CRC Risk Prediction: Cohort Creation Logic

What This Cohort Is

We build a dataset to predict which **unscreened patients** are most likely to be diagnosed with colorectal cancer (CRC) within the **next 6 months**. The model's purpose is to prioritize screening outreach — identifying the highest-risk unscreened patients so limited screening resources reach those who need them most.

The cohort covers **January 2023 through September 2024**, drawing from all patients with encounters in Mercy's integrated health system during that period.

Step 1: Patient Identification

Plain language: We start with every patient who had a completed visit — outpatient or inpatient — in our health system during the study window, and who meets basic eligibility criteria.

Encounter Sources

Source	Table	Qualifying Criteria
Outpatient	`PAT_ENC_ENH`	Appointment status = Completed or Arrived; department in our integrated system
Inpatient	`PAT_ENC_HSP_HAR_ENH`	Not a pre-admit; not canceled; has charges; not a combined/duplicate account; department in our integrated system

Eligibility Filters

Filter	Requirement	Why
Age	45–100 years at observation date	Matches USPSTF CRC screening guidelines
System history	At least 24 months of prior encounters anywhere in Mercy's EHR	Ensures enough historical data for meaningful feature engineering (lab trends, weight trajectories, visit patterns)
Data quality	Age plausible (0–100), first-seen date before observation date, system tenure not longer than patient's lifetime	Removes data entry errors and impossible records

PCP Status

For each patient-observation, we determine whether the patient has an **active Primary Care Provider within Mercy's integrated system** at that date. This is checked by joining against `pat_pcp` with effective/termination date ranges, restricted to providers flagged as Integrated or Integrated-Regional.

PCP status is critical for two reasons: 1. It is a **model feature** (patients with PCPs have different healthcare engagement patterns). 2. It determines **label confidence** for negative cases (Step 4 below).

Step 2: Medical Exclusions

Plain language: We remove any patient-observation where the patient has **already been diagnosed with or treated for CRC** before the observation date. These patients are not screening candidates.

Condition	ICD-10 Codes	Rationale
Prior CRC diagnosis	C18 (colon), C19 (rectosigmoid), C20 (rectum)	Already diagnosed — not a screening candidate
Prior colectomy	Z90.49	Colon partially or fully removed
Colostomy complications	K91.850	Implies prior colorectal surgical intervention
Hospice/palliative care	Z51.5x	End-of-life care — screening is inappropriate

These are checked against **all encounters on or before the observation date** — both outpatient diagnosis tables and inpatient diagnosis tables.

Step 3: Screening Exclusions

Plain language: The model targets **unscreened** patients. We need to identify and remove patients who are already up-to-date on CRC screening. We use **two independent data sources** because neither is complete on its own.

Important Data Limitations

Two constraints shape this entire step:

1. **Mercy's EHR procedure data only goes back to July 1, 2021.** Any screening procedures before that date are invisible to our internal procedure search. A patient who had a colonoscopy in 2019 (still valid for 10 years) would not be detected by the internal check alone.
2. **The VBC (Value-Based Care) screening registry is not timestamped.** It tells us whether a patient is *currently* flagged as screened, but not *when* they were screened. We

cannot determine whether a patient was screened as of their historical observation date (e.g., March 2023). We can only see their status as of today.

Because of these limitations, **our screening exclusions are intentionally over-broad.** We accept that we will exclude some patients who were actually unscreened at their observation date, because the alternative — including screened patients in the training data — would be worse. Screened patients have fundamentally different risk profiles, and including them would contaminate the model's ability to learn patterns specific to unscreened populations.

> **Note for other health systems:** If your screening registry includes a timestamp for when screening status was recorded or when the screening procedure occurred, you could apply **point-in-time** screening exclusions instead of current-status exclusions. This would reduce over-exclusion and broaden the eligible cohort — a patient screened in 2024 would no longer be retroactively excluded from their 2023 observations. The rest of the pipeline (medical exclusions, label tiers, etc.) would remain unchanged. The internal procedure check (Source 2 below) already operates on a point-in-time basis; only the registry-based exclusion (Source 1) would benefit from timestamping.

Source 1: VBC Screening Registry

- Administrative table tracking whether patients have met colon cancer screening requirements
- Provides a current `COLON_SCREEN_MET_FLAG` (Y/N)
- We exclude any patient where `COLON_SCREEN_MET_FLAG = 'Y'` and `COLON_SCREEN_EXCL_FLAG = 'N'`
- **This is a patient-level exclusion** — if the VBC table says a patient is currently screened, ALL of that patient's observations are excluded, regardless of observation date
- This is the primary source of screening exclusion and catches patients whose screening history predates July 2021

Source 2: Internal Procedure Records

We search Mercy's `ORDER_PROC_ENH` table for specific screening procedures, but **only from July 1, 2021 onward** (the boundary of reliable procedure data). Each screening type is checked against its own guideline-based validity window:

Screening Type	Valid For	CPT Codes / Keywords
Colonoscopy	10 years	45378, 45380–45398, "colonoscopy"
CT Colonography	5 years	74261–74263, "ct colonography", "virtual colonoscopy"
Flexible Sigmoidoscopy	5 years	45330–45350, "sigmoidoscopy"
FIT-DNA (Cologuard)	3 years	81528, "cologuard", "fit-dna"
FOBT/FIT	1 year	82270, 82274, G0328, "fobt", "fecal occult"

Per-modality validity checking: Each screening type is independently checked against its own window. A patient is excluded only if at least one modality has a procedure within its valid timeframe relative to the observation date. This avoids cross-contamination between modality windows (e.g., an expired FOBT being incorrectly validated by a colonoscopy's 10-year window).

Additional filters: only non-canceled orders from departments within our integrated system.

Combined Exclusion

A patient-observation is excluded if EITHER source indicates screening: - VBC registry says the patient is currently screened, **OR** - Internal procedure records show a valid (non-expired) screening procedure before the observation date

Approximate impact: ~49.5% of observations are excluded by screening filters, leaving the target unscreened population.

Step 4: Label Construction (Three-Tier Confidence System)

Plain language: For each remaining patient-observation, we determine: **was this patient diagnosed with CRC within the next 6 months?**

Positive Labels (Straightforward)

We search for CRC diagnosis codes (C18, C19, C20) in completed encounters during the 6-month window after the observation date. If found, the observation is labeled positive. The specific cancer subtype (colon, rectosigmoid, rectal) is recorded for analysis.

Negative Labels (The Challenge)

Just because we don't *see* a CRC diagnosis in the 6-month window doesn't definitively mean the patient doesn't have cancer. They may have: - Been diagnosed at a facility outside Mercy - Simply not returned to any healthcare provider - Had cancer that went undiagnosed during that period

Requiring perfect 6-month follow-up for every negative would discard massive amounts of training data. Instead, we assign **confidence tiers** based on how reliably we can confirm that no CRC diagnosis occurred.

The Three Tiers

Tier	Criteria	Share of Negatives	Confidence	Reasoning
Tier 1	Patient returned for a completed visit **after** the 6-month window (i.e., in months 7–12)	~47%	**High**	Their follow-up spans the entire prediction window. If CRC had been diagnosed during months 1–6, it would appear in their medical record. The return visit after month 6 confirms ongoing engagement with our system.
Tier 2	Patient returned in **months 4–6** AND has an **active PCP** in our system	~23%	**Medium**	The return visit covers most of the prediction window (months 4–6 out of 6). Combined with an active PCP relationship, there is strong evidence of continued system engagement. A cancer diagnosis — even from an external facility — would typically be communicated to the PCP.

Tier	Criteria	Share of Negatives	Confidence	Reasoning
Tier 3	Patient did **NOT return** within 12 months, BUT has an **active PCP** in our system	~30%	**Lower**	No return visit to directly confirm, but the active PCP relationship implies the patient is engaged with Mercy. If diagnosed anywhere, the PCP would typically receive notification. Additionally, 12 months have elapsed since the observation — enough time for a cancer diagnosis to surface through normal care channels.

Excluded (Not Used for Training)

Patients with **no return visit AND no active PCP** are excluded entirely. Without either form of follow-up confirmation, we cannot reliably label them as negative. Including uncertain labels would introduce noise that degrades model performance.

Why Include Tier 3?

Tier 3 patients are the most uncertain negatives. Including them is a deliberate tradeoff: - **Benefit:** Increases training data by ~30%, which is valuable when the positive rate is only ~0.41% (1 in 250). More training examples of negatives help the model learn what "normal" looks like. - **Risk:** Some Tier 3 patients may actually have undiagnosed CRC, introducing label noise. - **Justification:** In practice, the vast majority of patients with active PCPs who are not diagnosed with CRC within 12 months genuinely do not have CRC. The small amount of label noise is outweighed by the training data gain.

Overall Positive Rate

~0.41% of observations are positive (approximately 1 in 250), reflecting the rarity of CRC diagnosis within 6 months in this unscreened population.

Known Limitations

Limitation	Impact	Direction of Bias
Mercy procedure data begins July 1, 2021	Screenings before that date are invisible to the internal procedure check	Under-exclusion (mitigated by VBC table)
VBC screening registry is not timestamped	Cannot determine point-in-time screening status; a patient screened after their observation date may be retrospectively excluded	Over-exclusion (smaller but cleaner unscreened population). Systems with timestamped registries could apply point-in-time exclusions and retain a broader cohort.
Dual exclusion is intentionally over-broad	Some truly unscreened patients are excluded because VBC says they are currently screened	Over-exclusion (acceptable — conservative direction)
Tier 3 negatives are assumed negative	Some may have undiagnosed CRC	Label noise (~30% of negatives affected, but true mislabel rate is very low)
Single health system	Patients diagnosed at external facilities may not appear in our records	Under-labeling of positives (mitigated by PCP notification pathways)
Prevalent cases	Some "new" CRC diagnoses during the prediction window may be pre-existing undiagnosed tumors, not truly incident cases	Clinically acceptable — finding prevalent cases is still valuable for screening triage

---

Summary: The Cohort Pipeline

```
All patients with completed encounters in Mercy (Jan 2023 - Sept 2024)
    |
    +-- Filter: Age 45-100, 24+ months in system, data quality checks
    |
    +-- Exclude: Prior CRC diagnosis, colectomy, hospice
    |
    +-- Exclude: Currently screened (VBC registry OR internal procedures)
```

```
|       Note: Over-broad by design due to VBC lacking timestamps
|       and Mercy data starting July 2021
|
+-- Label: CRC diagnosis within 6 months?
|       Positives: CRC code (C18/C19/C20) in 6-month window
|       Negatives: Three-tier confidence system
|
+-- Filter: Keep only observations with reliable labels (Tiers 1-3)
|
Result: ~830K observations, ~0.41% positive rate
```