Supplemental Methods

Rudrapatna VA, Glicksberg BS, Avila P, Harding-Theobald E, Wang C, Butte AJ. Accuracy of Medical Billing Data Against the Electronic Health Record in the Measurement of Colorectal Cancer Screening Rates.

CPT Codes:
We excluded all CPT billing entries annotated as “cancelled.”

Colonoscopies: 45378, 45379, 45380, 45381, 45382, 45383, 45384, 45385, 45386, 45387, 45388, 45389, 45390, 45391, 45392, 45393, 45398, G0121
Sigmoidoscopy: 45330, 45331, 45332, 45333, 45334, 45335, 45337, 45338, 45340, 45341, 45342, 45346, 45347, 45349, G0104
Proctosigmoidoscopy: 45300, 45303, 45305, 45308, 45317, 45321, 45327

Fecal Occult Blood Test, Fecal Immunochemical Test, Computed Tomography – Colonography, and Double Contrast Barium Enema were identified from the billing table using regular expression-based string matching.

Coding Scheme:

1. Is their CRC screening up to date (screened at least once within 1/2016-3/2018 period)
   1=Y
   2=N
   3=Not eligible or empaneled

2) How screened?
0=No evidence of up-to-date screening or not applicable
10=UCSF colonoscopy
11=UCSF sigmoidoscopy
12=UCSF CT Colonography
13=UCSF FOBT or FIT
14=UCSF Barium enema
20=non-UCSF colonoscopy
21= non-UCSF sigmoidoscopy
22= non-UCSF CT Colonography
23= non-UCSF FOBT or FIT
24= non-UCSF Barium enema

3) Why not screened?
1=Was Screened
2a = Not eligible 2b = Not empaneled
3=Not or incorrectly documented by either PCP or GI
4=Patient declines
5a= FIT ordered, not complete
5b= Endoscopy ordered, not complete
5c= CT ordered not complete

6=Not enough time to address screening on this visit
7=Other

4) Why not eligible?
1 = Was eligible
2a= Poor life expectancy
2b= Risks>Benefits
3a = Prior polyp or colon cancer
3b = Family history of polyp or cancer
3c = Presence of IBD/Lynch/FAP/PJS
4 = Not living in 2017
5 = Not 50-75 in 2017
6 = Not empaneled
7 = Other

Chart Review protocol:

1. Assess whether the patient was or was not screened within 3 months of the last qualifying PCP or GI encounter within the January 2016 - December 2017 period using the following guidance:
    1. Check documentation under “Overdue Health Maintenance” Flag (located in top bar immediately under allergies)
    2. Use the “magnifying glass” global search function to search for: “colonoscopy”, “colon cancer”, “colorectal cancer”, “screening”, “CRC”, “flex sig”, “sigmoidoscopy”, ”virtual colonoscopy”, ”colonography”, ”FOBT”,”FIT”, “barium enema”, ”healthcare maintenance”, ”screening”
    3. Confirm no colonoscopy or sigmoidoscopy report under procedures
    4. Confirm no CT colonography/virtual colonoscopy/barium enema under imaging (load all, sort by Exam) (not requiring ‘double contrast’ or ‘air’ prefix to barium enema).
    5. Check the last primary care note under healthcare maintenance
    6. Check the last gastroenterology note
    7. Look at CareEverywhere in the following locations:
        (a) Other Results: “Endoscopy”, “Colonoscopy”, “Procedure Report”, “CT Colonoscopy/Colonography”
        (b) Labs: Fecal Immunochemical Test, Fecal Occult Blood Test
        (c) Documents, Description Field: “Colonoscopy”, “Endoscopy”
        (d) Documents, Dept. Specialty Field: “Gastroenterology”, “Surgery”, “Primary Care”
2. Confirm that the patient met all criteria for eligibility:
    1. Living in 2017 (presence of at least 1 encounter during this calendar year)
    2. Aged 50-75 during that year
    3. Absence of high-risk conditions:
        1. Use the “magnifying glass” global search function to search for: prior personal or family history of “colon polyp” or “colon cancer”. No history of “Lynch”,”HNPCC”,”FAP”,”polyposis”,”Crohn’s”,”Colitis”,”Peutz-Jehgers Syndrome”, ”IBD”
    4. Able to benefit from screening (e.g. Benefits>Risks), adequate life expectancy
        1. Use the “magnifying glass” global search function to search for: “life expectancy”, ”comorbid”, “comorbidity”, ”tolerate”, “tolerance”

2. Review last primary care note under screening

5. Had at least two PCP encounters (office or nurse visit) or two GI encounters or one of each in 2016-2017 matching the following locations:

"GI MZ", "GI OSTROFF PARN", "ZZZGI MZ","ZZZGI PARN"
"PK PRIMARY CARE", "PK GEN MED MZ 1545 1", "PK GEN MED MZ 1545 2"
"PRIMARY CARE CB BERRY", "PRIMARY CARE LV", "UCSF PRIMARY CARE",
"GEN MED MZ 1545 1", "GEN MED MZ 1545 2", "GEN MED MZ 1701", "GEN MED PARN MB1 4A", "GEN MED PARN MB1 4B", "360 POSITIVE HEALTH", "ZZZ360 POS HEALTH OLD", "EXEC HEALTH PARN OLD","EXEC HEALTH MB", "GERIATRIC CARE","ZZZGERIATRICS SVC","GERIATRICS HOME SVC", "HOUSECALLS","ZZZHOUSECALLS", "LAKESHORE","ZZZLAKESHORE HOME HEALTH-BIL","ZZZOL LAKESHORE LAGUNA GROVE", "WOMENS HEALTH MZ","ZZZWOMENS SPECIALTY OLD","WOMENS SPECIALTY","WOMENS HEALTH MZ3","ZZZWOMENS HEALTH 1725 MONT"

3. If the patient was not screened due to an apparent misunderstanding of the proper screening or surveillance interval by the primary care physician, please inform them via secured message.

Supplemental Table 1. Definition of healthcare data terminology used in this manuscript.

Term	**Definition**
Administrative Healthcare Data	Data collected over the course of healthcare operations and delivery. These data include routinely collected clinical data from the Electronic Health Records and medical claims data sent to healthcare payors. Although these data were captured for the purposes of healthcare operations and delivery, they are increasingly being repurposed to support other objectives, including healthcare research, improving quality, reducing cost, and streamlining operations.
Electronic Health Records (EHR)	Computerized health systems that capture data during the course of patient-provider encounters. These systems contain the data relevant to aid providers in the delivery of healthcare, such as clinical notes, test orders and results. They also contain data elements needed to submit medical claims/bills to healthcare payors (billing data).
Medical Billing Data	Data elements commonly needed to submit claims for (e.g. bill for) provided medical treatment. These data elements commonly include diagnosis and procedure codes, as well as other data about the provider, patient, and other resources needed to provide healthcare. Medical billing data is commonly captured and deposited into databases maintained by healthcare payors (i.e payor database), but may also be found in databases derived from the EHR systems that electronically transmit bills and supporting data to the payor. Although these sources of billing data are highly similar, they are not identical (see Discussion).
Structured Data	Data that is typically organized in a tabular format and readily queriable by common database languages such as SQL. In the setting of healthcare data, structured databases commonly contain medical billing data elements such as diagnosis and procedure codes but less commonly free-text elements such as clinical notes.
De-Identification	When repurposed for uses unrelated to direct healthcare delivery, patient-specific identifers in administrative healthcare data are commonly de-identified in order to protect the privacy of the patient.
De-linkage	The presence of mapping between disparate data streams on the basis of a common feature, such as patient identity. In the setting of de-identified healthcare databases such as payor claims databases repurposed for research purposes, the absence of such a map (‘linkage’) between these data and the full EHR (specifically the free-text) can significantly inhibit the ability to confirm findings using clinical notes. By extracting billing data elements from the EHR itself (rather than making use of entirely different, de-identified claims databases), the study design proposed in this work preserves data linkage and permits study validation via manual review.

Supplemental Table 2: Definition of clinical data science terms used in this manuscript.

Term	Definition
Data Harmonization	Because administrative healthcare data by definition is typically not collected with secondary reuse goals in mind, the raw data are commonly organized in a disorganized fashion not readily amenable for repurposing. Data harmonization refers to the task of re-organizing the data across disparate sources and formats to enable meaningful reuse.
Interoperability	The creation of standards to enable clinical data sharing across health systems. The primary objective of this data sharing is typically to facilitate clinical care; however, these data are increasingly being shared for other purposes including research, quality improvement, and others.
Natural Language Processing	A research field concerned with the computational processing and analysis of human language.
Optical Character Recognition	Methods that convert typed, handwritten, or printed text (such as found in scanned documents) into digitally-encoded text. Much healthcare data is found in the form of scanned documents and thus requires the use of these methods in order to make their content more useful.
Deep Learning	A subdomain of Machine Learning which uses complex artificial neural networks in order to train computer models to perform a variety of tasks including prediction. These models are characterized by significant flexibility in terms of their ability to handle wide varieties of data types from various sources, and are increasingly being used to perform natural language processing tasks.