CRC Model: Feature Pipeline by Notebook

This document traces every feature from raw clinical data through each reduction stage, with derivation details explaining how each feature is computed.

FHIR Availability Key: - **Yes** – Standard FHIR resource with well-defined codes (LOINC, SNOMED, ICD-10) - **Partial** – FHIR resource exists but implementation and data quality vary significantly across systems - **No** – No standard FHIR mapping; system-specific data

---

Book 0: Demographics & Cohort

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
Gender	PATIENT_ENH	Yes	Patient.gender
Birth date (for age)	PATIENT_ENH	Yes	Patient.birthDate
Marital status	PATIENT_ENH	Yes	Patient.maritalStatus
Race	PATIENT_ENH	Partial	US Core Patient.race extension
Encounter dates	PAT_ENC_ENH, PAT_ENC_HSP_HAR_ENH	Yes	Encounter.period
PCP assignment (with eff/term dates)	pat_pcp, clarity_ser_enh	Partial	CareTeam / PractitionerRole
Months since first encounter	PAT_ENC_ENH, PAT_ENC_HSP_HAR_ENH	Yes	Derived from Encounter.period

2. Features Engineered (13 features)

Feature	Derivation	FHIR-Derivable
AGE	Integer: `FLOOR(DATEDIFF(END_DTTM, BIRTH_DATE) / 365.25)`. Filtered to 45-100.	Yes
IS_FEMALE	Binary 1/0: 1 if `GENDER = 'Female'`	Yes
IS_MARRIED_PARTNER	Binary 1/0: 1 if `MARITAL_STATUS IN ('Married', 'Significant other')`	Yes
RACE_CAUCASIAN	Binary 1/0: 1 if RACE_BUCKETS = 'Caucasian'. Raw RACE mapped: 'Unknown/Refused'→NULL, small groups→'Other_Small'.	Partial
RACE_BLACK_OR_AFRICAN_AMERICAN	Binary 1/0: 1 if RACE_BUCKETS = 'Black or African American'	Partial
RACE_HISPANIC	Binary 1/0: 1 if RACE_BUCKETS = 'Hispanic'	Partial
RACE_ASIAN	Binary 1/0: 1 if RACE_BUCKETS = 'Asian'	Partial
RACE_OTHER	Binary 1/0: 1 if RACE_BUCKETS IN ('Other', 'Other_Small')	Partial

Feature	Derivation	FHIR-Derivable
OBS_MONTHS_PRIOR	Integer: `CAST(months_between(END_DTTM, first_seen_dt) AS INT)`. first_seen_dt = earliest encounter across outpatient + inpatient. Filtered to >= 24 months.	Yes
HAS_PCP_AT_END	Binary 1/0: 1 if patient has active PCP at snapshot date. Joins `pat_pcp` where `END_DTTM BETWEEN EFF_DATE AND COALESCE(TERM_DATE, '9999-12-31')`, PCP must be in integrated network (`RPT_GRP_ELEVEN_NAME IN ('Integrated-Regional', 'Integrated')`).	Partial
HAS_FULL_24M_HISTORY	Binary 1/0: 1 if `OBS_MONTHS_PRIOR >= 24`. Always 1 in final cohort (pre-filtered).	Yes
age_group	Categorical string: 'age_45_49' (45-49), 'age_50_64' (50-64), 'age_65_74' (65-74), 'age_75_plus' (75+). Aliased as AGE_GROUP in Book 8.	Yes
months_since_cohort_entry	Integer: `CAST(months_between(END_DTTM, first_obs_date) AS INT)` where first_obs_date = MIN(END_DTTM) for the patient within the cohort.	Yes

3. Features Sent to Book 9 (via Book 8 compilation)

Book 8 passes through all Book 0 columns except screening metadata. The columns Book 8 explicitly drops: `LABEL_CONFIDENCE`, `current_screen_status`, and all `vbc_*/last_*_date` screening audit columns.

Feature	Derivation	FHIR-Derivable
IS_FEMALE	Binary gender flag	Yes
IS_MARRIED_PARTNER	Binary marital status flag	Yes
HAS_PCP_AT_END	Active PCP at observation date	Partial
months_since_cohort_entry	Months since patient's first observation in cohort	Yes
RACE_CAUCASIAN	One-hot race encoding	Partial
RACE_BLACK_OR_AFRICAN_AMERICAN	One-hot race encoding	Partial
RACE_HISPANIC	One-hot race encoding	Partial
RACE_ASIAN	One-hot race encoding	Partial
RACE_OTHER	One-hot race encoding	Partial
HAS_FULL_24M_HISTORY	Always 1 (constant after filtering)	Yes
AGE_GROUP	Categorical age band (age_45_49, age_50_64, age_65_74, age_75_plus)	Yes

4. Features in Final Model (3 of 26)

Feature	Derivation	FHIR-Derivable
AGE_GROUP	Categorical age band	Yes
RACE_CAUCASIAN	Binary: 1 if Caucasian	Partial
months_since_cohort_entry	Months since first cohort observation	Yes

---

Book 1: Vitals

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource / LOINC
Systolic blood pressure	pat_enc_enh.BP_SYSTOLIC	Yes	Observation (85354-9)
Diastolic blood pressure	pat_enc_enh.BP_DIASTOLIC	Yes	Observation (85354-9)
Weight (ounces)	pat_enc_enh.WEIGHT	Yes	Observation (29463-7)
Pulse / heart rate	pat_enc_enh.PULSE	Yes	Observation (8867-4)
BMI	pat_enc_enh.BMI	Yes	Observation (39156-5)
Temperature	pat_enc_enh.TEMPERATURE	Yes	Observation (8310-5)
Respiratory rate	pat_enc_enh.RESP_RATE	Yes	Observation (9279-1)

Plausibility filters: Weight 50-800 lbs, BP systolic 60-280, BP diastolic 40-180, Pulse 20-250, BMI 10-100, Temperature 95-105 F, Resp rate 5-60. Measurements within 12-month lookback from END_DTTM (Jul 2021 data cutoff).

2. Features Engineered (~50 before reduction)

Latest Values (7): WEIGHT_OZ, WEIGHT_LB, BP_SYSTOLIC, BP_DIASTOLIC, PULSE, BMI, TEMPERATURE – most recent measurement via ROW_NUMBER() ordered by date DESC.

Recency (6): DAYS_SINCE_WEIGHT, DAYS_SINCE_SBP, DAYS_SINCE_PULSE, DAYS_SINCE_BMI, DAYS_SINCE_TEMPERATURE, DAYS_SINCE_RESP_RATE – DATEDIFF(END_DTTM, measurement_date).

Weight Change (6): WEIGHT_CHANGE_PCT_6M, WEIGHT_CHANGE_PCT_12M (% change from $_{180d/}$365d ago), MAX_WEIGHT_LOSS_PCT_60D (max consecutive loss in 60d window), WEIGHT_LOSS_5PCT_6M, WEIGHT_LOSS_10PCT_6M (binary flags), RAPID_WEIGHT_LOSS_FLAG (MAX_WEIGHT_LOSS_PCT_60D >= 5).

Weight Trajectory (3): WEIGHT_TRAJECTORY_SLOPE (REGR_SLOPE over 12mo), WEIGHT_TRAJECTORY_R2 (REGR_R2), WEIGHT_VOLATILITY_12M (STDDEV).

BP Derived (7): PULSE_PRESSURE (SBP-DBP), MEAN_ARTERIAL_PRESSURE ((2*DBP+SBP)/3), AVG_PULSE_PRESSURE_6M, SBP_VARIABILITY_6M (STDDEV over 6mo), DBP_VARIABILITY_6M, PULSE_PRESSURE_VARIABILITY_6M.

BMI Change (4): BMI_CHANGE_6M, BMI_CHANGE_12M, BMI_LOST_OBESE_STATUS, BMI_LOST_OVERWEIGH

Counts (2): WEIGHT_MEASUREMENT_COUNT_12M, BP_MEASUREMENT_COUNT_6M.

Clinical Flags (7): HYPERTENSION_FLAG (SBP>=140 or DBP>=90), SEVERE_HYPERTENSION_FLAG (SBP>=160 or DBP>=100), TACHYCARDIA_FLAG (Pulse>100), UNDERWEIGHT_FLAG (BMI<18.5), OBESE_FLAG (BMI>=30), FEVER_FLAG (Temp>100.4), TACHYPNEA_FLAG (Resp>20), BRADYPNEA_FLAG (Resp<12).

Composite Scores (1): CACHEXIA_RISK_SCORE (ordinal 0-2: 2=BMI<20+5%loss, 1=BMI<22+5%loss or BMI<20, 0=other).

3. Features Sent to Book 9 (24 after book-level reduction)

Reduction selects optimal features per correlated group, adds 5 composites. All carry `vit_` prefix after Book 8 compilation.

Feature	Derivation	FHIR
WEIGHT_OZ	Continuous (oz): most recent weight. ROW_NUMBER by date DESC. Plausibility [800, 12800] oz.	Yes
BP_SYSTOLIC	Continuous (mmHg): most recent systolic BP. Plausibility [60, 280].	Yes
BMI	Continuous (kg/m²): most recent BMI. Plausibility [10, 100].	Yes
PULSE	Continuous (bpm): most recent heart rate. Plausibility [20, 250].	Yes
PULSE_PRESSURE	Continuous (mmHg): latest BP_SYSTOLIC - BP_DIASTOLIC.	Yes
WEIGHT_CHANGE_PCT_6M	Continuous (%): `((latest_weight - weight_6mo_ago) / weight_6mo_ago) * 100`. Historical weight matched to ~180 days ± 30 days.	Yes
MAX_WEIGHT_LOSS_PCT_60D	Continuous (%): max pct loss between consecutive weight measurements within 60 days of snapshot.	Yes
WEIGHT_TRAJECTORY_SLOPE	Continuous: `REGR_SLOPE(WEIGHT_OZ, DAYS_BEFORE_END)` over all weights in 12-month window. Positive = gaining weight approaching snapshot.	Yes
WEIGHT_LOSS_10PCT_6M	Binary 1/0: 1 if 6-month weight change <= -10%.	Yes
RAPID_WEIGHT_LOSS_FLAG	Binary 1/0: 1 if MAX_WEIGHT_LOSS_PCT_60D >= 5%.	Yes
DAYS_SINCE_WEIGHT	Integer: days from most recent weight measurement to snapshot date.	Yes
SBP_VARIABILITY_6M	Continuous (mmHg): `STDDEV(BP_SYSTOLIC)` over 6-month window.	Yes
BMI_CHANGE_6M	Continuous (kg/m²): absolute BMI change over 6 months (latest BMI - BMI ~180 days ago).	Yes
HYPERTENSION_FLAG	Binary 1/0: 1 if SBP >= 140 or DBP >= 90.	Yes
TACHYCARDIA_FLAG	Binary 1/0: 1 if PULSE > 100.	Yes
FEVER_FLAG	Binary 1/0: 1 if TEMPERATURE > 100.4°F.	Yes
OBESE_FLAG	Binary 1/0: 1 if BMI >= 30.	Yes
UNDERWEIGHT_FLAG	Binary 1/0: 1 if BMI < 18.5.	Yes
CACHEXIA_RISK_SCORE	Ordinal 0-2: composite of low BMI + weight loss. 2=BMI<20 AND 6mo loss >= 5%; 1=BMI<22+loss or BMI<20 alone; 0=other.	Yes
weight_loss_severity	Ordinal 0-3: 3 if >=10% 6mo loss; 2 if >=5%; 1 if >2%; 0 otherwise.	Yes
vital_recency_score	Ordinal 0-3: 3 if weight measured <=30d ago; 2 if <=90d; 1 if <=180d; 0 if >180d/null.	Yes
cardiovascular_risk	Ordinal 0-2: 2 if HYPERTENSION + OBESE; 1 if either; 0 if neither.	Yes
abnormal_weight_pattern	Binary 1/0: 1 if MAX_WEIGHT_LOSS_PCT_60D > 5 OR WEIGHT_TRAJECTORY_SLOPE < -0.5.	Yes

Feature	Derivation	FHIR
bp_instability	Binary 1/0: 1 if SBP_VARIABILITY_6M > 15 mmHg.	Yes

4. Features in Final Model (8 of 26)

Feature	Derivation	FHIR
vit_WEIGHT_OZ	Most recent weight in ounces	Yes
vit_PULSE_PRESSURE	Latest SBP minus DBP	Yes
vit_WEIGHT_CHANGE_PCT_6M	6-month weight change percentage	Yes
vit_MAX_WEIGHT_LOSS_PCT_60D	Max consecutive weight loss % in 60 days	Yes
vit_WEIGHT_TRAJECTORY_SLOPE	Linear regression slope of weight over 12 months	Yes
vit_RECENCY_WEIGHT	Days since most recent weight measurement (renamed from DAYS_SINCE_WEIGHT)	Yes
vit_SBP_VARIABILITY_6M	Standard deviation of systolic BP over 6 months	Yes
vit_CACHEXIA_RISK_SCORE	Cachexia risk composite (BMI + weight loss)	Yes

---

Book 2: ICD-10 Diagnoses

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
Outpatient ICD-10 diagnosis codes	pat_enc_dx_enh (via pat_enc_enh)	Yes	Condition
Inpatient ICD-10 diagnosis codes	hsp_acct_dx_list_enh (via pat_enc_hsp_har_enh, using DISCH_DATE_TIME)	Yes	Condition
Problem list (active conditions)	problem_list_hx_enh (status='Active')	Yes	Condition (category=problem-list-item)
Structured family history	prod.clarity.family_hx (MEDICAL_HX_C codes)	Partial	FamilyMemberHistory

ICD-10 code families: GI bleeding (K62.5, K92.1, K92.2), abdominal pain (R10.*), bowel changes (R19.4, K59.0, R19.7), weight loss (R63.4), fatigue (R53.1, R53.83), anemias (D50-D53, D62-D64), iron deficiency anemia (D50.*), polyps (D12.*, K63.5), IBD (K50.*, K51.*), prior malignancy (Z85.*), diabetes (E10.*, E11.*), obesity (E66.*), diverticular (K57.*), constipation (K59.0), and 20+ additional categories.

Family history codes (structured FAMILY_HX table): 10404 (Colon Cancer), 20172 (Rectal Cancer), 103028 (Lynch Syndrome), 20103/20191 (Polyps). First-degree relatives: Mother, Father, Brother, Sister, Son, Daughter.

2. Features Engineered (116 before reduction)

All three diagnosis sources (outpatient, inpatient, problem list) are unioned into a curated conditions table. Features span:

Feature Category	Count	Time Window	Description
Symptom flags (12mo/24mo)	12	12/24 months	Binary: any occurrence of each symptom ICD-10 family
Symptom counts (12mo/24mo)	12	12/24 months	Integer: count of diagnosis records per symptom
Risk factor flags	13	Ever (lifetime)	Binary: polyps, IBD, malignancy, diabetes, obesity, etc.
Family history (structured + ICD)	9	Ever	Combined FAMILY_HX table + ICD Z-codes
Comorbidity scores	6	12/24 months	Weighted Charlson (17 conditions) and Elixhauser (10 conditions)
Recency features	5	Up to 24mo	Days since last occurrence of each symptom
Acceleration features	6	12mo windows	Is symptom frequency increasing over time?
Composite scores	12	Mixed	CRC_SYMPTOM_TRIAD, SYMPTOM_BURDEN, GI_COMPLEXITY, etc.
Additional condition flags	41	Mixed	Rare conditions, additional time windows, youngest onset age

Note: Comorbidity scores (Charlson, Elixhauser) use outpatient diagnoses only (pat_enc_dx_enh), not inpatient or problem list.

3. Features Sent to Book 9 (26 after book-level reduction)

All carry `icd_` prefix after the reduction step.

Feature	Derivation	FHIR
icd_BLEED_FLAG_12MO	Binary 1/0: 1 if any K62.5/K92.1/K92.2 (GI hemorrhage) in 12mo before snapshot. Sources: outpatient dx, inpatient dx, problem list.	Yes
icd_BLEED_CNT_12MO	Count: number of bleeding diagnosis records in 12mo. Each encounter/problem entry counted separately.	Yes
icd_ANEMIA_FLAG_12MO	Binary 1/0: 1 if any D50-D53/D62-D64 (nutritional/other anemias) in 12mo.	Yes
icd_IRON_DEF_ANEMIA_FLAG_12MO	Binary 1/0: 1 if any D50.* (iron deficiency anemia specifically) in 12mo. Subset of ANEMIA_FLAG.	Yes
icd_PAIN_FLAG_12MO	Binary 1/0: 1 if any R10.* (abdominal/pelvic pain) in 12mo.	Yes
icd_BOWELCHG_FLAG_12MO	Binary 1/0: 1 if any R19.4 (bowel habit change), K59.0 (constipation), or R19.7 (diarrhea) in 12mo.	Yes
icd_WTLOSS_FLAG_12MO	Binary 1/0: 1 if any R63.4 (abnormal weight loss) in 12mo.	Yes
icd_SYMPTOM_BURDEN_12MO	Score 0-6: sum of 6 symptom flags (BLEED + PAIN + BOWELCHG + WTLOSS + FATIGUE + ANEMIA) in 12mo.	Yes

Feature	Derivation	FHIR
icd_CRC_SYMPTOM_TRIAD	Binary 1/0: 1 if >=2 of 3 cardinal symptoms present (bleeding, pain, bowel change) in 12mo.	Yes
icd_IDA_WITH_BLEEDING	Binary 1/0: 1 if BOTH anemia AND bleeding present in 12mo (occult blood loss pattern).	Yes
icd_METABOLIC_SYNDROME	Binary 1/0: 1 if BOTH diabetes (E10-E11) AND obesity (E66) ever.	Yes
icd_severe_symptom_pattern	Binary 1/0: 1 if ANY of: (bleeding+anemia), or (symptom burden >= 3), or (weight loss+anemia) in 12mo.	Yes
icd_POLYPS_FLAG_EVER	Binary 1/0: 1 if any D12.*/K63.5 (polyps/benign colon neoplasm) ever.	Yes
icd_IBD_FLAG_EVER	Binary 1/0: 1 if any K50./*K51.* (Crohn's/ulcerative colitis) ever.	Yes
icd_MALIGNANCY_FLAG_EVER	Binary 1/0: 1 if any Z85.* (personal history of malignant neoplasm) ever.	Yes
icd_DIABETES_FLAG_EVER	Binary 1/0: 1 if any E10./*E11.* (diabetes) ever.	Yes
icd_OBESITY_FLAG_EVER	Binary 1/0: 1 if any E66.* (obesity) ever.	Yes
icd_HIGH_RISK_HISTORY	Binary 1/0: 1 if ANY of IBD, polyps, or prior malignancy ever present.	Yes
icd_chronic_gi_pattern	Binary 1/0: 1 if ANY of: IBD ever, diverticular disease (K57) in 24mo, or GI complexity score >= 2 (sum of malabsorption/IBS-D/hematemesis/bloating/abscess flags).	Yes
icd_FHX_CRC_COMBINED	Binary 1/0: GREATEST of structured FAMILY_HX (codes 10404/20172) and ICD Z80.0 (family hx digestive malignancy). Captures 21.4% of cohort vs 1.9% from ICD alone.	Partial
icd_FHX_FIRST_DEGREE_CRC	Binary 1/0: 1 if first-degree relative (Mother/Father/Brother/Sister/Son/Daughter) has CRC in structured FAMILY_HX.	Partial
icd_HIGH_RISK_FHX_FLAG	Binary 1/0: 1 if ANY of: Lynch syndrome in family, onset age < 50, 2+ relatives with CRC/polyps, or first-degree CRC.	Partial
icd_genetic_risk_composite	Binary 1/0: GREATEST of HIGH_RISK_FHX_FLAG and (FHX_CRC_COMBINED OR FHX_FIRST_DEGREE_CRC). Combines all family history sources.	Partial
icd_CHARLSON_SCORE_12MO	Weighted score: Charlson Comorbidity Index from 17 condition categories in 12mo outpatient diagnoses. Weights 1-6 per standard mapping (MI, CHF, COPD=1; renal, malignancy=2; severe liver=3; metastatic, AIDS=6).	Yes
icd_ELIXHAUSER_SCORE_12MO	Weighted score: simplified Elixhauser from 10 condition categories in 12mo outpatient diagnoses. Weights 1-2 (HTN, CHF, CAD, malignancy, diabetes, dementia, depression, substance use=1; renal, liver=2).	Yes
icd_COMBINED_COMORBIDITY_12MO	Integer: `GREATEST(CHARLSON_SCORE_12MO, ELIXHAUSER_SCORE_12MO)`.	Yes

4. Features in Final Model (3 of 26)

Feature	Derivation	FHIR
icd_BLEED_CNT_12MO	Count of GI bleeding diagnoses (K62.5, K92.1, K92.2) in 12 months	Yes
icd_MALIGNANCY_FLAG_EVER	Binary: any personal history of malignant neoplasm (Z85.*) ever	Yes
icd_SYMPTOM_BURDEN_12MO	Score 0-6: number of distinct CRC symptom categories present in 12 months	Yes

---

Book 3: Social Factors – SKIPPED

All features excluded during development due to data quality issues. No features from this book enter the pipeline.

Book 4: Labs

1. Raw Data Gathered

Two lab data paths (combined): - **Outpatient:** `order_results` → `clarity_component` (via order_proc_enh). Uses `RESULT_TIME` as date. - **Inpatient:** `order_proc_enh` → `clarity_eap` → `spec_test_rel` → `res_db_main` → `res_components` → `clarity_component`. Uses `COMP_VERIF_DTTM` as date.

Both paths filtered to `LAB_STATUS_C IN (3, 5)` (final/edited results) and `ORDER_STATUS_C` for completed orders.

Raw Lab Component	FHIR Available	FHIR LOINC
Hemoglobin	Yes	718-7
MCV	Yes	787-2
Platelets	Yes	777-3
Iron (serum)	Yes	2498-4
TIBC	Yes	2500-7
Ferritin	Yes	2276-4
Transferrin saturation	Yes	2502-3
CRP	Yes	1988-5
ESR	Yes	4537-7
Albumin	Yes	1751-7
ALT	Yes	1742-6
AST	Yes	1920-8
Alkaline phosphatase	Yes	6768-6
CA-125	Yes	10334-1

Removed from pipeline (circular reasoning): CEA (2039-6), CA 19-9 (24108-3), FOBT/FIT (29771-3). See `docs/book4_cea_fobt_removal_guide.md`.

Lookback windows: Routine labs = 2 years (730 days). Slow-changing markers (CA125, Ferritin, CRP, ESR) = 3 years (1095 days).

Temporal reference points for trends: current = most recent; 1mo/3mo/6mo/9mo/12mo prior = closest value to N days ago (±15 day tolerance windows).

2. Features Engineered (~80 before reduction)

Feature Category	Count	Description
Latest lab values	~18	Most recent value per component via ROW_NUMBER
6-month changes	~8	current_value - value_6mo_prior for key labs
12-month changes	~4	current_value - value_12mo_prior
Velocity (per month)	~4	Rate of change over 3-month windows
Acceleration flags	~4	Is the rate of change getting worse?
Drop/pattern flags	~6	HEMOGLOBIN_DROP_10PCT, THROMBOCYTOSIS, etc.
Anemia classification	~6	WHO-grade anemia, iron deficiency pattern
Calculated ratios	~4	ALT/AST ratio, iron saturation %
Trajectory categories	~2	HGB_TRAJECTORY (ordinal 0-3)

3. Features Sent to Book 9 (25 after book-level reduction)

All carry `lab_` prefix.

Feature	Derivation	FHIR
lab_HEMOGLOBIN_VALUE	Continuous (g/dL): most recent hemoglobin. Plausibility [3, 20].	Yes
lab_HEMOGLOBIN_6MO_CHANGE	Continuous (g/dL): current hemoglobin minus hemoglobin ~6mo prior. Negative = decline.	Yes
lab_HEMOGLOBIN_DROP_10PCT_FLAG	Binary 1/0: 1 if current HGB < 0.9 × HGB_6mo_prior.	Yes
lab_HEMOGLOBIN_ACCELERATING_DECLINE	Binary 1/0: 1 if HGB declining AND acceleration increasing. Requires recent velocity < -0.5 g/dL/month AND recent velocity more negative than prior velocity (0-3mo vs 3-6mo comparison).	Yes
lab_HGB_TRAJECTORY	Ordinal 0-3: 0=stable/rising, 1=mild decline (0 to -1 g/dL over 12mo), 2=moderate (-1 to -2), 3=rapid (< -2 g/dL).	Yes
lab_ANEMIA_GRADE	Ordinal 0-3: WHO classification. 0=normal (HGB>=12), 1=mild (11-12), 2=moderate (8-11), 3=severe (<8).	Yes
lab_ANEMIA_SEVERITY_SCORE	Ordinal 0-6: anemia_grade (0-3) + iron_deficiency×2 + microcytosis(MCV<80)×1.	Yes
lab_IRON_DEFICIENCY_ANEMIA_FLAG	Binary 1/0: 1 if HGB<12 AND MCV<80 AND (FERRITIN<30 OR iron sat<20%).	Yes
lab_IRON_SATURATION_PCT	Continuous (%, 0-100): calculated as `(IRON_VALUE / TIBC_VALUE) × 100`. Uses most recent paired IRON and TIBC from same encounter where both non-null and TIBC>0.	Yes
lab_PLATELETS_VALUE	Continuous (×10³/μL): most recent platelet count. Plausibility [10, 2000].	Yes
lab_PLATELETS_ACCELERATING_RISE	Binary 1/0: 1 if platelets > 450 AND rise accelerating (0-3mo velocity > 3-6mo velocity).	Yes
lab_THROMBOCYTOSIS_FLAG	Binary 1/0: 1 if most recent platelets > 450 ×10³/μL.	Yes
lab_ALBUMIN_VALUE	Continuous (g/dL): most recent albumin. Plausibility [1, 6].	Yes
lab_ALBUMIN_DROP_15PCT_FLAG	Binary 1/0: 1 if current albumin < 0.85 × albumin_6mo_prior.	Yes

Feature	Derivation	FHIR
lab_AST_VALUE	Continuous (U/L): most recent AST. Plausibility [0, 2000].	Yes
lab_ALK_PHOS_VALUE	Continuous (U/L): most recent alkaline phosphatase. Plausibility [0, 2000].	Yes
lab_ALT_AST_RATIO	Continuous (unitless): ALT_VALUE / AST_VALUE when AST > 0. NULL if AST null/zero.	Yes
lab_ESR_VALUE	Continuous (mm/hr): most recent ESR. 3-year lookback. Plausibility [0, 200].	Yes
lab_CRP_6MO_CHANGE	Continuous (mg/L): current CRP minus CRP ~6mo prior. Positive = increasing inflammation.	Yes
lab_FERRITIN_6MO_CHANGE	Continuous (ng/mL): current ferritin minus ferritin ~6mo prior. Negative = declining iron stores.	Yes
lab_CA125_VALUE	Continuous (U/mL): most recent CA-125 (ovarian marker, not CRC-specific). 3-year lookback. Plausibility [0, 50000].	Yes
lab_comprehensive_iron_deficiency	Binary 1/0 (composite): 1 if ANY of: IRON_DEFICIENCY_ANEMIA_FLAG=1, or (HGB<12 AND MCV<80), or (FERRITIN<30 AND HGB<13).	Yes
lab_metabolic_dysfunction	Binary 1/0 (composite): 1 if ANY of: ALT abnormal, AST abnormal, ALK_PHOS>150, or ALBUMIN_DROP_15PCT=1.	Yes
lab_inflammatory_burden	Binary 1/0 (composite): 1 if ANY of: CRP>10, THROMBOCYTOSIS=1, or ESR>30.	Yes
lab_progressive_anemia	Binary 1/0 (composite): 1 if ANY of: HGB_TRAJECTORY is RAPID_DECLINE or MODERATE_DECLINE, or HEMOGLOBIN_ACCELERATING_DECLINE=1.	Yes

4. Features in Final Model (7 of 26)

Feature	Derivation	FHIR
lab_ALBUMIN_VALUE	Most recent serum albumin (g/dL)	Yes
lab_HEMOGLOBIN_ACCELERATING_DECLINE	Hemoglobin declining with accelerating rate	Yes
lab_IRON_SATURATION_PCT	Calculated iron saturation: (Iron/TIBC)×100	Yes
lab_PLATELETS_ACCELERATING_RISE	Platelets rising with accelerating rate	Yes
lab_PLATELETS_VALUE	Most recent platelet count ($\times 10^3/\mu L$)	Yes
lab_THROMBOCYTOSIS_FLAG	Platelets > 450	Yes
lab_comprehensive_iron_deficiency	Composite: IDA or microcytic anemia or low ferritin+anemia	Yes

---

Book 5.1: Outpatient Medications

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
Outpatient medication orders (16 categories)	order_med_enh	Yes	MedicationRequest

Filters: `ORDERING_MODE_C <> 2` (exclude inpatient), `ORDER_STATUS_C IN (2, 5)` (Sent/Completed), `ORDER_CLASS <> 'Historical Med'`, `ORDER_START_TIME >= '2021-07-01'`. 24-month lookback from END_DTTM.

16 medication categories mapped via institutional grouper IDs and direct medication ID: Iron supplements, PPIs, NSAIDs/ASA, Statins, Metformin, Laxatives, Antidiarrheals, Antispasmodics, B12/Folate, IBD medications (mesalamine, biologics, immunosuppressants), Hemorrhoid/rectal meds, GI bleeding meds (tranexamic acid, vasopressin, octreotide), Chronic opioids, Broad-spectrum antibiotics, Hormone therapy, Chemotherapy agents.

2. Features Engineered (48 before reduction)

For each of 16 categories, 3 features: `{category}_use_flag` (binary 1/0), `{category}_use_days_since` (integer, NULL if never), `{category}_use_count_2yr` (count of distinct order-days). B12/folate and chemotherapy categories removed pre-selection (near-zero variance), leaving 42 for selection.

3. Features Sent to Book 9 (19 after book-level reduction)

All carry `out_med_` prefix. Selection rules: hemorrhoid = keep flag + days_since; iron/laxative/antidiarrheal = keep flag; PPI/statin/metformin = keep flag; others = best MI score feature.

Feature	Derivation	FHIR
out_med_iron_use_flag	Binary 1/0: any outpatient iron supplement order in 24mo	Yes
out_med_laxative_use_flag	Binary 1/0: any outpatient laxative order in 24mo	Yes
out_med_antidiarrheal_use_flag	Binary 1/0: any outpatient antidiarrheal order in 24mo	Yes
out_med_hemorrhoid_meds_flag	Binary 1/0: any outpatient hemorrhoid/rectal medication order in 24mo	Yes
out_med_hemorrhoid_meds_days_since	Integer/NULL: days from most recent hemorrhoid med order to snapshot	Yes
out_med_ppi_use_flag	Binary 1/0: any outpatient PPI order in 24mo	Yes
out_med_statin_use_flag	Binary 1/0: any outpatient statin order in 24mo	Yes
out_med_metformin_use_flag	Binary 1/0: any outpatient metformin order in 24mo	Yes
out_med_nsaid_asa_use_flag	Binary 1/0: any outpatient NSAID/ASA order in 24mo (best MI for NSAID category)	Yes
out_med_antispasmodic_use_flag	Binary 1/0: any outpatient antispasmodic order in 24mo (best MI)	Yes
out_med_ibd_meds_{type}	Best MI feature for IBD meds (data-dependent: flag or days_since)	Yes
out_med_gi_bleed_meds_{type}	Best MI feature for GI bleeding meds	Yes
out_med_opioid_use_{type}	Best MI feature for opioids	Yes
out_med_broad_abx_{type}	Best MI feature for broad-spectrum antibiotics	Yes
out_med_hormone_therapy_{type}	Best MI feature for hormone therapy	Yes
out_med_gi_symptom_meds	Binary 1/0 (composite): 1 if ANY of laxative/antidiarrheal/antispasmodic	Yes
out_med_alternating_bowel	Binary 1/0 (composite): 1 if BOTH laxative AND antidiarrheal (alternating bowel pattern)	Yes

Feature	Derivation	FHIR
out_med_gi_bleeding_pattern	Binary 1/0 (composite): 1 if BOTH iron AND PPI (GI bleeding management)	Yes
out_med_hemorrhoid_risk_score	Continuous (composite): `30 × exp(-hemorrhoid_meds_days_since / 30)`. Exponential decay; 0 if never prescribed.	Yes

4. Features in Final Model (1 of 26)

Feature	Derivation	FHIR
out_med_broad_abx_recency	Recency of broad-spectrum antibiotic prescription (days_since or binned variant)	Yes

---

Book 5.2: Inpatient Medications

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
Inpatient medication administration records (16 categories)	order_med_enh + mar_admin_info_enh	Yes	MedicationAdministration

Key difference from outpatient: Uses actual administration records (MAR `TAKEN_TIME`), not prescription orders. Filter: `ORDERING_MODE_C = 2` (inpatient), 16 accepted MAR action types (GIVEN, PUSH, BOLUS, NEW BAG, etc.), `TAKEN_TIME >= '2021-07-01'`. 24-month lookback.

2. Features Engineered (48 before reduction)

Same 16 categories × 3 feature types as outpatient, but from confirmed inpatient administration. Column names carry `inp_` prefix from SQL, then `inp_med_` prefix added during save (creating `inp_med_inp_` double prefix).

3. Features Sent to Book 9 (20 after book-level reduction)

All carry `inp_med_inp_` prefix. Selection rules: hemorrhoid = flag + days_since; GI bleeding = flag + days_since; iron/laxative/opioid = flag; broad ABX/antidiarrheal/PPI = flag; others = best MI.

Feature	Derivation	FHIR
inp_med_inp_iron_use_flag	Binary 1/0: any inpatient iron administration in 24mo	Yes
inp_med_inp_laxative_use_flag	Binary 1/0: any inpatient laxative administration in 24mo	Yes
inp_med_inp_antidiarrheal_use_flag	Binary 1/0: any inpatient antidiarrheal administration in 24mo	Yes
inp_med_inp_hemorrhoid_meds_flag	Binary 1/0: any inpatient hemorrhoid med administration in 24mo	Yes
inp_med_inp_hemorrhoid_meds_days_since	Integer/NULL: days from most recent inpatient hemorrhoid med to snapshot	Yes
inp_med_inp_gi_bleed_meds_flag	Binary 1/0: any inpatient GI bleeding med (tranexamic acid, vasopressin, octreotide) in 24mo	Yes

Feature	Derivation	FHIR
inp_med_inp_gi_bleed_meds_days_since	Integer/NULL: days from most recent inpatient GI bleeding med to snapshot	Yes
inp_med_inp_opioid_use_flag	Binary 1/0: any inpatient opioid administration in 24mo	Yes
inp_med_inp_ppi_use_flag	Binary 1/0: any inpatient PPI administration in 24mo	Yes
inp_med_inp_broad_abx_flag	Binary 1/0: any inpatient broad-spectrum antibiotic in 24mo	Yes
inp_med_inp_{remaining 5}	Best MI feature for antispasmodic, IBD, NSAID, statin, metformin (data-dependent)	Yes
inp_med_inp_acute_gi_bleeding	Binary 1/0 (composite): BOTH iron AND PPI administered inpatient (IV iron + IV PPI pattern)	Yes
inp_med_inp_obstruction_pattern	Binary 1/0 (composite): BOTH laxative AND opioid inpatient (obstruction/ileus)	Yes
inp_med_inp_severe_infection	Binary 1/0 (composite): BOTH broad ABX AND opioid inpatient (sepsis pattern)	Yes
inp_med_inp_any_hospitalization	Binary 1/0 (composite): ANY of iron/PPI/laxative/opioid/broad ABX administered inpatient	Yes
inp_med_inp_gi_hospitalization	Binary 1/0 (composite): ANY of laxative/antidiarrheal/GI bleeding meds administered inpatient	Yes

4. Features in Final Model

None. All 20 inpatient medication features were eliminated during Book 9 iterative SHAP winnowing.

Book 6: Visit History

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
Outpatient encounters (dates, status)	pat_enc_enh	Yes	Encounter
Appointment status (completed/no-show)	pat_enc_enh.APPT_STATUS_C	Yes	Encounter.status
Provider specialty	clarity_ser_enh.SPECIALTY_NAME	Yes	PractitionerRole.specialty
PCP visit identification	pe.VISIT_PROV_ID = pe.PCP_PROV_ID	Yes	Encounter.participant
ED encounters	pat_enc_hsp_har_enh (ACCT_CLASS='Emergency' or ED_EPISODE_ID IS NOT NULL)	Yes	Encounter (class=emergency)
Inpatient admissions & LOS	pat_enc_hsp_har_enh (ACCT_CLASS='Inpatient')	Yes	Encounter (class=inpatient)
GI symptom diagnoses on encounters	pat_enc_dx_enh / hsp_acct_dx_list_enh	Yes	Condition

GI Symptom ICD-10 codes on encounters: `^(K92|K59|R19|R50|D50|K62)` (GI bleeding, bowel changes, abdominal pain, fever, anemia, anal/rectal).

GI Specialty: `SPECIALTY_NAME IN ('GASTROENTEROLOGY', 'COLON AND RECTAL SURGERY')`.

No-shows: `APPT_STATUS_C = 4` only (not status 3, which is cancellation).

2. Features Engineered (41 before reduction)

Feature Category	Count	Description
ED counts (90d/12mo/24mo)	4	ED visits total and GI-symptom ED visits
Inpatient counts	4	Admissions, GI-symptom admissions, total inpatient days
Outpatient counts	5	Total visits, GI specialty, PCP, GI-symptom outpatient
No-shows	1	Appointment no-shows in 12mo
Recency	3	Days since last ED, inpatient, GI specialty visit
Binary flags	7	Frequent ED user, recent hospitalization, etc.
Composite scores	4	Healthcare intensity, primary care continuity, etc.
Combined counts	1	Total GI symptom visits across all settings

3. Features Sent to Book 9 (24 after book-level reduction)

19 surviving pre-reduction features + 5 new composites. All carry `visit_` prefix.

Feature	Derivation	FHIR
visit_ed_last_12_months	Count: ED encounters in 12mo. ED = ACCT_CLASS='Emergency' or ED_EPISODE_ID IS NOT NULL.	Yes
visit_ed_last_24_months	Count: ED encounters in 24mo.	Yes
visit_gi_ed_last_12_months	Count: ED encounters in 12mo with GI symptom diagnosis attached (K92/K59/R19/R50/D50/K62).	Yes
visit_inp_last_12_months	Count: inpatient encounters in 12mo.	Yes
visit_inp_last_24_months	Count: inpatient encounters in 24mo.	Yes
visit_outpatient_visits_12mo	Count: completed/arrived outpatient visits in 12mo.	Yes
visit_gi_visits_12mo	Count: completed outpatient visits to GI specialty (gastroenterology, colon/rectal surgery) in 12mo.	Yes
visit_pcp_visits_12mo	Count: completed outpatient visits where VISIT_PROV_ID = PCP_PROV_ID (patient saw own PCP) in 12mo.	Yes
visit_gi_symptom_op_visits_12mo	Count: completed outpatient visits with GI symptom ICD-10 code in 12mo.	Yes
visit_no_shows_12mo	Count: appointments with APPT_STATUS_C = 4 (no-show only) in 12mo.	Yes
visit_total_gi_symptom_visits_12mo	Sum: GI_ED + GI_INP + GI_SYMPTOM_OP visits in 12mo (all settings combined).	Yes
visit_days_since_last_gi	Integer/NULL: days since most recent completed GI specialty outpatient visit in 24mo. NULL if no visit (~97.9% missing).	Yes
visit_frequent_ed_user_flag	Binary 1/0: 1 if ED_LAST_12_MONTHS >= 3.	Yes
visit_high_inpatient_days_flag	Binary 1/0: 1 if TOTAL_INPATIENT_DAYS_12MO >= 10.	Yes

Feature	Derivation	FHIR
visit_engaged_primary_care_flag	Binary 1/0: 1 if PCP_VISITS_12MO >= 2.	Yes
visit_recent_ed_use_flag	Binary 1/0: 1 if any ED visit in last 90 days.	Yes
visit_recent_hospitalization_flag	Binary 1/0: 1 if days_since_last_inpatient <= 180. NULL treated as 9999.	Yes
visit_healthcare_intensity_score	Ordinal 0-4: (1 if ED>=3) + (1 if INP>=2) + (1 if inpatient_days>=10) + (1 if GI_visits>=2).	Yes
visit_primary_care_continuity_ratio	Ratio: PCP_VISITS_12MO / OUTPATIENT_VISITS_12MO. NULL if no outpatient visits.	Yes
visit_gi_symptoms_no_specialist	Binary 1/0 (composite): 1 if total_gi_symptom_visits > 0 AND gi_visits = 0. GI symptoms present but no specialist seen.	Yes
visit_frequent_ed_no_pcp	Binary 1/0 (composite): 1 if frequent_ed_user AND pcp_visits = 0. Frequent ED user with no PCP engagement.	Yes
visit_acute_care_reliance	Ratio (composite): (ED + inpatient) / outpatient when outpatient > 0; else raw ED + inpatient count. Higher = more acute care.	Yes
visit_complexity_category	Ordinal 0-3 (composite): 3=High (intensity>=3), 2=Moderate (intensity>=1), 1=Low (any visits), 0=None.	Yes
visit_recent_acute_care	Binary 1/0 (composite): 1 if ED in last 90 days OR hospitalization in last 180 days.	Yes

4. Features in Final Model (4 of 26)

Feature	Derivation	FHIR
visit_outpatient_visits_12mo	Count of completed outpatient visits in 12 months	Yes
visit_recency_last_gi	Days since last GI specialty visit (renamed from visit_days_since_last_gi)	Yes
visit_gi_symptoms_no_specialist	GI symptoms present but no specialist seen (care gap)	Yes
visit_recent_acute_care	Any recent ED (90d) or hospitalization (180d)	Yes

---

Book 7: Procedures

1. Raw Data Gathered

Raw Data Element	Source	FHIR Available	FHIR Resource
CT abdomen/pelvis (18 internal codes)	order_proc_enh (RESULT_TIME, ORDER_STATUS_C=5)	Yes	Procedure / ImagingStudy
MRI abdomen/pelvis (9 internal codes)	order_proc_enh	Yes	Procedure / ImagingStudy
Upper GI endoscopy/EGD (4 codes)	order_proc_enh	Yes	Procedure

Raw Data Element	Source	FHIR Available	FHIR Resource
Blood transfusion (4 codes)	order_proc_enh	Yes	Procedure
Anoscopy (1 code: PRO105)	order_proc_enh	Yes	Procedure
Hemorrhoid procedures (2 codes)	order_proc_enh	Yes	Procedure
Iron infusions (IV iron formulations)	mar_admin_info_enh (MAR_ACTION_C=1, medication name matching)	Yes	MedicationAdministration

Note: Colonoscopy deliberately excluded (screened patients already removed from cohort). Procedure date = `RESULT_TIME` (when performed), not `ORDERING_DATE`.

2. Features Engineered (28 before reduction)

Feature Category	Count	Description
Procedure counts (12mo)	7	Per-category counts in 12-month window
Procedure counts (24mo)	7	Per-category counts in 24-month window
Recency	6	Days since last procedure per category
Binary flags	7	High imaging intensity, transfusion history, etc.
Composite scores	1	procedure_intensity_count (0-4)

3. Features Sent to Book 9 (17 after book-level reduction)

14 surviving + 3 new composites. All carry `proc_` prefix.

Feature	Derivation	FHIR
proc_ct_abd_pelvis_count_12mo	Count: completed CT abdomen/pelvis orders (18 internal codes) in 12mo. COALESCE to 0.	Yes
proc_mri_abd_pelvis_count_12mo	Count: completed MRI abdomen/pelvis orders (9 internal codes) in 12mo. COALESCE to 0.	Yes
proc_upper_gi_count_12mo	Count: completed upper GI endoscopy/EGD orders (4 codes) in 12mo. COALESCE to 0.	Yes
proc_blood_transfusion_count_12mo	Count: completed blood transfusion orders (4 codes) in 12mo. COALESCE to 0.	Yes
proc_total_imaging_count_12mo	Count: CT + MRI abdomen/pelvis combined in 12mo.	Yes
proc_iron_infusions_12mo	Count: IV iron infusion administrations (MAR, MAR_ACTION_C=1, iron medication name patterns) in 12mo.	Yes
proc_iron_infusion_flag	Binary 1/0: 1 if any iron infusion in 24mo.	Yes
proc_high_imaging_intensity_flag	Binary 1/0: 1 if total_imaging_count_12mo >= 2.	Yes
proc_transfusion_history_flag	Binary 1/0: 1 if any blood transfusion in 24mo.	Yes
proc_anal_pathology_flag	Binary 1/0: 1 if any anal/hemorrhoid procedure in 24mo.	Yes
proc_comprehensive_gi_workup_flag	Binary 1/0: 1 if upper GI endoscopy AND abdominal imaging both in 12mo.	Yes
proc_severe_anemia_treatment_flag	Binary 1/0: 1 if any transfusion OR iron infusion in 24mo.	Yes
proc_recent_diagnostic_activity_flag	Binary 1/0: 1 if any CT or MRI in last 180 days.	Yes

Feature	Derivation	FHIR
proc_procedure_intensity_count	Ordinal 0-4: (1 if imaging>=2) + (1 if upper GI>=1) + (1 if transfusion>=1) + (1 if iron infusion>=1). Distinct procedure types active in 12mo.	Yes
proc_anemia_treatment_intensity	Ordinal 0-3 (composite): `LEAST(3, transfusion_count + iron_infusion_count)` in 12mo.	Yes
proc_diagnostic_cascade	Binary 1/0 (composite): 1 if imaging >= 2 AND upper GI >= 1 in 12mo (unresolved diagnostic workup).	Yes
proc_acute_bleeding_pattern	Binary 1/0 (composite): 1 if (transfusion in last 90 days) OR (2+ transfusions in 12mo). Active/recurrent bleeding.	Yes

4. Features in Final Model

None. All 17 procedure features were eliminated during Book 9 iterative SHAP winnowing.

Pipeline Summary

Feature Counts at Each Stage

Book	Domain	Raw Data	Engineered	After Book Reduction	Sent to Book 9	In Final Model
0	Demographics	7	13	–	11	3
1	Vitals	7	~50	24	24	8
2	ICD-10 Diagnoses	~50 code families	116	26	26	3
3	Social Factors	–	–	–	–	–
4	Labs	14 components	~80	25	25	7
5.1	Outpatient Meds	16 med categories	48	19	19	1
5.2	Inpatient Meds	16 med categories	48	20	20	0
6	Visit History	6 encounter types	41	24	24	4
7	Procedures	7 procedure types	28	17	17	0
Total			**~424**	**~166**	**~166**	**26**

Final 26 Features by Domain

Domain	Count	Features
Demographics (Book 0)	3	AGE_GROUP, RACE_CAUCASIAN, months_since_cohort_entry

Domain	Count	Features
Vitals (Book 1)	8	vit_WEIGHT_OZ, vit_PULSE_PRESSURE, vit_WEIGHT_CHANGE_PCT_6M, vit_MAX_WEIGHT_LOSS_PCT_60D, vit_WEIGHT_TRAJECTORY_SLOPE, vit_RECENCY_WEIGHT, vit_SBP_VARIABILITY_6M, vit_CACHEXIA_RISK_SCORE
ICD-10 (Book 2)	3	icd_BLEED_CNT_12MO, icd_MALIGNANCY_FLAG_EVER, icd_SYMPTOM_BURDEN_12MO
Labs (Book 4)	7	lab_ALBUMIN_VALUE, lab_HEMOGLOBIN_ACCELERATING_DECLINE, lab_IRON_SATURATION_PCT, lab_PLATELETS_ACCELERATING_RISE, lab_PLATELETS_VALUE, lab_THROMBOCYTOSIS_FLAG, lab_comprehensive_iron_deficiency
Outpatient Meds (Book 5.1)	1	out_med_broad_abx_recency
Visit History (Book 6)	4	visit_outpatient_visits_12mo, visit_recency_last_gi, visit_gi_symptoms_no_specialist, visit_recent_acute_care

FHIR Availability Summary

Stage	Total Features	FHIR-Derivable	Partial	Not Available
Raw data gathered	~80 elements	~75	~5 (PCP, family hx, race)	0
Sent to Book 9	~166	~159	~7 (PCP, family hx, race features)	0
Final 26 features	26	25	1 (RACE_CAUCASIAN)	0

Key finding: 25 of the 26 final model features can be fully derived from standard FHIR resources. The sole Partial feature is `RACE_CAUCASIAN`, which depends on US Core race extension implementation. All medication features except one (broad-spectrum antibiotic recency) were eliminated. All procedure features were eliminated. The model relies primarily on vitals, labs, diagnoses, and visit patterns.