PRECISE DATA USER GUIDE

V3.2.1

Contents

Version History

Version	Date	Notes
3.1.1	09/13/2022	• V3.1.1 finalized (3rd major data submission, 1st version of data, 1st version of user guide)
3.2.1	03/30/2023	• V3.2.1 finalized (3rd major data submission, 2nd version of data, 1st version of user guide) • Added documentation for PRECISE-only Diagnosis & Comorbidity tables • Made minor updates to Appendix G “Analytic Approaches” section • Added note to clarify analytic use of KPWA LargePolyp variable in the Procedure table documentation • Updated notes in the following locations to accompany corrected IMS/PCC data submission: ○ About the Data User Guide / Overarching Concepts ○ Participant / Variables / RaceMultipleNOS ○ CalendarYear / Variables / Medicare, Medicaid, Uninsured ○ CancerRegistry / Variables / Sequence, Rx*, NAACCR* ○ Provider / Variables / Provider_Race1 ○ SDoH / Variables / Yost_State_Quintile

About PRECISE

Optimizing Colorectal Cancer Screening PREcision and outcomes in CommunIty-baSEd populations (PRECISE) is an integrated PROSPR II Research Center (PRC). The long-term goal of PRECISE is to reduce colorectal cancer (CRC) morbidity, mortality, and disparities. All analyses either directly evaluate groups with disparities (e.g., African Americans and older patients) or will explore multiple underlying barriers to care. Our specific aims are to:

- **Aim 1:** Create a unified PRC with >10 years of comprehensive longitudinal CRC screening process and outcome data for >8.9 million persons across diverse, community-based settings.
- **Aim 2:** Conduct well-powered observational studies (Projects 1–4) to close evidence gaps regarding in whom and when CRC screening and surveillance should be performed, why failures occur, and how to improve test effectiveness.
- **Aim 3:** Design and pilot test potential intervention(s) likely to be effective, impactful, acceptable, and implemented using a multistep process based on findings from Projects 1–4, theory-based methods for designing effective interventions, modeling, and stakeholder engagement.
- **Aim 4:** Collaborate with the trans-PROSPR Coordinating Center and cervical cancer and lung cancer organ research centers to develop common measures of screening quality and system-level factors across organs and address trans-organ questions to improve cancer screening processes in the United States.

Participating Sites

Kaiser Permanente Washington (KPWA)

KPWA, a mixed-model health care system that became Kaiser Permanente's eighth region in 2017, is comprised of Kaiser Foundation Health Plan of Washington (KFHPW) and the Washington Permanente Medical Group (WPMG). KPWA provides coverage to approximately 700,000 members for health care received in the internal delivery system (IDS) and/or the external delivery system (EDS). As of August 2021, the IDS consists of 31 medical offices and other outpatient facilities that are owned and operated by KFHPW and provide primary care services. For services not provided by or exceeding the capacity of the IDS, KFHPW contracts with providers, medical groups, and hospitals—the EDS—to provide these services to KPWA members. About two-thirds of members receive primary care within the IDS from WPMG providers at KFHPW-owned medical centers in 17 cities in Washington State. However, many of these members receive their specialty care within the EDS, as most KFHPW-owned medical centers focus on primary care. About one-third of KPWA members receive all their care within the EDS. KPWA's mixed-model structure allows assessment of different payment models across a large region. KPWA's IDS has two geographically defined clinical operations units that oversee primary care and clinic operations. Each unit has regional operational and clinical directors, and each medical center has medical and administrative leaders. (Three additional care delivery units oversee specialty services, telehealth, and acute/post-acute care.) The organization-wide Quality and Safety division oversees including Quality Reporting (e.g., HEDIS and CAHPS metrics and dashboard reports), guidelines (available internally and publicly), and KPWA's Screening and Outreach Program. The Screening Program mails annual birthday letters advising patients of preventive care due dates, including cancer screening. Evidence-based clinical guidelines are developed with KPWA clinicians and researchers.

University of Texas Southwestern (UTSW)

UT Southwestern participates in PRECISE in collaboration with its longstanding clinical and research partner, Parkland Health & Hospital System (PHHS). PPHS is an integrated healthcare system that is one of the largest public hospital systems in the country. As the Dallas County tax-supported public safety-net system, PHHS is the primary provider of healthcare for all under- and un-insured county residents. PHHS is licensed for over 980 beds and is comprised of more than 110 specialty clinics, including 12 primary care clinics in areas of high need throughout Dallas County. Community-based clinics provide healthcare access to a diverse population. Services include financial counseling and linking individuals with appropriate federal- and state-funded insurance programs. Based on a PROSPR I study showing much higher initial compliance with mailed FIT kits vs. colonoscopy invitations, PHHS adopted a “FIT-first” screening strategy, rather than endoscopic screening only.

Kaiser Permanente Northern California (KPNC)

KPNC is one of the nation’s largest integrated healthcare delivery systems, with approximately 4.2 million current members across urban, suburban, and semi-rural areas from agricultural areas in North and Central California. It currently has 43 medical centers (which include clinics and hospitals), 127 medical offices or clinics, and 7 hospitals (separate from those at comprehensive medical centers). Medical centers are administratively distinct, providing variation in policies and approaches. Member insurance types include most U.S. options including Medicare; Medicaid; plans that are high-deductible, employer-paid, or individual prepaid; and safety net systems (e.g., Healthy San Francisco for uninsured low-income residents). By census data, KPNC members are similar to the region’s population in demographics, socioeconomic profiles, and other factors. Members include retirees, white- and blue-collar workers, service employees, government employees, under and unemployed residents, and patients in Covered California. Medical care is provided through The Permanente Medical Group, the nation’s largest multispecialty group. A distinct corporate entity owned and operated by its physician shareholders, the Medical Group uses a staff-model salaried system. It contracts annually with the Kaiser Foundation Health Plan insurance, a separate nonprofit entity, to care for Health Plan members. KPNC uses a comprehensive EHR system that captures detailed medical, procedure, medication, pathology, and outcome data. KPNC performs most CRC screening through a mailed annual FIT outreach program, delivering FIT to patients not up to date with screening by other methods. Screening colonoscopy is by patient or physician request. In daily operations, the Medical Group is like many large private and public multispecialty groups, including the >830 networks of doctors, hospitals, and other providers in Accountable Care Organizations by Medicare, which serve >28 million in the United States.

Kaiser Permanente Southern California (KPSC)

KPSC is an integrated healthcare delivery system serving over 4.6 million people in Southern California, spanning 12 geographic regions from agricultural Bakersfield in the north to San Diego in the south. It currently has 15 medical centers and 234 medical offices. Although KPSC and KPNC are Kaiser Permanente entities, KPSC is financially autonomous and organizationally different. Similar to KPNC, its medical group is separate from the insurer; however, unlike KPNC, some physicians (including gastroenterologists) use incentive-based payment systems (which may change screening behaviors). KPSC also has a greater emphasis on colonoscopy (vs. FIT) for CRC cancer screening. It has also developed different methods for performing CRC outreach and performance, including the use of high-volume endoscopy units and anesthesia-assisted procedures. The KPSC membership approximates the demographics of the Southern California population, including large numbers of U.S.-born and immigrant Hispanics, a rapidly growing demographic group with some of the lowest rates nationally for CRC screening.

About the Data User Guide

Overarching Concepts

The following notes apply throughout:

- Data on the PRECISE Virtual Data Center (VDC) will be stored as SAS® data sets.[1]
- Type (Length) documentation uses SAS data set concepts of SAS variable type (character or numeric) and number of bytes allotted to variable storage (1+ for character, 3–8 for numeric). Note that this does not necessarily correspond to the length of the variable when translated into CSV or other formats.
- Valid Values documentation covers the full range of values that are allowable in a given field—not necessarily the full range of *expected* values.
- IMS3 submission covers the 2010–2020 cohort period.
- “Reference date” refers to the participant’s date of birth. I.e., all days since reference (DSR) variables represent the participant’s age in days at the given timepoint.
- In all SAS date fields, UTSW imputes the day as the 15th of the given month/year; however, UTSW does provide exact DSR values.
- Data sources and availability differ significantly by site, year, table (a.k.a. file or data area), and variable (common data element [CDE]). Please note the overarching Data Sources section below and look for table- and variable-level notes throughout this document.
- KPNC and KPSC shared similar programming approaches for many tables and variables; hence the user will find many references to “KPSC = Same as KPNC” throughout.
- Cross-reference links within the guide will be formatted like this (cf. external hyperlinks). After clicking a cross-reference link, use [Alt]+[Left Arrow] or the Back button to return to the previous location.

For IMS Only

The following notes apply only to data prepared for and shared with Information Management Services, Inc. (IMS):

- Data will be released to IMS in CSV format; hence references in this document to SAS dates and variable type/length are not applicable.
- No full dates will be released to IMS, so all date variables will be set to the 15th of the given month and year. Full DSR will be provided.

Other IMS-related caveats and exclusions are documented in the relevant tables throughout this user guide.

Data Sources

Virtual Data Warehouse

As members of the Health Care Systems Research Network (HCSRN), the three Kaiser Permanente sites have access to common data elements in local implementations of the Virtual Data Warehouse (VDW) model. These data are actively maintained by local data warehouse teams and undergo semi-regular quality checks. Available data include details of member/patient demographics, health plan enrollment, health care utilization (encounters, diagnoses, procedures), selected laboratory results, prescription fills, etc.

Electronic Medical Records

All four sites use Epic as their primary electronic medical records (EMR) system. As such, programmers have access to the Clarity database that reflects much of the data available to and entered by clinicians as part of

---

[1] *SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.*

patient care. However, Clarity data are generally not as “shovel-ready” as data in the VDW (KP sites). KP site programmers are often able to work from data elements that have already been extracted into the VDW, whereas Clarity is the primary data repository available at UTSW.

Tumor Registries

All four sites have access to tumor registry data, although the sources vary by site.

KPWA has access to a full NAACCR-formatted extract from the SEER Seattle Puget-Sound central registry; a subset of the data elements serves as the basis for the VDW Tumor table.

UTSW has access to data extracts from both the Texas Cancer Registry as well as the Parkland Health & Hospital System registry.

At KPNC, data from both local (facility-based) registries and central (California state and SEER) registries are consolidated into a cancer registry file; a subset of those variables flow into the site’s VDW Tumor table.

Finally, KPSC receives an extract from their local cancer registry; these data are not incorporated into the VDW.

Deaths

At KPWA, fact of death data come from official state vital records data as well as a variety of internal sources (e.g., membership and billing, discharge status on claims). However, cause of death information is only available from the state. For the IMS3 submission, state data were only officially complete through 2019 and provisionally complete through Q3 2020. During 2010–2019, 96–97% of enrollee deaths could be ascertained from non-state sources. Therefore, while KPWA fact of death ascertainment should be >95% complete for the IMS3 data submission, information on cause of death was known to be incomplete for the final year of the cohort.

KPNC and KPSC also receive information on participant deaths from state vital records data as well as a variety of internal sources, with cause of death data again coming only from state records. Both fact and cause of death data were considered complete for the entire 2010–2020 cohort period.

UTSW derives death data from the Parkland Hospital EHR and Texas Cancer Registry. Deaths are documented in the EHR when any of the following events occurs: a patient dies at a Parkland-operated facility (e.g., inpatient hospital, emergency department); a patient dies within Dallas County and the primary care physician signs the death certificate; a payer notifies Parkland of the patient’s death (commercial payers, Medicare, or an ACA plan); or a family member notifies Parkland of the patient’s death. Deaths that occur outside of Dallas County or at non-Parkland facilities or are not reported by payers and/or family members may not be documented. Therefore, vital status may be under-ascertained in the Parkland Hospital EHR.

Summary

	KPWA	UTSW	KPNC	KPSC
VDW	✓		✓	✓
Epic/Clarity	✓	✓	✓	✓
State/central tumor registry	✓	✓	✓	
Local/hospital tumor registry		✓	✓	✓
State death certificate data	✓*		✓	✓

Data Dictionary

Participant Table

Overview

This table contains one record per eligible participant in the PRECISE cohort. For the IMS3 data delivery, all sites required each participant to spend at least one day in the cohort (i.e., exit date > entry date).

Participants were eligible to enter the cohort on the earliest date during 1/1/2010–12/30/2020 upon which all of the following criteria were met:

- Presumed living
- Aged 40–95 years
- Enrolled in health plan (KP sites) or with a visit to a primary care provider (UTSW)
- UTSW only:
    - Residing in Dallas County (per local payor assistance program criteria)
- KPWA only:
    - Assigned to Group Practice Division (GPD) (due to higher availability of both claims and EMR data)
    - Residing in SEER Seattle-Puget Sound Registry geographic area (to ensure capture of CRC occurrence)
    - Not insured via Medicaid (due to incomplete claims capture)
    - Not on any study exclusion lists

Participants exited the cohort on the earliest post-entry date during 1/2/2010–12/31/2020 upon which any of the following criteria were met:

- Date of death
- Day before 96th birthday
- End of continuous eligible enrollment (90-day gaps allowed; KP sites) or was known to have resided outside of Dallas County for >6 months (UTSW)
- UTSW only:
    - Lack of primary care utilization (>37 months)
- KPWA only:
    - Left GPD
    - Moved out of SEER Seattle-Puget Sound Registry catchment area
    - Insured by Medicaid
- End of cohort period (i.e., 12/31/2020)

Note: KPNC, KPSC, and UTSW allowed participants to exit and re-enter the cohort as long as the above criteria were met; KPWA restricted participants to a single cohort eligibility period. For more information, see the CohortEntryDate and CutoffDate variables in this table as well as the Enrollment Table documentation.

Data Sources

The three Kaiser Permanente sites pulled Participant data from their local VDW Death, Demographic, Enrollment, and Vital Signs tables. UTSW pulled data directly from EMR. See About the Data User Guide > Data Sources > Deaths for more information on availability of death data across the four sites.

Note: Due to KPWA's mixed model (described in the About PRECISE > Kaiser Permanente Washington (KPWA) section), participants who receive care in the external delivery system are less likely to have race/ethnicity (i.e., EMR-captured, not claims-based) data available for analysis.

Variables

PID

Definition	Participant ID
Type (Length)	Numeric (8)
Valid Values	Numeric values unique at the participant level within each site
General Notes	• PID + ProvidingSite = composite primary key. • A given PID should represent the same person: ○ across all PROSPR II organ sites (cervical [METRICS], colorectal [PRECISE], and/or lung [LOTUS]); and ○ in the previous round of PROSPR funding (PROSPR I).

UTSW Notes	• Site did not maintain the same participant IDs used in PROSPR I.
KPNC Notes	• First two digits represent site ID.
KPSC Notes	• Same as KPNC.

BIRTHYR

Definition	Year of birth
Type (Length)	Character (4)
Valid Values	YYYY-formatted years 1914–1980
General Notes	• Valid values reflect participant being 40–95 years of age at cohort entry.

BIRTHMTH

Definition	Month of birth
Type (Length)	Character (2)
Valid Values	MM-formatted months 01–12
UTSW Notes	• Site does not provide full birthdate, so all records have this variable set to 06.

BIRTHDAY

Definition	Day of birth
Type (Length)	Character (2)
Valid Values	DD-formatted days 01–31
General Notes	• Variable is excluded from IMS data submission.
UTSW Notes	• Site does not provide full birthdate, so all records have this variable set to 15.

HISPANIC

Definition	Whether participant is of Hispanic or Latino origin
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
KPNC Notes	• All records set to 01 (Yes) or 99 (Unknown), as “No” does not exist in source data.

RACEWHITE

Definition	Whether participant’s race is White
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes 99 = Unknown

RACEBLACK

Definition	Whether participant’s race is Black or African American
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

RACEASIAN

Definition	Whether participant’s race is Asian
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

RACEAIAN

Definition	Whether participant’s race is American Indian or Alaska Native
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

RACEPI

Definition	Whether participant’s race is Native Hawaiian or other Pacific Islander
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

RACEMULTIPLENOS

Definition	Whether participant is of multiple races, not otherwise specified
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

General Notes	• This variable is intended to accommodate the scenario in which a participant's race is recorded as simply "multiracial" with no information about the specific multiple races. • Additional multiracial participants can be identified via combining the other Race_ variables (e.g., RaceBlack = 01 and RaceAsian = 01).
UTSW Notes	• All records set to 99 (Unknown), because the concept of recording a participant as "multiracial, not otherwise specified" does not exist in EMR.
KPNC Notes	• The data capture scenario described under General Notes does not exist at this site, so there are no records set to 01 (Yes); all participants have known, specified race(s) (i.e., this variable coded as 00, i.e., participant's race is not "multiracial, NOS") or race is completely unknown (this variable coded as 99).

RACEOTHER

Definition	Whether participant is of another race not specified above
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• This variable is intended to accommodate the scenario in which a participant's race is recorded as a single, specific race other than White, Black, American Indian or Alaska Native, Asian, or Native Hawaiian or other Pacific Islander.
KPNC Notes	• No "other" race data are captured at this site.
KPSC Notes	• Same as KPNC.

SEX

Definition	Sex of participant
Type (Length)	Character (2)
Valid Values	01 = Male 02 = Female 03 = Other 99 = Unknown
UTSW Notes	• Site does not capture 03 (Other) at time of data pull.

HEIGHTINCH

Definition	Participant's height in inches
Type (Length)	Numeric (4)
Valid Values	Integers 48–90 (valid heights) or -99999 (missing or invalid heights)
General Notes	• Value represents median height calculated over a lookback period of 10 years prior to cohort entry.

UTSW Notes	• Lookback was limited by data availability; earliest year of availability is 2006.

COHORTENTRYDATE

Definition	Date of first cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/30/2020
General Notes	• See Overview for more information on cohort eligibility. • A variable with the same name exists in the Enrollment Table. • Upper bound of valid value range reflects the requirement that each participant spend at least one day in the cohort.
KPWA Notes	• Site did not allow participants to exit and re-enter the cohort.
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site allowed for multiple entry/exit periods; this value represents the participant’s earliest cohort entry.
KPNC Notes	• Site allowed participants to exit and re-enter the cohort; this value represents the participant’s earliest cohort entry.
KPSC Notes	• Same as KPNC.

COHORTENTRYFIRSTDSR

Definition	Date of first cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35062
General Notes	• See Overview for more information on cohort eligibility. • Valid values reflect participant being 40–95 years old at cohort entry as well as the requirement that each participant spend at least one day in the cohort.
KPWA Notes	• Site did not allow participants to exit and re-enter the cohort.
UTSW Notes	• Site allowed for multiple entry/exit periods; this value represents the participant’s earliest cohort entry.
KPNC Notes	• Same as UTSW.
KPSC Notes	• Same as UTSW.

CUTOFFDATE

Definition	Date of last cohort exit (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/02/2010–12/31/2020

General Notes	• See Overview for more information on cohort eligibility. • See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites. • A variable with the same name exists in the Enrollment Table. • Lower bound of valid value range reflects the requirement that each participant spend at least one day in the cohort.
KPWA Notes	• Site did not allow participants to exit and re-enter the cohort.
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site allowed for multiple entry/exit periods; this value represents the latest cohort exit.
KPNC Notes	• Site allowed for multiple entry/exit periods; this value represents the latest cohort exit.
KPSC Notes	• Same as KPNC.

CutoffDateLastDSR

Definition	Date of last cohort exit (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14611–35063
General Notes	• See Overview for more information on cohort eligibility. • See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites. • Valid values reflect participant being 40–95 years old at cohort exit as well as the requirement that each participant spend at least one day in the cohort.
KPWA Notes	• Site did not allow participants to exit and re-enter the cohort.
UTSW Notes	• Site allowed for multiple entry/exit periods; this value represents the latest cohort exit.
KPNC Notes	• Site allowed for multiple entry/exit periods; this value represents the latest cohort exit.
KPSC Notes	• Same as KPNC.

FamHxAnyCRCDate

Definition	Earliest available date on which health system recorded any affirmative indication of family history of CRC, regardless of relation or age at onset (as SAS date)
Type (Length)	Numeric (4)
Valid Values	10/08/2003–12/31/2020 or missing
KPWA Notes	• Limited to data collected when participant was 18+ per IRB approval; otherwise, data could have been collected any time through a participant's cohort exit date. • Sourced from Epic’s family history table where the type of family history being recorded was labeled as “Colon Cancer” or “Cancer - Colon.”
UTSW Notes	• When available, family history of colorectal cancer was ascertained from endoscopist’s selection of “Yes” in an Epic SmartForm. For more information, see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009919/.

KPNC Notes	• Site collected information from Epic’s family history table where the type of family history being recorded was labeled as “Colon Cancer”.
KPSC	• Same as KPNC.

FamHxAnyCRCDSR

Definition	Earliest available date on which health system recorded any affirmative indication of family history of CRC, regardless of relation or age at onset (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	8620–35063 or missing
General Notes	• See additional information under FamHxAnyCRCDate. • Upper bound of valid value range reflects the fact that participants must exit the cohort on the day before their 96th birthday.

FamHx1stDegCRCDate

Definition	Earliest available date on which health system recorded any affirmative indication of family history of CRC in a 1st degree relative, regardless of age at onset (as SAS date)
Type (Length)	Numeric (4)
Valid Values	10/08/2003–12/31/2020 or missing
General Notes	• First-degree relative = participant’s biological parent, full sibling, or biological child
KPWA Notes	• Limited to data collected when participant was 18+ per IRB approval; otherwise, data could have been collected any time through a participant's cohort exit date. • Sourced from Epic’s family history table where the type of family history being recorded was labeled as “Colon Cancer” or “Cancer - Colon” and relationship was labeled as Mother, Son, Biological parent, Biological sibling, Biological child, Father, Sister, Brother, or Daughter.
UTSW Notes	• Information was not available at this site; all records set to missing.
KPNC Notes	• Site collected information from Epic’s family history table where the type of family history being recorded was labeled as “Colon Cancer” and relationship indicated Brother, Daughter, Father, Mother, Sister, Son, Natural child, Natural parent, or Natural sibling.
KPSC Notes	• Site collected information from Epic’s family history table where the type of family history being recorded was labeled as “Colon Cancer” and relationship indicated Brother, Daughter, Father, Mother, Sister or Son.

FamHx1stDegCRCDSR

Definition	Earliest available date on which health system recorded any affirmative indication of family history of CRC in a 1st degree relative, regardless of age at onset (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)

Valid Values	8620–35063
General Notes	• See additional information under FamHx1stDegCRCDate. • Upper bound of valid value range reflects the fact that participants must exit the cohort on the day before their 96th birthday.

EXTRACTDATE

Definition	Date when record was generated at providing site
Type (Length)	Character (8)
Valid Values	YYYYMMDD-formatted dates

PROVIDINGSITE

Definition	Providing site ID
Type (Length)	Character (2)
Valid Values	71 = KPWA 72 = UTSW 73 = KPNC 76 = KPSC

CalendarYear Table

Overview

This table contains one record per participant per whole or partial calendar year of cohort eligibility. E.g., if a person enters the cohort on 04/01/2014 and exits on 9/30/2016, then they should have distinct Calendar Year records for 2014, 2015, and 2016:

- the 2014 record would cover 04/01/2014–12/31/2014;
- the 2015 record would cover 01/01/2015–12/31/2015; and
- the 2016 record would cover 01/31/2016–09/30/2016.

Different data elements within each record are assessed at different points in time: some closest to the start of the time period, others as of the end of the year looking back. See variables below for more details.

Data Sources

The three Kaiser Permanente sites pulled these data from their local VDW Enrollment, Diagnosis, Procedure, Social History, and Vital Signs tables. (N.B. Social History and Vital Signs tables are generally EMR derived.)

UTSW used EMR data.

Variables

PID

Same as above.

CALENDARYR

Definition	Calendar year of cohort enrollment (whole or partial)
Type (Length)	Character (4)
Valid Values	YYYY-formatted years 2010–2020

WEIGHTPOUND

Definition	Participant’s weight in pounds
Type (Length)	Numeric (4)
Valid Values	Integers 50–600 (valid weights) or -99999 (missing or invalid weights)

General Notes	• Value represents valid weight in pounds (50–600) obtained while enrolled, within the calendar year, as close to the end of the calendar year as possible. • For the participant’s first calendar year only: If no weight data were collected during that year, sites used the last available data point (if any) from prior to cohort entry.

WEIGHTDATE

Definition	Date on which WeightPound was recorded (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/07/1993–12/31/2020 (valid weights) or null (missing or invalid weights)
General Notes	• See additional information re: timing of allowed weights under WeightPound.
UTSW Notes	• All dates are set to day 15 of the given month/year.

WEIGHTDSR

Definition	Date on which WeightPound was recorded (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 8588–35063 (valid weights) or null (missing or invalid weights)
General Notes	• Valid values reflect possible use of prior-to-cohort-entry data as well as the participant being ≤95 years old during cohort. • See additional information re: timing of allowed weights under WeightPound.

HEIGHT

Definition	Participant’s height in inches
Type (Length)	Numeric (4)
Valid Values	Integers 48–90 (valid heights) or -99999 (missing or invalid heights)
General Notes	• Value represents valid height in inches (48–90) obtained while enrolled, within the calendar year, as close to the end of the calendar year as possible. • For the participant’s first calendar year only: If no height data were collected during that year, sites used the last available data point (if any) from prior to cohort entry. • Note: While this value may be used in conjunction with WeightPound, it does not have to be collected on the same date. • A similar variable (HeightInch) exists in the Participant table, although that value represents median height (if available) at or during 10 years prior to cohort entry.

HEIGHTDATE

Definition	Date on which height was recorded (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/07/1993–12/31/2020 (valid heights) or null (missing or invalid heights)

General Notes	• See additional information re: timing of allowed heights under Height.
UTSW Notes	• All dates are set to day 15 of the given month/year.

HEIGHTDSR

Definition	Date on which height was recorded (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 8520–35063 (valid heights) or null (missing or invalid heights)
General Notes	• Valid values reflect possible use of prior-to-cohort-entry data as well as the participant being ≤95 years old during cohort. • See additional information re: timing of allowed heights under Height.

PPTSTFIPRES

Definition	Participant's state of residence (as numeric FIPS code)
Type (Length)	Character (2)
Valid Values	##-formatted (leading zeros preserved) FIPS state numeric codes from the American National Standards Institute (ANSI) Codes for States or blank (unknown or not applicable, e.g., non-U.S. addresses)
General Notes	• Variable is excluded from IMS data submission.
KPWA Notes	• Site used health plan enrollment data to obtain the participant's place of residence as of the start of the calendar period or as close to it as possible, excluding any pre- or post-period information. • Site left this field blank if the enrollment FIPS code was not valid per the April 2021 SAS ZIP code data set. • The field may also be blank if the start of the period occurred during a tolerated ≤90-day gap in enrollment.
UTSW Notes	• Site used EMR address history within and closest to beginning of calendar year to obtain participant's place of residence and confirm residence within the state of Texas.
KPNC Notes	• Values correspond to participant's residential address as of the start of the calendar period or as close to it as possible.
KPSC Notes	• Site used participant address with an effective date closest to but before the beginning of calendar year, so that the address was effective at the start of calendar year.

PPTCOUNTYFIPRES

Definition	Participant's county of residence (as numeric FIPS code)
Type (Length)	Character (3)
Valid Values	###-formatted (leading zeros preserved) FIPS county codes from 2020 FIPS codes or blank (unknown or not applicable, e.g., non-U.S. addresses)
General Notes	• Variable is excluded from IMS data submission.

KPWA Notes	• See additional information under PPTStFIPRes.
UTSW Notes	• Site used EMR address history within and closest to beginning of calendar year to obtain participant’s place of residence and confirm residence within Dallas County.
KPNC Notes	• See additional information under PPTStFIPRes.
KPSC Notes	• See additional information under PPTStFIPRes.

PPTZipRes

Definition	Participant’s ZIP code of residence
Type (Length)	Character (5)
Valid Values	#####-formatted (leading zeros preserved) 5-digit ZIP codes or blank (unknown or not applicable, e.g., non-U.S. addresses)
General Notes	• Variable is excluded from IMS data submission.
KPWA Notes	• See additional information under PPTStFIPRes.
UTSW Notes	• Site used EMR address history within and closest to beginning of calendar year to obtain participant’s place of residence and confirm residence within Dallas County ZIP codes.
KPNC Notes	• See additional information under PPTStFIPRes.
KPSC Notes	• See additional information under PPTStFIPRes.

ScreeningIntervention

Definition	Whether participant was enrolled in an intervention that might affect screening behavior
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
KPWA Notes	• Information was unavailable as of extract date; all records set to 99.
UTSW Notes	• Site used prior study data to identify participants enrolled in site-specific colorectal cancer screening studies during the calendar year for PROSPR I Project 1 CRIS (August 2013 – January 2016), PROSPR I Project 2 (March 2013 – July 2016), and CPRIT FIT4ALL (August 2017 – present). More information on these studies can be found in the following publications: ○ https://www.ncbi.nlm.nih.gov/pubmed/26265201 ○ https://www.ncbi.nlm.nih.gov/pubmed/26535565 ○ https://www.ncbi.nlm.nih.gov/pubmed/28873161
KPNC Notes	• Information was unavailable as of extract date; all records set to 99.
KPSC Notes	• Same as KPNC.

Medicare

Definition	Whether participant was covered by Medicare at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately "Did KP record any Medicare coverage for this participant while enrolled in PRECISE during the calendar year period?" • At UTSW, the definition is "Did the participant either have Medicare coverage or use Medicare coverage to pay for any of their care at PHHS while enrolled in PRECISE during the calendar year period?"
KPWA Notes	• Site pulled all of a participant's VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Medicare variable as follows: ○ If <u>any</u> records provided evidence <u>for</u> Medicare coverage (INS_MEDICARE = Y or E) → Medicare = 01 (Yes). ○ Else if <u>all</u> records indicated <u>no</u> Medicare coverage (INS_MEDICARE = N) → Medicare = 00 (No). ○ Otherwise → Medicare = 99 (Unknown). • There was an apparent drop in commercial coverage among Medicare age group circa 2017. A change in KPWA source data around this time obscured/revealed varying aspects of coverage. For example, there were likely some people pre-2017 who just had Medicare but showed up as having commercial coverage due to the way medical market groups were coded. This issue is not "fixable" at present.
UTSW Notes	• Site pulled all of a participant's payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Medicare variable as follows: ○ If participant had Medicare coverage during the calendar year or Medicare was used to pay for ≥1 encounter → Medicare = 01 (Yes). ○ Else if participant had non-Medicare coverage during the calendar year or had ≥1 encounter, but Medicare was never used as payment → Medicare = 00 (No). ○ Otherwise → Medicare = 99 (Unknown).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

MEDICAID

Definition	Whether participant was covered by Medicaid at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

General Notes	• At the KP sites, the definition is more accurately "Did KP record any Medicaid coverage for this participant during the calendar year?" • At UTSW, the definition is "Did the participant either have Medicaid coverage or use Medicaid coverage to pay for any of their care at PHHS during the calendar year?"
KPWA Notes	• Site pulled all of a participant's VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Medicaid variable as follows: ○ If any records provided evidence for Medicaid coverage (INS_MEDICAID = Y or E) → Medicaid = 01 (Yes). ○ Else if all records indicated no Medicaid coverage (INS_MEDICAID = N) → Medicaid = 00 (No). ○ Otherwise → Medicaid = 99 (Unknown). • As mentioned in the Participant Table Overview, KPWA's cohort eligibility was restricted to periods of non-Medicaid enrollment. Therefore, any observed Medicaid coverage could only occur during ≤90-day tolerated gaps in cohort-eligible enrollment.
UTSW Notes	• Site pulled all of a participant's payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Medicaid variable as follows: ○ If participant had Medicaid coverage during the calendar year or Medicaid was used to pay for ≥1 encounter → Medicaid = 01 (Yes). ○ Else if participant had non-Medicaid coverage during the calendar year or had ≥1 encounter, but Medicaid was never used as payment → Medicaid = 00 (No). ○ Otherwise → Medicaid = 99 (Unknown).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INSOTHERGOV

Definition	Whether participant was covered by any other federal or state health insurance program at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately "Did KP record any other government insurance coverage for this participant during the calendar year?" • At UTSW, the definition is "Did the participant either have any other government insurance coverage or use other government insurance coverage to pay for any of their care at PHHS during the calendar year?"

KPWA Notes	• Site pulled all of a participant’s VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the InsOtherGov variable as follows: ○ If any records provided evidence for other government insurance coverage (INS_STATESUBSIDIZED = Y or E) → InsOtherGov = 01 (Yes). ○ Else if all records indicated no other government insurance coverage (INS_STATESUBSIDIZED = N) → InsOtherGov = 00 (No). ○ Otherwise → InsOtherGov = 99 (Unknown).
UTSW Notes	• Site pulled all of a participant’s payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the InsOtherGov variable as follows: ○ If participant had federal (other than Medicaid/Medicare) or state-subsidized coverage (e.g., TITLE V, X, XX, BCCS, FEMA, Ryan White) during the calendar year or federal/state-subsidized coverage was used to pay for ≥1 encounter → InsOtherGov = 01 (Yes). ○ Else if participant had non-federal/-state-subsidized coverage or participant had ≥1 encounter, but other government insurance coverage was never used as payment → InsOtherGov = 00 (No). ○ Otherwise → InsOtherGov = 99 (Unknown).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INSCOMMERC

Definition	Whether participant was covered by commercial and/or private health insurance at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record any commercial/private insurance coverage for this participant during the calendar year?” • At UTSW, the definition is “Did the participant either have any commercial/private insurance coverage or use commercial/private insurance coverage to pay for any of their care at PHHS during the calendar year?”
KPWA Notes	• Site pulled all of a participant’s VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the InsCommerc variable as follows: ○ If any of those enrollment records provided evidence for commercial/private insurance coverage (INS_COMMERCIAL = Y or E) → InsCommerc = 01 (Yes). ○ Else if all those enrollment records indicated no commercial/private insurance coverage (INS_COMMERCIAL = N) → InsCommerc = 00 (No). ○ Otherwise → InsCommerc = 99 (Unknown).

UTSW Notes	• Site pulled all of a participant's payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the InsCommerc variable as follows: ○ If participant had commercial/private insurance during the calendar year or commercial/private insurance coverage was used to pay for ≥1 encounter → InsCommerc = 01 (Yes). ○ Else if participant had non-commercial/private insurance during the calendar year or participant had ≥1 encounter, but commercial/private insurance coverage was never used as payment → InsCommerc = 00 (No). ○ Otherwise → InsCommerc = 99 (Unknown). • Note: Commercial/private coverage is uncommon at PHHS.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

MEDICALASSIST

Definition	Whether participant was covered by a medical assistance charity program for the uninsured at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately "Did KP record any medical assistance insurance coverage for this participant during the calendar year?" • At UTSW, the definition is "Did the participant have medical assistance insurance coverage or use medical assistance insurance coverage to pay for any of their care at PHHS during the calendar year?"
KPWA Notes	• Information was unavailable as of extract date; all records set to 99 (Unknown).
UTSW Notes	• Site pulled all of a participant's payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the MedicalAssist variable as follows: ○ If participant had medical assistance (Dallas County/Parkland) coverage during the calendar year or medical assistance was used to pay for ≥1 encounter → MedicalAssist = 01 (Yes). ○ Else if participant had non-medical assistance coverage during the calendar year or had ≥1 encounter, but medical assistance coverage was never used as payment → MedicalAssist = 00 (No). ○ Otherwise → Medical Assist = 99 (Unknown). • Note: Medical assistance at Parkland includes a variety of charity programs funded by Dallas County tax dollars. Most patients who receive medical assistance receive Parkland Financial Assistance (PFA). PFA is available for uninsured residents of Dallas County and based on a sliding income scale; patients must renew PFA every 6 months to remain eligible.
KPNC Notes	• Same as KPWA.

KPSC Notes	- Same as KPWA.

UNINSURED

Definition	Whether participant was uninsured at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record any gaps in insurance coverage for this participant during the calendar year?” • At UTSW, the definition is “Did the participant self-pay or have no payor/coverage documented for any of their care at PHHS during the calendar year?”
KPWA Notes	• Site pulled all of a participant’s VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Uninsured variable as follows: ○ If there were any gaps between the end date of one period and the start date of the next → Uninsured = 01 (Yes). ○ Otherwise → Uninsured = 00 (No). • As mentioned in the Participant Table Overview, cohort eligibility at the KP sites was restricted to periods of health plan enrollment; therefore, any evidence of uninsured status during the calendar year could only occur during ≤90-day tolerated gaps in cohort-eligible enrollment.
UTSW Notes	• Site pulled all of a participant’s payment/insurance data during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. • Site used these records to assign the Uninsured variable as follows: ○ If participant had ≥1 encounter marked self-pay or no payor/coverage → Uninsured = 01 (Yes). ○ Else if participant had some form of insurance coverage during the calendar year or ≥1 encounter, but some form of insurance coverage was documented for any encounters → Uninsured = 00 (No). ○ Otherwise → Uninsured= 99 (Unknown).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INSOTHER

Definition	Other type(s) of insurance coverage during the calendar year period
Type (Length)	Character (*)
Valid Values	See site-specific notes.

General Notes	- At the KP sites, the definition is more accurately “What not otherwise categorized type(s) of insurance coverage did KP record for this participant during the calendar year?” - At UTSW, the definition is “What not otherwise categorized type(s) of insurance coverage did the participant have or use to pay for any of their care at PHHS during the calendar year?”
KPWA Notes	- Site pulled all of a participant’s VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. - Site used these records to assess whether the participant had coverage from a private-pay, self-funded, and/or any other plan and created a semicolon-delimited string using the following logic: ○ If any records provided evidence for private pay coverage (INS_PRIVATEPAY = Y or E) → InsOther will contain the string “Private pay”. ○ If any records provided evidence for self-funded coverage (INS_SELFFUNDED = Y or E) → InsOther will contain the string “Self funded”. ○ If any records provided evidence for other insurance coverage (INS_OTHER = Y or E) → InsOther will contain the string “Other insurance, type unknown”. ○ Otherwise → InsOther will be blank.
UTSW Notes	- Site pulled all of a participant’s payment/insurance data for PHHS utilization during the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. - InsOther is populated for records that indicate either the financial application is pending or when a worker compensation or other injury management organization was the listed payor; otherwise, InsOther will be blank.
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as KPWA.

INSHIGHDEDUCTIBLE

Definition	Whether participant was covered by high-deductible insurance at any time during the calendar year period
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	- “High-deductible” refers to definition from U.S. IRS Pub 969. - At the KP sites, the definition is more accurately “Did KP record any high-deductible insurance coverage for this participant during the calendar year?” - At UTSW, the definition is “Did the participant have any high-deductible insurance coverage or use any high-deductible insurance coverage to pay for any of their care at PHHS during the calendar year?”

KPWA Notes	- Site pulled all of a participant's VDW Enrollment records that spanned the calendar year, truncating at start/end of PRECISE cohort enrollment if necessary. - Site used these records to assign the InsHighDeductible variable as follows: ○ If any of those enrollment records provided evidence for high-deductible insurance coverage (INS_HIGHDEDUCTIBLE = Y or E) → InsHighDeductible = 01 (Yes). ○ Else if all those enrollment records indicated no high-deductible insurance coverage (INS_DEDUCTIBLE = N) → InsHighDeductible = 00 (No). - Otherwise → InsHighDeductible = 99 (Unknown).
UTSW Notes	- Information not available; all records set to 99 (Unknown).
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as KPWA.

ENROLLEDMONTHS

Definition	Number of months participant spent enrolled in the health system during calendar year period
Type (Length)	Numeric (3)
Valid Values	Integers 0–12
General Notes	- To avoid potential confusion between this calculated value and information available in the Enrollment Table, this variable will NOT be included in data available on the PRECISE Virtual Data Center, nor will it be transferred to IMS for the IMS3 submission.

SMOKE

Definition	Participant's smoking status
Type (Length)	Character (2)
Valid Values	00 = Never smoked 01 = Ever smoked 99 = Unknown
General Notes	- Variable was called SmokingStatus in IMS1 data submission. - Calculated at end of calendar year period using the single data point collected as close to the end of the year as possible (while enrolled in the cohort). - For the participant's first calendar year only: If no smoking data were collected during that year, sites used the last available data point (if any) from prior to cohort entry. - At the end of each calendar year period: ○ If the last data point indicated that the participant was a current or former smoker, Smoke = 01 (Ever smoked). ○ Otherwise, if the last data point indicated that participant had never been a smoker, Smoke = 00 (Never smoked). ○ Otherwise (including if no smoking data collected during calendar year), Smoke = 99 (Unknown).

SMOKEDATE

Definition	Date on which participant's smoking status was recorded (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/02/1996–12/31/2020 (known smoking status) or null (missing smoking status)
General Notes	• See additional information re: timing of allowed smoking status data points under Smoke variable.
UTSW Notes	• All dates are set to day 15 of the given month/year.

SMOKEDSR

Definition	Date on which participant's smoking status was recorded (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 6053–35063 (known smoking status) or null (missing smoking status)
General Notes	• Valid values reflect possible use of prior to cohort entry data as well as the participant being ≤95 years old during cohort. • See additional information re: timing of allowed smoking status data points under Smoke variable.

CHARLSONINDEX

Definition	In-person encounter-based Charlson comorbidity score
Type (Length)	Numeric (4)
Valid Values	Integers 0–29 (valid scores) or -99999 (insufficient enrollment -or- no eligible encounters during window [UTSW] or CalendarYr = 2020 [UTSW]); see site-specific notes for more details
General Notes	• See notes below and Appendix C. Charlson Comorbidity Index for information on site-specific differences in calculating the Charlson score. • Observation window includes 365-day period spanning January 1 through December 31 (non-leap years) or January 2 through December 31 (leap years). • For the participant's first calendar year only: If the patient was enrolled in the cohort for less than the full 365-day observation window but *was* enrolled in the cohort on December 31—and if the patient also had prior-to-cohort-entry health plan enrollment/engagement to complete the 365-day observation window—then the prior health plan enrollment data were used when calculating this variable.
KPWA Notes	• Variable was set to 0 if the participant was enrolled in the health plan for the full 365-day observation window but had no inpatient or ambulatory encounters in which to observe relevant billing codes. • Variable was set to -99999 if the participant was not enrolled in the health plan for the full 365-day observation window—i.e., 365 days including December 31—in a given calendar year (allowing for 90-day gaps). • Site included events from in-person encounters only, namely VDW encounter types AV, ED, IP, IS, and OE.

UTSW Notes	• Site was unable to distinguish among encounter types and therefore calculated this score only for 2010–2019 calendar year periods (i.e., prior to the onset of the COVID-19 pandemic, when telehealth began to be utilized more heavily). • Therefore, variable was set to -99999 if participant had no encounters in which to observe relevant billing codes, if participant had insufficient observation time in cohort, and/or for 2020 calendar year periods.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

ComorbDiabetes

Definition	Diabetes component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

ComorbDiabetesComp

Definition	Diabetes with complications component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

ComorbCOPD

Definition	Chronic obstructive pulmonary disease component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

ComorbMI

Definition	Myocardial infarction component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

COMORBCHF

Definition	Congestive heart failure component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

COMORBMALIGNANCY

Definition	Malignancy component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

COMORBTUMOR

Definition	Metastatic solid tumor component of in-person encounter-based Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr = 2020 [UTSW])
General Notes	• See additional details under CharlsonIndex.

TCHARLSONINDEX

Definition	Telehealth-inclusive Charlson comorbidity score

Type (Length)	Numeric (4)
Valid Values	Integers 0–29 (valid scores) or -99999 (insufficient enrollment -or- no eligible encounters during window [UTSW] -or- CalendarYear < 2020 [UTSW]); see site-specific notes for more details
General Notes	• See notes below and Appendix C. Charlson Comorbidity Index for information on site-specific differences in calculating the Charlson score. • Observation window includes 365-day period spanning January 1 through December 31 (non-leap years) or January 2 through December 31 (leap years). • For the participant’s first calendar year only: If the patient was enrolled in the cohort for less than the full 365-day observation window but *was* enrolled in the cohort on December 31—and if the patient also had prior-to-cohort-entry health plan enrollment/engagement to complete the 365-day observation window—then the prior health plan enrollment data were used when calculating this variable.
KPWA Notes	• Score was calculated using the same in-person encounter types listed under CharlsonIndex during 2010–2020 as well as selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous chats) during 2019–2020 only. • Variable was set to 0 if the participant was enrolled in the health plan for the full 365-day observation window but had no inpatient or ambulatory encounters in which to observe relevant billing codes. • Variable was set to -99999 if the participant was not enrolled in the health plan for the full 365-day observation window—i.e., 365 days including December 31—in a given calendar year (allowing for 90-day gaps).
UTSW Notes	• Site was unable to distinguish among encounter types and therefore calculated this score only for 2020 calendar year periods (i.e., the year when the COVID-19 pandemic increased utilization of telehealth). • Therefore, variable was set to -99999 if participant had no encounters in which to observe relevant billing codes, if participant had insufficient observation time in cohort, and/or for 2010–2019 calendar year periods.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

TComorbDiabetes

Definition	Diabetes component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TComorbDiabetesComp

Definition	Diabetes with complications component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TComorbCOPD

Definition	Chronic obstructive pulmonary disease component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TComorbMI

Definition	Myocardial infarction component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TComorbCHF

Definition	Congestive heart failure component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TComorbMalignancy

Definition	Malignancy component of telehealth-inclusive Charlson comorbidity score

Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

TCOMORBTUMOR

Definition	Metastatic solid tumor component of telehealth-inclusive Charlson comorbidity score
Type (Length)	Character (2)
Valid Values	0 = No 1 = Yes 99 = Unknown (insufficient enrollment [KP sites] -or- no eligible encounters during window [UTSW] -or- CalendarYr < 2020 [UTSW])
General Notes	• See additional details under TCharlsonIndex.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

PriorToCohortEntry Table

Overview

This table contains one row per participant and summarizes available information from the period prior to cohort entry.

The lookback period for this table varied by site and variable as follows:

- KPWA:
    - Lookback was truncated at the latter of 01/01/1993 and the participant's 18th birthday, per IRB approval.
- UTSW:
    - Lookback for non-CRC variables extended to approximately 01/01/2006 regardless of when the participant entered the cohort.
        - Note: Start of uniform EMR data availability was 01/01/2009; however, pre-2009 data were imported into Epic in select circumstances.
    - Lookback for CRC diagnoses extended to 01/01/1995 per Texas Cancer Registry data availability.
    - For participants with multiple enrollment periods, information in the Prior file corresponds to the time before the first enrollment period.
- KPNC:
    - Lookback for enrollment, visits, and CRC screening tests extended to 01/01/1996 regardless of when the participant entered the cohort.
    - Lookback for CRC diagnoses and GI surgery extended beyond 01/01/1996 and used the full extent of information available in the VDW Tumor and Procedure tables.
    - For participants with multiple enrollment periods, information in the Prior file corresponds to the time before the first enrollment period.
- KPSC:
    - Lookback for all variables except FIT/gFOBT results extended as far back as data were available, the earliest being early 1989 for ambulatory outpatient visits. FIT/gFOBT result lookback extended through 1996.

Unless otherwise noted, lookback was not restricted to events that occurred during a period of continuous enrollment or engagement prior to cohort entry.

Data Sources

The three KP sites compiled this table using their local VDW Diagnosis, Encounter, Enrollment, Procedure, Provider Specialty, Social History, Tumor, and Vital Signs tables. KPNC and KPSC also used VDW Laboratory Results and EMR data to assess prior FIT/gFOBT; KPWA used FOBT data from EMR. KPWA incorporated adenoma results from NLP for the lower endoscopy LastResult variables, while KPNC and KPSC used SNOMED codes from pathology data to populate the LastResult variables.

UTSW used EMR data (incl. both discrete data fields and NLP of pathology reports) in addition to cancer registry data.

Variables

PID

Same as above.

MONTHSPRIORTOCOHORT

Definition	Months of continuous enrollment prior to cohort entry
Type (Length)	Numeric (4)
Valid Values	Integers 0–335 or -99999 (no prior enrollment)
General Notes	• Gap tolerance was ≤90 days. • Lookback period varied by site; see notes under Overview.
KPWA Notes	• Site included health plan coverage that occurred at any point during lookback, including ≤90-day gaps in eligible coverage (e.g., non-GPD but still insured).
UTSW Notes	• Site is utilization-based, so value is based on date of first encounter in the healthcare system as a proxy for enrollment prior to cohort entry.

PCPVISITLASTDATE

Definition	Date of last in-person primary care visit prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)

Valid Values	01/02/1993–12/30/2020 or null if not applicable
General Notes	• See Appendix A. Primary Care Visits for site-specific primary care visit definitions. • Lookback period varied by site; see notes under Overview. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPSC Notes	• Data needed to distinguish primary care from other outpatient clinic visits were not available prior to 2008.

PCPVisitLastDSR

Definition	Date of last in-person primary care visit prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	5605–35062 or null if not applicable
General Notes	• See additional information under PCPVisitLastDate. • Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.
KPSC Notes	• Data needed to distinguish primary care from other outpatient clinic visits were not available prior to 2008.

HCVisitLastDate

Definition	Date of last ambulatory outpatient visit within health care system prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	02/07/1980–12/30/2020 or null if not applicable
General Notes	• Lookback period varied by site; see notes under Overview. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

HCVisitLastDSR

Definition	Date of last ambulatory outpatient visit within health care system prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	4132–35062 or null if not applicable
General Notes	• See additional information under HCVisitLastDate. • Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.

FITPrior

Definition	Whether FIT/gFOBT was performed prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	• Lookback period varied by site; see notes under Overview. • Variable reflects last documented pre-entry FIT/gFOBT laboratory result regardless of outcome (e.g., including inadequate samples).

FITResultDate

Definition	Date of last FIT/gFOBT result prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/1996–12/30/2020 or null if not applicable
General Notes	• See additional information under FITPrior. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

FITResultDSR

Definition	Date of last FIT/gFOBT result prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	5662–35062 or null if not applicable
General Notes	• See additional information under FITPrior. • Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.

FITLastResult

Definition	Result of last FIT/gFOBT prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = Negative 01 = Positive 02 = No result 03 = Not performed 04 = Inadequate sample 95 = Other 99 = Unknown or not applicable

General Notes	• Logic used to populate this variable corresponds to the FITResult variable in the FITgFOBTResults table.

COLONOSCOPYPRIOR

Definition	Whether colonoscopy was performed prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	• Colonoscopies were identified using the harmonized CRC Screening & GI Surgery code list (where PRECISE_CAT = "Colonoscopy"). • If the actual procedure date was not available for an inpatient colonoscopy, the event was attributed to the inpatient admit date. • Lookback period varied by site; see notes under Overview.
KPWA Notes	• Site applied code list to in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

COLONOSCOPYDATE

Definition	Date of last colonoscopy prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/04/1990–12/30/2020 or null if not applicable
General Notes	• See additional information under ColonoscopyPrior. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

COLONOSCOPYDSR

Definition	Date of last colonoscopy prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	4961–35062 or null if not applicable
General Notes	• See additional information under ColonoscopyPrior. • Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.

CSPYLASTRESULT

Definition	Whether ≥1 adenoma was detected in the last colonoscopy prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• Variable is based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable in the Procedure Table.

SIGPRIOR

Definition	Whether sigmoidoscopy was performed prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	• Sigmoidoscopies were identified using the harmonized CRC Screening & GI Surgery code list (where PRECISE_CAT = "Sigmoidoscopy (incl. procto, rigid, flexible)"). • If the actual procedure date was not available for an inpatient sigmoidoscopy, the event was attributed to the inpatient admit date. • Lookback period varied by site; see notes under Overview.
KPWA Notes	• Site applied code list in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

SIGDATE

Definition	Date of last sigmoidoscopy prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/04/1990–12/30/2020 or null if not applicable
General Notes	• See additional information under SigPrior. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

SIGDSR

Definition	Date of last sigmoidoscopy prior to cohort entry (as days since reference, i.e., participant DOB)

Type (Length)	Numeric (4)
Valid Values	3888–35062 or null if not applicable
General Notes	• See additional information under SigPrior. • Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.

SIGLASTRESULT

Definition	Whether ≥1 adenoma was detected in the last sigmoidoscopy prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• Variable is based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable in the Procedure Table.
UTSW Notes	• Information was not available; all records set to 99 (Unknown).

LENDOPRIOR

Definition	Whether lower endoscopy NOS (not otherwise specified) was performed prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	• Lower endoscopies NOS were identified using the harmonized CRC Screening & GI Surgery code list (where PRECISE_CAT = “Lower endoscopy NOS”). • If the actual procedure date was not available for an inpatient lower endoscopy NOS, the event was attributed to the inpatient admit date. • Lookback period varied by site; see notes under Overview.
KPWA Notes	• Site applied code list to in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

LENDODATE

Definition	Date of last lower endoscopy NOS prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/10/1990–12/30/2020 or null if not applicable

General Notes	- See additional information under LEndoPrior. - Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	- All dates are set to day 15 of the given month/year.

LENDODSR

Definition	Date of last lower endoscopy NOS prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	4108–35062 or null if not applicable
General Notes	- See additional information under LEndoPrior. - Valid values reflect the extent of site-specific lookback periods through one day prior to end of age eligibility.

LENDOLASTRESULT

Definition	Whether ≥1 adenoma was detected in the last lower endoscopy NOS prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	- Variable is based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable in the Procedure Table.
UTSW Notes	- Information was not available; all records set to 99 (Unknown).

CRCINSITUEVER

Definition	Whether participant was diagnosed with in situ colorectal cancer prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	- Lookback period varied by site; see notes under Overview. - All data were sourced from tumor registries (as opposed to administrative or EMR data). - Variable captures in situ behavior in any tumor with primary site and histology that met the SEER site recode definition of CRC (recode values 21041–21052).

CRCPRIOR

Definition	Whether participant was diagnosed with colorectal cancer prior to cohort entry
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes
General Notes	• Lookback period varied by site; see notes under Overview. • All data were sourced from tumor registries (as opposed to administrative or EMR data). • Variable captures any tumor (regardless of malignant vs. in situ behavior) with primary site and histology that met the SEER site recode definition of CRC (recode values 21041–21052).

CRCPriorDate

Definition	Date of first CRC diagnosis prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	08/01/1955–12/30/2020 or null if not applicable
General Notes	• See additional information under CRCPrior. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

CRCPriorDSR

Definition	Date of first CRC diagnosis prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	6353–35062 or null if not applicable
General Notes	• See additional information under CRCPrior. • Valid values reflect extent of site-specific lookback periods through one day prior to end of age eligibility.

GISurgPrior

Definition	Whether participant ever had gastrointestinal (GI) surgery prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes
General Notes	• GI surgeries were identified using the harmonized CRC Screening & GI Surgery code list (where PRECISE_CAT = “Colectomy NOS,” “Colectomy partial,” “Colectomy total,” “Proctectomy”, or “Colectomy history”). ○ Note: The codes for “Colectomy history” are ICD-10-CM diagnosis codes; all others are procedure codes. • If the actual procedure or diagnosis date was not available for an inpatient code, the event was attributed to the inpatient admit date. • Lookback period varied by site; see notes under Overview.

KPWA Notes	• Site applied Px code list to in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters), whereas the Colectomy history Dx codes could be associated with both in-person encounters and selected virtual care encounters (i.e., scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

GISurg1stDate

Definition	Date of first GI surgery (or documentation thereof) prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/02/1990–12/30/2020 or null if not applicable
General Notes	• See additional information under GISurgPrior. • Valid values are based on maximum lookback through one day prior to end of cohort period.
UTSW Notes	• All dates are set to day 15 of the given month/year.

GISurgDSR

Definition	Date of first GI surgery (or documentation thereof) prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	5111–35062 or null if not applicable
General Notes	• See additional information under GISurgPrior. • Valid values reflect extent of site-specific lookback periods through one day prior to end of age eligibility.

CharlsonPrior

Definition	Charlson comorbidity score for the 365 days prior to date of cohort entry
Type (Length)	Numeric (4)
Valid Values	Integers 0–29 (valid scores) or -99999 (insufficient enrollment -or- no eligible encounters during window [UTSW]); see site-specific notes for more details
General Notes	• See notes below and Appendix C. Charlson Comorbidity Index for information on site-specific differences in calculating the Charlson score. • Variable was set to -99999 at all sites if MonthsPriorToCohort < 12.

KPWA Notes	• Site included codes from in-person encounters only. • Variable was set to 0 if the participant was enrolled for the full 365-day observation window but had no inpatient or ambulatory encounters in which to observe relevant billing codes.
UTSW Notes	• Site included codes regardless of encounter type. • Variable was additionally set to -99999 if the participant had no encounters in which to observe relevant billing codes during the 365 days prior to cohort entry.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

SMOKEPRIOR

Definition	Participant's last known smoking status prior to cohort entry
Type (Length)	Character (2)
Valid Values	00 = Never smoked 01 = Ever smoked 99 = Unknown
General Notes	• Calculated using the latest single data point collected prior to cohort entry. • If the last data point prior to cohort entry indicated that the participant was a current or former smoker, SmokePrior = 01 (Ever smoked). • If the last data point prior to cohort entry indicated that the participant had never been a smoker, SmokePrior = 00 (Never smoked). • Otherwise (including if no prior smoking information was available), SmokePrior = 99 (Unknown).

WEIGHTPRIOR

Definition	Participant's last known weight in pounds prior to cohort entry
Type (Length)	Numeric (4)
Valid Values	Integers 50–600 (valid weights) or -99999 (missing or invalid weights)
General Notes	• Value represents valid weight in pounds (50–600) obtained as close to but prior to cohort entry, if available.

INDIBDLASTDATE

Definition	Date when participant was last diagnosed with inflammatory bowel disease (IBD) or sequelae prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/02/1980–12/30/2020 or null if not applicable

General Notes	• This variable is related to rule 2 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • IBD is defined here as the appearance of any Dx codes where IBD = 1 in the harmonized Relevant Symptoms & Conditions code list. • If the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDIBDLASTDSR

Definition	Date when participant was last diagnosed with inflammatory bowel disease (IBD) or sequelae prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	82–35062 or null if not applicable
General Notes	• See additional information under IndIBDLastDate.

INDSXLASTDATE

Definition	Date when participant was last diagnosed with a relevant symptom prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	12/30/1980–12/30/2020 or null if not applicable
General Notes	• This concept is related to rule 3 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • If the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date. • See site-specific implementation notes below.
KPWA Notes	• “Relevant symptoms” are defined here as the appearance of any Dx codes where SYMPTOMS = 1 in the harmonized Relevant Symptoms & Conditions code list. • Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).

UTSW Notes	• "Relevant symptoms" are defined here as the appearance of any Dx codes where SYMPTOMS = 1 in the harmonized Relevant Symptoms & Conditions code list. • Site applied code list regardless of encounter type. • All dates are set to day 15 of the given month/year.
KPNC Notes	• Variable incorporates data from multiple sources: ○ Dx codes where SYMPTOMSNOANEMIA = 1 in the harmonized Relevant Symptoms & Conditions code list, *and/or* ○ laboratory-based diagnoses of iron-deficiency anemia. • For Dx codes, site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats). • Laboratory-based diagnoses were identified as follows: ○ Identify abnormally low hemoglobin or hematocrit laboratory test results. ○ Look for iron-deficiency anemia diagnostic tests with abnormal results (i.e., low ferritin, low iron, low transferrin saturation ratio, low transferrin % saturation, high total iron binding capacity) on or within 90 days before or after the low hemoglobin or hematocrit result. ○ If both abnormal results are found, IDA diagnosis is set to the date of the *diagnostic test* result.
KPSC Notes	• Same as KPNC.

INDSXLASTDSR

Definition	Date when participant was last diagnosed with a relevant symptom prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	493–35062 or null if not applicable
General Notes	• See additional information under IndSxLastDate.

INDPXLASTDATE

Definition	Date when participant last had a relevant non-endoscopic, non-surgical procedure prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	09/15/1981–12/30/2020 or null if not applicable

General Notes	• This concept is related to rule 4 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ Dx codes where RECENTPX = 1 in the harmonized Relevant Symptoms & Conditions code list, *and/or* ○ Px codes from the harmonized CRC Screening & GI Surgery code list where PRECISE_CAT = “Abdominal CT”, “Barium enema”, or “CT colonography”. • If the actual diagnosis or procedure date was not available for an inpatient event, the event was attributed to the inpatient admit date.
KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats), whereas Px codes were restricted to in-person encounter types only.
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied code lists regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IndPxLastDSR

Definition	Date when participant last had a relevant non-endoscopic, non-surgical procedure prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	3499–35062 or null if not applicable
General Notes	• See additional information under IndPxLastDate

IndCRCDxLastDate

Definition	Date when participant last had a “recent CRC” diagnosis prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	12/01/1952–12/30/2020 or null if not applicable
General Notes	• This concept is related to rule 5 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ ICD-9/-10-CM codes where RECENTCRC = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ ICD-O-3 codes where RECENTCRC = 1 column of the harmonized Relevant Symptoms & Conditions code list, to be obtained from cancer registry data. • When using ICD-9/-10-CM codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.

KPWA Notes	• Site applied ICD-9/-10-CM code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied ICD-9/-10-CM code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDCRCDXLASTDSR

Definition	Date when participant last had a “recent CRC” diagnosis prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	3996–35062 or null if not applicable
General Notes	• See additional information under IndCRCDxLastDate.

INDCRCHXLASTDATE

Definition	Date when participant last had a “history of CRC” diagnosis prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	12/01/1952–12/30/2020 or null if not applicable
General Notes	• This variable is related to rule 9 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ ICD-9/-10-CM codes where HXCRC = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ ICD-O-3 codes where HXCRC = 1 column of the harmonized Relevant Symptoms & Conditions code list, to be obtained from cancer registry data. • When using ICD-9/-10-CM codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied ICD-9/-10-CM code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied ICD-9/-10-CM code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDCRCHXLASTDSR

Definition	Date when participant last had a “history of CRC” diagnosis prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	3358–35062 or null if not applicable
General Notes	• See additional information under IndCRCHxLastDate.

INDPOLYPDXLASTDATE

Definition	Date when participant last had a “recent colorectal polyp” diagnosis prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/27/1981–12/30/2020 or null if not applicable
General Notes	• This concept is related to rule 6 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates information from multiple sources: ○ the appearance of any Dx codes where RECENTPOLYP = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ adenoma(s) detected at prior-to-entry colorectal procedures, based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable in the Procedure Table. • When using Dx codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied Dx code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDPOLYPDXLASTDSR

Definition	Date when participant last had a “recent colorectal polyp” diagnosis prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	2663–35062 or null if not applicable
General Notes	• See additional information under IndPolypDxLastDate.

INDPOLYPHXLASTDATE

Definition	Date when participant last had a “history of colorectal polyp” diagnosis prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/27/1981–12/30/2020 or null if not applicable
General Notes	• This concept is related to rule 11 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates information from multiple sources: ○ the appearance of any Dx codes where HXPOLYP = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ adenoma(s) detected at prior-to-entry colorectal procedures, based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable in the Procedure Table. • When using Dx codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied Dx code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDPOLYPHXLASTDSR

Definition	Date when participant last had a “history of colorectal polyp” diagnosis prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	2663–35062 or null if not applicable
General Notes	• See additional information under IndPolypHxLastDate.

INDHEREDLASTDATE

Definition	Date when participant last had a diagnosis of hereditary colorectal cancer syndrome prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	09/30/2005–12/30/2020 or null if not applicable

General Notes	- This concept is related to rule 10 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. - Conditions included here are FAP (familial adenomatous polyposis) and HNPCC (hereditary nonpolyposis colorectal cancer).
KPWA Notes	- Information not available at this site; all records set to missing.
UTSW Notes	- Same as KPWA.
KPNC Notes	- Site used local Clarity Dx codes to identify FAP and HNPCC. - This variable includes all historic tentative Lynch syndrome diagnoses, including those that were not confirmed or were later proved wrong by genetic testing. - Actual diagnosis dates were not available for inpatient diagnoses, which were therefore attributed to the encounter admit date.
KPSC Notes	- Same as KPWA.

INDHEREDLASTDSR

Definition	Date when participant last had a diagnosis of hereditary colorectal cancer syndrome prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	9704–35062 or null if not applicable
General Notes	- See additional information under IndHeredLastDate.

INDBACKOFCLASTDATE

Definition	Date when participant last had a negative back-office FOBT result prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	04/01/2005–12/30/2020 or null if not applicable
General Notes	- This concept is related to rule 7 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. - Variable should incorporate prior-to-entry FIT/gFOBT results where back-office setting and negative results are defined using the same logic employed in the FITgFOBTResults Table.
UTSW Notes	- Site was unable to identify back-office FIT/gFOBT results; all records set to missing.
KPSC Notes	- Variable needed to identify back-office procedures was not available in EMR prior to 2007.

INDBACKOFCLASTDSR

Definition	Date when participant last had a negative back-office FOBT result prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)

Valid Values	8953–35062 or null if not applicable
General Notes	• See additional information under IndBackOfcLastDate.

IBDOnlyLastDate

Definition	Date when participant was last diagnosed with inflammatory bowel disease (IBD), excluding sequelae and other/unspecified noninfective gastroenteritis/colitis, prior to cohort entry (as SAS date)
Type (Length)	Numeric (4)
Valid Values	12/20/1980–12/30/2020 or null if not applicable
General Notes	• IBD (excluding sequelae and other/unspecified gastroenteritis/colitis) is defined here as the appearance of any Dx codes where IBDONLY = 1 and code is NOT ICD-9-CM 558.9 and code is NOT ICD-10-CM K52.9 in the harmonized Relevant Symptoms & Conditions code list. • If the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IBDOnlyLastDSR

Definition	Date when participant was last diagnosed with inflammatory bowel disease (IBD), excluding sequelae and other/unspecified noninfective gastroenteritis/colitis, prior to cohort entry (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	3426–35062 or null if not applicable
General Notes	• See additional information under IBDOnlyLastDate.

ExtractDate

Same as above.

ProvidingSite

Same as above.

Enrollment Table

Overview

This table contains one row per period of continuous, cohort-eligible health plan enrollment (KP sites) or primary care utilization lasting at least 36 months (UTSW). Enrollment/utilization gaps of ≤90 days are patched, but all other eligibility criteria must be met; see the Participant Table Overview for more information.

At KPWA, as mentioned under CohortEntryDate, cohort enrollment was limited to a single continuous, cohort-eligible period per participant. As such, all KPWA participants have exactly one record in this table.

KPNC, KPSC, and UTSW allowed for cohort exit and re-entry; therefore, a participant from any of these sites may have more than one record in this table.

With respect to UTSW's cohort: Across each PRC, for members with primary care-based cohort enrollment, follow-up was truncated 3 years after the last primary care visit or screening event. This truncation was applied to limit inflation of the denominator for cohort members that are likely to no longer be receiving primary care within the system and for whom we may have incomplete poor capture of longitudinal outcomes. Primary care-based cohort members, whose follow-up has been truncated due to lack of primary care or screening utilization, may subsequently re-enter the cohort on the date of a new primary care visit or screening event. This is denoted as a new enrollment period for that member. Fewer than 5% of those members, who have follow-up truncated due to a 3-year lapse in primary care or screening utilization, have a second enrollment period.

Data Sources

The three KP sites used their local VDW Death, Demographic, and Enrollment tables to populate this table.

UTSW used EMR and cancer registry data.

Variables

PID

Same as above.

COHORTENTRYDATE

Definition	Start of enrollment period (as SAS date)

Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/30/2020
General Notes	• A variable with the same name exists in the Participant Table. • Valid values reflect requirement that each participant must spend at least one day in the cohort.
UTSW Notes	• All dates are set to day 15 of the given month/year.

COHORTENTRYDSR

Definition	Start of enrollment event (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35062
General Notes	• Valid values reflect participant being 40–95 years old during cohort eligibility as well as the requirement that each participant spend at least one day in the cohort.

CUTOFFREASON

Definition	Reason for end of enrollment period
Type (Length)	Character (2)
Valid Values	01 = Administrative cutoff 02 = Age out 03 = Disenrollment, utilization based (UTSW only) 04 = Disenrollment, membership based (KP sites only) 05 = Move out of coverage area 06 = Disenrollment due to hospice (optional) 07 = Death where underlying cause is CRC 08 = Death where underlying cause is not CRC or unknown 09 = Death, non-specific 95 = Other 99 = Unknown
General Notes	• See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites. • For records where CutoffReason indicates death, more information about cause of death can be found in the COD Table.
UTSW Notes	• Site coded 05 (Move out of coverage area) if EMR address history indicated that participant had lived outside of Dallas County for >6 months, as medical assistance programs at PHHS are only eligible to Dallas County residents.

CUTOFFREASONOTHER

Definition	Other reason for end of PRECISE enrollment
Type (Length)	Character (1)

Valid Values	Missing (blank) at all sites
General Notes	- Newly requested by PCC for IMS3, to be populated when CutoffReason = 95 (Other); not applicable for any PRECISE sites.

CUTOFFDATE

Definition	End of enrollment period (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/02/2010–12/31/2020
General Notes	- See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites. - A variable with the same name exists in the Participant Table. - Valid values reflect requirement that each participant must spend at least one day in the cohort.
UTSW Notes	- All dates are set to day 15 of the given month/year.
KPSC Notes	- A comparatively high proportion of cutoff dates occur on the 1st of the month due to how membership data is handled in local database ETL process.

CUTOFFDSR

Definition	End of enrollment period (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14611–35063
General Notes	- See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites. - Valid values reflect participant being 40–95 years old during cohort eligibility as well as the requirement that each participant spend at least one day in the cohort.

ENGAGETYPE

Definition	Type of engagement
Type (Length)	Char (1)
Valid Values	0 = Health plan membership 1 = Primary care utilization
KPWA Notes	- All records set to 0 (Health plan membership).
UTSW Notes	- All records set to 1 (Primary care utilization).
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as KPWA.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

FITgFOBTResults Table

Overview

This table contains one row per documented fecal occult blood test (FOBT) result for PRECISE participants during cohort eligibility. (Note: KPNC and KPSC also included results that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort events.) Two subtypes of FOBT are relevant to PRECISE:

- Fecal Immunochemical Test (FIT, sometimes called iFOBT), and
- Guaiac Fecal Occult Blood Test (gFOBT).

Some records may appear to be duplicates, because sites maintained these data at the level of the records extracted from laboratory source systems: i.e., we did not attempt to determine whether multiple same-day records were unique results from distinct specimens or merely duplicative. With this in mind, analysts may wish to deduplicate records in this file as appropriate for the given use case. For example, when looking at days with testing, analysts could deduplicate to the level of ProvidingSite-PID-FITResultDSR.

Data Sources

KPWA sourced FOBT results directly from EMR data. Only internally ordered and resulted tests are available in this system: i.e., FOBTs resulted by non-KPWA laboratories are not available.

UTSW used EMR data.

KPNC used their local VDW Laboratory Result and Procedure tables to build this table.

KPSC used data from EMR.

Variables

PID

Same as above.

TESTTYPE

Definition	FOBT type

Type (Length)	Character (2)
Valid Values	01 = FIT 02 = gFOBT
KPWA Notes	• Site identified EMR procedures associated with known FOBT procedure codes (CPT 82270, 82272, and 82274; HCPCS G0107, G0328, G0394, and G0464) as well as those with descriptions that referenced relevant keywords; more information available upon request. • Based on manual review of the code list, site assigned TestType = 01 (FIT) if the procedure code was G0328 or if the procedure name contained the strings “FIT,” “IFOB,” “POLYMEDCO,” or “INSURE.” All other results were assigned 02 (gFOBT).
UTSW Notes	• UTSW derived this variable from CPT and local Clarity* codes, as follows: ○ 01 = FIT (CPT 82274, LAB310111*, LAB310180*) ▪ LAB310111 is a Parkland-specific Clarity code to indicate a FIT completed through the site’s mailed outreach program. All FITs offered in this program are one-sample Polymedco OC-Auto FITs. ▪ LAB310180 is a Parkland-specific Clarity code to indicate Polymedco OC-Auto FITs. ○ 02 = gFOBT (CPT 82270, 82272, or 89055) ▪ 89055 indicates diagnostic gFOBTs.
KPNC Notes	• Site identified lab test results associated with known gFOBT codes (VDW TEST_TYPE "FOB_GUAI" and "FOB1_GUAI") and FIT codes (TEST_TYPE "FOB_IMMUN"). • Site assigned TestType = 01 (FIT) or 02 (gFOBT) according to the above list of codes.
KPSC Notes	• Tests with CPT code 82274 were FITs assigned TestType = 01, and tests with CPT codes 82270 or 82272 were gFOBTs assigned TestType = 02.

TestSetting

Definition	FOBT setting
Type (Length)	Character (2)
Valid Values	01 = Back-office 02 = Inpatient 95 = Other 99 = Unknown
KPWA Notes	• Site parsed the EMR procedure description to populate this variable as follows: ○ If procedure description contains “(OVL)” [reference to Overlake Hospital] or starts with “RN” [reference to RN-collected samples at Central Hospital] → 02 (Inpatient) ○ Else if procedure description contains “BACK” and “OFFICE” → 01 (Back-office) ○ Otherwise → 99 (Unknown) • KPWA did not map any results to 95 (Other) due to a lack of confidence that all inpatient and back-office tests are captured with the logic above.
UTSW Notes	• UTSW first used encounter location to assign values of 02 (Inpatient). All other tests were coded as 95 (Other). • Note: The 01 (Back office) concept does not apply at UTSW.

KPNC Notes	• Site examined FOBT lab test data in conjunction with same-day procedure records with CPT code 82272 (a diagnostic fecal occult blood test) to identify tests done at office visits or inpatient exams as follows: ○ If TestType = 01 (FIT), result assumed to be from screening outreach program → 95 (Other) ○ Else if patient location was hospital or emergency room setting → 02 (Inpatient) ○ Else if code 82272 was found → 01 (Back-office) ○ Otherwise → 99 (Unknown)
KPSC Notes	• EMR variables order_mode (inpatient or outpatient) and order_class_c (back office, clinic collected, normal) were used to assign to TestSetting as follows: ○ If order_mode is inpatient → 02 (Inpatient) ○ Else if order_mode is outpatient and order_class_c indicates back office or clinic collected → 01 (Back-office) ○ Otherwise (e.g., mailed tests) → 95 (Other)

FITorFOBTCodes

Definition	List of FIT/gFOBT codes associated with the result
Type (Length)	Character (*)
Valid Values	Space-delimited list of laboratory test codes
KPWA Notes	• In KPWA's EMR, each FOBT result is associated with a single original order with a single relevant procedure code. Therefore, KPWA records do not contain any space-delimited lists in this field; site provides only a single code for each record.
UTSW Notes	• Test codes include CPT codes 82274, 82272, 82270, and 89055 as well as LAB310111 and LAB310180 (Parkland-specific Clarity codes).
KPNC Notes	• Site provided multiple space-delimited codes if multiple tests were performed.
KPSC Notes	• Value represents the CPT code linked to the test from the original data source.

FITCardQuant

Definition	Number of cards tested
Type (Length)	Numeric (4)
Valid Values	Integers ≥1 (known quantities) or -99999 (unknown)
KPWA Notes	• Site quantified the number of tested cards that yielded a useable (i.e., positive or negative) result. If no cards yielded a useable result, this variable was set to -99999.
UTSW Notes	• Site assigned values based on those reported within the test order.
KPNC Notes	• Site determined the number of tested cards based on number of records with distinct component values for specimen ID. • If tests for more than one specimen were performed on the same day, the maximum number of tested cards was selected.

KPSC Notes	• For FOBTs, number of cards tested is based on the number of distinct component values (e.g., Occult blood 1, stool, Occult, blood 2, stool, etc.) for the same order ID. • For FITs, all tests have FITCardQuant = 1.

FITRESULTDATE

Definition	Date of FOBT result (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020
UTSW Notes	• All dates are set to day 15 of the given month/year. • Site determined result date based on a combination of result date associated with the order and date of service associated with the transaction.
KPNC Notes	• Site included FIT/gFOBT results that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

FITRESULTDSR

Definition	Date of FOBT result (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063
General Notes	• Valid values reflect participant being 40–95 years old during cohort eligibility.
KPNC Notes	• Site included FIT/gFOBT results that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

FITRESULT

Definition	Result of FOBT
Type (Length)	Character (2)
Valid Values	00 = Negative 01 = Positive 02 = No result 03 = Not performed 04 = Inadequate sample 95 = Other 99 = Unknown

KPWA Notes	• Site summarized specimen/card-level results as follows: ○ If ≥1 positive specimen → 01 (Positive) ○ Else if ≥1 negative specimen & 0 unknown results → 00 (Negative) ○ Else if ≥1 unknown result → 99 (Unknown) ○ Else if 0 results → 02 (No result)
UTSW Notes	• Site assigned test results as follows: ○ ≥1 card has positive result → 01 (Positive) ○ Else if ≥1 card has negative result → 00 (Negative) ○ Else if ≥1 card result is unsatisfactory → FITResult = 04 (Unsatisfactory) ▪ "Unsatisfactory" includes results described as broken or leaking container, incomplete specimen label, unsatisfactory specimen, specimen collection error, specimen too old, and expired cards ○ Otherwise → FITResult = 99 (Unknown)
KPNC Notes	• Site summarized specimen/card-level results as follows: ○ ≥1 positive result → 01 (Positive) ○ Else if ≥1 negative result → 00 (Negative) ○ Else if ≥1 "INADEQUATE" result & no positive or negative results → 04 (Inadequate sample) ○ Otherwise → 95 (Other)
KPSC Notes	• For gFOBT: ○ If at least 1 card is positive for the same test order → FITResult = 01 (Positive) ○ Else if all the cards are negative → FITResult = 00 (Negative) ○ Otherwise → FITResult = 95 (Other) • For FIT: ○ If the test result indicates positive → FITResult = 01 (Positive) ○ Else if the test result indicates negative → FITResult = 00 (Negative) ○ Otherwise → FITResult = 95 (Other)

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Procedure Table

Overview

The Procedure table contains one row per participant per relevant procedure type per day during cohort eligibility. (Note: KPNC and KPSC also included procedures that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.)

As of the IMS2+ data submissions, the procedures of interest are lower endoscopies; other colorectal cancer screening tests (barium enema, CT colonography); colectomies (and history thereof); proctectomies; and abdominal CT scans.

Note that some variables in this table apply only to colonoscopies, which are of particular interest for many PRECISE analyses.

ADDITIONAL INFORMATION RE: KPWA

KPWA's delivery system is unique among the PRECISE sites in that colonoscopies are performed by providers in three primary "buckets": 1) KPWA's internal delivery system (IDS) providers; 2) providers with DHS/WAGI (Digestive Health Specialists, 2010–2017; Washington Gastroenterology, 2018–2020)[2]; and 3) a variety of other external delivery system (EDS) providers. KPWA laboratories result all pathology for procedures performed by KPWA providers. During 2010–2017, most DHS-performed procedures were also resulted by KPWA laboratories. Otherwise, external laboratories result any pathology associated with externally performed colonoscopies. This mixed model (IDS + EDS) results in differentially available procedure- and pathology-level data for colonoscopies that occur within the KPWA cohort.

For example, for KPWA-performed colonoscopies (BucketKPWA = 1), 65% had a specimen sent to pathology and 35% did not. For colonoscopies performed at DHS/WAGI (BucketKPWA = 2), 56% had a specimen sent to pathology and 36% were unknown; only 8% of DHS/WAGI colonoscopies could be classified as "no specimen sent" (ascertained via limited chart abstraction efforts). For other external providers (BucketKPWA = 3), 62% of colonoscopies had a specimen sent to pathology, while the remaining 38% were unknown. This difference partially explains why the difference in adenoma quantities (particularly 0 vs. unknown) differs based on where the colonoscopy was performed.

Data Sources

KPWA procedures were primarily sourced from the VDW Procedure table, limited to relevant procedure codes and in-person encounters. For KPWA's internally performed colonoscopies (all years) and DHS-performed colonoscopies (2010–2017), information about indication, colon preparation, and extent of the procedure were extracted via NLP (Natural Language Processing) from the text of the procedure report obtained from EMR. Pathology report text was also available for KPWA-performed colonoscopies (all years) and a subset of DHS-performed (2010–2017) colonoscopies; NLP algorithms were used to translate this free text into structured variables for pathology outcomes. In addition, KPWA chart-abstracted DHS-performed (2010–2017) colonoscopies with evidence of a sample sent to an external pathology laboratory (defined as a same-day surgical pathology CPT code) as well as all WAGI-performed colonoscopies (2018–2020). The variables BucketKPWA, DataSourceKPWA, and ProcSampleSentPathology can be used to distinguish among these groups. See other variable details below for additional information.

UTSW procedures were sourced from EMR transaction, claims, and procedure data. For colonoscopy results, information including procedure indication, colon preparation, and extent of procedure were extracted via NLP

---

[2] Beginning with 2018 data, KPWA's bucket 2 "WAGI" procedures encompassed not just the practice formerly known as Digestive Health Specialists but also the former Northwest Gastroenterology Associates and Overlake Internal Medicine Associates – Gastroenterology.

from the procedure and/or pathology reports in the EMR. NLP algorithms were used to translate this free text into structured variables for pathology outcomes. See variable details below for additional information.

KPNC/KPSC procedures were primarily sourced from the VDW Procedure table, limited to relevant procedure codes and in-person encounters. For colonoscopies, information about indication was calculated based on the KP modified indication algorithm, which includes data from various data sources (see Appendix D. KP Indication Algorithm). For internally performed colonoscopies, information about colon preparation and extent of the procedure were extracted via NLP from the text of the procedure report in the EMR. Pathology report text was also available for internally performed colonoscopies, sigmoidoscopies, and lower endoscopies NOS; NLP algorithms were used to translate this free text into structured variables for pathology outcomes. See variable details below for additional information. When available, pathology outcomes were based on SNOMED (Systemized Nomenclature of Medicine) codes obtained from pathology data sources; see Appendix B. Code Lists for Pathology code lists.

Variables

PID

Same as above.

PROCTYPE

Definition	Type of procedure
Type (Length)	Character (2)
Valid Values	01 = Colonoscopy 02 = Sigmoidoscopy 03 = Barium enema 04 = CT colonography 05 = History of colectomy 06 = Colectomy, partial 07 = Colectomy, total 08 = Colectomy, NOS 09 = Proctectomy 10 = Lower endoscopy, NOS 11 = Abdominal CT
General Notes	• Procedure types above correspond directly to PRECISE_CAT values in the harmonized CRC Screening & GI Surgery code list.
KPWA Notes	• For all procedure types other than 05 (History of colectomy), the underlying procedure codes must have been associated with in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters). • For ProcType 05, the underlying diagnosis codes could be associated with both in-person encounters and selected virtual care encounters (i.e., scheduled telephone calls, scheduled video visits, and synchronous online chats). • Site deduplicated lower endoscopies to the level of one procedure per participant per day using the following hierarchy: 01 (Colonoscopy) > 02 (Sigmoidoscopy) > 10 (Lower endoscopy, NOS).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• For all procedure types other than 05 (History of colectomy), the underlying procedure codes must have been associated with in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters). • For ProcType 05, the underlying diagnosis codes could be associated with both in-person encounters and selected virtual care encounters (i.e., scheduled telephone calls, scheduled video visits, and synchronous online chats). • Site deduplicated non-surgical procedures to the level of one procedure per participant per day using the following hierarchy: 01 (Colonoscopy) > 02 (Sigmoidoscopy) > 10 (Lower endoscopy, NOS) > 03 (Barium enema) > 04 (CT colonography) > 11 (Abdominal CT). • Site deduplicated surgical procedures to the level of one procedure per participant per day using the following hierarchy: 07 (Colectomy, total) > 06 (Colectomy, partial) > 09 (Proctectomy) > 08 (Colectomy, NOS) > 05 (History of colectomy).
KPSC Notes	• Same as KPNC.

PROCCODE

Definition	Procedure code(s) observed for this procedure date/type

Type (Length)	Character (*)
Valid Values	Space-delimited list of procedure codes
General Notes	• All codes should appear in the harmonized CRC Screening & GI Surgery code list.

PROCDATE

Definition	Date of procedure (as SAS date value)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020
General Notes	• If the actual procedure or diagnosis date was not available for an inpatient code, the event was attributed to the inpatient admit date.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPNC Notes	• Site included procedures that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort procedures.
KPSC Notes	• Same as KPNC.

PROCDSR

Definition	Date of procedure (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063
General Notes	• If the actual procedure or diagnosis date was not available for an inpatient code, the event was attributed to the inpatient admit date. • Valid values reflect participant being 40–95 years old during cohort eligibility.
KPNC Notes	• Site included procedures that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort procedures.
KPSC Notes	• Same as KPNC.

PROCSETTING

Definition	Whether colonoscopy or lower endoscopy NOS was performed in outpatient setting
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• This variable only applies to colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records set to 99. • Variable reflects administrative or EMR data specific to the procedure, not inferred from other sources (e.g., concurrent hospitalization).

PROCINDIC

Definition	Indication for colonoscopy or lower endoscopy NOS
Type (Length)	Character (2)
Valid Values	02 = Diagnostic 03 = Surveillance 04 = Screening 06 = High-risk surveillance 07 = Polyp/adenoma surveillance 08 = Non-screening, NOS 99 = Unknown (indication not known) or not applicable (invalid ProcType)
General Notes	• This variable applies to colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • PRECISE recommends that analysts use this variable (instead of ProcIndicKPMod).
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); all other records were set to 99 (Unknown). See Appendix E. Natural Language Processing / NLP at KPWA for more information. • Site did not map any records to codes 06 or 07.
UTSW Notes	• This information was only available for colonoscopies; all other procedure records were set to 99. • Site applied NLP algorithms to text of colonoscopy procedure reports found in the EMR. Manuscript describing details of NLP is in preparation. • Site did not map any records to codes 06, 07, or 08.
KPNC Notes	• Site used the KP modified indication algorithm (described further under ProcIndicKPMod and Appendix D. KP Indication Algorithm) to populate this variable, with the following revision: When assessing history of inflammatory bowel disease, ICD-9-CM code 558.9 and ICD-10-CM code K52.9 were included only when they appeared within 365 days prior to the procedure in question. • Site did not map any records to codes 03 or 08.
KPSC Notes	• Same as KPNC.

INDANYECTOMY

Definition	KP indication algorithm flag: Any history of colectomy or proctectomy
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Not applicable (invalid ProcType)

General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant underwent either of the following gastrointestinal surgeries ***at any time prior to*** the current procedure: ○ Colectomy ○ Proctectomy • There is no minimum requirement for prior enrollment/engagement; therefore, all colonoscopies and lower endoscopies NOS should have this variable populated as 00 or 01. • This variable is related to rule 1 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Colectomy and proctectomy are defined here as the appearance of any codes in the harmonized CRC Screening & GI Surgery code list where PRECISE_CAT = "Colectomy total", "Colectomy partial", "Colectomy NOS", "Colectomy history", or "Proctectomy". ○ Note: The codes for "Colectomy history" are ICD-10-CM diagnosis codes; all others are procedure codes. • If the actual procedure or diagnosis date was not available for an inpatient code, the event was attributed to the inpatient admit date.
KPWA Notes	• Site applied Px code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters), whereas the Colectomy history Dx codes could be associated with both in-person encounters and selected virtual care encounters (i.e., scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDANYIBD

Definition	KP indication algorithm flag: Any history of inflammatory bowel disease (IBD) diagnosis, including sequelae
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Not applicable (invalid ProcType)

General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant had a history of inflammatory bowel disease (IBD) ***at any time prior to*** the current procedure. • There is no minimum requirement for prior enrollment/engagement; therefore, all colonoscopies and lower endoscopies NOS should have this variable populated as 00 or 01. • Inflammatory bowel disease includes Crohn's disease and ulcerative colitis (and their sequelae); other forms of colitis are not part of IBD. • This variable is related to rule 2 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • IBD is defined here as the appearance of any Dx codes where IBD = 1 in the harmonized Relevant Symptoms & Conditions code list. • If the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IndRecentSx

Definition	KP indication algorithm flag: Recent diagnosis of relevant symptom(s)
Type (Length)	Character (2)
Valid Values	00 = No (with ≥180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<180 days enrollment/engagement) or not applicable (invalid ProcType)

General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with any of the following symptoms ***≤180 days prior to*** the current procedure: ○ Abdominal pain ○ Anemia (including iron deficiency and some unspecified anemias) ○ GI bleeding or blood in stools ○ Diarrhea ○ Weight loss or underweight ○ Diverticulitis ○ Constipation ○ Abdominal mass ○ Change in bowel habits • This concept is related to rule 3 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • If the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date. • "Relevant symptoms" were defined somewhat differently across sites; see site-specific notes below.
KPWA Notes	• "Relevant symptoms" are defined here as the appearance of any Dx codes where SYMPTOMS = 1 in the harmonized Relevant Symptoms & Conditions code list. • Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• "Relevant symptoms" are defined here as the appearance of any Dx codes where SYMPTOMS = 1 in the harmonized Relevant Symptoms & Conditions code list. • Site applied code list regardless of encounter type.
KPNC Notes	• Variable incorporates data from multiple sources: ○ Dx codes where SYMPTOMSNOANEMIA = 1 in the harmonized Relevant Symptoms & Conditions code list, *and/or* ○ laboratory-based diagnoses of iron-deficiency anemia. • Site applied code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats). • Laboratory-based diagnoses were identified as follows: ○ Identify abnormally low hemoglobin or hematocrit laboratory test results. ○ Look for iron-deficiency anemia diagnostic tests with abnormal results (i.e., low ferritin, low iron, low transferrin saturation ratio, low transferrin % saturation, high total iron binding capacity) on or within 90 days before or after the low hemoglobin or hematocrit result. ○ If both abnormal results are found, IDA diagnosis is set to the date of the *diagnostic test* result.
KPSC Notes	• Same as KPNC.

INDRECENTPX

Definition	KP indication algorithm flag: Recent relevant procedure
Type (Length)	Character (2)
Valid Values	00 = No (with ≥180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<180 days enrollment/engagement) or not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant underwent either of the following procedures ***≤180 days prior to*** the current procedure: ○ Barium enema ○ CT colonography ○ Abdominal CT • This concept is related to rule 4 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ Dx codes where RECENTPX = 1 in the harmonized Relevant Symptoms & Conditions code list, *and/or* ○ Px codes from the harmonized CRC Screening & GI Surgery code list where PRECISE_CAT = “Abdominal CT”, “Barium enema”, or “CT colonography”. • If the actual diagnosis or procedure date was not available for an inpatient event, the event was attributed to the inpatient admit date.
KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats), whereas Px codes were limited to those associated with in-person encounter types only.
UTSW Notes	• Site applied code lists regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDRECENTCRC

Definition	KP indication algorithm flag: Recent diagnosis of colorectal cancer
Type (Length)	Character (2)
Valid Values	00 = No (with ≥180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<180 days enrollment/engagement) or not applicable (invalid ProcType)

General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with colorectal cancer ***≤180 days prior to*** the current procedure. • This concept is related to rule 5 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ ICD-9/-10-CM codes where RECENTCRC = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ ICD-O-3 codes where RECENTCRC = 1 column of the harmonized Relevant Symptoms & Conditions code list, to be obtained from cancer registry data. • When using ICD-9/-10-CM codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied ICD-9/-10-CM code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• Site applied ICD-9/-10-CM code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IndRecentPolyp

Definition	KP indication algorithm flag: Recent diagnosis of adenomatous or other colorectal polyp(s)
Type (Length)	Character (2)
Valid Values	00 = No (with ≥365 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<365 days enrollment/engagement) or not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with colorectal adenoma(s) or other colorectal polyp(s) ***≤365 days prior to*** the current procedure. • This concept is related to rule 6 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates information from multiple sources: ○ the appearance of any Dx codes where RECENTPOLYP = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ adenoma(s) detected at prior colorectal procedures, based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable described later in this table. • When using Dx codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.

KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats)
UTSW Notes	• Site applied Dx code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDRECENTNEGBO

Definition	KP indication algorithm flag: Recent negative FOBT result from back-office procedure
Type (Length)	Character (2)
Valid Values	00 = No (with ≥180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<180 days enrollment/engagement) or not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant had a negative gFOBT or FIT result from a back-office procedure ***≤180 days prior to*** the current procedure. • This concept is related to rule 7 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable should incorporate negative back-office FIT/gFOBT results defined using the same logic employed in the FITgFOBTResults Table.
KPWA Notes	• Site did not find any records with recent negative back-office FOBT result; all records set to 00 or 99.
UTSW Notes	• Site is unable to identify back-office FOBTs; all records set to 00 or 99.

IND1STCSPYAFTERPOS

Definition	KP indication algorithm flag: First colonoscopy or lower endoscopy NOS after positive FOBT result
Type (Length)	Character (2)
Valid Values	00 = No, patient had no prior positive FOBT result OR this is not the first colonoscopy or lower endoscopy NOS after a prior positive FOBT result 01 = Yes, patient had a prior positive FOBT result AND this is the first colonoscopy or lower endoscopy NOS after that result 99 = Not applicable (invalid ProcType)

General Notes	• This flag should be populated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether this is the *first* colonoscopy or lower endoscopy NOS after a prior positive gFOBT or FIT result, per the following rationale: ○ A positive FOBT result affects the indication of the next colonoscopy regardless of how long the colonoscopy occurs after the positive stool test. ○ However, that positive FOBT result only affects the next colonoscopy's indication—not all subsequent colonoscopies. • There is no minimum requirement for prior enrollment/engagement; therefore, all colonoscopies and lower endoscopies NOS should have this variable populated as 00 or 01. • This concept is related to rule 8 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • If the actual procedure date was not available for an inpatient event, it was attributed to the inpatient admit date. • Variable should incorporate positive FIT/gFOBT results defined using the same logic employed in the FITgFOBTResults Table.
KPWA Notes	• When looking for prior colonoscopies or lower endoscopies NOS, site applied relevant code list to in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters).
UTSW Notes	• When looking for prior colonoscopies or lower endoscopies NOS, site applied relevant code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDEARLIERCRC

Definition	KP indication algorithm flag: Non-recent history of colorectal cancer diagnosis
Type (Length)	Character (2)
Valid Values	00 = No (with >180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (≤180 days enrollment/engagement) or not applicable (invalid ProcType)

General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with colorectal cancer ***>180 days prior to*** the current procedure. • This variable is related to rule 9 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates data from multiple sources: ○ ICD-9/-10-CM codes where HXCRC = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* ○ ICD-O-3 codes where HXCRC = 1 column of the harmonized Relevant Symptoms & Conditions code list, to be obtained from cancer registry data. • When using ICD-9/-10-CM codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied ICD-9/-10-CM code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats).
UTSW Notes	• Site applied ICD-9/-10-CM code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IndAnyHered

Definition	KP indication algorithm flag: Any history of diagnosis of hereditary disorder that increases CRC risk
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with any of the following ***at any time prior to*** the current procedure: ○ Lynch syndrome (a.k.a. hereditary nonpolyposis colorectal cancer [HNPCC]) ○ Familial adenomatous polyposis (FAP) • There is no minimum requirement for prior enrollment/engagement or data availability; therefore, all colonoscopies and lower endoscopies NOS should have this variable populated as 00 or 01. • This concept is related to rule 10 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information.
KPWA Notes	• Information not available at this site; all records set to 00 (No), as in “not observed.”
UTSW Notes	• Same as KPWA.

KPNC Notes	• Site used local Clarity Dx codes to identify FAP and HNPCC. • This variable includes all historic tentative Lynch syndrome diagnoses, including those that were not confirmed or were later proved wrong by genetic testing. • Actual diagnosis dates were not available for inpatient diagnoses, which were therefore attributed to the encounter admit date.
KPSC Notes	• Same as KPWA.

IndEarlierPolyp

Definition	KP indication algorithm flag: Non-recent history of adenomatous or other colorectal polyp(s)
Type (Length)	Character (2)
Valid Values	00 = No (with >365 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (≤365 days enrollment/engagement) or not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was diagnosed with colorectal adenoma(s) or other colorectal polyp(s) ***>365 days prior to*** the current procedure. • This concept is related to rule 11 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Variable incorporates information from multiple sources: o the appearance of any Dx codes where HXPOLYP = 1 in the harmonized Relevant Symptoms & Conditions code list, to be obtained from health care utilization data; *and/or* o adenoma(s) detected at colorectal procedures, based on the same data sources (e.g., SNOMED codes, NLP) and logic used for the Adenoma variable later in this table. • When using Dx codes, if the actual diagnosis date was not available for an inpatient event, the diagnosis was attributed to the inpatient admit date.
KPWA Notes	• Site applied Dx code list to in-person encounter types (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters) and selected virtual care encounters (scheduled telephone calls, scheduled video visits, and synchronous online chats)
UTSW Notes	• Site applied Dx code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

IndRecentCspy

Definition	KP indication algorithm flag: Recent prior ~~non-screening~~ colonoscopy or lower endoscopy NOS
Type (Length)	Character (2)

Valid Values	00 = No (with ≥180 days of health system enrollment/engagement in which to observe) 01 = Yes (irrespective of duration of enrollment/engagement) 99 = Unknown (<180 days enrollment/engagement) or not applicable (invalid ProcType)
General Notes	• This flag is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant had a previous colonoscopy or lower endoscopy NOS ***≤180 days prior to*** the current procedure. • This concept is related to rule 12 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information. • Previous colonoscopies and lower endoscopies NOS were defined here as the appearance of any codes in the harmonized CRC Screening & GI Surgery code list where PRECISE_CAT = "Colonoscopy" or "Lower endoscopy NOS". • If the actual date of the previous procedure was not available for inpatient events, the procedure was attributed to the inpatient admit date.
KPWA Notes	• Site applied code list to in-person encounter types only (e.g., ambulatory visits, acute inpatient hospital stays, emergency department encounters).
UTSW Notes	• Site applied code list regardless of encounter type.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INDRECENTKPMOD

Definition	KP algorithm-derived indication of most recent prior colonoscopy or lower endoscopy NOS within prior 180 days
Type (Length)	Character (2)
Valid Values	02 = Diagnostic 04 = Screening 06 = High-risk surveillance 07 = Polyp/adenoma surveillance 99 = Unknown (indication not known) or not applicable (invalid ProcType)
General Notes	• This variable is calculated only for colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10) **where the variable IndRecentCspy = 01**; all other records will be set to 99. • Flag represents the KP modified algorithm determination of indication for the most recent prior colonoscopy or lower endoscopy NOS **if one occurred ≤180 days prior to** the current procedure. • This concept is related to rule 12 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information.

INDENROLL

Definition	KP indication algorithm flag: Adequate prior enrollment
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes 99 = Not applicable (invalid ProcType)
General Notes	• This flag is calculated as 00 (No) or 01 (Yes) for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • Flag represents whether participant was continuously enrolled or engaged in the health system for ***≥365 days (allowing for 90-day gaps) prior to*** the current procedure. • This concept is related to rule 13 of the KP indication algorithm. See Appendix D. KP Indication Algorithm for more information.

PROCINDICKPMOD

Definition	Indication for colonoscopy or lower endoscopy NOS derived from KP algorithm
Type (Length)	Character (2)
Valid Values	02 = Diagnostic 04 = Screening 06 = High-risk surveillance 07 = Polyp/adenoma surveillance 99 = Unknown (indication not known) or not applicable (invalid ProcType)
General Notes	• This variable is calculated for both colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); all other procedure records will be set to 99. • PRECISE recommends that analysts use ProcIndic instead of this variable. • Variable calculated as follows: ○ If IndAnyEctomy, IndAnyIBD, IndRecentSx, IndRecentPx, IndRecentCRC, IndRecentPolyp, IndRecentNegBO, or Ind1stCspyAfterPos = 01 (Yes) → ProcIndicKPMod = 02 (Diagnostic) ○ Else if IndAnyEctomy, IndAnyIBD, IndRecentSx, IndRecentPx, IndRecentCRC, IndRecentPolyp, IndRecentNegBO, or Ind1stCspyAfterPos = 99 (Unknown) → ProcIndicKPMod = 99 (Unknown) ○ Else if IndEarlierCRC or IndAnyHered = 01 (Yes) → ProcIndicKPMod = 06 (High-risk surveillance) ○ Else if IndEarlierCRC or IndAnyHered = 99 (Unknown) → ProcIndicKPMod = 99 (Unknown) ○ Else if IndEarlierPolyp = 01 (Yes) → ProcIndicKPMod = 07 (Polyp/adenoma surveillance) ○ Else if IndEarlierPolyp = 99 (Unknown) → ProcIndicKPMod = 99 (Unknown) ○ Else if IndRecentCspy = 01 (Yes) → ProcIndicKPMod = IndRecentKPMod ○ Else if IndRecentCspy = 99 (Unknown) → ProcIndicKPMod = 99 (Unknown) ○ Else if IndEnroll = 01 (Yes) → ProcIndicKPMod = 04 (Screening) ○ Otherwise → ProcIndicKPMod = 99 (Unknown) • See Appendix D. KP Indication Algorithm for more information.

COLOPREP

Definition	Quality of bowel preparation

Type (Length)	Character (2)
Valid Values	00 = Inadequate 01 = Adequate 99 = Unknown or not applicable
General Notes	• This variable applies to colonoscopies only (ProcType = 01); all other procedure records will be set to 99. • In NLP approaches to populating this variable, the term “fair” is interpreted as inadequate bowel preparation. • When using Boston Bowel Preparation Scale (BBPS) values to populate this variable, a total score <6 or any colorectal segment score <2 (regardless of total score) is interpreted as inadequate bowel preparation.
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information. • In short, standard clinical terms were mapped to DRP categories as follows: ○ 01 (Adequate) = Excellent, very good, good, well-prepared, moderate, optimal (ideal), adequate ○ 00 (Inadequate) = Poor, fair, suboptimal, very poor, inadequate
UTSW Notes	• Site applied NLP algorithms to text of colonoscopy procedure reports found in the EMR. Manuscript describing details of NLP is in preparation.
KPNC Notes	• Site applied NLP algorithms to text of colonoscopy procedure reports found in the EMR; see Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information. Mapping was similar to KPWA.
KPSC Notes	• Same as KPNC.

COLOEXTENT

Definition	Extent of procedure
Type (Length)	Character (2)
Valid Values	00 = Incomplete 01 = Complete to cecum 99 = Unknown or not applicable
General Notes	• This variable applies to colonoscopies only (ProcType = 01); all other procedure records will be set to 99.
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Site applied NLP algorithms to text of colonoscopy procedure reports found in the EMR. Manuscript describing details of NLP is in preparation.
KPNC Notes	• Site applied NLP algorithms to text of colonoscopy procedure reports found in the EMR; see Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information. Mapping was similar to KPWA.

KPSC Notes	- Same as KPNC.

PROVIDPERFORM

Definition	Performing provider ID
Type (Length)	Numeric (8)
Valid Values	Provider IDs from the Provider Table (known) or -99999 (unknown)
KPSC Notes	- Performing provider information is highly missing in source data.

FACILITYIDPERFORM

Definition	Performing facility ID
Type (Length)	Numeric (8)
Valid Values	Facility IDs from the Facility Table (known) or -99999 (unknown)
KPSC Notes	- Performing facility information is highly missing in source data.

PROVIDERIDPCP

Definition	Primary care provider ID
Type (Length)	Numeric (8)
Valid Values	Provider IDs from the Provider Table (known) or -99999 (unknown)
General Notes	- Corresponds to the participant's primary care provider at the time of the procedure.

FACILITYIDPCP

Definition	Primary care facility ID
Type (Length)	Numeric (8)
Valid Values	Facility IDs from the Facility Table (known) or -99999 (unknown)
General Notes	- This variable should be used with caution; please refer to the Facility Table documentation for more information.
KPWA Notes	- Corresponds to the participant's primary care facility at the time of the procedure.
UTSW Notes	- Corresponds to the facility with the highest frequency of encounters associated with ProviderIDPCP in year of procedure.
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as KPWA.

PROVIDPATH

Definition	Pathology provider ID
Type (Length)	Numeric (8)

Valid Values	Provider IDs from the Provider Table (known) or -99999 (unknown or not applicable)
General Notes	- Corresponds to the pathologist associated with the procedure, where applicable.
KPWA Notes	- Site was only able to populate this variable for colonoscopies (ProcType = 01) with pathology resulted by KPWA laboratories (i.e., DataSourceKPWA = 2); all other records set to -99999.
UTSW Notes	- Data were obtained for all colonoscopies with records in pathology database with collection date within ±7 days of procedure date.
KPNC Notes	- Data were obtained for sigmoidoscopy, colonoscopy, and lower endoscopy NOS procedures with records in pathology database with collection date on the same day as procedure date.
KPSC Notes	- Data were obtained for sigmoidoscopy, colonoscopy, and lower endoscopy NOS procedures with pathology accession date within ±7 days of procedure date.

FACILITYIDPATH

Definition	Pathology facility ID
Type (Length)	Numeric (8)
Valid Values	Facility IDs from the Facility Table (known) or -99999 (unknown or not applicable)
General Notes	- Corresponds to the pathologist associated with the procedure, where applicable.
KPWA Notes	- Information was not available; all records set to -99999.
UTSW Notes	- Data were obtained for all colonoscopies with records in pathology database with collection date within ±7 days of procedure date.
KPNC Notes	- Data were obtained for sigmoidoscopy, colonoscopy, and lower endoscopy NOS procedures with records in pathology database with collection date on the same day as procedure date.
KPSC Notes	- For all procedures with non-missing ProvIDPath, this variable was set to the facility assigned to or associated with that provider on the procedure date.

PROCSAMPLESENTPATHOLOGY

Definition	Whether sample from procedure was sent to pathology
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable

KPWA Notes	• Site populated this variable according to the following logic: ○ If non-colonoscopy procedure → 99 (N/A) ○ Else if DataSourceKPWA = 1 (Chart abstraction) → populate according to abstractor findings ○ Else if DataSourceKPWA = 2 (NLP) → 01 (Yes) ○ Else if procedure was KPWA-performed (BucketKPWA = 1) → 00 (No) ○ If procedure was not performed by KPWA (BucketKPWA = 2, 3) → ▪ If procedure was temporally isolated (i.e., no other lower endoscopy or GI surgery within 4 days on either side) and had evidence of same-day surgical pathology (via CPT codes) → 01 (Yes) ▪ Otherwise → 99 (Unknown) • Based on a small validation, the site believes it would not be unreasonable to recode 99 (Unknown) as 00 (No) for analyses. Please consult KPWA for further information.
UTSW Notes	• Where available, variable was populated based on NLP outcomes for all of the colonoscopy procedures with reports. Manuscript detailing NLP is in preparation. • Variable was mapped as follows: ○ If procedure report indicated that specimen was collected → 01 (Yes) ○ Else if procedure report indicated that no specimen was collected → 00 (No) ○ Otherwise (e.g., colonoscopy report not available, non-colonoscopy procedure) → 99 (Unknown or not applicable)
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, variable was mapped as follows: ○ If *no* pathology report with colorectal specimen(s) confirmed via SNOMED T (Topography) codes for specimen(s) collected on the same day as the procedure → 00 (No) ○ Else if pathology report was found for specimen(s) collected on the same day as the procedure *and* colorectal specimen(s) confirmed via SNOMED T (Topography) codes → 01 (Yes) • See Appendix B. Code Lists / Pathology for more information on how site defined colorectal location.
KPSC Notes	• Same as KPNC, except site only obtained data for colonoscopies and lower endoscopies NOS and allowed ±7 days around the procedure date when looking for relevant pathology records.

PATHCONTAINERQUANT

Definition	Number of containers sent to pathology
Type (Length)	Numeric (4)
Valid Values	Integers > 0 (known) or -99999 (unknown/not applicable)
KPWA Notes	• This information was not available; all records set to -99999.
UTSW Notes	• Same as KPWA.

KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to -99999. • Where applicable, group designator from pathology SNOMED database was used to determine number of containers sent to pathology.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMLQUANT

Definition	Number of adenomas (or containers with adenomas) detected in the left colon
Type (Length)	Character (2)
Valid Values	00 = 0 adenomas found 02 = 1–2 adenomas found 03 = 3+ adenomas found 04 = Adenoma(s) found; number unknown 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMRQUANT

Definition	Number of adenomas (or containers with adenomas) detected in the right colon
Type (Length)	Character (2)
Valid Values	00 = 0 adenomas found 02 = 1–2 adenomas found 03 = 3+ adenomas found 04 = Adenoma(s) found; number unknown 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.

KPSC Notes	- Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMRECTQUANT

Definition	Number of adenomas (or containers with adenomas) detected in the rectum
Type (Length)	Character (2)
Valid Values	00 = 0 adenomas found 02 = 1–2 adenomas found 03 = 3+ adenomas found 04 = Adenoma(s) found; number unknown 99 = Unknown or not applicable
KPWA Notes	- Information about adenoma location was not available; all records set to 99.
UTSW Notes	- Same as KPWA.
KPNC Notes	- Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. - Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	- Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMUNKQUANT

Definition	Number of adenomas (or containers with adenomas) detected in unknown location(s)
Type (Length)	Character (2)
Valid Values	00 = 0 adenomas found 02 = 1–2 adenomas found 03 = 3+ adenomas found 04 = Adenoma(s) found; number unknown 99 = Unknown or not applicable
KPWA Notes	- Information about adenoma location was not available; all records set to 99.
UTSW Notes	- Where available, variable was populated based on NLP outcomes. Manuscript detailing details of NLP is in preparation. - Site was not able to distinguish exact quantities of adenoma, so no records were mapped to 02 or 03.
KPNC Notes	- Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. - Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.

KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENTOTQUANT

Definition	Total number of adenomas (or containers with adenomas) detected
Type (Length)	Character (2)
Valid Values	00 = 0 adenomas found 02 = 1–2 adenomas found 03 = 3+ adenomas found 04 = Adenoma(s) found; number unknown 99 = Unknown or not applicable
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation. • Site was not able to distinguish exact quantities of adenoma, so no records were mapped to 02 or 03.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMA

Definition	Whether ≥1 adenoma was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation.

KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMVTL

Definition	Whether adenoma(s) with (tubulo)villous histology were detected in the left colon
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Villous and tubulovillous adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMVTR

Definition	Whether adenoma(s) with (tubulo)villous histology were detected in the right colon
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Villous and tubulovillous adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMVTRECT

Definition	Whether adenoma(s) with (tubulo)villous histology were detected in the rectum
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Villous and tubulovillous adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMVTUNK

Definition	Whether adenoma(s) with (tubulo)villous histology were detected in unknown location(s)
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• Information about adenoma location was not available; all records set to 99.
UTSW Notes	• Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Villous and tubulovillous adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENVTANY

Definition	Whether adenoma(s) with (tubulo)villous histology were detected in any location(s)
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • Villous and tubulovillous adenomas were identified using SNOMED M (Morphology) codes, while location was identified using SNOMED T (Topography) codes. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

LARGEADENOMA

Definition	Whether adenoma(s) with size ≥10 mm were detected in any location
Type (Length)	Character (2) or not applicable
Valid Values	00 = No 01 = Yes 97 = Maybe (large polyp & adenoma detected; unclear whether large polyp is adenoma) 99 = Unknown or not applicable
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Information was not available; all records set to 99.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used a combination of polyp size (from NLP) and presence of adenoma (from SNOMED M [Morphology] codes) to populate variable. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC and Appendix B. Code Lists / Pathology for more information on NLP and SNOMED codes, respectively.
KPSC Notes	• Same as UTSW.

LARGEPOLYP

Definition	Whether polyp(s) with size ≥10 mm were detected in any location
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• This variable includes both adenomas and non-adenomatous polyps.
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information. • Even in these data sources, site only had chart-abstracted size information for adenomas (both), SSPs (abstraction only), and hyperplastic polyps (abstraction only); therefore, capture of any large polyp may be incomplete. • Site took a conservative approach and did not include “maybes” from the LargeAdenoma variable when setting this variable; analysts may wish to set this variable to 01 (Yes) when LargeAdenoma = 97 (Maybe).
UTSW Notes	• Information was not available; all records set to 99.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used NLP to assess polyp size. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information on NLP.
KPSC Notes	• Same as UTSW.

PROXIMALHP

Definition	Whether ≥1 hyperplastic polyp was detected in the proximal colon
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 97 = Maybe (≥1 hyperplastic polyp found but location unknown) 99 = Unknown or not applicable
General Notes	• The proximal colon refers to any location above the splenic flexure: i.e., cecum, ascending colon, hepatic flexure, or transverse colon.
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction).
UTSW Notes	• Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation. • Site was only able to classify procedures as 97 (Maybe) or 99 (Unknown).
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used SNOMED M (Morphology) codes to detect presence of hyperplasia at a proximal location. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

ADENOMAHGD

Definition	Whether ≥1 adenoma with high-grade dysplasia was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 97 = Maybe (high-grade dysplasia & adenoma detected; unclear whether high-grade dysplasia occurred in adenoma) 99 = Unknown or not applicable
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Information was not available; all records set to 99.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used a combination of presence of high-grade dysplasia (from NLP) and presence of adenoma (from SNOMED M [Morphology] codes) to populate variable. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC and Appendix B. Code Lists / Pathology for more information on NLP and SNOMED codes, respectively.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

POLYPHGD

Definition	Whether ≥1 polyp with high-grade dysplasia was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• This variable includes both adenomas and non-adenomatous polyps.
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information. • Where DataSourceKPWA = 1 (Chart abstraction), high-grade dysplasia was only assessed among adenomas and SSPs; therefore, capture of *any* polyp with high-grade dysplasia may be underrepresented.
UTSW Notes	• Information was not available; all records set to 99.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used NLP to detect high-grade dysplasia. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information.

KPSC Notes	- Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

SSP

Definition	Whether ≥1 sessile serrated polyp (SSP) or sessile serrated adenoma (SSA) was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	- This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	- Where available, variable was populated based on NLP outcomes from pathology reports linked to colonoscopies. Manuscript detailing details of NLP is in preparation.
KPNC Notes	- Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. - For aforementioned procedure types, site used NLP to detect sessile serrated polyps or sessile serrated adenomas. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information.
KPSC Notes	- Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

TSA

Definition	Whether ≥1 traditional serrated adenoma (TSA) was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
KPWA Notes	- This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction).
UTSW Notes	- Information not available; all records set to 99.
KPNC Notes	- Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. - For aforementioned procedure types, site used NLP to detect traditional serrated adenomas. See Appendix E. Natural Language Processing / NLP at KPNC/KPSC for more information.
KPSC Notes	- Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

CRCDETECTED

Definition	Whether invasive or in situ colorectal cancer was detected
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown or not applicable
General Notes	• This variable is populated based on information specific to the procedure, such as the associated pathology report—*not* from other sources like tumor registries. • Cancer was defined using the SEER site recode ICD-O-3/WHO 2008 definition of colorectal cancer, which is limited to the histology types listed below (for tumors with either malignant or in situ behavior): o Neoplasm o Tumor cells o Tumor o Carcinoma o Epithelioma o Adenocarcinoma o Carcinoid tumor o Enterochromaffin-like cell tumor o Adenocarcinoid tumor o Cystadenocarcinoma o Sarcoma o Desmoplastic small round cell tumor o Fibrosarcoma o Fibromyxosarcoma o Solitary fibrous tumor o Liposarcoma o Leiomyosarcoma o Angiomyosarcoma o Myosarcoma o Carcinofibroma o Carcinosarcoma o Myoepithelioma o Melanoma ***(rectum only)***
KPWA Notes	• This information was only available for colonoscopies where DataSourceKPWA = 1 (Chart abstraction) or 2 (NLP); see Appendix E. Natural Language Processing / NLP at KPWA for more information.
UTSW Notes	• Information not available; all records set to 99.
KPNC Notes	• Site had pathology outcomes available for colonoscopies, sigmoidoscopies, and lower endoscopies NOS; all other procedures set to 99. • For aforementioned procedure types, site used SNOMED M (Morphology) codes to identify colorectal cancer. See Appendix B. Code Lists / Pathology for more information.
KPSC Notes	• Same as KPNC except that information was computed only for colonoscopies and lower endoscopies NOS.

BUCKETKPWA

Definition	KPWA data availability bucket
Type (Length)	Character (1)
Valid Values	1 = KPWA-performed procedure 2 = DHS-/WAGI-performed procedure 3 = Procedure performed by other external provider (blank) = Not applicable (non-KPWA site)
General Notes	• This is a KPWA-specific variable; all other sites' records will be left blank.
KPWA Notes	• While this variable was populated for all procedure types, it is primarily of use when interpreting information about colonoscopy indication, prep, extent/completeness, and pathology outcomes. • Bucket 1 refers to procedures performed in KPWA-owned internal delivery system facilities. ○ Full text of colonoscopy procedure reports and result letters should be available in EMR. ○ Pathology report text should be available via administrative feed. • Bucket 2 refers to procedures performed by DHS/WAGI; see Additional Information re: KPWA for more information on these external providers. ○ DHS-era (2010–2017) colonoscopy procedure reports were only available as scanned documents in EMR; these were processed via optical character recognition (OCR) and sent through NLP pipeline where available. ○ Most DHS colonoscopy pathology was performed by KPWA during 2010–2017, meaning that full pathology report text was available in KPWA's administrative data for a large proportion of these procedures. Remaining DHS-performed colonoscopies with administrative evidence of biopsy (i.e., same-day surgical pathology CPT codes) were chart-abstracted. ○ All WAGI colonoscopies (2018–2020) were chart-abstracted. • Bucket 3 refers to procedures performed by various other contracted providers in KPWA's external delivery system. ○ Administrative data were used to extrapolate whether a sample was collected from some externally performed colonoscopies; see ProcSampleSentPathology for more information. ○ Otherwise, very little structured data were available about these procedures.
UTSW Notes	• Not applicable; all records left blank.
KPNC Notes	• Same as UTSW.
KPSC Notes	• Same as UTSW.

DATASOURCEKPWA

Definition	KPWA data source for colonoscopy procedure/pathology info
Type (Length)	Character (1)

Valid Values	1 = Chart abstraction 2 = NLP 3 = Administrative data 9 = Not applicable (i.e., non-colonoscopy procedure) (blank) = Not applicable (non-KPWA site)
General Notes	• This is a KPWA-specific variable; all other sites' records will be left blank.
KPWA Notes	• See the Procedure Table Data Sources section as well as BucketKPWA above for more details about the overlap between where/by whom a colonoscopy was performed and the types of data that were available.
UTSW Notes	• Not applicable; all records left blank.
KPNC Notes	• Same as UTSW.
KPSC Notes	• Same as UTSW.

BUCKETKPNC

Definition	KPNC data availability bucket
Type (Length)	Character (1)
Valid Values	1 = Procedure performed at KP-owned facility 2 = Procedure performed at external facility (blank) = Facility relationship is unknown (KPNC) or not applicable (non-KPNC site)
General Notes	• This is a KPNC-specific variable; all other sites' records will be left blank.
KPWA Notes	• Not applicable; all records left blank.
UTSW Notes	• Same as KPWA.
KPNC Notes	• This variable is particularly important when interpreting information about colonoscopy indication, prep, extent/completeness, and pathology outcomes. • Bucket 1 refers to procedures performed in KPNC-owned facilities. ○ Full text of colonoscopy procedure reports should be available in EMR. ○ Pathology report text should be available via laboratory database. • Bucket 2 refers to procedures performed in external contracting facilities. ○ Very little data are available about these procedures or pathology outcomes.
KPSC Notes	• Same as KPWA.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Encounter Table

Overview

This table contains one record per **in-person primary care encounter**, per provider, per day for PRECISE participants during cohort eligibility. (Note: KPNC and KPSC also included encounters that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment file to restrict to during-cohort events.) KPNC and KPSC further deduplicated to one in-person primary care encounter per day using a provider hierarchy to pick the "top" provider per day. See detailed site-specific primary care encounter definitions in Appendix A. Primary Care Visits.

The table also includes one record per **overnight inpatient or institutional stay** during cohort eligibility, per distinct combination of admit and discharge dates, for PRECISE participants during cohort eligibility. Encounters that either begin or end during cohort eligibility are included (with admit and discharge dates truncated at start or end of cohort eligibility if necessary) when the length of the stay during cohort eligibility is at least one day (i.e., discharge date must be after admit date).

Finally, beginning in IMS3, this table contains one record per completed **virtual primary care encounter**, per provider, per day for PRECISE participants during cohort eligibility *in 2019 and 2020 only*.

Data Sources

KP sites pulled the data from their local VDW Enrollment and Utilization tables.

UTSW used EMR data.

Variables

PID

Same as above.

ENCID

Definition	Unique encounter ID
Type (Length)	Character (1)
Valid Values	Blank
General Notes	• PCC requested this variable for IMS3, but PRECISE is unable to provide it.

ENCTYPE

Definition	Encounter type
Type (Length)	Character (1)
Valid Values	1 = Primary care, in-person visit 2 = Primary care, scheduled phone call 3 = Primary care, scheduled video visit 4 = Primary care, synchronous online chat 7 = Inpatient stay 8 = Institutional stay
General Notes	• Per the PCC, telehealth encounters (codes 2–4) only need to be included for completed encounters during 2019–2020.
UTSW Notes	• Institutional stay data were not available at this site.

ENCADMITDATE/PCPVISITDATE

Definition	Encounter admit date (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020

General Notes	• Variable was called PCPVisitDate in IMS1 data submission. In IMS2, PRECISE changed the name to EncAdmitDate when we decided to include inpatient and institutional stay data in this table along with primary care visits. For IMS3, the variable will be called EncAdmitDate in the PRECISE VDC data; however, it will be renamed to PCPVisitDate in files sent to IMS. • Where EncType = 1–4 (PC), this variable reflects the date of the primary care visit. • Where EncType = 7 (IP) or 8 (IS), this variable reflects the date on which the participant was admitted to the facility, adjusted to start of cohort eligibility where applicable. (For this reason, exact duration of inpatient/institutional stays cannot always be calculated with this file.) • Note: The “overnight stay” requirement was applied *after* adjusting for start of cohort eligibility: i.e., the discharge date must be later than the adjusted admit date.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPNC Notes	• Site included encounters that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

ENCADMITDSR/PCPVISITDSR

Definition	Encounter admit date (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063
General Notes	• Variable was called PCPVisitDSR in IMS1 data submission. In IMS2, PRECISE changed the name to EncAdmitDSR when we decided to include inpatient and institutional stay data in this table along with primary care visits. For IMS3, the variable will be called EncAdmitDSR in the PRECISE VDC data; however, it will be renamed to PCPVisitDSR in files sent to IMS. • Where EncType = 1–4 (PC), this variable reflects the date of the primary care visit. • Where EncType = 7 (IP) or 8 (IS), this variable reflects the date on which the participant was admitted to the facility, adjusted to start of cohort eligibility where applicable. (For this reason, exact duration of inpatient/institutional stays cannot always be calculated with this file.) • Note: The “overnight stay” requirement was applied *after* adjusting for start of cohort eligibility: i.e., the discharge date must be later than the adjusted admit date. • Valid values reflect participant being 40–95 years old during cohort eligibility.
KPNC Notes	• Site included encounters that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

ENCDISCHDATE

Definition	Encounter discharge date (as SAS date)
Type (Length)	Numeric (4)

Valid Values	01/02/2010–12/31/2020 or null (when EncType = 1–4 [PC])
General Notes	• This variable is only applicable to inpatient and institutional stays, for which discharge date must be after admit date (i.e., only overnight stays are included in this file). • Where EncType = 7 (IP) or 8 (IS), this variable reflects the date on which the participant was discharged from the relevant facility, adjusted to end of cohort eligibility where applicable. (For this reason, exact duration of inpatient/institutional stays cannot always be calculated with this file.) • Valid values reflect "overnight stay" requirement, i.e., the discharge date must be later than the admit date.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPNC Notes	• Site included encounters that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

ENCDISCHDSR

Definition	Encounter discharge date (as SAS date)
Type (Length)	Numeric (4)
Valid Values	Integers 14611–35063 or null (when EncType = 1–4 [PC])
General Notes	• This variable is only applicable to inpatient and institutional stays, for which discharge date must be after admit date (i.e., only overnight stays are included in this file). • Where EncType = 7 (IP) or 8 (IS), this variable reflects the date on which the participant was discharged from the relevant facility, adjusted to end of cohort eligibility where applicable. (For this reason, exact duration of inpatient/institutional stays cannot always be calculated with this file.) • Valid values reflect participant being 40–95 years old during cohort eligibility as well as "overnight stay" requirement, i.e., the discharge date must be later than the admit date.
KPNC Notes	• Site included encounters that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

PROVIDERIDENC

Definition	ID of primary care encounter provider
Type (Length)	Numeric (8)
Valid Values	Provider IDs from the Provider Table (known) or -99999 (unknown or not applicable)
General Notes	• This variable is only applicable where EncType = 1–4 (PC). • Encounters with EncType = 7 (IP) or 8 (IS) will have this variable set to -99999, as will PC encounters for which the provider is unknown.

FACILITYIDENC

Definition	ID of primary care encounter facility
Type (Length)	Numeric (8)
Valid Values	Facility IDs from the Facility Table (known) or -99999 (unknown or not applicable)
General Notes	• This variable is only applicable where EncType = 1–4 (PC). • Encounters with EncType = 7 (IP) or 8 (IS) will have this variable set to -99999, as will PC encounters for which the facility is unknown.

MEDICARE

Definition	Whether participant was covered by Medicare at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record any Medicare coverage for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is “Did the participant use Medicare coverage to pay for at least part of this encounter?” This differed from the approach used in CalendarYear, which incorporated both coverage and payor data.
KPWA Notes	• There was an apparent drop in commercial coverage among Medicare age group circa 2017. A change in KPWA source data around this time obscured/revealed varying aspects of coverage. For example, there were likely some people pre-2017 who just had Medicare but showed up as having commercial coverage due to the way medical market groups were coded. This issue is not “fixable” at present.

MEDICAID

Definition	Whether participant was covered by Medicaid at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record any Medicaid coverage for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is “Did the participant use Medicaid coverage to pay for at least part of this encounter?” This differed from the approach used in CalendarYear, which included both coverage and payor data.

KPWA Notes	• Site has a very low number of during-cohort encounters at which a participant had Medicaid coverage. This is because KPWA's cohort eligibility was restricted to periods of non-Medicaid enrollment. Therefore, primary care encounters at which the participant *did* have Medicaid coverage could only occur during ≤90-day tolerated gaps in cohort-eligible enrollment.

INSOTHERGOV

Definition	Whether participant was covered by any other federal or state health insurance program at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately "Did KP record any other government insurance coverage for this participant at the time of this encounter?" These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is "Did the participant use any other government insurance coverage to pay for at least part of this encounter?" This differed from the approach used in CalendarYear, which included both coverage and payor data.

INSCOMMERC

Definition	Whether participant was covered by commercial and/or private health insurance at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately "Did KP record any commercial/private insurance coverage for this participant at the time of this encounter?" These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is "Did the participant use any commercial/private insurance coverage to pay for at least part of this encounter?" This differed from the approach used in CalendarYear, which included both coverage and payor data.

MEDICALASSIST

Definition	Whether participant was covered by a medical assistance charity program for the uninsured at this encounter
Type (Length)	Character (2)

Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record any medical assistance insurance coverage for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is “Did the participant use any medical assistance insurance coverage to pay for at least part of this encounter?” This differed from the approach used in CalendarYear, which included both coverage and payor data.
KPWA Notes	• Information was not available at time of data extract; all records set to 99 (Unknown).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

UNINSURED

Definition	Whether participant was uninsured at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown
General Notes	• At the KP sites, the definition is more accurately “Did KP record a gap in insurance coverage for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. • At UTSW, the definition is “Did the participant self-pay or have no payor/coverage documented for this encounter?” This differed from the approach used in CalendarYear, which included both coverage and payor data.
KPWA Notes	• Primary care encounters at which the patient was uninsured could only occur during ≤90-day tolerated gaps in cohort-eligible enrollment.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

INSHIGHDEDUCTIBLE

Definition	Whether participant was covered by high-deductible insurance at this encounter
Type (Length)	Character (2)
Valid Values	00 = No 01 = Yes 99 = Unknown

General Notes	- “High-deductible” refers to definition from U.S. IRS Pub 969. - At the KP sites, the definition is more accurately “Did KP record any high-deductible insurance coverage for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. - At UTSW, the definition would be, “Did the participant use any high-deductible coverage to pay for at least part of this encounter?” However, this information was not available.
UTSW Notes	- Information not available at this site; all records set to 99.

INSOTHER

Definition	Other type(s) of insurance coverage at this encounter
Type (Length)	Character (*)
Valid Values	See site-specific notes for the InsOther variable in the CalendarYear Table.
General Notes	- At the KP sites, the definition is more accurately “What not otherwise categorized type(s) of insurance coverage did KP record for this participant at the time of this encounter?” These sites assessed insurance coverage using logic similar to that described for the variable with the same name in the CalendarYear Table. - At UTSW, the definition is “What not otherwise categorized type(s) of insurance coverage did the participant use to pay for at least part of this encounter?” This differed from the approach used in CalendarYear, which included both coverage and payor data.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

CancerRegistry Table

Overview

This table contains one record per tumor registry sequence number per participant for colorectal tumors diagnosed during cohort eligibility. (Note: KPNC and KPSC also included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.) "Colorectal" refers to tumors located at primary sites that could be associated with colorectal cancer as defined by SEER site recode (ICD-O-3/WHO 2008 definition). The Behavior, SEER_CRC, Appendix, and CRC_Adenoca variables may be useful in restricting analyses to various commonly used definitions of colorectal cancer.

Death due to CRC, cardiovascular disease, other cause, or unknown cause can be ascertained from the COD Table.

Data Sources

All sites used data from locally available tumor registries.

KPWA used NAACCR-formatted data provided via its relationship with the SEER Seattle-Puget Sound Registry, which abstracts cancer cases for residents of a 13-county western Washington catchment area.

UTSW used NAACCR-formatted data provided by the Texas Cancer Registry and Parkland Hospital tumor registry.

KPNC used its VDW Tumor table, which contains data sourced from KPNC Cancer Registry, which reports directly to SEER (National Cancer Institute).

KPSC used a data file received from the KPSC Cancer Registry, which reports directly to SEER (National Cancer Institute).

Variables

PID

Same as above.

SEQUENCE

Definition	Tumor sequence within the relevant reporting facility (central or hospital registry)
Type (Length)	Character (2)
Valid Values	##-formatted values 00, 01, 02, etc.
General Notes	• Variable was called SequenceNumber in IMS1 data submission. • Concept corresponds to NAACCR item #380 (central) or #560 (hospital).
UTSW Notes	• Set to unknown (99) or blank if value not provided by cancer registry; this was a data quality issue on behalf of the cancer registry and unfortunately not fixable at the PRECISE site level.

PRIMARYSITE

Definition	Tumor primary site
Type (Length)	Character (5)
Valid Values	C18.0–C18.9, C19.9, C20.9, C26.0 (i.e., ICD-O-3 topography codes that could be associated with colorectal cancer as defined by SEER site recode ICD-O-3/WHO 2008 definition)
General Notes	• Variable was called PrimarySiteICD in IMS1 data submission. • Concept corresponds to NAACCR item #400.
UTSW Notes	• No tumors with primary site C26.0 were identified in either cancer registry for IMS3 data submission.

DXDATE

Definition	Tumor diagnosis date (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020
General Notes	• Concept corresponds to NAACCR item #390.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPNC Notes	• Site included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

DXDSR

Definition	Tumor diagnosis date (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063
General Notes	• Concept corresponds to NAACCR item #390. • Valid values reflect participant being 40–95 years old during cohort eligibility.
KPNC Notes	• Site included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment Table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

BEHAVIOR

Definition	Tumor behavior
Type (Length)	Character (1)

Valid Values	ICD-O-3 behavior codes for neoplasms, i.e.: 1 = Uncertain whether benign or malignant 2 = In situ 3 = Malignant, primary site
General Notes	• Variable was called TumorBehavior in IMS1 data submission. • Concept corresponds to NAACCR item #523.

SEERSTAGE

Definition	Best available SEER summary stage
Type (Length)	Character (1)
Valid Values	0 = In situ 1 = Localized 2 = Regional, direct extension only 3 = Regional, regional lymph nodes only 4 = Regional, direct extension and regional lymph nodes 5 = Regional, NOS 7 = Distant 8 = Not applicable 9 = Unstaged
General Notes	• To accommodate 2018+ data, variable was populated as a coalesce of the following NAACCR items, in this order: ○ 764 SUMMARY STAGE 2018 (2018+ diagnoses) ○ 762 DERIVED SUMMARY STAGE 2018 (2018+ diagnoses) ○ 759 SEER SUMMARY STAGE 2000 (pre-2018 diagnoses) ○ 3020 DERIVED SS2000 (pre-2018 diagnoses)
UTSW Notes	• Variable was set to 9 (Unstaged) if value stage was not available from registry.

HISTOLOGY

Definition	Tumor histology
Type (Length)	Character (4)
Valid Values	####-formatted values that represent first four digits of morphology codes listed in ICD-O-3 (1st rev).
General Notes	• Variable was called HistologyICD02or03 in IMS1 data submission. • Concept corresponds to NAACCR item #522.

SEERSITERECODE

Definition	SEER site recode
Type (Length)	Character (5)
Valid Values	#####-formatted values from ICD-O-3/WHO 2008 SEER site recode definition

General Notes	• Variable maps a given tumor’s topography (PrimarySite) + histology (Histology) to SEER’s definition of cancer type.

SEER_CRC

Definition	Whether tumor meets SEER definition of colorectal cancer
Type (Length)	Character (1)
Valid Values	Y = Yes, SEER site recode is in range 21041–21052 N = No, SEER site recode is available but not in range 21041–21052
General Notes	• Variable serves as a quick means of determining whether SEER publications would include the given tumor in colorectal cancer analyses.

Appendix

Definition	Whether tumor is in the appendix
Type (Length)	Character (1)
Valid Values	Y = Yes, PrimarySite is C18.1 N = No, PrimarySite is available but not equal to C18.1
General Notes	• Variable serves as a quick way to determine whether the given tumor is an appendiceal cancer; appendiceal cancers are excluded from some definitions of colorectal cancer.

CRC_Adenoca

Definition	Whether tumor is a colorectal adenocarcinoma
Type (Length)	Character (1)

Valid Values	Y = Yes
	N = No

General Notes	• Set to Y if: ○ PrimarySite is C18.0, C18.2–C18.9, C19.9, or C20.9; <u>and</u> ○ Histology is 8000, 8010, 8020, 8140, 8143, 8144, 8210, 8211, 8215, 8220, 8221, 8230, 8244, 8245, 8255, 8260, 8261, 8262, 8263, 8323, 8480, 8481, 8490, 8510, 8560, 8570, 8571, 8572, 8573, or 8574. • Set to N if: ○ PrimarySite is C18.1 (Appendix) or C26.0 (Intestinal tract, NOS); <u>or</u> ○ Histology not in list above.

RxSummTxStatus

Definition	Treatment status summary
Type (Length)	Character (1)
Valid Values	As provided by cancer registry in NAACCR item #1285 or blank if unavailable

General Notes	• Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

RXSEERDATE1

Definition	Date therapy initiated, SEER definition (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020 or null if unavailable or outside participant’s cohort enrollment
General Notes	• Concept corresponds to NAACCR item #1260. • Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

RXSEERDSR1

Definition	Date therapy initiated, SEER definition (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063 or null if unavailable or outside participant’s cohort enrollment
General Notes	• Concept corresponds to NAACCR item #1260. • Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry. • Valid values reflect participant being 40–95 years old during cohort eligibility.

RXCOCDATE1

Definition	Date therapy initiated, CoC definition (as SAS date)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020 or null if unavailable or outside participant’s cohort enrollment
General Notes	• Concept corresponds to NAACCR item #1270. • Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

RXCOCDSR1

Definition	Date therapy initiated, CoC definition (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063 or null if unavailable or outside participant’s cohort enrollment

General Notes	• Concept corresponds to NAACCR item #1270. • Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry. • Valid values reflect participant being 40–95 years old during cohort eligibility.

RXSEQUENCE

Definition	Systemic/surgery sequence
Type (Length)	Character (1)
Valid Values	As provided by cancer registry in NAACCR item #1639 or blank if unavailable
General Notes	• Variable was marked “optional” by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR910

Definition	TNM Path Stage Group
Type (Length)	Character (4)
Valid Values	As provided by cancer registry in NAACCR item #910 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR970

Definition	TNM Clin Stage Group
Type (Length)	Character (4)
Valid Values	As provided by cancer registry in NAACCR item #970 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR3430

Definition	Derived AJCC-7 Stage Grp
Type (Length)	Character (3)
Valid Values	As provided by cancer registry in NAACCR item #3430 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR1004

Definition	AJCC TNM Clin Stage Group
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #1004 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR1014

Definition	AJCC TNM Path Stage Group
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #1014 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR1024

Definition	AJCC TNM Post Therapy Stage Group
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #1024 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry. • Note: The description of the underlying NAACCR item changed to “AJCC TNM Post Therapy Path (yp) Stage Group” beginning in V21.

NAACCR3645

Definition	NPCR Derived AJCC 8 TNM Clin Stg Grp
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #3645 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR3646

Definition	NPCR Derived AJCC 8 TNM Path Stg Grp
Type (Length)	Character (15)

Valid Values	As provided by cancer registry in NAACCR item #3646 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR3647

Definition	NPCR Derived AJCC 8 TNM Post Therapy Stg Grp
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #3647 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR818

Definition	Derived EOD 2018 Stage Group
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #818 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR772

Definition	EOD Primary Tumor
Type (Length)	Character (3)
Valid Values	As provided by cancer registry in NAACCR item #772 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR774

Definition	EOD Regional Nodes
Type (Length)	Character (3)
Valid Values	As provided by cancer registry in NAACCR item #774 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR776

Definition	EOD Mets
Type (Length)	Character (2)
Valid Values	As provided by cancer registry in NAACCR item #776 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR820

Definition	Regional Nodes Positive
Type (Length)	Character (2)
Valid Values	As provided by cancer registry in NAACCR item #820 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR756

Definition	Tumor Size Summary
Type (Length)	Character (3)
Valid Values	As provided by cancer registry in NAACCR item #756 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

NAACCR1067

Definition	AJCC TNM Post Therapy Clin (yc) Stage Group
Type (Length)	Character (15)
Valid Values	As provided by cancer registry in NAACCR item #1067 or blank if unavailable
General Notes	• Variable was requested by the PCC for IMS3 data submission with the understanding that this information may be unavailable, messy, or sparsely populated depending on the diagnosis year and/or cancer registry.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Provider Table

Overview

This table contains one record per provider ID that appears in any of the tables described above. IDs should be unique within each Providing Site.

Data Sources

KPWA provider information is sourced from the VDW Provider Specialty table. Per underlying business rules, a KPWA provider is “an entity that delivers care to a KPWA member.” In other words, some KPWA “providers” may in fact be facilities; see the ProviderIsPerson variable for more information. Providers employed by KPWA can typically be identified by determining whether the accompanying encounter or procedure facility is owned and operated by KPWA, as only KPWA providers work in KPWA-owned facilities. In general, KPWA has more robust information capture for providers employed by KPWA.

UTSW pulled this data from EMR.

KPNC provider information is sourced from the VDW Provider Specialty table and local Epic/Clarity provider database. The rules are similar to rules described above for KPWA.

KPSC provider information is sourced from various KPSC-specific department files. This file is considered the department master provider file used for most studies. Some variables were separately obtained from EMR.

Variables

PROVIDERID

Definition	Provider ID
Type (Length)	Numeric (8)
Valid Values	Numeric values unique at the provider level within each site
General Notes	• ProviderID + ProvidingSite = composite provider-level primary key • A given ProviderID should represent the same provider: ○ across all PROSPR II organ sites (cervical [METRICS], colorectal [PRECISE], and/or lung [LOTUS]); and ○ in the previous round of PROSPR funding (PROSPR I).
KPNC Notes	• Site did not re-use PROSPR I ProviderIDs.

PROVIDERSPECIALTY

Definition	Provider medical specialty
Type (Length)	Character (*)
Valid Values	21 = Anesthesiology 22 = Emergency medicine 23 = Family medicine 24 = Internal medicine, general internal medicine 25 = Internal medicine, geriatrics 26 = Internal medicine, gastroenterology 27 = Internal medicine, oncology 28 = Internal medicine, other 29 = Internal medicine, sub-specialty unknown or no sub-specialty 30 = Midwifery 31 = Nursing 32 = Obstetrics and gynecology 33 = Pathology 34 = Pediatrics 35 = Radiology 36 = Surgery 40 = Internal medicine, pulmonary disease 97 = Other, specify 99 = Unknown
General Notes	• Multiple specialties are represented by pipe-delimited strings (e.g., 23|25).

KPWA Notes	• Site mapped coded VDW-based provider specialty as follows (up to three per provider): ○ Anesthesiology → 21 (Anesthesiology) ○ Emergency Medicine → 22 (Emergency medicine) ○ Family Medicine → 23 (Family medicine) ○ General Internal Medicine → 24 (Internal medicine, general internal medicine) ○ Gerontology → 25 (Internal medicine, geriatrics) ○ Gastroenterology → 26 (Internal medicine, gastroenterology) ○ Oncology → 27 (Internal medicine, oncology) ○ Radiation Oncology → 27 (Internal medicine, oncology) + 35 (Radiology) ○ Gynecologic Oncology → 27 (Internal medicine, oncology) + 32 (Obstetrics and gynecology) ○ Surgical Oncology → 27 (Internal medicine, oncology) + 36 (Surgery) ○ Nurse → 31 (Nursing) ○ Obstetrics – Gynecology → 32 (Obstetrics and gynecology) ○ Pathology → 33 (Pathology) ○ Pediatrics, Pediatric Subspecialty → 34 (Pediatrics) ○ Radiology → 35 (Radiology) ○ Cardiovascular Surgery, Colon & Rectal Surgery, Hand Surgery, Neurosurgery, Oral Surgery, Plastic Surgery, Surgery, Transplant Surgery, Vascular Surgery → 36 (Surgery) ○ Pulmonary Medicine → 40 (Internal Medicine, Pulmonary disease) ○ Unknown → 99 (Unknown) ○ Other → 97 (Other, specify) • Note that site did not map any sub-specialties to 28 (Internal medicine, other); however, sub-specialties of interest may be identified using the ProviderSpecialtyOther variable.
UTSW Notes	• Providers with multiple specialties were prioritized (i.e., mapped to a single specialty) as follows: Internal Medicine, Gastroenterology set to 26, OB/Gyn set to 32; Family Medicine set to 23, Internal Medicine, General Internal Medicine set to 24, and Internal Medicine, Geriatrics set to 25; Emergency Medicine set to 22, Internal Medicine, Oncology set to 27, and Surgery set to 36; Nursing set to 31; Anesthesiology set to 21, Midwifery set to 30, and Pediatrics set to 34; and Internal Medicine, Other set to 28, Pathology set to 33, Radiology set to 35, and Other set to 97. • Site mapped the following specialties to 28 (Internal Medicine, Other): Advanced Heart Failure And Transplant Cardiology, Allergy & Immunology, Cardiology, Interventional Cardiology, Cardiovascular Disease, Clinical Cardiac Electrophysiology, Critical Care Medicine, Diabetes, Endocrinology, Endocrinology - Diabetes & Metabolism, Genetics, Hematology, Hematology & Oncology, Infection Control, Infectious Disease, Infectious Diseases, Internal Medicine/Psychiatry, Interventional Cardiology, Medical Genetics - Clinical Biochemical Genetics, Medical Genetics - Clinical Genetics (M.D.), Nephrology, Rheumatology, Sleep Medicine, and Transplant Hepatology.

KPNC Notes	• Information was obtained from VDW Provider Specialty table where available. • Site set ProviderSpecialty = 99 (Unknown) when ProviderIsPerson = 00 (No) or 99 (Unknown). • Same mapping as KPWA with the following exceptions: ○ The following VDW-based specialties were mapped to 28 (Internal medicine, other): Allergy, Cardiology, Endocrinology, Infectious Disease, Nephrology ○ VDW-based specialty Gynecologic Oncology was mapped only to 27 (Internal medicine, oncology)
KPSC Notes	• This variable was obtained from the EMR specialty field using descriptions that corresponded to valid values listed above. • The following EMR specialty descriptions were mapped to 28 (Internal medicine, other): ○ Int Med - Hospice & Palliative ○ Internal Med - Allergy & Immunology ○ Internal Med - Sleep Medicine ○ Internal Medicine, Biliary Endoscopy ○ Internal Medicine, Cardiology ○ Internal Medicine, Hematology ○ Internal Medicine, Nephrology Chronic ○ Internal Medicine, Sports Medicine ○ Internal Medicine, Transplant Hepatology ○ Internal Medicine, Adolescent Medicine ○ Internal Medicine, Hepatology

PROVIDERSPECIALTYOTHER

Definition	Other provider specialty
Type (Length)	Character (*)
Valid Values	See general and site-specific notes below.
General Notes	• Populated when ProviderSpecialty contains 97 (Other, specify). • Multiple other provider specialties are represented by pipe-delimited strings (e.g., CARDIOLOGY|PSYCHIATRY).
UTSW Notes	• This field is also populated when ProviderSpecialty = 28 (Internal Medicine, Other).

PROVIDERTYPE

Definition	Provider type
Type (Length)	Character (2)

Valid Values	01 = Administrative staff 02 = Fellows (includes MDs and DOs) 03 = Licensed practical nurse 04 = Medical assistant 05 = Nurse practitioner 06 = Physician (includes MDs and DOs) 07 = Physician assistant 08 = Registered nurse 09 = Resident physician (includes MDs and DOs) 97 = Other, specify 99 = Unknown
KPWA Notes	• Site mapped coded VDW-based provider type as follows: ○ Fellow → 02 (Fellows) ○ Medical Assistant → 04 (Medical assistant) ○ Nurse Practitioner → 05 (Nurse practitioner) ○ Osteopath, Physician → 06 (Physician) ○ Ortho Phy Asst, Physician Assistant → 07 (Physician assistant) ○ Registered Nurse → 08 (Registered nurse) ○ Resident → 09 (Resident physician) ○ Unknown → 99 (Unknown) ○ All other coded provider types → 97 (Other, specify)
KPNC Notes	• Same as KPWA, with the following additions: ○ Certified Nurse Specialist, Certified Reg Nurse Anesthetist, Clinical Nurse Specialist, Diabetic Nurse, Nurse → 08 (Registered nurse)

PROVIDERTYPEOTHER

Definition	Other provider type
Type (Length)	Character (*)
Valid Values	See general and site-specific notes below.
General Notes	• Populated when ProviderType = 97 (Other, specify).

PROVIDERYEAR

Definition	Year when provider started working in health system
Type (Length)	Numeric (4)
Valid Values	Integers 1950–2020 or -99999 (unknown)
KPWA Notes	• Information was not available at this site; all records set to -99999.
KPNC Notes	• Site obtained this information from Epic/Clarity, although not all providers have records in this database. • Records always set to -99999 where ProviderIsPerson = 00 (No) or 99 (Unknown).
KPSC Notes	• Employment start date was highly missing in source data.

PROVIDER_BIRTH_YEAR

Definition	Year when provider was born
Type (Length)	Numeric (4)
Valid Values	Integers 1915–2000 or -99999 (unknown)
KPWA Notes	• Information was obtained from VDW Provider Specialty table where available. • Records always set to -99999 where ProviderIsPerson = 00 (No) or 99 (Unknown).
KPNC Notes	• Same as KPWA with the exception that birth years >2000 were ignored out of an abundance of caution in case of source data entry errors.
KPSC Notes	• Same as KPWA. Note that provider birth year was highly missing in source data.

PROVIDER_GENDER

Definition	Provider gender and/or sex
Type (Length)	Character (1)

Valid Values	M = Male
	F = Female
	O = Other, including transgender
	(blank) = Unknown

General Notes	• If both gender and sex are known, this variable should represent gender.
KPWA Notes	• Information was obtained from VDW Provider Specialty table. • Records always left blank where ProviderIsPerson = 00 (No) or 99 (Unknown).
UTSW Notes	• Records always left blank where ProviderIsPerson = 00 (No) or 99 (Unknown).
KPNC Notes	• Same as KPWA.

PROVIDER_RACE1

Definition	Provider race 1
Type (Length)	Character (1)

Valid Values	A = Asian
	B = Black or African American
	H = Native Hawaiian or Pacific Islander
	I = American Indian or Alaskan Native
	W = White
	M = Multiple races, not otherwise specified
	O = Other (values that do not fit well in any other category)
	(blank) = Unknown

KPWA Notes	• Information not available; all records left blank.
UTSW Notes	• Same as KPWA.

KPNC Notes	• Information was obtained from VDW Provider Specialty table where available. • Records always left blank where ProviderIsPerson = 00 (No) or 99 (Unknown).
KPSC Notes	• Same as KPWA.

PROVIDER_RACE2

Definition	Provider race 2
Type (Length)	Character (2)
Valid Values	Same as Provider_Race1
General Notes	• Information not available at any site; all records left blank.

PROVIDER_RACE3

Definition	Provider race 3
Type (Length)	Character (2)
Valid Values	Same as Provider_Race1
General Notes	• Information not available at any site; all records left blank.

PROVIDER_RACE4

Definition	Provider race 4
Type (Length)	Character (2)
Valid Values	Same as Provider_Race1
General Notes	• Information not available at any site; all records left blank.

PROVIDER_RACE5

Definition	Provider race 5
Type (Length)	Character (2)
Valid Values	Same as Provider_Race1
General Notes	• Information not available at any site; all records left blank.

PROVIDER_HISPANIC

Definition	Whether provider is of Hispanic or Latino origin
Type (Length)	Character (1)

Valid Values	Y = Yes
	N = No
	(blank) = Unknown

KPWA Notes	• Information not available; all records left blank.

UTSW Notes	- Same as KPWA.
KPNC Notes	- Information was obtained from VDW Provider Specialty table where available. - Records always left blank where ProviderIsPerson = 00 (No) or 99 (Unknown).
KPSC Notes	- Same as KPWA.

YEAR_GRADUATED

Definition	Year when provider graduated from medical school
Type (Length)	Numeric (4)
Valid Values	Integers 1953–2020 or -99999 (unknown)
KPWA Notes	- Information was obtained from VDW Provider Specialty table where available. - Records always left blank where ProviderIsPerson = 00 (No) or 99 (Unknown).
UTSW Notes	- Information not available at this site; all records left blank.
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as UTSW.

PROVIDERISPERSON

Definition	Whether the provider is a person (vs administrative entity)
Type (Length)	Character (2)
Valid Values	00 = No, provider is not a person 01 = Yes, provider is a person 99 = Unknown whether provider is a person
KPWA Notes	- Site defines a provider as an entity that delivers care; this entity may or may not be a person. - The 00 (No) values are those providers for which the site knows that the entity is really a clinic. All other records are set to 99. - Known non-person entities may still have associated specialty information documented in ProviderSpecialty or ProviderSpecialtyOther; however, their ProviderType is always 99 (Unknown).
UTSW Notes	- Known non-person entities (i.e., resources) were assigned 00 for this variable, although they may still have an associated specialty noted in ProviderSpecialty / ProviderSpecialtyOther or associated ProviderType as appropriate.
KPNC Notes	- Site obtained this information from Epic/Clarity, although not all providers have records in this database. - Known non-person entities were assigned 00 for this variable. Similarly, ProviderType was assigned 97 (Other); ProviderTypeOther was set to “Non-person provider, resource” or “Non-person provider, class”; and ProviderSpecialty was assigned 99 (Unknown) for all non-person entities.
KPSC Notes	- Site only included providers known to be people; all records set to 01.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Facility Table

Overview

This table contains one record per facility ID that appears in any of the tables described above. IDs should be unique within each Providing Site.

For the IMS3 data submission, facilities were defined as follows:

- KPWA: Facility ID is at the level of the medical center campus (owned/operated facilities) or street address (external facilities).
- UTSW: Facility ID is at the level of department or clinic (but can be rolled up to campus/building level using the FacilityIDRelatedPhys variable).
- KPNC: Facility ID is at the level of the medical center, hospital, or medical office (owned/operated facilities) or street address (external facilities). (Facilities owned/operated by KP can be identified using Relationship= O.)
- KPSC: Facility ID is at the level of the medical center, hospital, or medical office; only owned/operated facilities were included.

Data Sources

KPWA facility data is primarily sourced from the VDW Facility table, which contains a record for each facility that appears in claims or EMR records, and an internal administrative facility database, which provides supplemental information (e.g., facility type) for KPWA-owned and some network/contract facilities. When the facility type was not available from the internal administrative database, KPWA used place of service codes from related claims to extrapolate this information.

UTSW pulled facility data from the EMR.

KPNC facility data were sourced from the VDW Facility table.

KPSC unique facilities were identified and defined based on the VDW Facility table, with additional variables obtained from other internal tables.

Variables

FACILITYID

Definition	Facility ID
Type (Length)	Numeric (8)
Valid Values	Numeric values unique at the facility level within each site
General Notes	• A facility refers to a building with a distinct street address and represents a distinct physical location. • FacilityID + ProvidingSite = composite facility-level primary key • A given FacilityID should represent the same facility: ○ across all PROSPR II organ sites (cervical [METRICS], colorectal [PRECISE], and/or lung [LOTUS]); and ○ in the previous round of PROSPR funding (PROSPR I).
KPWA Notes	• KPWA did not re-use the PROSPR I FacilityIDs.
UTSW Notes	• UTSW facility ID is at the level of the clinic or specialty and thus may be physically co-located in the same building/campus with other facilities; however, facility IDs represent distinct resources (e.g., office staff).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

FACILITYSTFIP

Definition	State in which facility is located (as numeric FIPS code)
Type (Length)	Character (2)
Valid Values	##-formatted (leading zeros preserved) FIPS state numeric codes from the American National Standards Institute (ANSI) Codes for States or blank (unknown or not applicable, e.g., non-U.S. addresses)
KPWA Notes	• For owned and operated facilities (Relationship = O), this represents the state where the care delivery facility is located. • For other facilities (Relationship = E), this may represent the state where care was delivered, or it may represent the location of administrative offices.
UTSW Notes	• 48 (Texas) for all records.
KPNC Notes	• Same as KPWA.
KPSC Notes	• 06 (California) for all records.

FACILITYCOUNTYFIP

Definition	County in which facility is located (as numeric FIPS code)
Type (Length)	Character (3)

Valid Values	###-formatted (leading zeros preserved) FIPS county codes from [2020 FIPS codes]() or blank (unknown or not applicable, e.g., non-U.S. addresses)
KPWA Notes	• For KPWA-owned/-operated facilities (Relationship = O), this represents the county where the care delivery facility is located. • For other facilities, this may represent the county where care was delivered, or it may represent the location of administrative offices.
UTSW Notes	• 113 (Dallas County) for all records
KPNC Notes	• Same as KPWA.

FACILITYZIP

Definition	ZIP code in which facility is located
Type (Length)	Character (5)
Valid Values	#####-formatted (leading zeros preserved) 5-digit ZIP codes or blank (unknown or not applicable, e.g., non-U.S. addresses)
KPWA Notes	• For KPWA-owned/-operated facilities (Relationship = O), this represents the ZIP code of the location where care was delivered. • For other facilities, this may represent the ZIP code where care was delivered, or it may represent the location of administrative offices.
UTSW Notes	• A subset of youth and family clinics that are operationally linked but located across multiple buildings/campuses are rolled up to the same facility ID. The ZIP code for this facility ID is set to missing because the component clinics are located across multiple ZIP codes.
KPNC Notes	• Same as KPWA.

ADDRESS_FACILITY_TYPE

Definition	Facility address type
Type (Length)	Character (1)

Valid Values	C = Clinical facility
	B = Billing facility
	(blank) = Unknown

General Notes	• Requested in PCC’s PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx
KPWA Notes	• Variable was sourced from VDW Facility variable of the same name.
UTSW Notes	• All facilities are C (Clinical).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

FACILITYTYPE

Definition	Type of facility
Type (Length)	Character (2)
Valid Values	21 = Medical Center 22 = Hospital 23 = Emergency Room – Hospital 24 = Urgent Care Facility 25 = Ambulatory Surgical Center 26 = Office or Clinic 27 = Public Health Clinic 28 = Rural Health Clinic 29 = Federally Qualified Health Center 30 = Indian Health Service facility 31 = Tribal Facility 32 = Mobile Unit 33 = Laboratory 97 = Other, specify 99 = Unknown
KPWA Notes	• Internal facilities: ○ Most KPWA-owned/-operated facilities are medical centers (21), except for a few standalone walk-in clinics (set to 26 [Office or Clinic]), a very small inpatient facility (set to 22 [Hospital]), and a few administrative services locations (set to 97 [Other]). • External facilities: ○ If a facility exists in KPWA's internal administrative database, metadata are used to assign facility type of 21 (Medical Center), 22 (Hospital), 23 (Emergency Room – Hospital), or a handful of administrative services [97 (Other)]. ○ Other facilities were assigned to facility types above based on CMS place of service (POS) codes most frequently associated with incoming claims among facilities that appear in PRECISE cohort data. Note: 21 (Medical Center) does not correspond to an available place of service code.
UTSW Notes	• Parkland includes a hospital (22), an ambulatory surgical center (25), community-oriented primary care (COPC) sites (26), a homeless outreach program that includes a mobile unit (32), and labs and radiologic locations (33).
KPNC Notes	• KPNC-owned facilities were coded as medical centers (21), hospitals (22), offices or clinics (26), or other (97) based on the facility name. • External facilities were set to 99 (Unknown).
KPSC Notes	• Obtained from local facility name/description; most KPSC-owned facilities are described as medical centers or offices/clinics.

FacilityTypeOther

Definition	Other facility type
Type (Length)	Character (*)
Valid Values	See general and site-specific notes below

General Notes	- Populated when FacilityType = 97 (Other, specify).
KPWA Notes	- Valid values at this site are ADMINISTRATIVE FACILITY, DIALYSIS CENTER, RETAIL STORE, and SKILLED NURSING FACILITY.
UTSW Notes	- Not applicable; all records left blank.
KPNC Notes	- Valid values at this site are AACC (call center), Administration, Health Center, Other Health Care, and Providers Network.

FACILITYPC

Definition	Whether facility provides primary care
Type (Length)	Character (1)
Valid Values	0 = No 1 = Yes (blank) = Unknown
General Notes	- Requested in PCC's PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx with the following guidance: "Please define a primary care facility in a way that's most relevant to cancer screening at your site and include details in data documentation."
KPWA Notes	- Site populated this variable as follows: ○ If FacilityID appeared in the Encounter file (as FacilityIDEnc) → FacilityPC = 1 (Yes) ○ Otherwise → FacilityPC = blank (Unknown) ▪ Rationale: We cannot say that primary care does not happen at these facilities; we just did not observe any primary care at those facilities for PRECISE cohort.
UTSW Notes	- Site populated this variable as follows: ○ If facility has a community health, family practice, internal medicine, women's health, or geriatric specialty → FacilityPC = 1 (Yes) ○ Otherwise → FacilityPC = 0 (No)
KPNC Notes	- Site populated this variable as follows: ○ If FacilityType = 21 (Medical Center) or 26 (Office or Clinic) → FacilityPC = 1 (Yes) ○ Else if Relationship = O → FacilityPC = 0 (No) ○ Otherwise → FacilityPC = blank (Unknown)
KPSC Notes	- Site populated this variable based on internal facility code and description.

RELATIONSHIP

Definition	Relationship between facility and health care organization
Type (Length)	Character (1)
Valid Values	O = Owned and/or operated by health care organization E = External (incl. contracted) (blank) = Unknown
General Notes	- Requested in PCC's PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx
KPWA Notes	- Variable was sourced from the VDW Facility variable of the same name.

UTSW Notes	- All records set to O.
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as UTSW.

RELATIONSHIP_HISTORY

Definition	History of relationship between facility and health care organization
Type (Length)	Character (1)
Valid Values	O = Always owned and/or operated by health care organization E = Always external (incl. contracted) 1 = Was owned, most recently external 2 = Was external, most recently owned (blank) = Unknown
General Notes	- Requested in PCC's PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx
KPWA Notes	- Variable was sourced from the VDW Facility variable of the same name.
UTSW Notes	- All records set to O.
KPNC Notes	- Same as KPWA.
KPSC Notes	- Same as UTSW.

FACILITYIDRELATEDPHYS

Definition	Facility ID of associated larger physical facility
Type (Length)	Numeric (8)
Valid Values	Site-specific (see below) or -99999 (missing)
General Notes	- Requested in PCC's PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx with the following guidance: "If applicable, provide the Facility ID of a larger facility in the facility file that is physically associated with this facility based on location. For example, if 2 facilities are located within a hospital, there would be 1 record for each facility in the Facility file, and 1 record for the hospital in the Facility file for a total of 3 records—this physical relationship could be described by reporting the facility ID for the hospital under 'FacilityIDRelatedPhys' for the 2 facilities." - On 4/8/2021, the PRECISE DAU agreed that we would use this variable to capture Medical Center Area at KPNC/SC or hospital/facility at UTSW. (The concept does not apply at KPWA.)
KPWA Notes	- All records set to -99999.
UTSW Notes	- For this site, variable represents a distinct medical center, campus, or building for Parkland-UTSW and can be used to "roll up" distinct Facility IDs (representing clinic/specialty locations) to the same level at which Facility IDs were defined for the KP sites.

KPNC Notes	• For owned/operated facilities, this variable corresponds to one of 24 KPNC “medical center areas” with which the facility is associated. • For external facilities, variable is set to -99999.
KPSC Notes	• Variable corresponds to internal KPSC MSA (medical service area).

FACILITYIDRELATEDAGGR

Definition	Facility aggregating value
Type (Length)	Numeric (8)
Valid Values	Site-specific (see below) or -99999 (missing)
General Notes	• Requested in PCC’s PROSPR II Facility DRP v2.0 FINAL - 2021_02_12.xlsx with the following guidance: “If applicable, provide the Facility ID of a facility that is grouped with this facility. Intended for other types of facility aggregation, not based solely on location, with implications for the cancer screening process (e.g., organizational relationships, common policies/protocols, retired facilities, etc.). Include an explanation of aggregation categories.” • On 4/8/2021, the DAU agreed that we would use this variable to capture the Service Area for KPNC/SC. KPWA will use this variable to provide context about the type of facility. UTSW will not populate this variable.
KPWA Notes	• Valid values: ○ 1 = Primary or specialty care facility ○ 2 = Administrative facility ○ 3 = Care Clinic or urgent care facility ○ 4 = Facility outside study service area ○ 5 = Retired facility ○ 9 = Unknown (i.e., external facility)
UTSW Notes	• All records set to -99999.
KPNC Notes	• For owned/operated facilities, this variable corresponds to one of 15 KPNC “service areas” with which the facility is associated. • For external facilities, variable is set to -99999.
KPSC Notes	• Like FacilityIDRelatedPhys, this variable corresponds to internal KPSC MSA (medical service area).

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

SDoH Table

Overview

The Social Determinants of Health (SDoH) table contains one row per participant and provides characteristics of the participant's place of residence, ideally at time of cohort entry. Most of the variables in this table were provided by IMS in a 09/14/2020 file called prospr.sdoh.census.tract.level.data.csv (henceforth referred to as "the IMS file"). This file contains a row for every U.S. census tract along with accompanying characteristics of that tract. The IMS-provided variables will not be fully described here.

KPWA and UTSW are members of the METRICS cervical cancer PRECISE research center in addition to PRECISE; as such, some of the same participants exist in both PRECISE and METRICS cohorts. However, these people will not necessarily have the same SDoH data in both PRC data sets, given that they may have entered the cohorts at different times, while living at different addresses.

Data Sources

The three KP sites used VDW Census Tract Location table to obtain previously geocoded, point-in-time residential location data for participants (where available). There were a few instances in which a participant had a location available in the VDW, but that VDW geocode did not match any GeoIDs (census tracts) in the file provided by IMS. KPSC also used an internal historical address file to identify P.O. box addresses.

UTSW used EMR address history to identify address at cohort entry.

Variables

PID

Same as above.

GeoAddrPOBox

Definition	Whether address used for geocoding is a post office box
Type (Length)	Character (1)
Valid Values	0 = No 1 = Yes 8 = Not applicable (i.e., no address available for this participant)
General Notes	• If 1 (Yes), all subsequent data fields will be missing except for GeoAddrBegin variables, ExtractDate, and ProvidingSite. • If 8 (N/A), all subsequent data fields will be missing except for ExtractDate and ProvidingSite.

GeoAddrBeginMth

Definition	Month component of date used for geocoding
Type (Length)	Character (2)
Valid Values	MM-formatted months 01–12 or missing
General Notes	• Per PCC guidance, GeoAddrBegin variables should reflect the date of cohort entry if participant's address at that time was known. • Otherwise, GeoAddrBegin variables should reflect the date at which the participant's address was known. ○ Note: The PRECISE team decided that, if participant's address was known at multiple other time periods, we would select the address date that is closest to cohort entry in absolute terms, regardless of whether it occurred prior to or after entry (as long as it did not occur after cohort exit). • Variable will always be set to missing when GeoAddrPOBox = 8.

GeoAddrBeginDay

Definition	Day component of date used for geocoding
Type (Length)	Character (2)
Valid Values	DD-formatted days 01–31 or missing

General Notes	• Per PCC guidance, GeoAddrBegin variables should reflect the date of cohort entry if participant’s address at that time was known. • Otherwise, GeoAddrBegin variables should reflect the date at which the participant’s address was known. o The PRECISE team decided that, if participant’s address was known at multiple other time periods, we would select the address date that is closest to cohort entry in absolute terms, regardless of whether it occurred prior to or after entry (as long as it did not occur after cohort exit). • Variable will always be set to missing when GeoAddrPOBox = 8.
UTSW Notes	• Site does not provide full dates, so all records have this variable set to 15.

GeoAddrBeginYr

Definition	Year component of date used for geocoding
Type (Length)	Character (4)
Valid Values	YYYY-formatted years 1974–2020 or missing
General Notes	• Per PCC guidance, GeoAddrBegin variables should reflect the date of cohort entry if participant’s address at that time was known. • Otherwise, GeoAddrBegin variables should reflect the date at which the participant’s address was known. o The PRECISE team decided that, if participant’s address was known at multiple other time periods, we would select the address date that is closest to cohort entry in absolute terms, regardless of whether it occurred prior to or after entry (as long as it did not occur after cohort exit). • Variable will always be set to missing when GeoAddrPOBox = 8.

GeoAddrDSR

Definition	Date used for geocoding (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	1961–35062 or missing
General Notes	• Per PCC guidance, GeoAddrBegin variables should reflect the date of cohort entry if participant’s address at that time was known. • Otherwise, GeoAddrBegin variables should reflect the date at which the participant’s address was known. o The PRECISE team decided that, if participant’s address was known at multiple other time periods, we would select the address date that is closest to cohort entry in absolute terms, regardless of whether it occurred prior to or after entry (as long as it did not occur after cohort exit). • Variable will always be set to missing when GeoAddrPOBox = 8.

GeoCertainty

Definition	Level of geocoding certainty

Type (Length)	Character (1)
Valid Values	1 = Census tract based on complete and valid address 2 = Census tract based on ZIP+4 3 = Census based on residence ZIP+2 4 = Census tract based on ZIP only 5 = Census tract based on ZIP of P.O. box 6 = Census based on residence city/ZIP, city/ZIP has only one census tract 7 = Other 9 = Unknown (missing)
General Notes	• Per PCC: Use 7 (Other) if software's geocoding certainty output cannot be mapped to one of the options 1–6. Use 9 (Unknown) if geocoding certainty is unavailable. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.

GEOCERTAINTYOTHER

Definition	Other geocoding certainty level
Type (Length)	Character (37)
Valid Values	Site-specific values (see below) or missing
General Notes	• Per PCC: If GeoCertainty = 7 (Other), provide values for geocoding certainty here and describe what those values mean under Valid Values. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Not applicable for this site; all records set to missing.
UTSW Notes	• Valid values "Point of Interest", "Street only (no address number match)", or missing
KPNC Notes	• Valid values are "Street Intersection" or missing
KPSC Notes	• Same as KPWA.

GEOMATCHSCORE

Definition	Geocoding match score
Type (Length)	Numeric (4)
Valid Values	0–100 (valid match scores) or -99999 (missing/unknown)
General Notes	• Per PCC: Geocoding match score represents how well the candidate address string (from geocoding software) matches the participant's address string. • Variable will always be set to -99999 when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Information not available at this site; all records set to -99999.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

GEOSOFTWARENAME

Definition	Geocoding software name
Type (Length)	Character (29)
Valid Values	Software names or missing
General Notes	• Variable will always be set to missing when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Variable sourced from VDW Census Location variable GEOCODE_APP.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

GEOSOFTWAREVERSION

Definition	Geocoding software version
Type (Length)	Character (13)
Valid Values	Software versions or missing
General Notes	• Variable will always be set to missing when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Information not available; all records set to missing.
KPNC Notes	• Site coordinated with Strategic Programming Group to determine the version of GEMS/ESR software used; version information was not available for SAS.
KPSC Notes	• Same as KPWA.

[TRACT CHARACTERISTICS]

Definition	43 IMS-provided variables that contain characteristics of each census tract
Type (Length)	Numeric (8)
Valid Values	As provided by IMS or -99999 where missing
General Notes	• Variables will always be set to missing when GeoAddrPOBox = 1 or 8.
Variable List	• Race_NHisp_White • Race_NHisp_Black • Race_NHisp_AIAN • Race_NHisp_Asian • Race_NHisp_NatHaw • Race_NHisp_Other • Race_NHisp_Multi • Race_Hisp_White • Race_Hisp_Black • Race_Hisp_AIAN • Race_Hisp_Asian • Race_Hisp_NatHaw • Race_Hisp_Other • Race_Hisp_Multi

	• Educ_Less9th • Educ_9th_12th • Educ_HSGrad • Educ_SomeColl • Educ_AssocDeg • Educ_BachDeg • Educ_MastProfDeg • Educ_DoctDeg • Person_Below_Pov • RUCA4A • Flag_Tract_Pop_Zero • LQ_White_Alone • LQ_Black_Alone • LQ_API_Alone • LQ_Hispanic • LQ_NH_White_Alone • ICE_Black_Alone_White_Alone • ICE_API_Alone_White_Alone • ICE_Hispanic_NH_White_Alone • Lex_Is_Black_Alone_White_Alone • Lex_Is_API_Alone_White_Alone • Lex_Is_Hispanic_NH_White_Alone • LIS_White_Alone • LIS_Black_Alone • LIS_API_Alone • LIS_Hispanic • LIS_NH_White_Alone • Yost_Overall_Quintile • Yost_State_Quintile
UTSW Notes	• Yost_State_Quintile was erroneously left blank instead of set to -99999 for a subset of records; they should be treated as missing/-99999 for analysis.

ENC_ST

Definition	Encrypted state code
Type (Length)	Character (3)
Valid Values	A00–C99 or missing

General Notes	• Per PCC: "Provide an encrypted state code associated with this address of total variable length of 3 digits. METRICS and PRECISE will prepend A, B, C, or D to denote individual encryption keys, and LOTUS will prepend E. The remaining 2 digits will be pseudocoded by sites/PRCs." • PRECISE decided that KPNC and KPSC would use the same encryption crosswalk, while KPWA and UTSW would use site-specific crosswalks across both PRECISE and METRICS. • With these guidelines in mind, PRECISE encrypted state/territory FIPS codes as described in the site-specific notes below. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Site mapped each value in the IMS file State column to another list of two-digit codes, with the crosswalk to be maintained at KPWA (and used for both PRECISE and METRICS). • Each code was then prefaced by "A" (alpha equivalent of the 1 from ProvidingSite code 71) to prevent to prevent cross-site/-PRC misclustering.
UTSW	• Site mapped each value in the IMS file State column to another list of two-digit codes, with the crosswalk to be maintained at UTSW (and used for both PRECISE and METRICS). • Each code was then prefaced by "B" (alpha equivalent of the 2 from ProvidingSite code 72) to prevent cross-site/-PRC misclustering.
KPNC	• Site mapped each value in the IMS file State column to another list of two-digit codes, with the crosswalk to be maintained at KPNC for use by both KPNC and KPSC (given the potential for geographic overlap). • Each code was then prefaced by "C" (alpha equivalent of the 3 from ProvidingSite code 73) to prevent cross-site/-PRC misclustering.
KPSC	• See KPNC notes above.

ENC_COUNTY

Definition	Encrypted county code
Type (Length)	Character (4)
Valid Values	A000–C999 or missing
General Notes	• Per PCC: "Provide an encrypted county code associated with this address of total variable length of 4 digits. METRICS and PRECISE will prepend A, B, C, or D to denote individual encryption keys, and LOTUS will prepend E. The remaining 3 digits will be pseudocoded by sites/PRCs." • PRECISE decided that KPNC and KPSC would use the same encryption crosswalk, while KPWA and UTSW would use site-specific crosswalks across both PRECISE and METRICS. • With these guidelines in mind, PRECISE encrypted county FIPS codes as described in the site-specific notes below. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.

KPWA Notes	• Site mapped each value in the IMS file County column to another list of three-digit codes, with the crosswalk to be maintained at KPWA (and used for both PRECISE and METRICS). • Each code was then prefaced by "A" (alpha equivalent of the 1 from ProvidingSite code 71) to prevent cross-site/-PRC misclustering.
UTSW	• Site mapped each value in the IMS file County column to another list of three-digit codes, with the crosswalk to be maintained at UTSW (and used for both PRECISE and METRICS). • Each code was then prefaced by "B" (alpha equivalent of the 2 from ProvidingSite code 72) to prevent cross-site/-PRC misclustering.
KPNC	• Site mapped each value in the IMS file County column to another list of three-digit codes, with the crosswalk to be maintained at KPNC for use by both KPNC and KPSC (given the potential for geographic overlap). • Each code was then prefaced by "C" (alpha equivalent of the 3 from ProvidingSite code 73) to prevent cross-site/-PRC misclustering.
KPSC	• See KPNC notes above.

ENC_TRACT

Definition	Encrypted census tract
Type (Length)	Character (7)
Valid Values	A000000–C999999 or missing
General Notes	• Per PCC: "Provide an encrypted census tract code associated with this address of total variable length of 7 digits. METRICS and PRECISE will prepend A, B, C, or D to denote individual encryption keys, and LOTUS will prepend E. The remaining 6 digits will be pseudocoded by sites/PRCs." • PRECISE decided that KPNC and KPSC would use the same encryption crosswalk, while KPWA and UTSW would use site-specific crosswalks across both PRECISE and METRICS. • With these guidelines in mind, PRECISE encrypted census tract FIPS codes as described in the site-specific notes below. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.
KPWA Notes	• Site mapped each value in the IMS file Tract column to another list of six-digit codes, with the crosswalk to be maintained at KPWA (and used for both PRECISE and METRICS). • Each code was then prefaced by "A" (alpha equivalent of the 1 from ProvidingSite code 71) to prevent cross-site/-PRC misclustering.
UTSW	• Site mapped each value in the IMS file Tract column to another list of six-digit codes, with the crosswalk to be maintained at UTSW (and used for both PRECISE and METRICS). • Each code was then prefaced by "B" (alpha equivalent of the 2 from ProvidingSite code 72) to prevent cross-site/-PRC misclustering.

KPNC	• Site mapped each value in the IMS file Tract column to another list of six-digit codes, with the crosswalk to be maintained at KPNC for use by both KPNC and KPSC (given the potential for geographic overlap). • Each code was then prefaced by “C” (alpha equivalent of the 3 from ProvidingSite code 73) to prevent cross-site/-PRC misclustering.
KPSC	• See KPNC notes above.

ENC_FIPS

Definition	Encrypted combined FIPS code
Type (Length)	Character (14)
Valid Values	A00A000A000000–C99C999C999999 or missing
General Notes	• Per PCC: “Provide an encrypted 14-digit ID representing the combined FIPS code (3-char Enc_St + 4-char Enc_County + 7-char Enc_Tract).” • PRECISE decided that KPNC and KPSC would use the same encryption crosswalk, while KPWA and UTSW would use site-specific crosswalks across both PRECISE and METRICS. • With these guidelines in mind, each PRECISE site: ○ Mapped each combination of State, County, and Tract in the IMS-provided file to the corresponding encrypted state, county, and tract codes described above. ○ Created Enc_FIPS by concatenating the three encrypted values, Enc_St + Enc_County + Enc_Tract. • Variable will always be set to missing when GeoAddrPOBox = 1 or 8.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

COD Table

Overview

The COD (Cause of Death) table contains one row for every participant who exited the cohort due to death, as documented in the Enrollment Table (i.e., where CutoffReason = 07–09 [Death-related categories]).

Data Sources

KPWA used the VDW Cause of Death table, which incorporates information from multiple sources. The primary source of cause of death data is the Washington state vital records department. Cause of death data were fully complete through 2019 and provisionally complete through Q3 2020 for the IMS3 data submission.

UTSW's source for cause of death data was the Texas Cancer Registry: i.e., only cohort members who died after being diagnosed with colorectal cancer in Texas would have their cause of death information captured here.

KPNC also used the VDW Cause of Death table, with California state death data considered complete through 2020 for the IMS3 data submission.

KPSC obtained cause of death information from state death files as well as internal hospital discharge codes. Cause of death data were complete through 2020 for the IMS3 data submission.

See About the Data User Guide > Data Sources > Deaths for more information on death data availability across sites.

Variables

PID

Same as above.

DEATHDATE

Definition	Date of death (as SAS date value)
Type (Length)	Numeric (4)
Valid Values	01/02/2010–12/31/2020

<table>
<tr><td>General Notes</td><td>
<ul>
<li>Start of valid value range reflects the fact that participant cannot enter and exit the cohort on the same day.</li>
<li>Valid values reflect requirement that each participant must spend at least one day in the cohort.</li>
<li>Value in this field must match CutoffReasonOther</li>
</ul>
<table>
<tr><td>Definition</td><td>Other reason for end of PRECISE enrollment</td></tr>
<tr><td>Type (Length)</td><td>Character (1)</td></tr>
<tr><td>Valid Values</td><td>Missing (blank) at all sites</td></tr>
<tr><td>General Notes</td><td><ul><li>Newly requested by PCC for IMS3, to be populated when CutoffReason = 95 (Other); not applicable for any PRECISE sites.</li></ul></td></tr>
</table>
<ul>
<li>CutoffDate in the Enrollment Table: i.e., this table should not include any deaths not already documented elsewhere.</li>
</ul>
</td></tr>
<tr><td>UTSW Notes</td><td><ul><li>All dates are set to day 15 of the given month/year.</li></ul></td></tr>
</table>

DeathDSR

Definition	Date of death (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14611–35063
General Notes	• Valid values reflect participant being 40–95 years old during cohort eligibility as well as the requirement that each participant spend at least one day in the cohort. • Value in this field must match CutoffDSR in the Enrollment Table: i.e., this table should not include any deaths not already documented elsewhere.

DeathCause

Definition	Underlying cause of death (ICD-10 code)
Type (Length)	Character (8)
Valid Values	ICD-10 mortality codes (known) or blank (unknown)
General Notes	• Valid ICD-10 mortality codes contain at least three characters and begin with A00–Z99; include decimal points where applicable. • This field captures only the underlying cause of death, not other causes (e.g., immediate, contributory).
UTSW Notes	• Cause of death data were only available for participants with colorectal cancer records in the Texas Cancer Registry.

DeathType

Definition	Underlying cause of death (categorical)
Type (Length)	Character (2)

Valid Values	01 = Colorectal cancer 02 = Cardiovascular disease 03 = Other 99 = Unknown
General Notes	• Valid values were assigned as follows: ○ If DeathCause starts with C18, C19, or C20 → 01 (CRC) ○ Else if DeathCause equals C26.0, C78.5, D01.0, D01.1, or D01.2 → 01 (CRC) ○ Else if DeathCause starts with I00 through I78 → 02 (CVD) ○ Else if DeathCause is not missing → 03 (Other) ○ Otherwise → 99 (Unknown)
UTSW Notes	• Cause of death data were only available for participants with colorectal cancer records in the Texas Cancer Registry.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Diagnosis Table

Overview

The Diagnosis table will be prepared after the original IMS3 submission for use in PRECISE-only analyses; **it will not be transferred to IMS**.

The table will contain one row per participant per relevant diagnosis code per day during cohort eligibility. (Note: KPNC and KPSC also included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.) The diagnosis codes included in this table are related to potential harms of CRC screening and/or indications that can trigger lower endoscopy. Therefore, this table can be used for the following purposes:

- To identify events that may be considered harms of CRC screening; and
- To identify participants with colorectal symptoms or other findings that would exclude them from analyses of screening (which, by definition, occurs in patients without signs or symptoms of cancer). Doing so will also require use of other tables. See Appendix F. Identifying Screening-Eligible Participants for more information.

For the build covering the 2010–2020 IMS3 cohort, this table will include codes associated with in-person encounters (specifically ambulatory visits, emergency department visits, inpatient stays, and institutional stays) as well as codes from selected virtual care encounters (namely scheduled telephone calls, scheduled video visits, and synchronous online chats). Additionally, laboratory-only encounters may be included *only for iron-deficiency anemia diagnoses at KPNC/KPSC*.

Data Sources

KPWA sourced these data from the VDW Diagnosis table.

KPNC used both the VDW Diagnosis and Laboratory Results tables along with Epic/Clarity problem list data.

KPSC used the VDW Diagnosis table along with a non-VDW internal laboratory results data feed.

UTSW used EMR data.

Variables

PID

Same as above.

DIAGDATE

Definition	Date of diagnosis (as SAS date value)
Type (Length)	Numeric (4)
Valid Values	01/01/2010–12/31/2020

General Notes	• For diagnoses associated with inpatient or institutional stays, this field represents the admit date.
UTSW Notes	• All dates are set to day 15 of the given month/year.
KPNC Notes	• Site included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

DIAGDSR

Definition	Date of diagnosis (as days since reference, i.e., participant DOB)
Type (Length)	Numeric (4)
Valid Values	Integers 14610–35063
General Notes	• For diagnoses associated with inpatient or institutional stays, this field represents the admit date. • Valid values reflect participant being 40–95 years old during cohort eligibility.
KPNC Notes	• Site included diagnoses that occurred between PRECISE-eligible enrollment periods; analysts should use the Enrollment table to restrict to during-cohort events.
KPSC Notes	• Same as KPNC.

DIAGCODE

Definition	Event code
Type (Length)	Character (8)
Valid Values	ICD-9-CM and ICD-10-CM diagnosis codes from the Relevant Symptoms & Conditions list or “LOCAL” for relevant laboratory-based diagnoses of iron-deficiency anemia (KPNC/KPSC only) and hereditary disorders that increase CRC risk (KPNC only)

DIAGCODETYPE

Definition	Type of code
Type (Length)	Character (2)
Valid Values	09 = ICD-9-CM diagnosis 10 = ICD-10-CM diagnosis LR = Laboratory result (KPNC/KPSC only) OT = Other local Dx code (KPNC only, for hereditary disorders)

DIAGAV

Definition	Whether code was associated with one or more ambulatory visit encounters on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Includes outpatient clinic visits, same-day surgeries, observation beds, urgent care visits, and same-day ambulatory hospital encounters. • Excludes emergency department encounters.
KPWA Notes	• Site populated this variable as Y (Yes) when the VDW Diagnosis ENCTYPE variable = AV, otherwise N (No).
UTSW Notes	• All records left blank (not applicable). Site was unable to distinguish whether a diagnosis was associated with an ambulatory visit.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

DIAGED

Definition	Whether code was associated with one or more emergency department encounters on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Excludes urgent care encounters.
KPWA Notes	• Site populated this variable as Y (Yes) when the VDW Diagnosis ENCTYPE variable = ED, otherwise N (No).
UTSW Notes	• All records left blank (not applicable). Site was unable to distinguish whether a diagnosis was associated with an emergency department visit.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

DIAGIP

Definition	Whether code was associated with one or more acute inpatient encounters in which the participant was admitted on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Includes inpatient stays, same-day hospital discharges, hospital transfers when patient was admitted into hospital, acute inpatient pysch, and detox stays.

KPWA Notes	• Site populated this variable as Y (Yes) when the VDW Diagnosis ENCTYPE variable = IP, otherwise N (No).
UTSW Notes	• Site populated this variable as Y (Yes) when the diagnosis occurred during an inpatient stay, otherwise N (No).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

DIAGIS

Definition	Whether code was associated with one or more non-acute institutional stays in which the participant was admitted on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Includes hospice, skilled nursing facility, rehab, nursing home, residential, overnight non-hospital dialysis, and other non-hospital stays.
KPWA Notes	• Site populated this variable as Y (Yes) when the VDW Diagnosis ENCTYPE variable = IS, otherwise N (No).
UTSW Notes	• Site does not have information on institutional stays; all records left blank (Not applicable).
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

DIAGVC

Definition	Whether code was associated with one or more virtual care encounters on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Pertains to the following types of virtual care encounter: scheduled telephone calls, scheduled video visits, and synchronous online chats
UTSW Notes	• All records left blank. Site was unable to distinguish whether a diagnosis was associated with a virtual visit.

DIAGLO

Definition	Whether code was associated with one or more lab-only encounters on this date
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes
	(blank) = Unknown or not applicable

General Notes	• Includes lab only encounters that cannot be matched to another encounter. • This variable only applies to KPNC and KPSC, which will be supplementing their iron-deficiency anemia diagnoses with laboratory results.
KPWA Notes	• Not applicable at this site; all records left blank (Not applicable).
UTSW Notes	• Same as KPWA.
KPNC Notes	• Site populated this variable as Y (Yes) when laboratory-based diagnosis of iron-deficiency anemia was detected using VDW laboratory results data (also coded as DiagCodeType = LR and DiagCode = LOCAL in this table); otherwise, site coded this variable as N (No). • See DiagIronDef for information on how iron-deficiency anemia diagnosis was defined.
KPSC Notes	• Same as KPNC.

DIAGHARM

Definition	Whether code denotes any potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• This flag indicates that the code is any of the following specific potential harms, each of which has its own flag listed below: Splenic injury, colonic perforation, some forms of GI bleed, some forms of diverticulitis, acute appendicitis, bacteremia, stroke, myocardial infarction, and shock due to anesthesia. • Corresponds to Relevant Symptoms & Conditions codes where DIAGHARM = 1.

DIAGSPLINJ

Definition	Whether code denotes splenic injury, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where SPLENINJ = 1.

DIAGPERF

Definition	Whether code denotes colonic perforation, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where PERFORATION = 1.

DIAGAPPEND

Definition	Whether code denotes acute appendicitis, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where APPENDICITIS = 1.

DIAGBACT

Definition	Whether code denotes bacteremia, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where BACTEREMIA = 1.

DIAGSTROKE

Definition	Whether code denotes a stroke, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where STROKE = 1.

DIAGMI

Definition	Whether code denotes myocardial infarction, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where MI = 1.

DIAGSHOCK

Definition	Whether code denotes anesthetic shock, a potential harm of CRC screening
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where ANESTHESIASHOCK = 1.

DIAGBLEED

Definition	Whether code denotes GI bleed, which can be both a harm of CRC screening and an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where GIBLEED = 1. • Some GI bleeds are harms only, others are lower endoscopy indications only, and others are both. The DiagSx and DiagHarm flags can be used to distinguish these associations.

DIAGDIVER

Definition	Whether code represents diverticulitis, a symptom that may be a harm of CRC screening as well as an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where DIVERTICULITIS = 1. • All included diverticulitis diagnoses are considered lower endoscopy indications; some of them are also classified as harms. The subset can be distinguished using the DiagHarm flag.

DIAGSX

Definition	Whether code denotes a symptom (including anemia) that is a possible indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• This flag summarizes the presence of any of the following symptoms that may be indications for lower endoscopy: GI bleeding, diverticulitis, abdominal pain, abdominal mass, weight loss, change in bowel habits, constipation, iron-deficiency and other relevant anemias, diarrhea, and other colonic disorders. • Note the inclusion of anemias in this variable. • Each symptom included in this summary has its own specific flag below. • Flag was populated somewhat differently across sites; see site-specific notes below for more information.
KPWA Notes	• Set to Y for Dx codes in the Relevant Symptoms & Conditions list where SYMPTOMS = 1.
UTSW Notes	• Same as KPWA.

KPNC Notes	- Set to Y for: ○ Dx codes in the Relevant Symptoms & Conditions list where SYMPTOMSNOANEMIA = 1, *and/or* ○ Laboratory-based diagnoses of iron-deficiency anemia (coded as DiagCodeType = LR and DiagCode = LOCAL in this table). See DiagIronDef for information on how iron-deficiency anemia diagnosis was defined.
KPSC Notes	- Same as KPNC.

DIAGABDPAIN

Definition	Whether code denotes abdominal pain, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	- Corresponds to Relevant Symptoms & Conditions codes where ABDPAIN = 1.

DIAGABDMASS

Definition	Whether code denotes an abdominal mass, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	- Corresponds to Relevant Symptoms & Conditions codes where ABDMASS = 1.

DIAGWTLOSS

Definition	Whether code denotes weight loss or being underweight, a finding that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

DRP Details	- Corresponds to Relevant Symptoms & Conditions codes where WEIGHTLOSS = 1.

DIAGCHANGE

Definition	Whether code denotes a change in bowel habits, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where BOWELHABITS = 1.

DIAGCONST

Definition	Whether code denotes constipation, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where CONSTIPATION = 1.

DIAGIRONDEF

Definition	Whether code denotes iron-deficiency anemia, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Flag was populated somewhat differently across sites; see site-specific notes below for more information. • KPNC & KPSC should also set this flag to Y when DiagCodeType = LR and DiagCode = LOCAL, i.e., laboratory results that correspond to diagnoses of iron-deficiency anemia.
KPWA Notes	• Corresponds to Relevant Symptoms & Conditions codes where ANEMIAIDA = 1.
UTSW Notes	• Same as KPWA.
KPNC Notes	• Same as KPWA with additional laboratory result-based diagnoses identified as follows: ○ Identify abnormally low hemoglobin or hematocrit laboratory test results. ○ Look for iron-deficiency anemia diagnostic tests with abnormal results (i.e., low ferritin, low iron, low transferrin saturation ratio, low transferrin % saturation, high total iron binding capacity) on or within 90 days *before or after* the low hemoglobin or hematocrit result. ○ If both abnormal results are found, IDA diagnosis is set to the date of the *diagnostic test* result.
KPSC Notes	• Iron-deficiency anemia diagnosis was defined as follows: ○ Identify abnormally low hemoglobin or hematocrit laboratory test results. ○ Look for iron-deficiency anemia diagnostic tests with abnormal results (i.e., low ferritin, low iron, low transferrin saturation ratio, low transferrin % saturation, high total iron binding capacity) within 90 days *before or after* the low hemoglobin or hematocrit result. ○ If both types of abnormal results are found, IDA diagnosis was set to the date of the *diagnostic test result*.

DIAGOTHANEMIA

Definition	Whether code denotes other relevant (i.e., not iron-deficiency) anemia, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where ANEMIAOTHER = 1.

DIAGDIARR

Definition	Whether code denotes diarrhea, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where DIARRHEA = 1.

DIAGOTHCOL

Definition	Whether code denotes another colonic disorder not otherwise specified above, a symptom that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where OTHERCOLODISORDER = 1.

DIAGIBD

Definition	Whether code denotes inflammatory bowel disease (including sequelae), a possible indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where IBD = 1.

DIAGIBDONLY

Definition	Whether code denotes inflammatory bowel disease (excluding sequelae), a possible indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where IBDONLY = 1. • This variable can be used when a more specific (but possibly less sensitive) measure of IBD is desired.

DIAGCRC

Definition	Whether code denotes a diagnosis of colorectal cancer, which is a possible indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where CODE_TYPE = DX and HXCRC = 1.

DIAGPOLYP

Definition	Whether code denotes colorectal polyp(s), a finding that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where RECENTPOLYP = 1.

DIAGABNIMG

Definition	Whether code denotes abnormal finding(s) on diagnostic imaging, a finding that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to Relevant Symptoms & Conditions codes where RECENTPX = 1.

DIAGHERED

Definition	Whether code denotes diagnosis of a hereditary disorder that increases CRC risk, a finding that may be an indication for lower endoscopy
Type (Length)	Character (1)

Valid Values	N = No
	Y = Yes

General Notes	• Corresponds to diagnoses of Lynch syndrome (a.k.a. hereditary nonpolyposis colorectal cancer [HNPCC]) or Familial adenomatous polyposis (FAP).
KPWA Notes	• Information not available at this site; all records set to N.

UTSW Notes	• Same as KPWA.
KPNC Notes	• Site used local Clarity problem list codes to identify active diagnoses of FAP and HNPCC. For HNPCC, this implies that the diagnosis was confirmed by genetic testing.
KPSC Notes	• Same as KPWA.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Comorbidity Table

Overview

The Comorbidity table will be prepared after the original IMS3 submission for use in PRECISE-only analyses; **it will not be transferred to IMS**.

This table will contain one record per participant per calendar quarter during which the participant was enrolled in the cohort. E.g., if a person enters the cohort on 03/01/2014 and exits on 11/30/2016, then they should have distinct Comorbidity records for 2014 Q1–Q4, 2015 Q1–Q4, and 2016 Q1–Q4.

See Appendix C. Charlson Comorbidity Index for information on site-specific differences in calculating the Charlson score. Of particular note: The PCC's requested addition of telehealth-inclusive Charlson measures (as seen in the CalendarYear Table) was *not* implemented here; KP sites included codes from in-person encounters only, whereas UTSW included codes regardless of encounter type.

Data Sources

The KP sites used VDW Diagnosis and Procedure tables. UTSW used diagnosis and procedure codes from EMR in addition to HIV laboratory results and HIV clinic visits, also from EMR.

Variables

PID

Same as above.

CALENDARYR

Definition	Calendar year
Type (Length)	Character (4)
Valid Values	YYYY-formatted years 2010–2020

CALENDARQT

Definition	Calendar quarter
Type (Length)	Character (1)
Valid Values	1 = January 1 through March 31 2 = April 1 through June 30 3 = July 1 through September 30 4 = October 1 through December 31
General Notes	• Calculations for Q1–Q3 of participant's first year in the cohort will utilize prior-to-entry data if available, similar to how CharlsonPrior was calculated in the PriorToCohortEntry Table.

CHARLSONQT

Definition	Charlson comorbidity score for the 365 days including the last day of the given quarter
Type (Length)	Numeric (4)
Valid Values	Integers 0–29 (valid scores) or -99999 (insufficient enrollment -or- no eligible encounters during window; UTSW only); see site-specific notes for more details
General Notes	• Observation window includes 365-day period that varies by CalendarQt as follows: ○ 1: April 1 (common years) or April 2 (leap years) of previous year through March 31 ○ 2: July 1 (common years) or July 2 (leap years) of previous year through June 30 ○ 3: October 1 (common years) or October 2 (leap years) of previous year through September 30 ○ 4: January 1 (common years) or January 2 (leap years) through December 31 of the same year • Variable was set to -99999 if the participant did not have full 365-day observation window (possibly including prior-to-entry data) for a given quarter. • For information on site-specific differences, see Appendix C. Charlson Comorbidity Index.
KPWA Notes	• Variable was set to 0 if the participant had full 365 days of observation but had no inpatient or ambulatory encounters in which to observe relevant billing codes.

UTSW Notes	• Variable was set to -99999 if participant had no encounters in which to observe relevant billing codes or if participant had insufficient observation time.
KPNC Notes	• Same as KPWA.
KPSC Notes	• Same as KPWA.

COMORBMYOCARDIAL

Definition	Myocardial infarction flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBCONGHEART

Definition	Congestive heart disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBVASPERIPH

Definition	Peripheral vascular disorder flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBVASCEREBRO

Definition	Cerebrovascular disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBDEMENTIA

Definition	Dementia flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBPULMCHRONIC

Definition	Chronic pulmonary disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBRHEUMATIC

Definition	Rheumatologic disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBPEPTICULCER

Definition	Peptic ulcer disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBLIVERMILD

Definition	Mild liver disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBDIABETES

Definition	Diabetes flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBDIABETESCOMP

Definition	Diabetes with chronic complications flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBHEMIPLEGIA

Definition	Hemiplegia or paraplegia flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBRENAL

Definition	Renal disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBMALIGNANCY

Definition	Malignancy (including leukemia and lymphoma) flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBLIVER

Definition	Moderate or severe liver disease flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBTUMOR

Definition	Metastatic solid tumor flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonQt = -99999)

COMORBHIV

Definition	HIV/AIDS flag
Type (Length)	Character (1)
Valid Values	1/0 (Yes/No) or blank (when CharlsonIndex = -99999)
UTSW Notes	• Site used additional information beyond diagnosis codes to implement this variable. See Appendix C. Charlson Comorbidity Index for more information.

EXTRACTDATE

Same as above.

PROVIDINGSITE

Same as above.

Appendices

Index

Appendix A. Primary Care Visits

The definition of what constitutes a primary care visit is relevant to both the PriorToCohortEntry Table and Encounter Table. Definitions varied by site and are described below.

KPWA

In-person primary care visits at KPWA were identified using data in the VDW Encounter and Provider Specialty tables. Potential visits were defined as outpatient clinic visits (excluding urgent care) in which the provider was a physician, osteopath, physician assistant, or nurse practitioner.

KPWA then limited to visits where either department or provider specialty indicated possible primary care delivery:

- Department: Family Medicine, Gerontology/Geriatrics, Internal Medicine, OB/GYN, Preventive Medicine, Primary Care, Women’s Health, or Unknown; and/or
- Provider specialty: Family Medicine, Gerontology/Geriatrics, Internal Medicine, OB/GYN, Preventive Medicine, or Unknown.

Finally, the following combinations of department and provider specialty were then excluded:

- Department and provider specialty are both OB/GYN, or
- Department and provider specialty are both unknown.

When incorporating virtual primary care visits in 2019–2020 as requested by the PCC, the definition above was simply expanded to include not just outpatient clinic visits but also scheduled video visits, scheduled telephone calls, and synchronous chat encounters.

UTSW

In-person primary care visits are completed outpatient office visits/appointment encounters at a community health, family practice, internal medicine, women’s health or geriatric department with a physician, resident, attending, fellow, nurse practitioner, physician assistant, or PGY1–PGY9.

When incorporating virtual primary care visits in 2019–2020 as requested by the PCC, the definition above was expanded to include scheduled telephone visits and video visits.

KPNC/KPSC

In-person primary care visits at KPNC and KPSC were identified using data in the VDW Encounter table. These visits were defined as outpatient clinic visits (excluding urgent care) in any of the following departments: Community Health, Family Practice, Gerontology/Geriatrics, Internal Medicine, Obstetrics/Gynecology, or Primary Care.

KPNC took the additional step of deduplicating in-person primary care visits to the person-day level and selected the “top” daily provider by hierarchy: Physician > Resident > Fellow MD > Ph. Assistant > Registered nurse > Licensed practical nurse > Nurse practitioner > Medical assistant > Administrative staff. (Rationale: Site did not limit providers by specialty, and data can have multiple provider records for the same visit.)

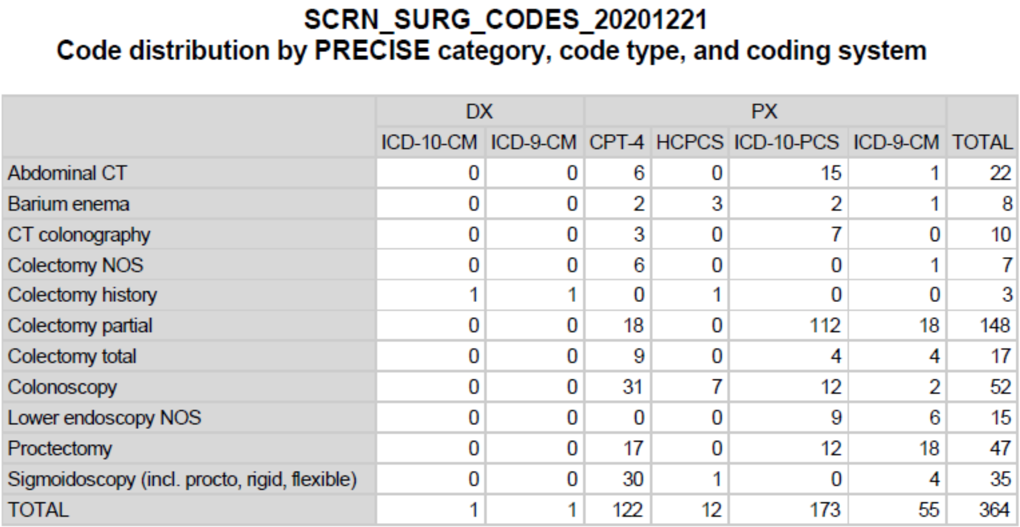
When incorporating virtual primary care visits in 2019–2020 as requested by the PCC, both KPNC and KPSC simply expanded the definition above to include not just outpatient clinic visits but also scheduled video visits, scheduled telephone calls, and synchronous chat encounters.

Appendix B. Code Lists

Three main types of harmonized code list were developed for the IMS2+ PRECISE data deliveries; see additional details below. The general approach to code list development was as follows: identify a broad range of potential codes (e.g., chapters in a given coding system); ask reviewers to independently decide whether each code in that range should be included; meet to compare and adjudicate selections (incl. comparing against prior study lists where applicable); and decide upon a final list of codes to define each concept, using systematic rules wherever possible.

CRC Screening & GI Surgery

For IMS+, PRECISE used a list of 364 billing codes to identify the following categories of CRC screening and GI surgery procedures:



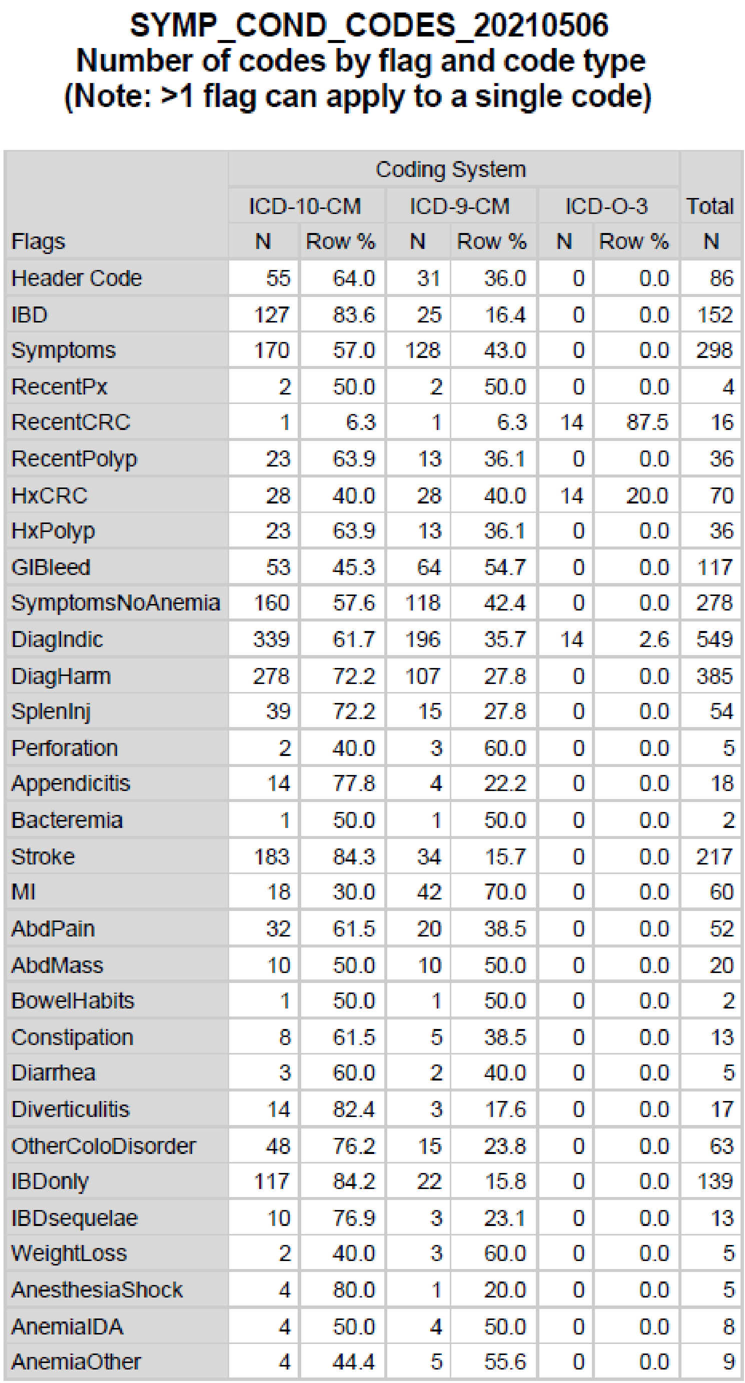
SCRN_SURG_CODES_20201221
Code distribution by PRECISE category, code type, and coding system

	DX		PX				
	ICD-10-CM	ICD-9-CM	CPT-4	HCPCS	ICD-10-PCS	ICD-9-CM	TOTAL
Abdominal CT	0	0	6	0	15	1	22
Barium enema	0	0	2	3	2	1	8
CT colonography	0	0	3	0	7	0	10
Colectomy NOS	0	0	6	0	0	1	7
Colectomy history	1	1	0	1	0	0	3
Colectomy partial	0	0	18	0	112	18	148
Colectomy total	0	0	9	0	4	4	17
Colonoscopy	0	0	31	7	12	2	52
Lower endoscopy NOS	0	0	0	0	9	6	15
Proctectomy	0	0	17	0	12	18	47
Sigmoidoscopy (incl. procto, rigid, flexible)	0	0	30	1	0	4	35
TOTAL	1	1	122	12	173	55	364

The list was compiled at KPWA and distributed for use at the other three sites. To facilitate complete data capture, both UTSW and KPNC had to include additional local codes, which were mapped by each site to the same set of standard categories shown above. The full SCRN_SURG_CODES list is available upon request.

Relevant Symptoms & Conditions

PRECISE developed a list of 913 billing codes that could be used identify various symptoms and conditions relevant to the KP indication algorithm and/or planned PRC-specific harms analyses:



SYMP_COND_CODES_20210506
Number of codes by flag and code type
(Note: >1 flag can apply to a single code)

	Coding System						
	ICD-10-CM		ICD-9-CM		ICD-O-3		Total
Flags	N	Row %	N	Row %	N	Row %	N
Header Code	55	64.0	31	36.0	0	0.0	86
IBD	127	83.6	25	16.4	0	0.0	152
Symptoms	170	57.0	128	43.0	0	0.0	298
RecentPx	2	50.0	2	50.0	0	0.0	4
RecentCRC	1	6.3	1	6.3	14	87.5	16
RecentPolyp	23	63.9	13	36.1	0	0.0	36
HxCRC	28	40.0	28	40.0	14	20.0	70
HxPolyp	23	63.9	13	36.1	0	0.0	36
GIBleed	53	45.3	64	54.7	0	0.0	117
SymptomsNoAnemia	160	57.6	118	42.4	0	0.0	278
DiagIndic	339	61.7	196	35.7	14	2.6	549
DiagHarm	278	72.2	107	27.8	0	0.0	385
SplenInj	39	72.2	15	27.8	0	0.0	54
Perforation	2	40.0	3	60.0	0	0.0	5
Appendicitis	14	77.8	4	22.2	0	0.0	18
Bacteremia	1	50.0	1	50.0	0	0.0	2
Stroke	183	84.3	34	15.7	0	0.0	217
MI	18	30.0	42	70.0	0	0.0	60
AbdPain	32	61.5	20	38.5	0	0.0	52
AbdMass	10	50.0	10	50.0	0	0.0	20
BowelHabits	1	50.0	1	50.0	0	0.0	2
Constipation	8	61.5	5	38.5	0	0.0	13
Diarrhea	3	60.0	2	40.0	0	0.0	5
Diverticulitis	14	82.4	3	17.6	0	0.0	17
OtherColoDisorder	48	76.2	15	23.8	0	0.0	63
IBDonly	117	84.2	22	15.8	0	0.0	139
IBDsequelae	10	76.9	3	23.1	0	0.0	13
WeightLoss	2	40.0	3	60.0	0	0.0	5
AnesthesiaShock	4	80.0	1	20.0	0	0.0	5
AnemiaIDA	4	50.0	4	50.0	0	0.0	8
AnemiaOther	4	44.4	5	55.6	0	0.0	9

The list was compiled at KPWA and distributed for use at the other three sites. No local modifications were required. The full SYMP_COND_CODES list is available upon request.

Pathology

COLORECTAL LOCATION

The following SNOMED T (Topography) codes were used to identify colorectal tissue in KPNC and KPSC pathology data:

Code	Location	Left/Right/Rectum
T67000	Colon, NOS	
T67010	Colonic mucous membrane	
T67012	Colonic gland, NOS	
T67015	Colonic epithelium	
T67016	Colonic lamina propria	
T67030	Colonic submucosa	
T67040	Colonic muscularis propria	
T67050	Colonic solitary lymphoid nodule	
T67060	Appendix epiploica	
T67080	Colonic subserosa	
T67090	Colonic serosa	
T67100	Cecum	Right
T67200	Ascending colon	Right
T67300	Right colic flexure	Right
T67400	Transverse colon	Right
T67500	Left colic flexure	Left
T67600	Descending colon	Left
T67700	Sigmoid colon	Left
T67800	Mesentery of colon, NOS	
T67810	Mesentery of ascending colon	Right
T67820	Transverse mesocolon	Right
T67830	Mesentery of descending colon	Left
T67840	Mesentery of sigmoid colon	Left
T67860	Pericolic tissue	
T67910	Colon and abdominal wall, CS	
T67920	Colon and rectum, CS	
T67950	Colon and skin, CS	
T67965	Cecum and ascending colon, CS	Right
T67995	Descending colon and sigmoid colon, CS	Left
T68000	Rectum, NOS	Rectum
T68010	Rectal mucous membrane	Rectum
T68012	Rectal gland, NOS	Rectum
T68040	Rectal muscularis propria	Rectum
T68050	Rectal solitary lymphoid follicle	Rectum
T68060	Perirectal tissue	Rectum
T68070	Rectovaginal septum	
T68200	Rectosigmoid junction	
T68920	Rectum and vagina, CS	
T68950	Rectum and sigmoid colon, CS	

ADENOMA

The following SNOMED M (Morphology) codes were used to identify adenomas and the (tubulo)villous subset thereof in KPNC and KPSC pathology data:

Code	Description	(Tubulo) Villous
M74000	Dysplasia, NOS	
M74001	Dysplasia, focal	
M74005	Dysplasia, atypical	
M74006	Dysplasia, mild	
M74007	Dysplasia, moderate	
M74008	Dysplasia, severe	

Code	Description	(Tubulo) Villous
M74009	Dysplasia, pre-cancerous	
M81236	Basaloid adenoma, metastatic	
M81400	Adenoma, NOS	
M814001	Adenoma, NOS, well differentiated	
M814002	Adenoma, NOS, moderately differentiated	
M814003	Adenoma, NOS, poorly differentiated	
M82100	Adenomatous polyp, NOS	
M821001	Adenomatous polyp, NOS, well differentiated	
M82102	Adenomatous polyp, NOS, in-situ	
M82103	Adenocarcinoma in adenomatous polyp	
M821031	Adenocarcinoma in adenomatous polyp, well differentiated	
M821032	Adenocarcinoma in adenomatous polyp, moderately differentiated	
M821033	Adenocarcinoma in adenomatous polyp, poorly differentiated	
M82106	Adenomatous polyp, NOS, metastatic	
M821062	Adenocarcinoma in adenomatous polyp, moderately differentiated, m.	
M82110	Tubular adenoma, NOS	
M821101	Tubular adenoma, NOS, well differentiated	
M821102	Tubular adenoma, NOS, moderately differentiated	
M821103	Tubular adenoma, NOS, poorly differentiated	
M82111	Tubular adenoma, NOS, uncertain whether benign or malignant	
M82112	Tubular adenoma, NOS, in-situ	
M82200	Adenomatous polyposis coli	
M82610	Villous adenoma, benign	Y
M82611	Villous adenoma, NOS	Y
M826111	Villous adenoma, well differentiated	Y
M826112	Villous adenoma, NOS, moderately differentiated	Y
M826113	Villous adenoma, NOS, poorly differentiated	Y
M82612	Adenocarcinoma in villous adenoma, benign, in-situ	Y
M82613	Adenocarcinoma in villous adenoma	Y
M826131	Adenocarcinoma in villous adenoma, well differentiated	Y
M826132	Adenocarcinoma in villous adenoma, moderately differentiated	Y
M826133	Adenocarcinoma in villous adenoma, poorly differentiated	Y
M826134	Adenocarcinoma in villous adenoma, undifferentiated	Y
M82616	Villous adenoma, benign, metastatic	Y
M826161	Adenocarcinoma in villous adenoma, well differentiated	Y
M82630	Tubulovillous adenoma	Y
M826301	Tubulovillous adenoma, well differentiated	Y
M826302	Tubulovillous adenoma, moderately differentiated	Y
M826303	Tubulovillous adenoma, poorly differentiated	Y
M82632	Tubulovillous adenoma, in-situ	Y
M82633	Tubulovillous adenoma, malignant	Y
M826331	Adenocarcinoma in tubulovillous adenoma, malignant, well differentiated	Y
M826332	Adenocarcinoma in tubulovillous adenoma, malignant, moderately differentiated	Y
M826333	Adenocarcinoma in tubulovillous adenoma, malignant, poorly differentiated	Y
M826334	Adenocarcinoma in tubulovillous adenoma, malignant, undifferentiated	Y
M82636	Adenocarcinoma in tubulovillous adenoma, metastatic	Y
M83300	Follicular adenoma	
M83330	Microfollicular adenoma	
M84400	Cystadenoma, NOS	
M844002	Cystadenoma, NOS, moderately differentiated	
M84401	Cystadenoma, NOS, uncertain whether benign or malignant	

Code	Description	(Tubulo) Villous
M84410	Serous cystadenoma, NOS	
M84800	Mucinous adenoma	
M84809	Mucinous adenocarcinoma uncertain primary tumor / metastasis	
M90100	Fibroadenoma, NOS	

HYPERPLASIA

The following SNOMED M (Morphology) codes were used to identify hyperplasia in KPNC and KPSC pathology data:

Code	Description
M72000	Hyperplasia, NOS
M72001	Hyperplasia, focal
M72003	Hyperplasia, diffuse
M72005	Hyperplasia, atypical
M72020	Hyperplasia, secondary
M72021	Hyperplasia, secondary, focal
M72025	Atypical hyperplasia, secondary
M72030	Hyperplasia, nodular, NOS
M72031	Hyperplasia, nodular, focal
M72032	Hyperplasia, nodular, multifocal
M72035	Hyperplasia, nodular, atypical
M72040	Hyperplasia, polypoid
M72041	Hyperplasia, polypoid, focal
M72045	Hyperplasia, polypoid, atypical
M72050	Hyperplasia, papillary
M72060	Hyperplasia, cystic, NOS
M72061	Hyperplasia, cystic, focal
M72065	Hyperplasia, cystic, atypical
M72100	Hyperplasia, lobular
M72101	Hyperplasia, lobular, focal
M72105	Hyperplasia, atypical lobular
M72170	Hyperplasia, intraductal
M72171	Hyperplasia, intraductal, focal
M72175	Hyperplasia, atypical intraductal
M72420	Hyperplasia, glandular
M72421	Hyperplasia, glandular, focal
M72425	Hyperplasia, glandular, atypical
M72480	Hyperplasia, microglandular
M72481	Focal hyperplasia, microglandular
M72485	Hyperplasia, microglandular, atypical
M72490	Hyperplasia, adenomatoid

COLORECTAL CANCER

A list of 398 SNOMED M (Morphology) codes was used to identify CRC detected in KPNC and KPSC pathology data; full code list is available upon request.

Appendix C. Charlson Comorbidity Index

The KP sites used the HCSRN VDW Charlson standard macro (as of 12/01/2020) to calculate Charlson comorbidity scores for IMS3. This macro uses the weights from the 1992 Deyo publication, although the algorithm has been modified to include ICD-10 codes and to allow codes from all in-person encounters (and, where applicable in the CalendarYear table for 2019–2020, selected virtual care encounters, namely scheduled telephone calls, synchronous chats, and video visits). In VDW terms, the in-person encounter types were AV, ED, IP, IS, and OE; the virtual care (VC) subtype codes were TS (telephone, scheduled), CH (synchronous chat), and VV (video visit). The encounter admit date was used as the diagnosis or procedure date for events associated with inpatient stays. Exact SAS code utilized for IMS3 data submission is available upon request.

UTSW used diagnosis/procedure codes and weights from the HCSRN VDW Charlson macro, and they used encounter admit date for events associated with inpatient stays; however, the site was unable to determine whether events occurred at in-person or telehealth encounters and employed an alternative method of implementing the ComorbHIV flag. To ascertain HIV comorbidity, a combination of HIV billing codes, lab (HIV RNA or CD4 lab test) results, and HIV clinic visits were used, wherein HIV diagnosis was confirmed once two of the preceding items appeared in the EMR. More information on this method can be found in the following publication: Barnes, Artia et al. Cervical cancer screening among HIV-infected women in an urban, United States safety-net healthcare system, AIDS: August 24, 2018 - Volume 32 - Issue 13 - p 1861-1870 doi: 10.1097/QAD.0000000000001881. Again, SAS code is available upon request.

Appendix D. KP Indication Algorithm

Version: 06/17/2021

Implementation

1. **Everywhere that "colonoscopy" is referenced, please also include lower endoscopy NOS.** I.e., 1) run the algorithm on lower endoscopy NOS procedures as well as colonoscopies, and 2) consider occurrence and results of lower endoscopy NOS whenever colonoscopy is mentioned in steps below (e.g., step 12).
2. Create flag for each rule for each colonoscopy of interest.
3. Colonoscopies needs to be assigned sequentially moving forward in time, since earlier colonoscopies' indication may affect the indication of later colonoscopies.

Rules

Rule	Condition	Timing	Indication (main algorithm)	Rationale
1	Colectomy and proctectomy • Colectomy • Proctectomy	codedate < cspydate (i.e., surgery date precedes colonoscopy date)	Diagnostic	To keep the screening group pure and average risk, we don't want people with histories of these surgeries to be included.
2	IBD history (specifically ulcerative colitis and Crohn's, but not other kinds of colitis) Includes sequelae of IBD	codedate < cspydate (i.e., condition date precedes colonoscopy date)	Diagnostic	IBD patients often come in for a "diagnostic" colonoscopy due to worsening symptoms (i.e., abdominal pain, rectal bleeding, etc.) or disease assessment after medication adjustment/initiation. However, teasing whether the indication is for surveillance versus diagnostic for IBD patients may not be that easy.
3	Symptoms: • Abdominal pain • Iron-deficiency anemia, and some unspecified anemias • GI bleeding or blood in stools • Diarrhea • Weight loss or underweight • Diverticulitis • Constipation • Abdominal mass • Change in bowel habits	(cspydate−180) ≤ codedate AND codedate < cspydate (i.e., symptom date is in the 6 mos. before colonoscopy date)	Diagnostic	Symptoms are defining features of diagnostic exams. While they may not have high PPV for cancer, keeping them as diagnostic indication keeps the screening indication pure.
4	Recent procedure: • Barium enema • Abdominal CT	(cspydate−180) ≤ codedate AND	Diagnostic	The rationale for not including a colonoscopy or sigmoidoscopy in the past 6 months is if there

Rule	Condition	Timing	Indication (main algorithm)	Rationale
	- CT colonography	codedate < cspydate (i.e., procedure date is in the 6 mos. before colonoscopy date)		was no polyp or adenoma diagnosis, we assume the colonoscopy / sigmoidoscopy was incomplete or had an inadequate bowel preparation. Therefore, the colonoscopy following that recent colonoscopy (without a polyp or adenoma diagnoses) would be for "screening" purposes. Rationale for 6 months: usually f/up is within this time for diagnostic evaluation (as opposed to early f/up at one year).
5	Recent diagnosis of CRC Any cancer located at sites C180–C189, C199, C209, C218, C260, regardless of histology.	(cspydate−180) ≤ codedate AND codedate < cspydate (i.e., CRC date is in the 6 mos. before colonoscopy date)	Diagnostic	Index colonoscopy is likely a work-up Any cancer type in the colorectum could trigger a diagnostic exam.
6	Recent colorectal adenoma, polyp, benign neoplasm, or neoplasm of uncertain behavior	(cspydate−365) ≤ codedate AND codedate < cspydate (i.e., polyp date is in the 1 year before colonoscopy date)	Diagnostic	Patients with incompletely resected polyps/adenomas or polyps that are too large to resect or uncertain if fully resected during the index colonoscopy are often asked to come back within a few months for another attempt.
7	Negative back-office gFOBT/FIT	(cspydate−180) ≤ codedate AND codedate < cspydate (i.e., negative back office gFOBT/FIT is in the 6 mos. before colonoscopy date)	Diagnostic	A person who had an in-clinic FOBT and is coming back within 6 mos. likely has symptoms.
8	Positive gFOBT/FIT	1) codedate < cspydate AND 2) cspydate[n-1] < codedate OR cspydate[n-1] is missing (i.e., positive gFOBT/FIT affects the indication of the next colonoscopy regardless of how long before the	Diagnostic	The first colonoscopy after a positive gFOBT/FIT is considered diagnostic, regardless of how long after the positive gFOBT/FIT it occurs. People don’t go back to true screening until they’ve had follow-up.

Rule	Condition	Timing	Indication (main algorithm)	Rationale
		colonoscopy it occurred. But it affects only the next colonoscopy's indication and not subsequent ones, unless those subsequent ones are within a short interval, which is handled by rule 12.)		
9	History of CRC	codedate < (cspydate−180) (i.e., CRC date is more than 6 mos. before colonoscopy date)	High-risk surveillance	This person is not average risk
10	High-risk condition FAP or HNPCC (personal or family history)	codedate < cspydate (i.e., condition date precedes colonoscopy date)	High-risk surveillance	This person is not average risk
11	History of colorectal adenoma, polyp, benign neoplasm, or neoplasm of uncertain behavior	codedate < (cspydate−365) (i.e., polyp date is more than 1 year before colonoscopy date)	Polyp/adenoma surveillance	This person is likely under a surveillance regimen but does not have a high-risk condition
12	Colonoscopy in 6 mos. prior	(cspydate−180) ≤ cspydate[n-1] AND cspydate[n-1] < cspydate	Same indication as prior colonoscopy	If a person had a recent colonoscopy, another one within a short period of time is probably because the first one wasn't complete. The second one is an extension of the first. Note that if a polyp/adenoma was diagnosed on the first colonoscopy, the second one has already been classified as diagnostic.
13	At least 365 days of prior enrollment or evidence of participation in the health system (allowing for 90-day gaps)		Screening	If we don't have adequate lookback, we are less certain of the screening indication.
14	Otherwise		Unknown	

Appendix E. Natural Language Processing

NLP at KPWA

The 2021mayC NLP algorithm described below was used to extract colorectal findings from colonoscopy procedure report and pathology report text for the KPWA data described previously in this user guide.

INTRODUCTION

In building datasets to support research into colorectal cancer risk, vital information is contained in predominantly free-text colonoscopy and pathology reports. At KPWA, the data present in these documents are not provided in a structured format, thereby requiring the use of manual methods to obtain this information. Rather than physically review the text via chart abstraction, we have opted to develop a Natural Language Processing (NLP) system to consistently and scalably extract a handful of variables from the clinical reports. This document describes the current progress and implementation details for the IMS3 data submission.

The following is only intended to provide a general understanding of the NLP algorithm and is not intended to be complete. Complete word lists and logic are available within the git repository, please see

Source Code for more details.

SHARED PREPROCESSING

All the algorithms described rely on a series of initial preprocessing to load the data into a more structured form from which the individual queries can be applied.

COLONOSCOPY

Document Sectioning

For certain variables, a specific section within the colonoscopy is required. The colonoscopy text is separated by capitalized words followed by a colon (e.g., “Impression:”) so that these sections can be used by the algorithm.

Polyp Model

One of the goals of the post-IMS1 2020aug round of algorithm improvements was to better identify large adenomas. This required a more robust approach for the identification of polyps and their sizes in the colonoscopy report.[3] Previously, all possible polyp locations were identified along with any neighboring sizes without any attempt at disambiguation. If a location was repeated in the text, it was treated as a separate polyp (e.g., “removed polyps from ascending and sigmoid: sigmoid polyp was 2mm” would result in two sigmoid polyps being recorded). Thus, the number of polyps was not only usually exaggerated (i.e., every location mentioned was treated independently), but no information was retained to directly link a polyp in the colonoscopy with a polyp in the pathology report. While this simple approach performed rather well, it was unable to discriminate between polyps (since locations rather than polyps were recorded) and therefore failed at various edge cases. For example, if two descending polyps were described in the colonoscopy procedure—one large and one small—and only the small one was noted by the pathologist to be adenomatous, it would be considered a large adenoma.

---

[3] These are occasionally referred to as “findings.” A “finding” consists of the polyp along with details (e.g., size, depth, number, etc.) relating to that polyp/those polyps.

With this context in mind, a finite state machine[4] was developed to better capture polyps in the colonoscopy report along with associated details (e.g., size, depth/location, and count). The basic process consists of following a series of steps, summarized as follows:

1. Look for mention of “polyp” in the colonoscopy report. Exclusionary language was applied if the context consisted of words like “diverticulosis”, “WNL” (within normal limits), or “not evaluated”, which could cause false positives. If any of these were triggered, the “polyp” mention was ignored.
2. Next, the polyp size is identified. Since size and depth (i.e., depth in the colon, e.g., “80 cm”) can appear in the same context as a polyp’s size and are only distinguishable by the value (e.g., an “80 cm” polyp is unlikely), the algorithm alternates between attempting to identify size and depth. If multiple sizes were identified (e.g., “Removed descending polyps: 3mm, 4mm, 10mm”), the number of polyps was incremented accordingly.
3. The location of the polyp is next considered. First, depth patterns which might be mistaken for sizes are looked at (e.g., “polyp at 5cm removed”). Second, the more common cases of the colonic section/location are identified. If multiple locations are enumerated, the number of polyps was incremented accordingly.
4. Finally, additional details are extracted to prevent too many polyps from being identified. This includes identifying both polyp boundaries (e.g., “removed”) and added descriptions (“Removed descending polyp. The polyp was …”) which previously resulted in an exaggerated number of polyps being reported.

Once the polyps have been identified, they are merged according to some general heuristics, including:

1. After a first polyp has been identified, subsequent polyps missing features (e.g., location) inherit locations from that predecessor or are merged.
2. Polyps with compatible features across different sections (e.g., “findings” and “impressions”) are merged into a single polyp model.
3. Polyps missing key information are removed. This is required since the finite state machine analyzes every mention of “polyp” in the text.

While some difficult cases still do leak through due to the varied language used in colonoscopy reports, these do not appear to negatively impact the large adenoma variable. Some additional work would be required if the goal was to report the number of polyps in the colonoscopy report.

PATHOLOGY

Jars

The first stage in processing the pathology report is to separate the individual jars. This is done by keying on a letter (or, rarely, a number) followed by a close parenthesis or period, and then associating the subsequent text with that letter (or number).[5]

The letter may sometimes be given in an “and”-separated list (e.g., “A & B)”) or even in a range (“A-C.”), and in these cases, there are two possible meanings. First, the same text is meant to be applied individually to each jar (e.g., “A,B,C) POLYP: NO ADENOMA”). This is treated as the default case and the same text is individual assigned to each jar. Second, the jars are all described in one section and are not meant to be multiply applied (e.g., “A,B,C) DESC, ASC, & SIGMOID POLYPS: HYPERPLASTIC”). This less typical meaning appears mostly in

---

[4] A finite state machine is a series of states, and transitions between those states. This allows better management of steps to identify the components of a polyp. See code for precise implementation: https://github.com/kpwhri/precise_nlp/blob/master/src/precise_nlp/extract/cspy/finding_builder.py#L125

[5] Earlier versions (pre-2021mayB) required each pathology report to start with the letter ‘A’, but this has been fixed, so any letter may appear first (and no letter is required to appear).

older data and is identified and handled as a single jar containing three polyps. (In this case, the relevant heuristic has to do with the presence of multiple locations within the same jar.)

Most of the variables are obtained from the first (“diagnosis”) section, though a polyp location and high-grade dysplasia are sometimes only noted in the following section.

The process for teasing apart the jars, however, is imperfect. Most of the issues arise from extra text that contains comments or context which do not belong to one of the jars. Sometimes, these bits of extra text contain references to a jar in a manner which makes them appear like a new jar is being discussed (e.g., “COMMENT: <some text> in A) and in B)…”). The parser assumes that “A)” and “B)” are introducing the “A” and “B” jars rather than just describing them.

The Polyp Model

Several of the variables are concerned with the type and location of polyps. Instead of independently calculating these for each variable (i.e., calculating once for the total count of adenomas and then a second time for the count of proximal tubular adenomas), an underlying representation of a polyp is developed and the text only read through a single time. As each jar in the pathology text is parsed (see Jars, above), a polyp is eventually identified (usually from the keywords in Polyp, or is implied from the existence of the jar). The parser then looks in the immediate context of the identified polyp, looking for information like number (e.g., “COLON POLYP **x 2**”), location (e.g., “POLYP, **ASCENDING**”), adenoma-ness, and histology.

The most complex of these descriptors is number. On the one hand, number can be well defined such as in “COLON POLYP x 2” and “DESCENDING POLYP” which have 2 and 1, respectively. On the other, a phrase such as “ASCENDING POLYPS: ADENOMA FRAGMENTS” is replete with ambiguity. “POLYPS” could represent, in theory, any number greater than or equal to 2. We will thus represent the number as ≥2.

This example is further complicated by the equally ambiguous “ADENOMA FRAGMENTS”. Are the fragments from a single polyp? Two polyps? More? Rather than make a decision at this point, the internal representation would retain as much precision as possible, marking the polyp as having ≥1 adenoma.

This internal representation (the “polyp model”) can then be queries for the number of polyps bearing certain features (e.g., number of tubular adenomas in the rectum, etc.).

Not all variables are collected at the polyp-level. For example, high grade dysplasia does not need to be connected to a particular polyp for the purposes of the study, so it is collected at the jar-level. When the results of a particular variable are desired, the controller (i.e., the main program) will look for the information from the internal representation of the pathology report, which will in turn ask each jar contained therein, and each jar (in the case of adenoma counts) will calculate the information from the various polyp models it contains.

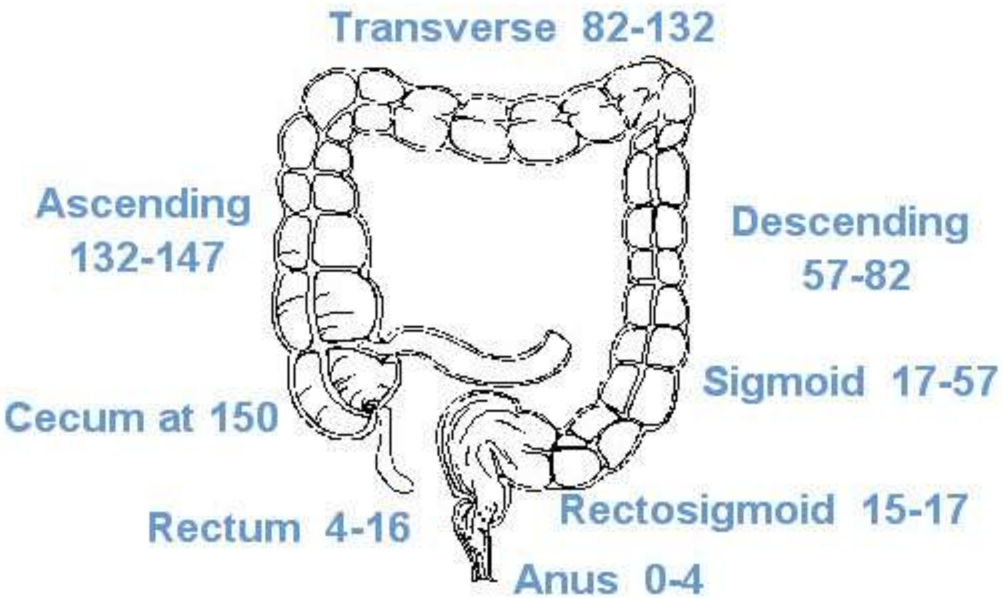
Location

Polyps are removed from a particular location within the colon. The colonoscopist records the source location on the jars, and the pathologist reference these in their report. Location provides our only opportunity to link individual polyps between the colonoscopy and pathology text.[6] Locations can either be described by the section of the colon or the distance from the anus (in cm). For the latter, the exact numbers will obviously vary by the individual, but we are relying on the NIH’s SEER Training Modules as our starting point.[7]

---

[6] This is used explicitly in the large adenoma algorithm.

[7] https://training.seer.cancer.gov/colorectal/anatomy/figure/figure1.html



The primary divergences from this model are:

1. Rectosigmoid is treated as both rectum and sigmoid.
2. Hepatic flexure added (80–84cm)[8]
3. Splenic flexure added (130–134cm)[9]
4. Cecum is treated as anything ≥ 147cm.

All lengths used are listed under Rectal Location, Distal Location, and Proximal Location.

Depending on the label and depth, multiple locations can be assigned. The simplest example is “rectosigmoid,” which describes a region incorporating both the rectal and sigmoid segments of the colon. The polyp (let’s say it’s adenomatous: “RECTOSIGMOID ADENOMA x1”) will then have two locations (rectal, sigmoid) which can make counting somewhat error prone. The total number of adenomas for this example would be 1. If you asked for the number of rectal adenomas, the answer would also be 1 (there is a single polyp/adenoma with the location “rectal”). Problematically, when you ask how many sigmoid adenomas were present, the result would also be 1 (there is a single polyp/adenoma with the location “sigmoid”). The same sorts of variability can occur in other areas as the depths overlap, particularly in the case of the hepatic and splenic flexures. It is important to look at the total number of polyps and not just add polyps at different locations together.

Location in the pathology text is assigned at the polyp level.

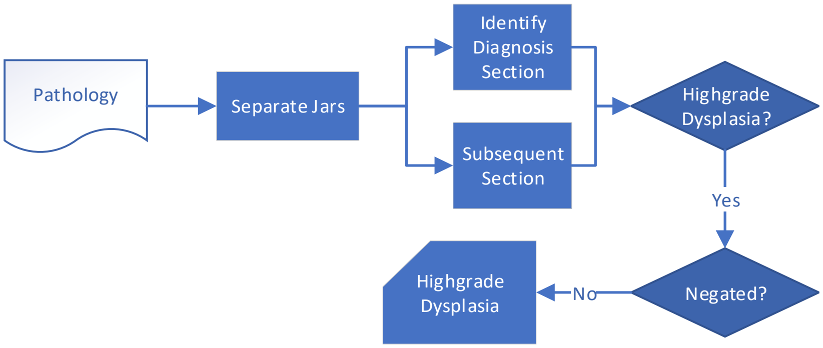
VARIABLES

HIGH-GRADE DYSPLASIA

To determine high-grade dysplasia, the diagnostic section is considered, looking for the keyword Dysplasia. The term must be preceded by a High-Grade keyword and not preceded by a Dysplasia Negation term. If the diagnostic section does not contain any high-grade dysplasia, the first non-diagnostic section is then considered with the same process. This is assessed separately for each jar.

---

[8] I made these numbers up based on including some small portion of the transverse and descending sections.
[9] I made these numbers up based on including some small portion of the transverse and ascending sections.



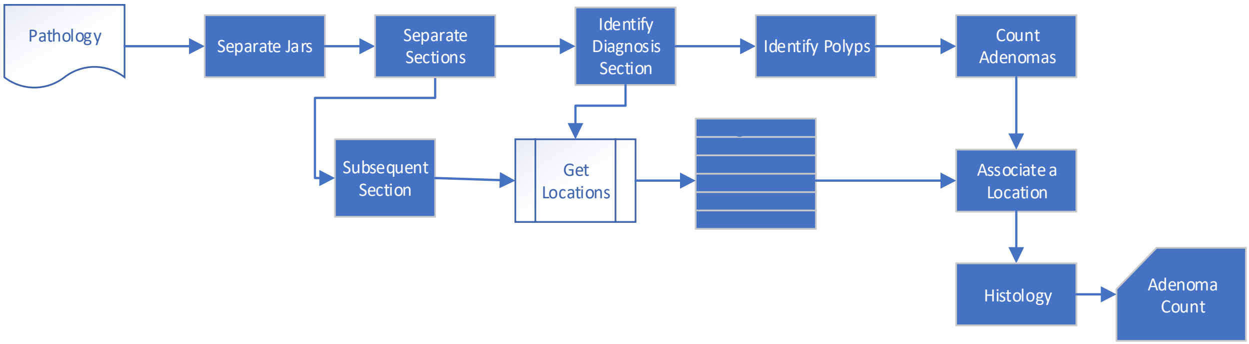
ADENOMA

The adenoma variable is obtained from the pathology report and attempts to capture the number of adenomas in each jar. It depends on the process for identifying individual polyps. While in the course of identifying the number of polyps within each jar (see description of the polyp model), if an un-negated Adenoma term is encountered, the adenoma is associated with a particular polyp. In the simple case of a solitary polyp in a jar, the adenoma count for the jar would be 1. The complexity increases when multiple polyps are placed within a single jar, often breaking apart into smaller “fragments.” The pathologist leaves the precise origin of the adenoma ambiguous (quite likely because it is also unclear to them), perhaps reporting “1 fragment of adenoma” or just “adenoma”. The approach to handle these is to be as conservative as possible regarding the number of adenomas, while still encoding the uncertainty. For example, if the pathologist were to report:

> POLYP X 2, DESCENDING: ADENOMA

The most conservative position would be associate the “adenoma” with only one of the polyps (i.e., 1 adenoma). To encode the uncertainty, however, we add a “greater than or equal” sign (i.e., ≥1 adenoma).

One final stage links the polyp with its histology: Tubular, Tubulovillous, or Villous.



LARGE ADENOMA

A large adenoma is defined by the presence of two characteristics. First, the largest dimension of the extracted polyp must be equal to or exceed 10mm. This data is usually only present in the colonoscopy report, though it is sometimes reported by the pathologist.[10] Second, the polyp must be determined to be adenomatous by the pathologist. The most challenging component of this is determining how to link a polyp in the colonoscopy report with a polyp from the jars described by the pathologist.

---

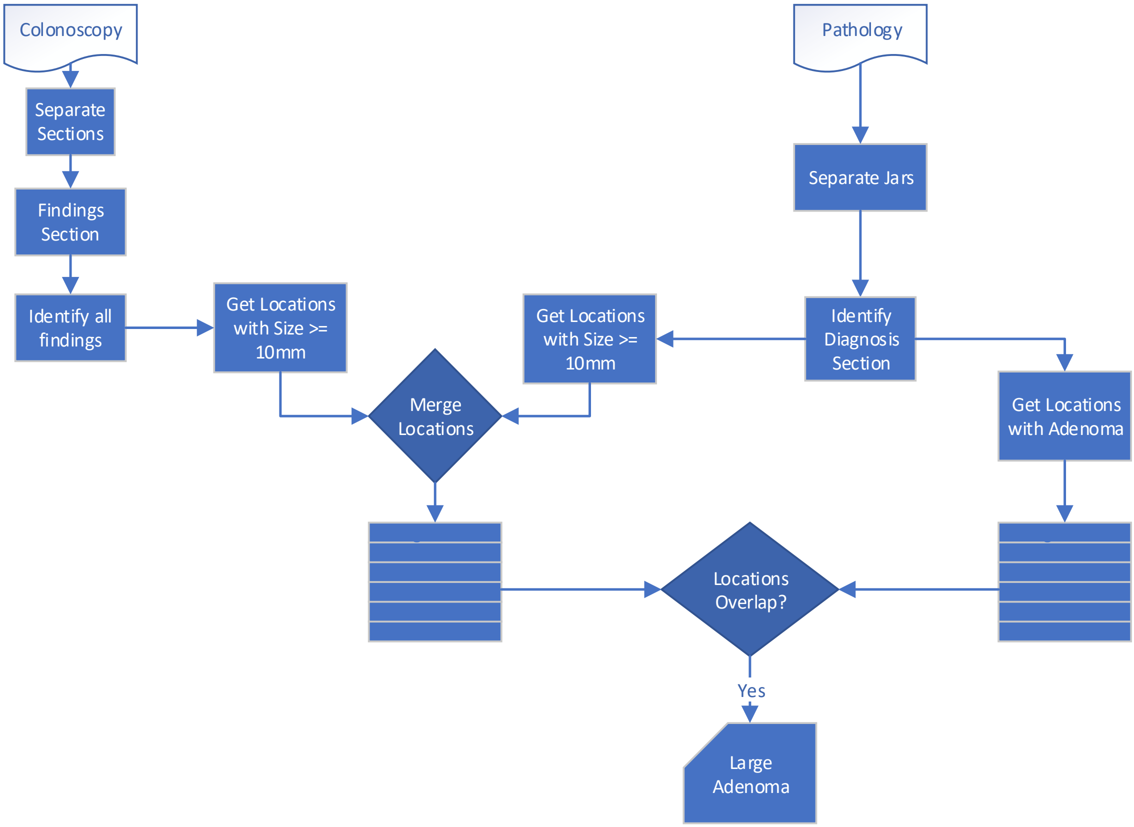
[10] The pathologist’s report is not generally relied on since the shape of the polyp can change during transit.

We made improvements to the approach taken for the 2019jun algorithm (July 2019 IMS1 submission). Not only did we improve the colonoscopy polyp model, but we also updated the algorithm logic to more finely handle multiple polyps from the same location. This results in having three outputs:

1. Large adenoma present
2. No large adenoma present
3. Large adenoma might be present (small and large polyp from the same location, but only one of them is adenomatous, and it is unclear which)

First, we identify all locations with an adenoma from the pathology to get a list of locations with adenomas along with their associated sizes (sizes are rarely reported reliably in the pathology report). Second, polyps described in the colonoscopy report, particularly in the “findings” and “impression” sections, are identified along with their associated attributes. Third, the polyps are matched using all available attributes to the adenomas and polyps in the pathology report. Also, if a large polyp or adenoma lacked an identifiable location, it was allowed to match any polyp that was not already unambiguously matched. If any polyp pair contained both a large polyp (≥10mm) and an adenoma, it is considered a large adenoma. However, if multiple polyps could match an adenoma and not all were large (i.e., some were <1cm in size), “possible” large adenoma was output.

This method permits the handling of more complex cases, such as if three descending polyps of sizes 3cm, 4cm, and 10cm are described in the colonoscopy. If one jar (“A”) in the pathology report is described as “descending polyps: adenomas” but another (“B”) is “descending polyp: hyperplastic”, then “possible large adenoma” is output. If, however, the second jar (“B”) instead says “descending polyp 3cm: hyperplastic”, the algorithm will work out that this implies that the large polyp is in jar “A”. Since the first jar (“A”) stipulates multiple adenomas, then there is a “large adenoma.”



INDICATION

The indication variable looks to capture the reason that the colonoscopy is being performed. It is obtained from the “indications” section of the colonoscopy report. Regular expressions are used to determine whether the indication is “diagnostic,” “surveillance,” or “screening.” If no regular expression matches, the indication is given as “unknown.” The complete regular expressions can be found in the code itself.

Diverticulitis was removed from definition of diagnostic indication for IMS2+ submissions due to our determination that it is insufficient for a diagnostic indication.

EXTENT OF COLONOSCOPY

The extent variable captures whether the colonoscopist was able to reach the cecum and thus “complete” the procedure. A regular expression is used to identify if the colon was reached. If that is unable to find anything, a second regular expression looks to see if the procedure’s extent was discussed. If the extent of the procedure was discussed, but there is no mention of the cecum being reached, the extent is marked as “incomplete.” Otherwise, it is considered “unknown.”

BOWEL PREP

The prep variable looks at how well prepared/clear the colon is, which impacts the quality of the procedure. Individual terms are identified describing the bowel prep as either adequate or inadequate.

SESSILE SERRATED POLYP/SESSILE SERRATED ADENOMA

The sessile serrated polyp/adenoma variable (SSP/SSA)[11] captures mentions of sessile serrated polyps and sessile serrated adenomas in the pathology text. Previously (2019junQ), “sessile” was used as a negation term for “adenoma.” Now, however, the algorithm looks for “sessile” and “serrated” in a position before “adenoma” or “polyp” and records the number of SSP/SSA, otherwise “sessile” is still used as a negation term. In addition, “SSP” and “SSA” are directly extracted. The number of occurrences of these is counted at the jar level (i.e., a maximum of 3 SSPs/SSAs if the pathology contains three jars).

CANCER

The cancer variable[12] identifies cancers identified by the pathologist. The algorithm identifies words like “carcinoma”, “adenocarcinoma”, “cytoadenocarcinoma”, “malignant neoplasm”, and “adenomatoid tumor”. The entire “cancer phrase” (i.e., any associated cancer-related modifiers, e.g., “NOS”, “fusiform”, “tubular”, “in situ”, etc.) is identified. If any of these cancer-related modifiers is “in situ”, the cancer is labeled as “in situ” and output as a separate variable.

Once the cancer phrase (including in situ cancers) has been identified, its context is inspected for a limited set of qualifiers to determine if the cancer is being directly attested, negated, or described ambiguously. If negated, the term is ignored. If the term appears in the context of ambiguous language (suggesting some uncertainty regarding the cancer), it will be reported in one of two categories that we have called “maybe.”

First, cancer terms may be qualified by SEER ambiguous terms. These are defined in the SEER Program Coding and Staging Manual,[13] which is used to determine which qualifying ambiguous terms, when applied to cancer, are

---

[11] Implementation details: https://github.com/kpwhri/precise_nlp/blob/master/src/precise_nlp/extract/path.py#L606; see also previous lines containing “add_ssp” and “add_ssa.”

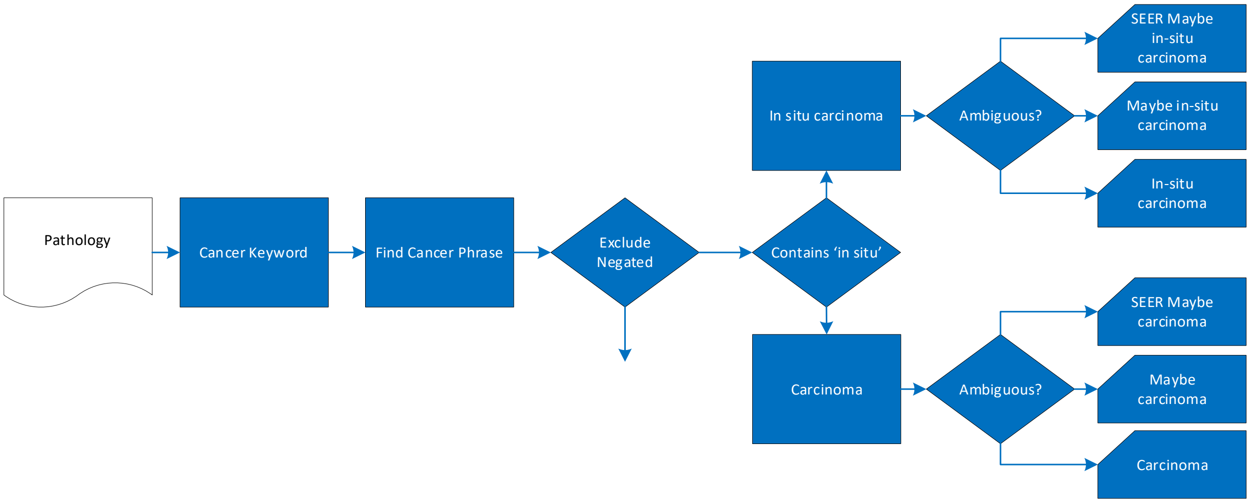
[12] Implementation details: https://github.com/kpwhri/precise_nlp/blob/master/src/precise_nlp/extract/path.py#L612

[13] Version 2021 was used; the ambiguous terms for reportability are discussed on page 10 of the document (pg 17 of the PDF) here: https://seer.cancer.gov/manuals/2021/SPCSM_2021_MainDoc.pdf

still reportable. The list was modified for our implementation.[14] When these terms appear, the cancer terminology is captured in a separate variable. (The full list of terms used is available below.)

Second, cancer terms may be qualified by other ambiguous terminology which SEER would consider non-reportable. These cancer terms are included separately. (See the full list below.)

The number of carcinomas and in situ carcinomas is counted at the jar level (i.e., a maximum of 3 cancers if the pathology contains three jars). Maybe carcinomas and maybe in situ carcinomas are similarly counted at the jar level.



LIMITATIONS AND ASSUMPTIONS

This document focuses on the NLP component of a larger system developed to identify colonoscopy text, locate the relevant pathology report, and extract the variables discussed herein. Here are some of the limitations and institutional assumptions which we have made (some of these have already been discussed above).

1. Each colonoscopy report is assumed to have a single corresponding pathology report.
2. If multiple pathology reports are available, the most complete one is given to the NLP.
3. Polyps removed during a colonoscopy can be biopsied and placed in a jar.
4. Information about these polyps (and related colonoscopy information) is included in a colonoscopy report.
5. Each jar may contain one or more polyps.
6. A single polyp is only contained in one jar.
7. All jars are sent to pathology.
8. The polyps can break apart and change shape during transit to pathology.
9. The pathologists record their findings in a pathology report.
10. Consistent language and structure are used in both the colonoscopy and pathology reports.

ADDITIONAL KPWA DETAILS

SOURCE CODE

Full implementation details, including source code, are available on the KPWHRI GitHub page: https://github.com/kpwhri/precise_nlp. The particular release described in this document can be obtained here:

---

[14] The NLP was limited by looking for a single term (so multi-term constructions were reduced). We added 'or' due to similar instructions given for chart abstraction. For complete list, see appendix (https://github.com/kpwhri/precise_nlp/blob/master/src/precise_nlp/extract/path/terms.py#L146).

https://github.com/kpwhri/precise_nlp/releases/tag/2021mayC (or by checking out the “2021mayC” tag from the repository within Git). N.B. The 2021mayB release was used for KPWA’s IMS2 data submission.

ALGORITHM PERFORMANCE

A complete report of the performance of the algorithm on Validation 02 is available in a document entitled “PRECISE Validation 02 Summary Report”[15].

HANDLING SCANNED RECORDS

Some of the colonoscopy/pathology comes from scans which must first be converted to machine-recognizable characters using optical character recognition (OCR).

OCR Process

The obtained scans consist of JPGs, TIFFs, and PDFs. The latter are mostly PDFs containing an array of TIFF images and are extracted, ordered, and output to that format. These images are then optimized for machine readability using the Python Pillow library[16], before being fed into Tesseract OCR (version 4.1.0)[17] which outputs text files.

Spelling Correction

The OCR process is far from perfect, producing a lot of noise (random glyphs trying to “identify” random marks and lines on scans—particularly, low quality ones). We initially relied on a probabilistic spell corrector, but the runtime of such a model was too high. For this, we used a targeted spelling correction approach.[18] This spelling correction module takes as input a set of target words to be corrected, a vocabulary, and the OCRed documents. The module will only attempt to spell correct the input set of targeted words.

The module presents as candidate spellings all those words an edit distance of 2 from target words. If these words match a set of criteria (e.g., they do not appear in the vocabulary and are thus an unknown word), they are converted to the target word.

Our only source of OCR data for IMS2+ was colonoscopy reports. While the spelling correction did make positive changes, none of these affected the algorithm itself.

This is the list of words identified for targeted spell correction:

- indications
- indication
- findings
- finding
- surveillance
- colonoscopy
- polyp
- polyps
- diverticulosis
- not evaluated
- rectum
- rectosigmoid
- rectosig
- sigmoid
- descending
- ascending
- hepatic
- transverse
- splenic
- cecum
- cecal
- terminal ileum
- terminal
- ileum
- ileocecal
- anorectum
- bowel prep
- bowel preparation
- colon prep
- colon preparation
- good
- excellent
- malignant neoplasm
- asa grade
- colonoscope
- propofol

---

[15] This can be found internally at: G:\CTRHS\PRECISE\DATA_COLLECTION\Validations\Validation02\PRECISE Validation 02 Report 20200915.docx

[16] https://github.com/python-pillow/Pillow/

[17] https://github.com/tesseract-ocr/tesseract/releases/tag/4.1.0

[18] The Python package precise_spelling_corrector was used (https://github.com/kpwhri/precise_spelling_corrector/releases/tag/202008_PRECISE_1).

- problem list
- hemoccult
- positive
- abnormal
- blood
- bleed
- hematochezia
- melena
- tarry
- anemia
- anemic
- constipated
- constipation
- urgency
- incontinent
- bowel
- irritable
- anorexia
- metastatic
- colitis
- suspect
- diverticulitis
- diverticulosis
- family history
- personal history
- ulcerative
- crohn
- inflammatory
- barrett
- procedure
- extent
- proximal
- location
- identified
- reached
- visualization
- prepared
- diarrhea
- quality of prep
- quality of preparation
- hepatic flexure
- descending
- ascending
- appendiceal
- orifice
- transverse
- appendectomy

KEYWORDS

Polyp

- Segment(s)
- Piece(s)
- Fragment(s)
- Adenoma(s)
- Polyp(s)
- Biopsy/Biopsies

Dysplasia

- Dysplasia
- Dysplastic

Dysplasia Negation

- No evidence
- No
- Without
- Low

High-Grade

- High grade
- Grade
- Severe

Adenoma

- Adenoma
- Adenomatoid
- Adenomatous
- Adenomat
- Adenom

Adenomas

- Adenomas

Adenoma Negation

- No
- History
- Hx
- Sessile
- Without

Colon

- Colon
- Rectum
- Rectal
- Cecal
- Cecum
- Colonic

Fragment

- Segment
- Fragment
- Piece

Fragments

- Segments
- Fragments
- Pieces

Number

- Numeric or character representation of numbers 1 through 9.

Tubular

- Tubular

Tubulovillous

- Tubulovillous
- Tubulovil
- Villotubular

Villous

- Villous
- Villiform

Histology Negation

- No
- Or
- Evidence
- Residual

Rectal Location

- Rectum (4-17cm)

Distal Location

- Descending (57-82cm)
- Sigmoid (15-57cm)
- Distal
- Splenic (flexure) (130-134cm)
- Left
- Rectum (4-17cm) – included because it improved performance against the gold standard

Proximal Location

- Proximal
- Ascending (132-147cm)
- Transverse (82-132cm)
- Cecum (147+cm)
- Hepatic (flexure) (80-84cm)
- Right

Adequate Bowel Prep

- Excellent
- Well
- Good
- Moderate
- Adequate
- Optimal

Inadequate Bowel Prep

- Fair
- Poor
- Inadequate
- Suboptimal

Diagnostic Indication

- Positive hemoccult
- “Abnormal”
- Blood-related (e.g., anemia)
- Bowel Changes/Diarrhea/Constipation
- Diverticulitis
- IBS
- “Suspect”
- “Mass”/”Metastatic”

Surveillance Indication

- IBD/UC
- Personal history
- Genetic
- “Follow-up”

Screening Indication

- Screening
- Family history

SSP/SSA

- Sessile serrated polyp(s)
- Sessile serrated adenoma(s)
- SSP(s)
- SSA(s)

Cancer

- Carcinoma(s)
- Adenocarcinoma(s)
- Cytoadenocarcinoma(s)
- Malignant neoplasm(s)
- Adenomatoid tumor(s)

SEER Maybe Cancer

Terms not strictly SEER are marked with an asterisk (*).

- Suspicious
- Apparent
- Apparently
- Appears
- Consistent
- Compatible
- Comparable
- Favor
- Favors
- Or*
- Appearing
- Likely*
- Presumed
- Presumptive
- Presumably
- Presumedly
- Favored*
- Suspecting
- Probable
- Suspect
- Suspected
- Typical
- Typically

Non-SEER Maybe Cancer

- Possible
- Possibly
- Suggestive
- Versus
- Questionable
- Vs
- V
- Surmise
- Suggesting
- Possibility

Cancer Negation

- No
- Not
- Unlikely
- Improbable
- Improbably
- Doubt
- Doubtful
- Unclear
- Preclude(s)
- Cannot

NLP at KPNC/KPSC

The NLP algorithms described below were used to extract colonoscopy characteristics and colorectal findings from pathology/colonoscopy text for the KPNC/KPSC data described previously in this user guide.

INTRODUCTION

NLP algorithms were developed using SAS text search functions and SAS PERL regular expressions; they were validated using manual chart review and Linguamatics I2E NLP software. Complete SAS code is available by request.

SHARED PREPROCESSING

All the algorithms described rely on a series of preprocessing steps to identify the most recent of a procedure or pathology report, extract text data, concatenate multiple text lines into a single string, and load the reports into a more structured form from which the individual queries can be applied.

COLONOSCOPY

Colonoscopy report text is obtained from Clarity provider notes. The colonoscopy reports are preselected by restricting by date and limiting to notes categorized as “Procedure” type with text containing the string “Colonoscopy Report”. The reports were further preprocessed to allow subsequent NLP analysis (see preprocessing description above).

PATHOLOGY

Pathology report text is obtained from the research data warehouse. Pathology reports are preselected by limiting to colorectal specimens (vis SNOMED T [Topography] codes) where collection date is on or within ±7 days of the procedure date. The reports were further preprocessed to allow subsequent NLP analysis (see preprocessing description above).

VARIABLES

EXTENT OF COLONOSCOPY

The extent of colonoscopy variable captures whether the procedure reached the cecum, which qualifies it as a “complete” procedure. A regular expression is used to identify individual terms describing the exam extent as complete. If the extent of the procedure is described, but there is no mention of the cecum being reached, the extent is marked as “incomplete.” Otherwise, it is considered “unknown.” The algorithm was validated on 3,000 charts (PMID: 32737597[19]).

BOWEL PREPARATION

The bowel preparation variable captures how well prepared or clear the colon is for visualizing colorectal abnormalities. This is a measure of the quality of the procedure. Individual terms are identified which lead to the bowel preparation being classified as either adequate or inadequate. The algorithm’s performance for this variable was validated on 3,000 charts (PMID: 32737597[20]).

LARGE POLYP

A large polyp is defined by the largest dimension of the extracted polyp, which must be equal to or greater than 10mm.

---

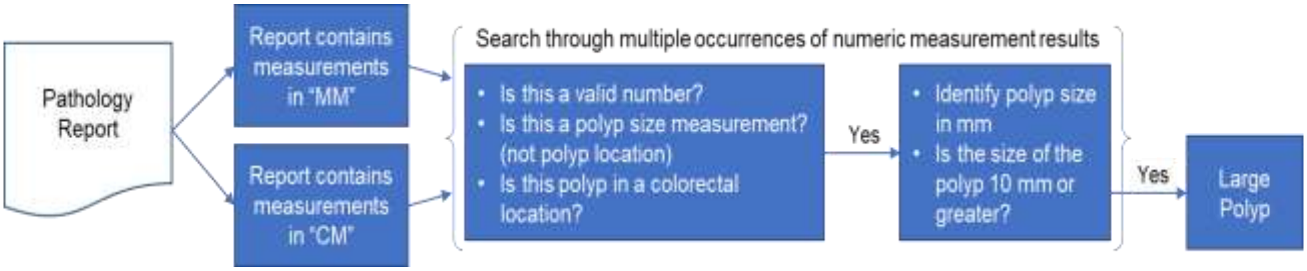
[19] "A Transparent and Adaptable Method to Extract Colonoscopy and Pathology Data Using Natural Language Processing". Helene B Fevrier, Liyan Liu, Lisa J Herrinton, Dan Li. J Med Syst. 2020 Jul 31;44(9):151. doi: 10.1007/s10916-020-01604-8.

[20] Ibid.

KPSC did not assess the presence of large polyps for IMS2+ due to questionable rates.

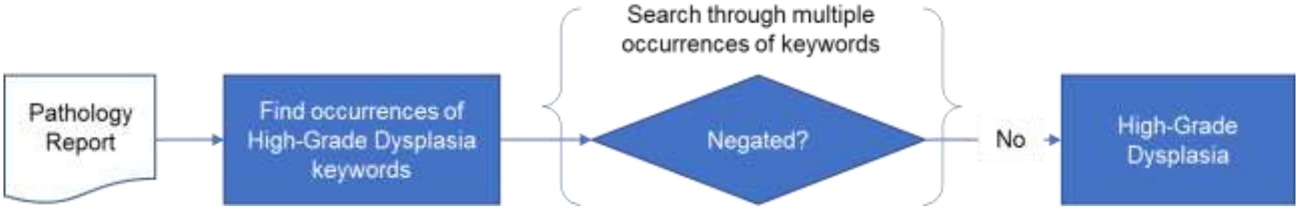
At KPNC, polyp size data is often present in colonoscopy and pathology reports. Pathology reports were consistently available through all years. For IMS2+, KPNC used pathology reports as the only data source because validation of the algorithm applied to colonoscopy reports has not yet been completed.

The presence of a ≥10 mm polyp detected during colonoscopy was identified at KPNC only by an NLP algorithm using SAS text search functions applied to the text of pathology reports; this was validated for polyp size findings on over 90,000 colonoscopies using findings from implementation of Linguamatics I2E NLP software as the gold standard (PMID: 31589872[21]). The algorithm identifies within pathology reports the largest polyp dimension by searching through multiple occurrences of numeric polyp size measurements in mm and cm, with subsequent analysis of characteristics of each measurement and processing negations within the vicinity of each occurrence of a numeric term.


Pathology Report
Report contains measurements in "MM"
Report contains measurements in "CM"
Search through multiple occurrences of numeric measurement results
• Is this a valid number?
• Is this a polyp size measurement? (not polyp location)
• Is this polyp in a colorectal location?
Yes
• Identify polyp size in mm
• Is the size of the polyp 10 mm or greater?
Yes
Large Polyp

HIGH-GRADE DYSPLASIA (HGD)

The presence of a polyp with high-grade dysplasia was identified by NLP using SAS text searches applied to the text of pathology reports; this was validated on 300 charts (PMID: 31589872[22]). The keywords for high-grade dysplasia were used to identify term occurrence and position in the pathology report text, with subsequent processing of negations within the vicinity of each occurrence.


Search through multiple occurrences of keywords
Pathology Report
Find occurrences of High-Grade Dysplasia keywords
Negated?
No
High-Grade Dysplasia

SESSILE SERRATED POLYP/SESSILE SERRATED ADENOMA

The presence of a sessile serrated polyp (SSP) or sessile serrated adenoma (SSA) was identified by SAS text searches applied to the text of pathology reports; this was validated on 300 charts (PMID: 31589872[23]). The keywords for SSP/SSA were used to identify term occurrence and position in the pathology report text, with the subsequent processing of negations within the vicinity of each occurrence.

---

[21] "Long-term Risk of Colorectal Cancer and Related Death After Adenoma Removal in a Large, Community-based Population". J K Lee, C D Jensen, T R Levin, C A Doubeni, A G Zauber, J Chubak, A S Kamineni, J E Schottinger, N R Ghai, N Udaltsova, W K Zhao, B H Fireman, C P Quesenberry, E J Orav, C S Skinner, E A Halm, D A Corley. Gastroenterology. 2020 Mar;158(4):884-894.e5. doi: 10.1053/j.gastro.2019.09.039. Epub 2019 Oct 4.

[22] Ibid.

[23] Ibid.

TRADITIONAL SERRATED ADENOMA

The presence of a traditional serrated adenoma (TSA) was SAS text searches applied to the text of pathology reports. The keywords for TSA were used to identify term occurrence and position in the pathology report text, with the subsequent processing of negations within the vicinity of each occurrence.

LIMITATIONS AND ASSUMPTIONS

This document focuses on the NLP component of a larger system developed to identify colonoscopy report text, locate the relevant pathology report, and extract the variables discussed herein. Here are some of the limitations and institutional assumptions we have made:

1. Each colonoscopy procedure is assumed to have a single corresponding colonoscopy report
2. The colonoscopists record their findings in a colonoscopy report.
3. Polyp removed during a colonoscopy can be biopsied and placed in a jar sent to pathology.
4. The pathologists record their findings in a pathology report.
5. Each procedure with biopsy is assumed to have a single corresponding pathology report
6. If multiple pathology reports are available, the most complete one is given to the NLP.
7. Consistent language in used in all colonoscopy and pathology reports.

ADDITIONAL KPNC/KPSC DETAILS

SOURCE CODE

Complete SAS code is available upon request.

KEYWORDS

Complete Colonoscopy

- CECUM
- ILEUM
- ILEO-COLIC ANASTOMOSIS
- ANASTOMOSIS
- ILEO-COLIC

Adequate Bowel Prep

- Excellent
- Well
- Good
- Moderate
- Adequate
- Optimal

Inadequate Bowel Prep

- Fair
- Poor
- Inadequate
- Suboptimal

Dysplasia

- Dysplasia
- Dysplastic

High-Grade

- High grade
- High-grade
- Severe

Dysplasia Negation

- No {…} evidence {…}
- Not identified
- Without
- Negative for
- No {definite, definitive, obvious, …}
- Rather than true
- Does not show
- None of the {polyps, biopsies} {demonstrate, show}
- No foci of

SSP/SSA

- Sessile & Serrated & {Adenoma or Polyp}

SSP/SSA Negation

- No
- R/o
- Rule out
- ?Hyperplastic vs sessile

TSA

- Traditional & Serrated & Adenoma

TSA Negation

- No evidence of
- No serrated
- History of
- R/o
- Rule out

Appendix F. Identifying Screening-Eligible Participants

Participants who undergo a screening colonoscopy are, by definition, free of certain signs and symptoms of colorectal cancer at the time of that procedure. For example, the KP modified indication algorithm will classify a colonoscopy as screening only in the absence of past colectomy and proctectomy, IBD, symptoms (e.g., abdominal pain, rectal bleeding), as well as other factors. See Appendix D. KP Indication Algorithm for more information.

In analyses that compare screened to unscreened persons, it will often be important to ensure that unscreened persons are also free of signs and symptoms of colorectal cancer at a given point in time. Failing to do so could lead to bias.

This appendix provides guidance on how to define sign- and symptom-free (i.e., screening-eligible) status. This status can be defined on any “index” date of interest by excluding people who meet any of the conditions in the table below.

Important notes:

- Many conditions (rules) must be assessed using information from two or more DRP tables. Since the KP modified indication algorithm relies on an index date as anchor for the lookback period, data from events both during the cohort period (e.g., Procedure, Diagnosis, CancerRegistry, FITgFOBTResults, and Enrollment tables) and prior to cohort entry (PriorToCohortEntry table) must be evaluated.
- DSR variables must be used for UTSW, which did not provide exact dates in Date variables. I.e., all references to Date variables in the table below would be replaced with the corresponding DSR variables if working with UTSW data.
- Index date must necessarily occur during cohort eligibility to enable complete data availability for subsequent follow-up/events.

If a participant meets any of the following conditions as of a given date during cohort eligibility (temporary variable IndexDate), they should **NOT** be considered screening-eligible at that point in time, as they are **NOT** free of signs or symptoms of CRC.

Condition [Rule[24]]	**DRP Table(s)**	**Variables/SAS Logic**
Any history of colectomy or proctectomy [1]	PriorToCohortEntry	if GISurgPrior = '01'
	OR	
	Procedure	if ProcType in ('05', '06', '07', '08', '09') and ProcDate < IndexDate
Any history of IBD [2]	PriorToCohortEntry	if IndIBDLastDate is not null
	OR	
	Diagnosis	if DiagIBD = 'Y' and DiagDate < IndexDate
Recent symptoms [3]	PriorToCohortEntry	if IndSxLastDate is not null and 1 <= (IndexDate - IndSxLastDate) <= 180
	OR	
	Diagnosis	if DiagSx = 'Y' and 1 <= (IndexDate - DiagDate) <= 180
Recent procedure (barium enema, abdominal CT, or CT colonography) [4]	PriorToCohortEntry	if IndPxLastDate is not null and 1 <= (IndexDate - IndPxLastDate) <= 180
	OR	
	Procedure	if ProcType in ('03', '04', '11') and 1 <= (IndexDate - ProcDate) <= 180

---

[24] From Rules table in Appendix D. KP Indication Algorithm.

Condition [Rule[24]]	**DRP Table(s)**	**Variables/SAS Logic**
	OR	
	Diagnosis	if DiagAbnImg = 'Y' and 1 <= (IndexDate - DiagDate) <= 180
Any history of CRC [5, 9]	PriorToCohortEntry	if CRCPrior = '01' or IndCRCDxLastDate is not null or IndCRCHxLastDate is not null
	OR	
	CancerRegistry	if DxDate < IndexDate
	OR	
	Diagnosis	if (DiagCRC = 'Y' and (IndexDate - DiagDate) > 180) or (DiagCode in ('C78.5', '197.5') and 1 <= (IndexDate - DiagDate) <= 180)
Any history of colorectal polyp [6, 11]	PriorToCohortEntry	if CspyLastResult = '01' or SigLastResult = '01' or LEndoLastResult = '01' or IndPolypDxLastDate is not null or IndPolypHxLastDate is not null
	OR	
	Procedure	if Adenoma = '01' and ProcDate < IndexDate
	OR	
	Diagnosis	if DiagPolyp = 'Y' and DiagDate < IndexDate
Recent negative back-office FIT/gFOBT [7]	PriorToCohortEntry	if IndBackOfcLastDate is not null and 1 <= (IndexDate - IndBackOfcLastDate) <= 180
	OR	
	FITgFOBTResults	if FITResult = '00' and TestSetting = '01' and 1 <= (IndexDate - FITResultDate) <= 180
Positive FIT/gFOBT without an intervening colonoscopy or lower endoscopy NOS [8]	FITgFOBTResults, PriorToCohortEntry, and Procedure used in combination	1. Obtain date of last positive FIT/gFOBT result prior to index (and store in temporary variable LastPosFOBT) by taking the later of: a. Date of last FIT/gFOBT+ during cohort eligibility (FITgFOBTResults: FITResult = '01' and FITResultDate < IndexDate) b. Date of last FIT/gFOBT+ prior to cohort entry (PriorToCohortEntry: FITLastResult = '01') 2. If no prior positive FIT/gFOBT was found, stop here; the condition at left was not met. 3. Otherwise, look for one or more colonoscopies or lower endoscopies NOS on or after LastPosFOBT and before IndexDate: a. Colonoscopy or lower endoscopy NOS during cohort eligibility (Procedure: ProcType in ('01', '10') and LastPosFOBT <= ProcDate < IndexDate) b. Colonoscopy or lower endoscopy NOS prior to cohort entry (PriorToCohortEntry: LastPosFOBT <= ColonoscopyDate or LastPosFOBT <= LEndoDate) 4. If a colonoscopy or lower endoscopy NOS is found on or after LastPosFOBT and prior to IndexDate, the condition was not met, i.e., participant already had a procedure to follow up on positive FIT/gFOBT result. 5. If a colonoscopy or lower endoscopy NOS was not found on or after LastPosFOBT and prior to IndexDate, the condition

Condition [Rule[24]]	**DRP Table(s)**	**Variables/SAS Logic**
		was met, and participant should not be considered free of signs/symptoms as of IndexDate.
Any history of high-risk condition (FAP or HNPCC) [10]	PriorToCohortEntry	if IndHeredLastDate is not null
	OR	
	Diagnosis	if DiagHered = 'Y' and DiagDate < IndexDate
Recent non-screening colonoscopy or lower endoscopy NOS [12]	PriorToCohortEntry	1. if ColonoscopyPrior = '01' and 1 <= (IndexDate - ColonoscopyDate) <= 180 → use Ind*Date[25] variables to calculate KP modified indication (temporary variable IndCspyPrior) for this prior colonoscopy per logic in Appendix D. 2. if IndCspyPrior <> '04' → condition at left has been met
		OR
		1. if LEndoPrior = '01' and 1 <= (IndexDate - LEndoDate) <= 180 → use Ind*Date variables to calculate KP modified indication (temporary variable IndLEndoPrior) for this prior lower endoscopy NOS per logic in Appendix D. 2. if IndLEndoPrior <> '04' → condition at left has been met
	OR	
	Procedure	if ProcType in ('01','10') and ProcIndicKPMod <> '04' and 1 <= (IndexDate - ProcDate) <= 180
Less than 365 days of prior enrollment (or evidence of prior participation in the health system), allowing for 90-day gaps [13]	Enrollment	if not (CohortEntryDate <= IndexDate <= CutoffDate and (IndexDate - CohortEntryDate) >= 365)

---

[25] Ind*Date refers to the suite of variables in PriorToCohortEntry Table with names that begin in "Ind" and end in "Date."

Appendix G. Identifying Screening Tests

Background

- Colonoscopies, sigmoidoscopies, and lower endoscopy NOS (described collectively here as “lower endoscopies”) can be done for indications other than screening. Stool tests (“FIT/gFOBTs”) are usually performed for screening purposes but can also be done for other reasons.
- The Procedure table contains two indication variables, ProcIndicKPMod and ProcIndic, which are populated only for colonoscopies (ProcType = 01) and lower endoscopies NOS (ProcType = 10); PRECISE recommends that analysts use ProcIndic for analyses, as ProcIndicKPMod does not perform as well as ProcIndic at KPWA and UTSW. Indication for sigmoidoscopy is not available.
- Indication for FIT/gFOBT can be inferred by the setting in which the test occurred, although this approach may result in some misclassification.

Below is guidance on how to identify screening tests for various analytic purposes.

Analytic Approaches for Various Study Questions

1. **To determine who is <u>not</u> in need of screening at a point in time or when someone is next due for screening**
    - These analyses do not depend on knowing the indication of a FIT/gFOBT or lower endoscopy.
    - Use the Procedure and PriorToCohortEntry files to identify colonoscopies, sigmoidoscopies, and lower endoscopies NOS. Ignore indication information from ProcIndicKPMod and ProcIndic variables.
    - Use the FITgFOBTResults file to identify FIT/gFOBT results.
        - KPWA Note: Stool tests occurring in our external delivery system are not captured. We assume this number is small.
    - Results of FIT/gFOBTs and lower endoscopies will affect when next screening/surveillance test should be done. Other factors may also affect when screening is needed but are not discussed in this document.
2. **To condition on a screening FIT/gFOBT or lower endoscopy in order to identify a pure screening group (e.g., to look at what happens after a screening test)**
    - Do not use data prior to cohort entry to identify the index screening FIT/gFOBT or lower endoscopy. Tests prior to cohort entry are, by definition, collected only on people who eventually entered the cohort.
    - Use the Procedure file to identify colonoscopy and lower endoscopy NOS for which ProcIndic = 04 (Screening).
        - Note that indication is not available for sigmoidoscopy. One option is to conduct a sensitivity analysis assuming all sigmoidoscopies were truly screening and then again assuming all sigmoidoscopies were truly *not* screening.
    - Use the FITgFOBTResults file to identify records where TestSetting ≠ 01 (Back office) or TestSetting ≠ 02 (Inpatient).
3. **To calculate screening rates (where indication is important)**
    - Do not calculate rates prior to cohort entry for the reason provided above.
    - Consider how to handle people with a non-screening FIT/gFOBT or lower endoscopy before a screening FIT/gFOBT or lower endoscopy (e.g., censor).
    - For colonoscopies and lower endoscopies NOS: Use Procedure file to identify records where 1) ProcType = 01 or 10, and 2) ProcIndic = 04 or (ProcIndic = 99 and ProcIndicKPMod = 04)
        - For KPWA and UTSW: Conduct a sensitivity analysis assuming all colonoscopies and lower endoscopies NOS with unknown ProcIndic were truly screening and then again assuming all colonoscopies and lower endoscopies NOS with unknown ProcIndic were truly *not* screening.
        - For KPNC and KPSC: Sensitivity analyses are not generally necessary, because both ProcIndic and ProcIndicKPMod are algorithm-based.

- For sigmoidoscopies: Conduct a sensitivity analysis assuming all sigmoidoscopies were truly screening and then again assuming all sigmoidoscopies were truly *not* screening.
- For FIT/gFOBTs: Use FITgFOBTResults file to identify records where TestSetting ≠ 01 (Back office) or TestSetting ≠ 02 (Inpatient).
    - KPWA Note: Stool tests occurring in our external delivery system are not captured. We assume this number is small.