske L, Xia Q, *et al*. Hospitalization rates among people with HIV/AIDS in New York City, 2013. *Clin Infect Dis* 2017;65:469–76.
- 2 Bachhuber MA, Southern WN. Hospitalization rates of people living with HIV in the United States, 2009. *Public Health Rep* 2014;129:178–86.
- 3 Davy-Mendez T, Napravnik S, Hogan BC, *et al*. Hospitalization rates and causes among persons with HIV in the United States and Canada, 2005–2015. *J Infect Dis* 2021;223:2113–23.
- 4 Sun R, Zeynal K, Wong HS. Trends in Hospital Inpatient Stays by Age and Payer, 2000–2015: Statistical Brief #235. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, 2018. Available: <https://pubmed.ncbi.nlm.nih.gov/29578670/> [Accessed April 12, 2021].
- 5 Commers T, Swindells S, Sayles H, *et al*. Antiretroviral medication prescribing errors are common with hospitalization of HIV-infected patients. *J Antimicrob Chemother* 2014;69:262–7.
- 6 Eginger KH, Yarborough LL, Inge LD, *et al*. Medication errors in HIV-infected hospitalized patients: a pharmacist's impact. *Ann Pharmacother* 2013;47:953–60.
- 7 Daniels LM, Raasch RH, Corbett AH. Implementation of targeted interventions to decrease antiretroviral-related errors in hospitalized patients. *Am J Health Syst Pharm* 2012;69:422–30.
- 8 Yehia BR, Mehta JM, Ciuffetelli D, *et al*. Antiretroviral medication errors remain high but are quickly corrected among hospitalized HIV-infected adults. *Clin Infect Dis* 2012;55:593–9.
- 9 Nijhawan AE, Bhattatiry M, Chansard M, *et al*. HIV care cascade before and after hospitalization: impact of a multidisciplinary inpatient team in the US South. *AIDS Care* 2020;32:1343–52.
- 10 Metsch LR, Bell C, Pereyra M, *et al*. Hospitalized HIV-infected patients in the era of highly active antiretroviral therapy. *Am J Public Health* 2009;99:1045–9.
- 11 Parikh AR, Rajurkar M, Van Seventer EE, *et al*. Phase II study of lamivudine in p53 mutant metastatic colorectal cancer (mCRC). *JCO* 2020;38:149 https://doi.org/10.1200/JCO2020384_suppl149
- 12 Ruffner AH, Ancona RM, Hamilton C, *et al*. Identifying ED patients with previous abnormal HIV or hepatitis C test results who may require additional services. *Am J Emerg Med* 2020;38:1831–3.
- 13 Ridgway JP, Lee A, Devlin S, *et al*. Machine learning and clinical informatics for improving HIV care continuum outcomes. *Curr HIV/AIDS Rep* 2021;18:229–36.
- 14 Ridgway JP, Almirol E, Schmitt J, *et al*. A clinical informatics approach to Reengagement in HIV care in the emergency department. *J Public Health Manag Pract* 2019;25:270–3.
- 15 Shade SB, Steward WT, Koester KA, *et al*. Health information technology interventions enhance care completion, engagement in HIV care and treatment, and viral suppression among HIV-infected patients in publicly funded settings. *J Am Med Inform Assoc* 2015;22:e104–11.

© 2022 Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Operationalising fairness in medical algorithms

Sonali Parbhoo,¹ Judy Wawira Gichoya,² Leo Anthony Celi ,^{3,4}
Miguel Ángel Armengol de la Hoz ,⁵ for MIT Critical Data

To cite: Parbhoo S, Wawira Gichoya J, Celi LA, *et al*. Operationalising fairness in medical algorithms. *BMJ Health Care Inform* 2022;**29**:e100617. doi:10.1136/bmjhci-2022-100617

Received 20 May 2022
Accepted 24 May 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Harvard Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA

²Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, Georgia, USA

³Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA

⁴Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

⁵Big Data Department, Regional Ministry of Health of Southern Spain, Sevilla, Spain

Correspondence to
Dr Leo Anthony Celi;
LCeli@mit.edu

The world is abuzz with applications of machine learning and data science in almost every field: commerce, transportation, banking, and more recently, health-care. Breakthroughs in these areas are a result of newly created algorithms, improved computing power and, most importantly, the availability of bigger and increasingly reliable data with which to train these algorithms. For healthcare specifically, machine learning is at the juncture of moving from the pages of conference proceedings to clinical implementation at the bedside. Yet, succeeding in this endeavour requires synthesising insights from both the algorithmic perspective as well as the healthcare domain to ensure that the unique characteristics of machine learning methods can be leveraged to maximise benefits and minimise risks.

While progress has recently been made in establishing certain guidelines or best practices for the development of machine learning models for healthcare as well as protocols for the regulation of such models, these guidelines and protocols tend to overlook important considerations such as fairness, bias and unintended disparate impact.^{1,2} Nevertheless, it is widely recognised in other domains that many of the machine learning models and tools may have discriminatory effect by inadvertently encoding and perpetuating societal biases.³

In this special issue, we highlight that machine learning algorithms should not be focused solely on accuracy but should be evaluated with respect to how they might impact disparities in patient outcomes. Our special issue aims to bring together the growing community of healthcare practitioners, social scientists, policymakers, engineers and computer scientists to design and discuss practical solutions to address algorithmic fairness and accountability. We invited papers that explore ways to reduce machine learning bias in healthcare or explain how

to create algorithms that specifically alleviate inequalities.

To prevent artificial intelligence (AI) from encoding the disparities that exist, algorithms should predict an outcome as if the world were fair. If designed well, AI may even provide a way to audit and improve the way care is being delivered across populations. There is growing community momentum towards not just detecting bias but operationalising fairness, but this is a monumental task. Some of the encouraging developments that we have seen have been incorporating patients' voices in AI. Patient engagement is crucial if algorithms are to truly benefit everyone.

The papers in this special issue cover a variety of topics that addressed the objectives laid out in the call, these were:

- ▶ Identifying Undercompensated Groups Defined by Multiple Attributes in Risk Adjustment⁴
- ▶ A Proposal for Developing a Platform That Evaluates Algorithmic Equity and Accuracy⁵
- ▶ Can medical algorithms be fair? Three ethical quandaries and one dilemma⁶
- ▶ Resampling to Address Inequities in Predictive Modeling of Suicide Deaths⁷
- ▶ Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation⁸
- ▶ Operationalizing fairness in medical AI adoption: Detection of early Alzheimer's Disease with 2D CNN⁹
- ▶ Global disparity bias in ophthalmology artificial intelligence applications¹⁰
- ▶ Investigating for bias in healthcare algorithms: A sex stratified analysis of supervised machine learning models in liver disease prediction¹¹

It has been more than 5 years since the ProPublica investigative report on machine bias was published. The report detailed how a software used in judicial courts across the



USA to inform decisions around parole was prejudiced against black people. Everything we have achieved since then has always been geared towards understanding how difficult it is to prevent AI from perpetuating societal biases in algorithms.

There is a long road ahead before we can leverage the zettabytes of data that are routinely collected in the process of care. We should not only invest in storage and compute technologies, federated learning platforms, GPTs, GRUs and NTFs. Machine learning in healthcare is not just about predicting something for the sake of prediction. The most important task is to augment our capacity to make decisions, and that requires understanding how those decisions are made.

Twitter Judy Wawira Gichoya @judywawira, Leo Anthony Celi @MITCriticalData and Miguel Ángel Armengol de la Hoz @miguearmengol

Contributors Initial conceptions and design—SP, JWG, LAC and MAAAdIH. Drafting of the paper—SP, JWG, LAC and MAAAdIH. Critical revision of the paper for important intellectual content—SP, JWG, LAC and MAAAdIH.

Funding LAC is funded by the National Institute of Health through the NIBIB R01 EB017205.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; internally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Leo Anthony Celi <http://orcid.org/0000-0001-6712-6626>

Miguel Ángel Armengol de la Hoz <http://orcid.org/0000-0002-7012-2973>

REFERENCES

- 1 Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289.
- 2 McCoy LG, Banja JD, Ghassemi M, *et al*. Ensuring machine learning for healthcare works for all. *BMJ Health Care Inform* 2020;27:e100237.
- 3 Sarkar R, Martin C, Mattie H, *et al*. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. *Lancet Digit Health* 2021;3:e241–9.
- 4 Zink A, Rose S. Identifying undercompensated groups defined by multiple attributes in risk adjustment. *BMJ Health Care Inform* 2021;28:e100414.
- 5 Cerrato P, Halamka J, Pencina M. A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health Care Inform* 2022;29:e100423.
- 6 Bærøe K, Gundersen T, Henden E, *et al*. Can medical algorithms be fair? three ethical quandaries and one dilemma. *BMJ Health Care Inform* 2022;29:e100445.
- 7 Reeves M, Bhat HS, Goldman-Mellor S. Resampling to address inequities in predictive modeling of suicide deaths. *BMJ Health Care Inform* 2022;29:e100456.
- 8 Foryciarz A, Pfohl SR, Patel B, *et al*. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health Care Inform* 2022;29:e100460.
- 9 Heising L, Angelopoulos S. Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN. *BMJ Health Care Inform* 2022;29.
- 10 Nakayama LF, Kras A, Ribeiro LZ, *et al*. Global disparity bias in ophthalmology artificial intelligence applications. *BMJ Health Care Inform* 2022;29:e100470.
- 11 Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform* 2022;29.

© 2022 Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

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



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

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 12903 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.

INTRODUCTION

Globally, 1.93 million people were diagnosed with colorectal cancer in 2020.¹ Further,

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.

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 Observatory³ 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 Hub⁶ (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 health-care 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 Sets⁹ and the National Cancer Registration and Analysis Service data sets including Cancer Outcomes and Services Data Set,¹⁰ Systemic Anti-Cancer Therapy Data Set¹¹ and National Radiotherapy Data Set¹² 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 Spine¹³ 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 LabKey¹⁴ 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 Server¹⁵ 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 package¹⁷ 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 12903 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

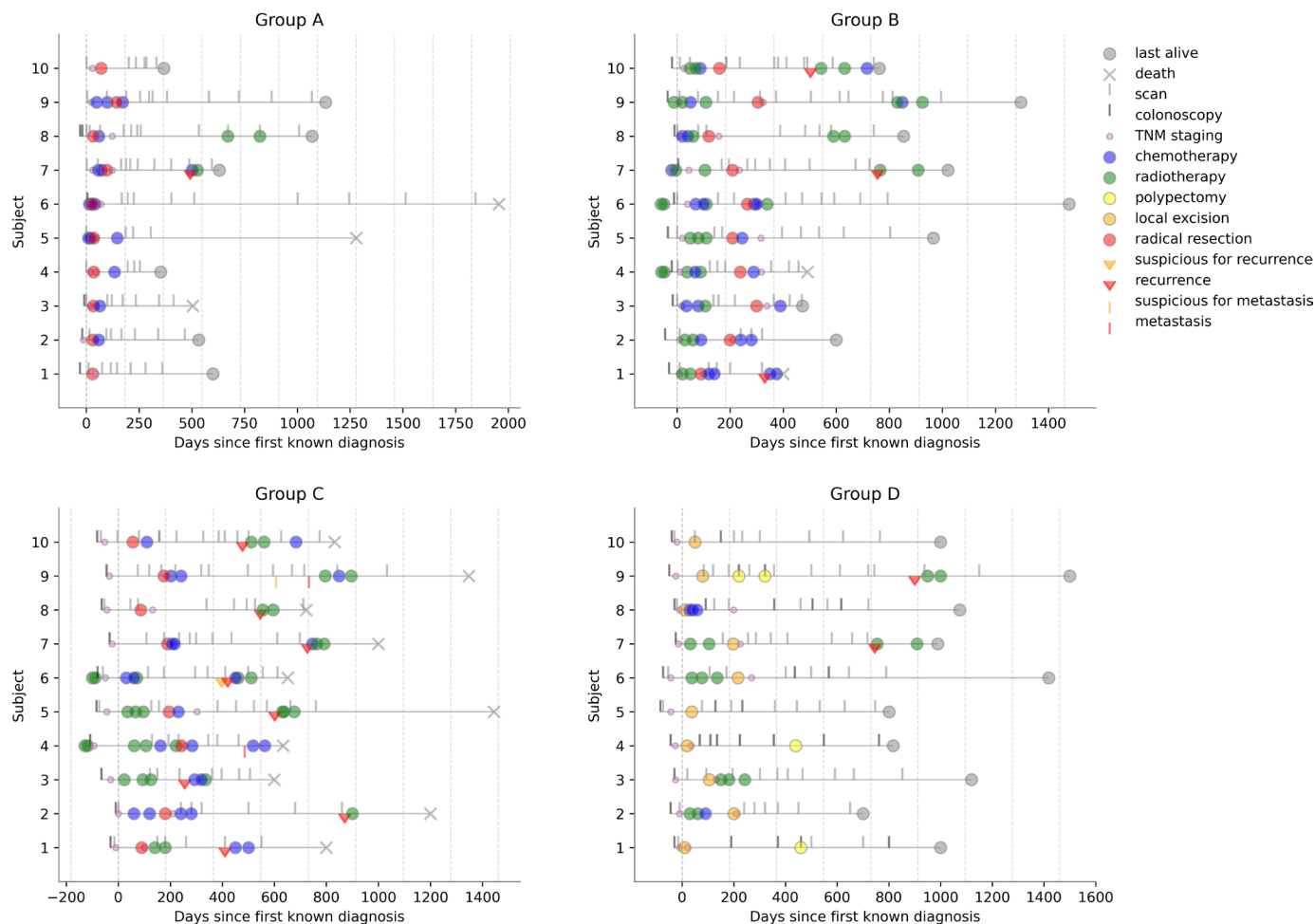


Figure 1 Hypothetical patient timelines that show specific treatment and surveillance patterns. Group A: Timelines of patients with colon cancer that follow the pattern ‘diagnosis, scan, surgery, scan’. Group B: Patients with rectal cancer with ‘diagnosis, scan, chemoradiotherapy, radical resection, chemo(radio)therapy, scan’. Group C: Patients 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 from 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.

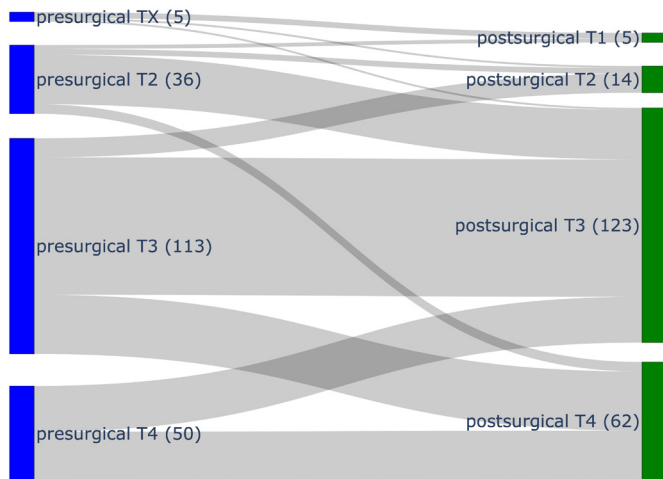


Figure 2 Presurgery and postsurgery T staging for patients with colon cancer (C18) who had a major resection, determined by natural language processing (NLP) of imaging reports (presurgery) and histopathology reports (postsurgery). Number of patients is given in brackets.

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.

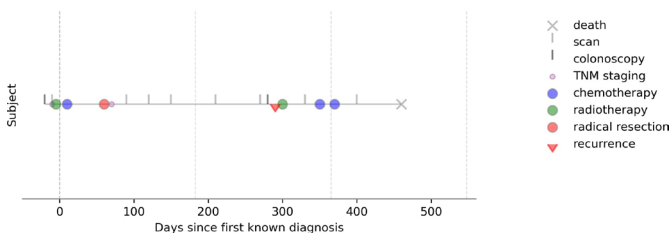


Figure 3 Longitudinal pathway plot of a hypothetical patient with rectal cancer treated with neoadjuvant therapy then radical resection. After a colonoscopy and around the time of diagnosis the patient had neoadjuvant radiotherapy and chemotherapy as identified by the green and blue circles. They then 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.

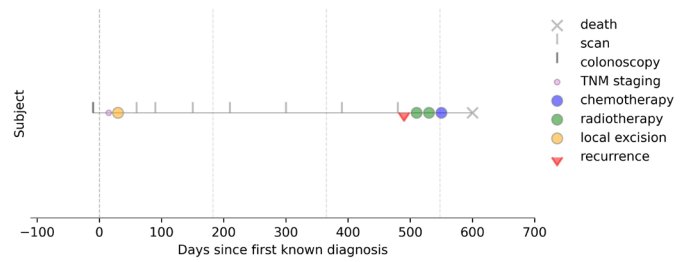


Figure 4 Longitudinal pathway plot of a hypothetical patient with rectal cancer who underwent local excision. Rectal cancer was picked up on colonoscopy as indicated by the dark grey line, and treated by local excision as indicated by the orange circle. After a disease-free surveillance period of approximately 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

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

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

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

- 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].
- 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].
- World Health Organization. Global Health Observatory. Geneva: World Health Organization, 2020. Available: <https://www.who.int/data/gho/>
- Institute for Health Metrics and Evaluation (IHME). GBD. Seattle, WA: IHME, University of Washington. Available: <http://www.healthdata.org> [Accessed Jan 2020].
- 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].
- 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.
- National Institute of Health Research. Health informatics collaborative, 2020. Available: <https://hic.nihr.ac.uk/>
- 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.
- 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].
- 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
- 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
- National Cancer Registration and Analysis Service (NCRAS) datasets. National radiotherapy dataset (RTDS), 2021. Available: http://www.ncin.org.uk/collecting_and_using_data/rtds
- NHS Digital. Spine. Available: <https://digital.nhs.uk/services/spine> [Accessed Nov 2021].
- 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.
- Microsoft Corporation. Microsoft SQL server 2016, 2016. Available: <https://www.microsoft.com/en-us/sql-server/sql-server-2016> [Accessed Mar 2022].
- Chang W, Cheng J, Allaire J. Shiny: web application framework for R, 2021. Available: <https://CRAN.R-project.org/package=shiny> [Accessed Oct 2021].
- R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing, 2021. Available: <https://www.R-project.org/>
- RStudio Team. RStudio: integrated development environment for R. RStudio, PBC, 2021. Available: <http://www.rstudio.com>
- pyodbc Development Team. pyodbc 4.0.32(v4.0.32), 2021. Available: <https://github.com/mkleehammer/pyodbc/> [Accessed Oct 2021].
- McKinney W. Data structures for statistical computing in python. In: *Proceedings of the 9th python in science conference*. , 2010: 445, 51–6.
- 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>
- 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.

© 2022 Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. <http://creativecommons.org/licenses/by-nc/4.0/>This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.