Colorectal Cancer Risk Prediction Model

Cohort Inclusion and Exclusion Criteria

Document Version: 1.0 **Date:** January 2026 **Purpose:** Formal documentation of cohort construction methodology for CRC risk prediction model

---

Executive Summary

This document describes the inclusion and exclusion criteria used to construct a training cohort for a colorectal cancer (CRC) risk prediction model targeting unscreened populations. The cohort uses a patient-month observation structure to maximize training samples from rare events (0.41% positive rate) while implementing rigorous label quality controls.

Final Cohort Size: ~831,000 patient-month observations from ~233,000 unique patients

---

1. Study Period and Temporal Parameters

Parameter	Value	Justification
Study Start Date	January 1, 2023	Limits prevalent case proportion; ensures contemporary clinical practice patterns
Study End Date	September 30, 2024	Data collection cutoff minus follow-up requirement
Prediction Window	6 months	Clinical standard for CRC risk prediction; balances actionability with lead time
Minimum Follow-up	12 months	Required to confirm negative labels with adequate observability
Data Collection Date	September 30, 2025	Date through which outcome data is available

Temporal Structure: Each patient contributes one observation per calendar month they meet inclusion criteria during the study period. A deterministic hash-based algorithm assigns a specific day within each month for reproducibility.

---

2. Inclusion Criteria

2.1 Age Requirement

Criterion: Age 45-100 years at observation date

Justification: - Age 45 is the current USPSTF recommendation for CRC screening initiation (updated 2021) - Upper bound of 100 years excludes implausible data quality errors - Age calculated as: `FLOOR(DATEDIFF(observation_date, birth_date) / 365.25)`

2.2 Healthcare System Engagement

Criterion: Patient had at least one completed encounter at the integrated health system during the study period

Definition of Encounter: - **Outpatient:** Appointment status = Completed (2) or Arrived (6), at health system locations (RPT_GRP_SIX IN '116001','116002') - **Inpatient:** Hospital admission with non-zero charges, not preadmit, not canceled, at health system locations

Justification: Ensures patient is actively engaged with the health system and would be reachable for screening outreach intervention.

2.3 Observability Requirement

Criterion: Minimum 24 months of prior system contact before observation date

Definition: Months between patient's first recorded encounter (outpatient or inpatient, anywhere in EHR) and the observation date

Justification: - Ensures sufficient historical data for feature engineering (lab trends, visit patterns, etc.) - Establishes patient as "established" rather than new to system - **Important Limitation:** This does NOT confirm absence of undiagnosed cancer; a patient could have 36 months of diabetes visits while harboring an undiagnosed colon tumor

2.4 Label Quality Requirement

Criterion: Observation must meet one of the following label quality tiers:

Tier	Name	Criteria	Confidence	Coverage
Positive	Positive Case	CRC diagnosis (C18/C19/C20) within 6 months of observation	Definitive	~0.4%
Tier 1	High Confidence Negative	Return visit over 6 months after observation, no CRC diagnosis	High	~47%
Tier 2	Medium Confidence Negative	Return visit 4-6 months after observation + has PCP	Medium	~23%
Tier 3	Assumed Negative	No return visit but has active PCP relationship	Lower	~30%

Justification: This tiered approach maximizes training data while maintaining clinical validity. Observations without adequate follow-up or PCP relationship are excluded due to inability to confirm negative status.

2.5 Data Quality Flag

Criterion: `data_quality_flag = 1`

Definition: Excludes observations with impossible combinations: - Age > 100 years - Age < 0 years - Observability months exceeding patient's lifetime in months - First seen date after observation date

---

3. Exclusion Criteria

3.1 Prior Colorectal Cancer Diagnosis

Criterion: Exclude all observations from patients with any CRC diagnosis code at any time BEFORE the observation date

ICD-10 Codes: - C18.x (Malignant neoplasm of colon) - C19 (Malignant neoplasm of rectosigmoid junction) - C20 (Malignant neoplasm of rectum)

Note: C21 (anus) is NOT included in exclusion or outcome definition

Justification: Patients with prior CRC diagnosis are not candidates for primary screening outreach. Excluding prior diagnosis at any historical date (not time-windowed) is conservative and appropriate.

3.2 History of Colectomy

Criterion: Exclude all observations from patients with colectomy history at any time BEFORE the observation date

ICD-10 Codes: - Z90.49 (Acquired absence of other specified parts of digestive tract) - K91.850 (Colectomy, history)

Justification: Patients without a colon cannot develop colon cancer and are not appropriate screening candidates.

3.3 Hospice Care

Criterion: Exclude all observations from patients with hospice care documented at any time BEFORE the observation date

ICD-10 Codes: - Z51.5x (Encounter for palliative care)

Justification: Patients receiving hospice/palliative care are not appropriate targets for cancer screening outreach due to prognosis and goals of care.

3.4 Currently Screened (VBC Table)

Criterion: Exclude all observations from patients where `COLON_SCREEN_MET_FLAG = 'Y'` in the VBC colon cancer screening table

CRITICAL LIMITATION: > The VBC screening table (`vbc_colon_cancer_screen`) lacks temporal fields. This creates a systematic limitation where exclusion is based on CURRENT screening status at the data snapshot date, NOT the patient's screening status at each observation date. > > **Consequences:** > - Patients who were genuinely unscreened during the observation period but subsequently completed screening are excluded from the entire cohort > - The training cohort is biased toward "persistently unscreened" patients > - Model may overestimate risk for patients who eventually comply with screening > > **Mitigation:** This limitation is documented and supplemental internal screening checks (see 3.5) provide partial temporal correction. Post-deployment calibration monitoring is recommended.

3.5 Recent Screening (Internal ORDER_PROC_ENH)

Criterion: Exclude observation if patient had screening procedure within the guideline-recommended interval, as detected in internal procedure records

Screening Modalities and Validity Windows:

Modality	CPT Codes	Validity Window
Colonoscopy	45378, 45380, 45381, 45382, 45384, 45385, 45386, 45388, 45389, 45390, 45391, 45392, 45393, 45398	10 years
CT Colonography	74261, 74262, 74263	5 years
Flexible Sigmoidoscopy	45330-45350 (various)	5 years
FIT-DNA (Cologuard)	81528	3 years
FOBT/FIT	82270, 82274, G0328	1 year

CRITICAL LIMITATION: > Internal procedure data is only trusted from **July 1, 2021 onward**. Procedures before this date may have data quality issues and are not used for internal screening exclusion. > > **Consequence:** Patients screened before July 2021 are only excluded if captured in the VBC table.

Justification for Validity Windows: Based on USPSTF CRC screening guidelines and standard clinical practice intervals.

---

4. Summary of Known Limitations

4.1 VBC Table Temporal Limitation (HIGH IMPACT)

The VBC screening table exclusion operates on current status rather than point-in-time status. This creates a training population biased toward persistently non-compliant patients. The direction of bias is toward OVERESTIMATION of risk for eventually compliant patients, which is clinically preferable (safer) for a screening outreach model.

4.2 Internal Screening Data Cutoff (MEDIUM IMPACT)

Screening procedures before July 1, 2021 are not reliably captured in internal ORDER_PROC_ENH data. Patients screened externally or before this date may only be excluded via VBC table.

4.3 External Screening Not Captured (MEDIUM IMPACT)

Screening performed at facilities outside the health system may not be captured in either VBC table or internal procedure records. These patients may incorrectly remain in the "unscreened" cohort.

4.4 Prevalent vs. Incident Cases (LOW IMPACT)

The cohort includes both: - **Prevalent cases:** Undiagnosed CRC present at observation date - **Incident cases:** CRC developing after observation date

This is appropriate for a screening model targeting unscreened populations, as prevalent undiagnosed cases ARE the target for detection.

4.5 Medical Exclusion Timing (NO TIME WINDOW)

Medical exclusions (prior CRC, colectomy, hospice) use **any historical date** rather than a lookback window. This is intentionally conservative: - A patient with CRC 15 years ago is still excluded - A patient with colectomy at any time is excluded - This may exclude some patients who could theoretically benefit from screening (e.g., partial colectomy), but the conservative approach is clinically appropriate

---

5. Cohort Construction Flow

```
Step 1: Identify all patients with encounters during study period
        → ~X million patient-months

Step 2: Apply age filter (45-100 years)
        → Removes ~Y% (patients outside screening age)

Step 3: Apply observability filter (24+ months prior contact)
        → Removes ~Z% (insufficient history for features)

Step 4: Apply medical exclusions (prior CRC, colectomy, hospice)
        → Removes <1% (inappropriate screening candidates)

Step 5: Apply label quality filter (tiered observability)
        → Removes ~W% (cannot confirm negative label)

Step 6: Apply VBC screening exclusion
        → Removes ~40-50% (currently screened per VBC)

Step 7: Apply internal screening exclusion (2021-07-01+)
        → Removes additional ~5-10% (recent screening in ORDER_PROC_ENH)

FINAL COHORT: ~831,000 observations, ~233,000 patients, 0.41% positive rate
```

---

6. ICD-10 Code Reference

6.1 Outcome Definition (CRC Diagnosis)

Code	Description	Included
C18.0	Malignant neoplasm of cecum	Yes
C18.1	Malignant neoplasm of appendix	Yes
C18.2	Malignant neoplasm of ascending colon	Yes
C18.3	Malignant neoplasm of hepatic flexure	Yes
C18.4	Malignant neoplasm of transverse colon	Yes
C18.5	Malignant neoplasm of splenic flexure	Yes
C18.6	Malignant neoplasm of descending colon	Yes
C18.7	Malignant neoplasm of sigmoid colon	Yes
C18.8	Malignant neoplasm of overlapping sites of colon	Yes
C18.9	Malignant neoplasm of colon, unspecified	Yes
C19	Malignant neoplasm of rectosigmoid junction	Yes
C20	Malignant neoplasm of rectum	Yes
C21.x	Malignant neoplasm of anus	**No**

6.2 Medical Exclusion Codes

Code	Description	Exclusion Type
C18.x, C19, C20	Prior CRC diagnosis	Medical
Z90.49	Acquired absence of digestive tract parts	Colectomy
K91.850	Colectomy history	Colectomy

Code	Description	Exclusion Type
Z51.5x	Encounter for palliative care	Hospice

---

7. References

1. US Preventive Services Task Force. Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. JAMA. 2021;325(19):1965-1977.
2. American Cancer Society. Colorectal Cancer Screening Guidelines.
3. Multi-Society Task Force on Colorectal Cancer. Recommendations for Follow-up After Colonoscopy and Polypectomy.

---

8. Document Control

Version	Date	Author	Changes
1.0	January 2026	[Your Name]	Initial version

This document describes the cohort construction methodology for research and quality improvement purposes. All patient data handling complies with institutional IRB protocols and HIPAA requirements.