BMJ Open Quality

Accuracy of medical billing data against the electronic health record in the measurement of colorectal cancer screening rates

Vivek A Rudrapatna ,[1,2] Benjamin S Glicksberg ,[1,3,4] Patrick Avila,[2] Emily Harding-Theobald,[5,6] Connie Wang ,[2,6] Atul J Butte [1,7,8]

To cite: Rudrapatna VA, Glicksberg BS, Avila P, *et al*. Accuracy of medical billing data against the electronic health record in the measurement of colorectal cancer screening rates. *BMJ Open Quality* 2020;**9**:e000856. doi:10.1136/bmjoq-2019-000856

VAR and BSG contributed equally.

the Amercan College of Gastroenterology Annual Meeting (2018), and Digestive Diseases Week (2019)

Received 13 October 2019
Revised 21 February 2020
Accepted 5 March 2020

For numbered affiliations see end of article.

Correspondence to
Professor Atul J Butte;
atul.butte@ucsf.edu

ABSTRACT

Objective Medical billing data are an attractive source of secondary analysis because of their ease of use and potential to answer population-health questions with statistical power. Although these datasets have known susceptibilities to biases, the degree to which they can distort the assessment of quality measures such as colorectal cancer screening rates are not widely appreciated, nor are their causes and possible solutions.

Methods Using a billing code database derived from our institution's electronic health records, we estimated the colorectal cancer screening rate of average-risk patients aged 50–74 years seen in primary care or gastroenterology clinic in 2016–2017. 200 records (150 unscreened, 50 screened) were sampled to quantify the accuracy against manual review.

Results Out of 4611 patients, an analysis of billing data suggested a 61% screening rate, an estimate that matches the estimate by the Centers for Disease Control. Manual review revealed a positive predictive value of 96% (86%–100%), negative predictive value of 21% (15%–29%) and a corrected screening rate of 85% (81%–90%). Most false negatives occurred due to examinations performed outside the scope of the database—both within and outside of our institution—but 21% of false negatives fell within the database's scope. False positives occurred due to incomplete examinations and inadequate bowel preparation. Reasons for screening failure include ordered but incomplete examinations (48%), lack of or incorrect documentation by primary care (29%) including incorrect screening intervals (13%) and patients declining screening (13%).

Conclusions Billing databases are prone to substantial bias that may go undetected even in the presence of confirmatory external estimates. Caution is recommended when performing population-level inference from these data. We propose several solutions to improve the use of these data for the assessment of healthcare quality.

INTRODUCTION

Colorectal cancer (CRC) screening is a high priority for public health in the USA and abroad. Although CRC remains the second leading cause of cancer-related death in the USA,[1] screening via modalities such as colonoscopy have the potential to reduce the mortality rate by 60% or more.[2] Despite its potential for such impact, screening uptake as estimated by the Centers for Disease Control (CDC) has remained stagnant at 60% for at least a decade.[3 4] These findings have prompted multiple calls for action such as the 80% by 2018 campaign led by the National Colorectal Cancer Roundtable.

The traditional benchmark for measuring CRC screening rates in the USA has been the National Health Interview Survey—an annual survey of the civilian and non-institutionalised population. Although these data are considered the gold standard, they suffer from a number of shortcomings including low participation rates (55%),[4] recall bias, lack of confirmation with the medical record, uneven health literacy and social desirability bias.[5]

An alternative source that avoids many of the aforementioned pitfalls is administrative healthcare data (see online supplementary table 1 for definition of all healthcare data-related terms as used in this article, adapted with permission from Rudrapatna and Butte[6]). Although these data were originally collected to support operations and financial objectives, they could potentially be useful for many other purposes: tracking the effectiveness of screening outreach measures, providing clinical decision support and rewarding providers and health systems for value-based care.[7]

However, precisely because these data were originally assembled for other reasons, they are prone to measurement bias.[8] More concerning, many large structured datasets such as those derived from medical claims can be difficult to validate, in part due to disconnection from the underlying medical context. Therefore, even though transparent and repeated benchmarking is a critical step for any valid data repurposing endeavour, this is rarely done.

Although it can be difficult to benchmark the accuracy of claims data from payor databases, billing data derived from the electronic health records (EHR) may represent a good proxy for two reasons: 1) much of claims data are derived from bills generated by EHR software over the course of clinical operations and 2) algorithms based on these data may be validated against the full clinical context captured in the EHR.

In this analytical study, we attempt to answer the question: how accurately do medical billing data capture the CRC screening rates within a healthcare system? Here, we perform an informatics-based estimation of the period prevalent screening rate using billing data derived from the EHR. We then review a random sample of charts in order to identify the reasons for algorithmic misclassification and missed screening. We conclude by proposing strategies to enhance future clinical informatics efforts and improve the primary prevention of CRC.

METHODS

Clinical data

EHR data were extracted from the University of California, San Francisco (UCSF) *Epic* system using Clarity and Caboodle tools.[9] To perform analysis on a dataset closely resembling typical payor claims databases in terms of constituent elements, we extracted the following structured fields: age, gender, 'alive' status, race, primary language, ethnicity, insurance, department, diagnosis code, procedure code, and encounter date.

Prior to being used for this study, the data was de-identified to comply with the US Department of Health and Human Services 'Safe Harbor' guidance. Temporal imprecision was introduced into the dataset via a random negative date offset (0–364 days).

Study population

We included patients aged 50–74 years who had at least two primary care visits, two gastroenterology clinic visits or one of each between January 2016 and December 2017 (figure 1). This criteria was used in order to exclude patients who had sought care for an isolated 'sick visit', and specifically identify patients with a clearly established primary care or gastroenterology relationship. These patients would be expected to be considered for colon cancer screening during the office visit. We included patients with only gastroenterology visits because may patients receive regular gastroenterology care at our institution, and these gastroenterologists counsel and refer many patients for CRC screening. Most of the patients in our cohort were included on the basis of being empanelled in primary care rather than gastroenterology care.

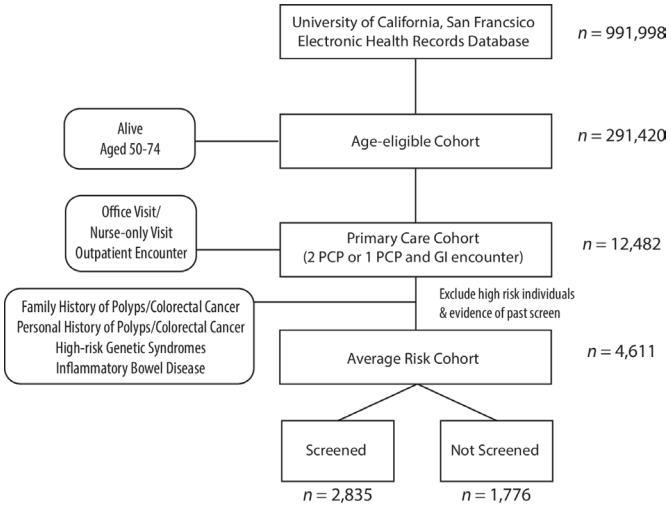

Figure 1 Cohort selection schematic.

We excluded charts bearing the following International Classification of Disease, Tenth Revision, Clinical Modification (ICD-10-CM) codes reflecting an elevated risk of CRC: family history of colon polyps (Z83.71, Z83.79), family history colon cancer (Z80.0, Z80.9), personal history of colon polyps (Z86.010, K63.5, D12, K63.5), personal history of colon cancer (C18-C21), hereditary non-polyposis CRC/Lynch syndrome (Z15.09 Z14.8, Z80.0, Z84.81), familial adenomatous polyposis (D12.6, Z14.8), juvenile polyposis syndrome (D12.6), Peutz-Jehgers syndrome (Q85.8, L81) and inflammatory bowel disease (K50, K51). We reviewed records corresponding to code Z98.89; patients who were annotated as having either a history of prior lower endoscopy or colectomy and lacked an order for a screening examination were excluded.

Classification algorithm

We identified charts with a prior history of lower endoscopy using Current Procedural Terminology (CPT) codes (see online supplementary methods). Additionally, we used regular expression-based string matching to identify billed-for procedures corresponding to colonography-protocoled CT (CT colonography), double contrast barium enema and faecal immunochemical test. Capsule colonoscopies, guaiac-based stool testing and faecal DNA tests are not performed at our facility.

We used the following schedule to determine the presence or absence of a qualifying screening examination: colonoscopy within the prior 10 years, sigmoidoscopy within the prior 5 years, faecal immunochemical test (FIT) in 2016, CT colonography within the last 5 years, double contrast barium enema within the last 5 years. Patients were classified as screened if they had been screened according to this schedule as of March 2018.

Database querying and analysis

All queries required several rounds of iterative refinement done in close collaboration between the clinical and bioinformatics teams. Identification and verification of CPT codes were performed in close consultation with gastroenterology billing specialists. ICD-10 codes were selected by manual review. Encounter names corresponding to primary care visits were identified by discussion with primary care physicians. Data extraction was performed using *MySQL* (V.5.6.10). Further refinement

and analysis was performed in the *R* programming environment[10] (V.3.4.1) using the *RMySQL*[11] and *data.table*[12] packages. Agresti-Coull binomial CIs[13] were calculated for all estimates derived from random samples. Coverage probabilities of the 95% CI for prevalent screening rates were confirmed via Monte-Carlo simulation using 10 000 replicates.

Manual chart review

After Institutional Review Board approval to proceed, we performed a stratified random sample of charts (50 classified positive, 150 classified negative). This ratio of charts was intentionally weighted towards negative charts because we anticipated a higher false-negative classification rate; as such, reviewing more negative charts was anticipated to be more informative. Two hundred charts were selected in total in order to achieve a reasonable balance of statistical precision with the effort required for chart review. A formal power calculation was not performed as this study was intended as an estimation study rather than one intending to test prespecified statistical hypotheses.

Chart annotation criteria were serially developed and agreed on by all reviewers after each completing a test set of 10 charts independent of the above set. Charts were annotated by the reasons for screening or the lack thereof where appropriate (see online supplementary methods). Clinician documentation of a history of prior screening outside the institution was counted as evidence of screening. Charts were each independently reviewed and annotated by one internist and one gastroenterologist each, with all disagreements discussed and resolved. In scenarios where screening appeared to have not been performed due to a misunderstanding of the proper screening or surveillance interval, direct communication was made with the primary care provider.

Patient and public involvement statement

There were no funds or time allocated for patient and public involvement (PPI) for this retrospective chart review, so we were unable to involve patients. However, this study was approved by a review board that includes PPI.

RESULTS

The population of patients aged 50–74 years with EHR data at our institution consisted of 291 420 patients, nearly a third of the total database population (figure 1 and table 1). Within this cohort, we identified a subcohort of 4611 average risk patients empanelled in the primary care or gastroenterology clinics in the 2016–2017 period. Ninety-nine per cent of these patients met the inclusion criteria on the basis of primary care visits within the study period. Nearly 60% of the cohort was female with an average age of 62 years. The racial makeup of this cohort was 42% white, 28% Asian and 11% black. Eighty-eight per cent were of non-Hispanic or Latino ethnicity, and 87% declared a primary language of English. Insurance coverage was collected but 95% of this information was missing.

Table 1 Demographics of primary care cohort at average risk for colorectal cancer at the University of California, San Francisco

	Average risk
	Primary care cohort
N (%)	4611
Age (years; mean±SD)	62±7
Sex N (%)	
Male	1877 (41)
Female	2734 (59)
Ethnicity N (%)	
Hispanic or Latino	415 (9)
Non-Hispanic or Latino	4076 (88)
Race N (%)	
Asian	1278 (28)
Black or African-American	524 (11)
White or Caucasian	1949 (42)
Preferred language N (%)	
English	4015 (87)
Spanish	70 (2)
Chinese—Cantonese	176 (4)
Russian	19 (0)
Chinese—Mandarin	84 (2)

Other, unknown and unspecified values were excluded.

We classified these patients by screened status based on the presence of antecedent procedure codes and calculated a screening rate of 61%.

We then performed manual review of 150 medical records lacking evidence of timely screening in the structured database (table 2). Thirty-one patients were correctly classified as unscreened, corresponding to a negative predictive value of 21%. Most of these patients had examinations that were ordered but not completed, lacked documentation for unscreened status or were incorrectly documented by the responsible physician (eg, misunderstanding of the screening interval). One hundred and four patients (69%, 95% CI 62% to 76%) had positive evidence of screening on manual review. Half of these underwent screening outside of our institution and 28% underwent screening prior to the implementation of the *Epic* EHR in June 2012. The remaining 22 false-negative records (21%) were associated with screening examinations otherwise expected to occur within the theoretical scope of the database. Some of these errors were related to screening examinations performed around the time of database creation (March 2018) or *Epic* software installation (June 2012). The reasons identified for other errors were multifactorial but include errors of database creation and structure and at the level of querying. Lastly,

Table 2 Reasons for true and false classifications identified by manual chart review

Reasons for true negative classification	**n=31, 21% (15% to 28%)**
Examinations ordered but not completed	15, 48% (32% to 65%)
Colonoscopy ordered but not completed	8
Faecal immunochemical test ordered but not completed	6
CT colonography ordered but not completed	1
Lack of documentation or incorrect documentation	9, 29% (16% to 47%)
Declined screening	4, 13% (5% to 29%)
Insufficient time to discuss	3, 10% (3% to 26%)
Reasons for false-negative misclassification	**n=104, 69% (62% to 76%)**
Screening outside of UCSF	53, 51% (41% to 60%)
Screening prior to *Epic* EHR implementation	29, 28% (20% to 37%)
Database and query errors	22, 21% (14% to 30%)
Misclassified as eligible for screening	**n=15, 10% (6% to 16%)**
Poor life expectancy, or risks outweighing benefits	8, 53% (30% to 75%)
Above risk (personal or family history of polyps)	6, 40% (20% to 64%)
Not primary care empanelled	1, 7% (0% to 32%)
Reasons for false-positive misclassification	**n=2, 4% (0% to 14%))**
Ordered but incomplete faecal immunochemical test	1, 50% (9% to 91%)
Performed colonoscopy revealed inadequate bowel preparation	1, 50% (9% to 91%)

The second column lists the number of charts and associated percentage of the group with 95% CIs.
EHR, electronic health records; UCSF, University of California, San Francisco.

15 patients (10%) were not actually eligible for screening, primarily due to the risks outweighing the benefits or otherwise being categorised as above-average risk.

Lastly, manual review of 50 records suggested to be up-to-date with screening indicated a positive predictive value of 96% (95% CI 86% to 100%) (table 2). Two patients (4%) were unscreened—one ordered but incomplete FIT and one with a prior colonoscopy but inadequate bowel preparation.

Using the aforementioned positive and negative predictive values, we calculated a corrected period prevalent screening rate of 85% (81%–90%). The most common screening modality used was colonoscopy. Other notable global findings include four charts with incorrectly documented surveillance intervals. For example, one chart with a negative FIT in 2014 was incorrectly flagged for follow-up screening in 2024. We identified one patient who screened positive by FIT and was referred for colonoscopy, but the referral expired. We noted occasional discrepancies between surveillance intervals proposed by the gastroenterologist and primary care physician (eg, 5-year vs 10-year follow-up).

DISCUSSION

Medical billing databases available from either healthcare payors or from the EHR (as used in this study) are attractive sources of secondary data analysis for research, operations and quality improvement for a variety of reasons. They are increasingly accessible and relatively easy to query using common database languages. They are far easier to use for analysis compared with free-text data within the EHR such as in clinical notes. Because these databases tend to cover large patient cohorts (1.2 million in our EHR database, tens to hundreds of millions in many commercially available databases derived from claims data), they are accompanied by considerable statistical power and the potential for population-level inference.

A shortcoming of these data is that they were not collected specifically for research purposes, and thus are intrinsically prone to measurement bias. This is especially a problem for datasets (such as those from healthcare payors) that cannot be validated against a ground source of truth due to de-identification and de-linkage from the EHR. Although a common practice in the field of secondary data analysis involves the 'external validation' of study results against that obtained by unrelated datasets and independent investigators, our study underscores the fact that this is no substitute for the internal validation of data quality. Our assessment of EHR-derived billing data resulted in a screening rate that precisely matches that of the CDC using different methods; yet, this estimate was substantially incorrect. A study relying on a de-identified and unvalidatable claims database might have come to a similarly incorrect conclusion without any possibility of uncovering the truth.

Our work suggests the importance of caution when interpreting studies using data that cannot be subjected to internal checks of validity. The practice of data repurposing intrinsically represents a trade-off between feasibility/statistical power and accuracy. Although we would not argue that accuracy is the be-all and end-all of clinical research endeavours, research designs that propose to trade-off one for the other should ideally incorporate some semi-quantitative notion as to how accuracy is being sacrificed. Studies for which this cannot be done can be misleading and can bear adverse consequences for public health policy and impede efforts to improve healthcare quality.

Our study highlights at least one simple approach for confirming data quality—sampled record review. More complex approaches such as natural language processing and machine learning might eventually be able to

Table 3 Potential solutions to improve informatic classification and CRC screening

Reasons for true negative classification	Potential solutions
Examinations ordered but not completed	
Colonoscopy ordered but not completed	► More transparent documentation of referral status and outcome ► Clinic-based patient outreach
Faecal immunochemical test ordered but not completed	► Clinic-based patient outreach
CT colonography ordered but not completed	► More transparent documentation of referral status and outcome ► Clinic-based patient outreach
Lack of documentation or incorrect documentation	► Improved primary care education ► Improved gastroenterologist-primary care communication
Declined screening	► Improved patient education
Insufficient time to discuss	► Clinic-based strategies to encourage follow-up
Reasons for false-negative misclassification	
Screening outside of UCSF	► Patient-approved data sharing, harmonisation and interoperability ► Natural language processing ► Optical character recognition ► Deep learning
Screening prior to *Epic* EHR implementation	► Institutional investment in clinical data integration and harmonisation
Database and query errors	► Recruitment, training and funding for more clinical informaticians, especially clinician-investigators ► Institutional investment in clinical data integration and harmonisation
Misclassified as eligible for screening	
Poor life expectancy, or risks outweighing benefits	► Deep learning with natural language processing
Above risk (personal or family history of polyps)	► Natural language processing ► Improved family history taking practices ► Patient consent for EHR data-sharing, chart-linkage by familial relationship
Not primary care empanelled	► Deep learning with natural language processing
Reasons for false-positive misclassification	
Ordered but incomplete faecal immunochemical test	► EHR flag/reminders to repeat screening
Performed colonoscopy revealed inadequate bowel preparation	► Natural language processing ► EHR flag/reminders to repeat screening

EHR, electronic health records; UCSF, University of California, San Francisco.

perform this task using EHR data in a scalable way. We highlight some of these solutions in table 3 (see online supplementary methods table 2 for a definition of terms used in this manuscript, adapted with permission from Rudrapatna and Butte[6]). However, for the immediate future, we see manual review as being an integral part of any study relying on sources of routinely collected clinical data.

How do our results compare with previously published estimates? To our knowledge, only one study from Petrik *et al*[14] has directly reported the accuracy of EHR billing codes in identifying screened and unscreened patients. They too reported a high positive predictive value, consistent with our findings here. By contrast, they reported an 88% (85%–91%) negative predictive value, compared with 21% in our study.

We note several potential explanations. First, there were important differences in underlying cohorts: patients receiving preventative care within a safety-net system (eg, those under study by the Petrik *et al*) may be less likely to 'shop around' and receive fragmented care at different systems, unlike the patients at UCSF. Another potential explanation is that their study aimed to identify patients in need of screening, whereas this study aimed to accurately capture the prevalent screening rate. Our study excluded from the denominator any patient lacking a primary care relationship as well as those for whom the risks of screening outweigh the benefits. Unlike the study by Petrik *et al*, we did not informatically exclude patients with significant comorbid diagnoses or compute a Charlson Comorbidity Index; doing so would have introduced bias in our tertiary-care centre where many

sick patients undergo cancer screening prior to organ transplantation. We also did not treat referrals alone as positive evidence of screening; our protocol included the review of endoscopic reports to confirm the adequacy of the examination.

Common reasons for misclassification across both studies include note-based evidence of a qualifying screening examination. Half of the false-negative charts we reviewed had evidence of screening elsewhere, and a quarter of the charts had evidence of screening within our institution but generated by legacy EHR software prior to June 2012. Although 21% of the false negatives involved examinations performed within the expected scope of our database, some of these examinations occurred at the end of 2012 or in March 2018 (the month the database was queried), suggesting errors due to incomplete data migration. However, we also noted a variety of other idiosyncratic errors inherent to the data itself. In our view, these errors are a common consequence of the complex processes involved in data capture and transformation. The identification and correction of insidious errors of this nature requires a significant degree of institutional investment in data engineering; it also requires the sustained involvement of many clinical experts optimally positioned to identify these errors early and provide corrective feedback (table 3).

The challenges inherent to obtaining accurate estimates from administrative healthcare data raise the important question: is the very enterprise of clinical informatics cost-effective (and is further investment justifiable)? Would research funds be better spent on improving the methods employed in the National Health Interview Survey (NHIS)? It is difficult to answer the first question in a rigorous way—how does one quantify or estimate the total future benefits of increasingly accessible health information? Nonetheless, our view is that the answer is clearly 'yes'. EHR systems have already been paid for (at a sizeable cost) and widely implemented; reverting back to paper charts is not a viable option. In the setting of this existing 'buy-in', ongoing improvements to the capture and quality of healthcare data are inevitable because they underlie the capacity of health systems to continuously innovate in a competitive environment. Secondary use cases, such as research, will benefit as well. These sources of 'real-world data' serve as important confirmations of (or challenges to) the results of prospective studies such as the NHIS, and are broadly generalisable to virtually all domains in healthcare beyond CRC screening.

A key strength of this work lies in the study methodology. This study used a comprehensive list of diagnosis and procedure codes developed in close collaboration with proceduralists, billing staff and members of the quality improvement and accountable care division. Study investigators simultaneously contributed to both the query development and the chart review process, and improved both as a result. All charts independently examined by one internist and gastroenterologist each. We reported robust binomial CIs and tested the coverage probability of the corrected prevalence estimate with Monte-Carlo simulation. This work was able to identify, at a fairly granular level, reasons for errors at all levels, from clinical informatics to the provision of primary care. This audit led to the identification of several clinical care errors, with clinicians informed and education provided where appropriate.

We acknowledge several limitations. First, the chart review process was challenging. Interpreting clinical notes is inherently a subjective process, and we encountered many edge cases that required discussion, criteria refinement and imperfect resolution. The nuances of balancing of competing agenda, incorporating values and weighing risk-benefits within a time-limited clinic visit frequently do not make it to the written page. We also note other potential sources of measurement bias. We decided to accept note-based documentation as sufficient evidence that screening was performed, rather than having required the full screening report in the chart. We also suspect that relevant family history (eg, interval diagnoses of advanced polyps) are not regularly rechecked and updated at each visit, contributing to some mismeasurement. Lastly, the specific CRC screening rates at our institution may not generalise to other primary care clinic populations.

Although billing data derived from the EHR and claims data from healthcare payors are similar, they are not identical. Claims data may capture healthcare utilisation across multiple sites. EHR structured data captures local patient data irrespective of insured status or changes to insurance carrier. The EHR also carries the potential to explore a richer dataset including test results and unstructured data in the form of clinical notes. Both systems are subject to breaks and discontinuities as patients leave and enter (or re-enter), as well as the errors inherent to intrinsically complex, non-research grade data.

Our results indicate that the primary care apparatus at our institution is effective at performing CRC screening. Nevertheless, we see several potential areas of improvement. Improved documentation of the CRC screening decision and the disposition of screening referrals, regular updating of family history and greater communication between gastroenterologists and internists will help all healthcare institutions improve their screening rates. They may also improve informatic ascertainment of screened status in combination with technologies such as natural language processing, optical character recognition and deep learning (table 3).

However, the greatest challenge to the future of clinical informatics lies in the problem of bias in observational data. Identifying and managing bias is fundamentally a task that requires humility, vigilance and the collaborative engagement of diverse stakeholders and domain experts who understand the provenance and meaning of the data. It requires that we stress test our data openly and often before drawing conclusions or taking action.

Author affiliations
[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, United States
[2]Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA, United States
[3]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[5]Division of Gastroenterology and Hepatology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States
[6]Department of Medicine, University of California, San Francisco, California, USA
[7]Department of Pediatrics, University of California, San Francsico, CA, United States
[8]Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA, United States

Acknowledgements The authors would like to thank Marlene Herrera and Sara Coleman-Hernandez for useful technical input. The authors would like to thank Boris Oskotsky and Dana Ludwig for database creation and management. The authors would also like to thank members of the UCSF Gastroenterology Division as well as Biostatistics and Epidemiology Department for valuable discussion.

Contributors VAR and BSG conceived the project, performed data extraction and analysis and drafted this manuscript. VAR, PA, EH-T and CW performed the chart review. AJB supervised the project and critically edited this manuscript.

Funding UCSF Bakar Computational Health Sciences Institute and the National Centre for Advancing Translational Sciences of the National Institutes of Health under award number UL1 TR001872. VAR was supported by the National Institute of Diabetes and Digestive and Kidney Disease of the National Institutes of Health grant under award number T32 DK007007-42.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval This study was approved by the University of California, San Francisco Institutional Review Board (#18–25166).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. No data from this study are available for reuse.

ORCID iDs
Vivek A Rudrapatna http://orcid.org/0000-0003-1789-3004
Benjamin S Glicksberg http://orcid.org/0000-0003-4515-8090
Connie Wang http://orcid.org/0000-0002-6621-8112
Atul J Butte http://orcid.org/0000-0002-7433-2740

REFERENCES

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* 2017;67:7–30.
2 Kahi CJ, Pohl H, Myers LJ, *et al*. Colonoscopy and colorectal cancer mortality in the Veterans Affairs health care system: a case-control study. *Ann Intern Med* 2018;168:481–8.
3 Centers for Disease Control and Prevention (CDC). Use of colorectal cancer tests--United States, 2002, 2004, and 2006. *MMWR Morb Mortal Wkly Rep* 2008;57:253–8.
4 White A, Thompson TD, White MC, *et al*. Cancer Screening Test Use - United States, 2015. *MMWR Morb Mortal Wkly Rep* 2017;66:201–6.
5 Newell SA, Girgis A, Sanson-Fisher RW, *et al*. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. *Am J Prev Med* 1999;17:211–29.
6 Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest* 2020;130:565–74.
7 Ward JC. Oncology reimbursement in the era of personalized medicine and big data. *J Oncol Pract* 2014;10:83–6.
8 Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013;51:S30–7.
9 EPIC. Electronic Health Record Software [Internet]. Available: www.epic.com [Accessed 2 Jan 2018].
10 Vienna, Austria: R Foundation for Statistical Computing. R: A language and environment for statistical computing, 2013. Available: http://www.r-project.org/
11 Database Interface and “MySQL” Driver for R [R package RMySQL version 0.10.17]. Available: https://cran.r-project.org/web/packages/RMySQL/index.html [Accessed 30 Jun 2019].
12 Extension of “data.frame” [R package data.table version 1.12.2]. Available: https://cran.r-project.org/web/packages/data.table/index.html [Accessed 30 Jun 2019].
13 Agresti A, Coull BA. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *Am Stat* 1998;52:119.
14 Petrik AF, Green BB, Vollmer WM, *et al*. The validation of electronic health records in accurately identifying patients eligible for colorectal cancer screening in safety net clinics. *Fam Pract* 2016;33:639–43.