Study Protocol Overview and Objectives

Official Title: Retrospective single-center cohort study of early-onset colorectal neoplasia: clinical and pathological differences between EO-CRC and HGIN.

NCT number: Not yet assigned

Document date: 2026-01-22

Sponsor / Responsible party: Shenzhen Hospital, Southern Medical University

Contact:

Name: Jian Lin; Email: szyyec@163.com

Study Title Retrospective single-center cohort study of early-onset colorectal neoplasia: clinical and pathological differences between EO-CRC and HGIN.

Primary Objective Identify clinical, endoscopic, laboratory, and pathological factors associated with **advanced tumor stage** among patients aged ≤50 years with pathologically confirmed colorectal cancer or high-grade intraepithelial neoplasia.

Secondary Objectives

- Describe demographic, clinical, endoscopic, and molecular characteristics of EO-CRC and HGIN.
- Compare tumor location, histologic grade, lymphovascular invasion, and molecular markers (if available) between EO-CRC and HGIN.
- Evaluate treatment patterns and short-term outcomes (30-day complications, recurrence within available follow-up).

Study Design

Design Type Retrospective observational cohort (single cohort with analytic subgrouping).

Study Population Patients aged ≤50 years with pathology-confirmed colorectal cancer (any T stage) or high-grade intraepithelial neoplasia diagnosed at Shenzhen Hospital, Southern Medical University between 2016-01-01 and 2025-09-30 who meet inclusion/exclusion criteria.

Data Sources De-identified electronic medical records, endoscopy reports, imaging archives, and pathology reports.

Cohort Definition and Grouping Single retrospective cohort. Analyses will compare subgroups by **diagnosis** (EO-CRC vs HGIN) and by **stage** (early: HGIN/Tis/T1; late: T2 and above).

Outcomes

Primary Outcome Measure Tumor Stage at Diagnosis — proportion of subjects with early stage (HGIN/Tis/T1) versus late stage (T2–T4). **Time Frame**: Baseline (date of pathology report).

Key Secondary Outcome Measures

- **Tumor Location Distribution** — right colon, left colon, rectum at baseline.
- **Pathologic High-Risk Features** — presence of poor differentiation, lymphovascular invasion, perineural invasion at baseline.

- **Molecular Marker Frequency** — MSI status, KRAS/NRAS/BRAF mutations if available.
- **Treatment and Early Outcomes** — type of initial treatment (endoscopic resection, surgery, chemoradiation) and 30-day postoperative complications. **Time Frame**: Baseline and available follow-up up to last recorded visit.

Sample Size and Power

Planned Sample All eligible records from 2016-01-01 to 2025-09-30 will be included. Expected sample size will be reported after case ascertainment; no prospective enrollment.

Power Considerations Because the study is retrospective and includes all available cases, formal sample size calculation is not required for enrollment. For hypothesis testing, detectable effect sizes will be calculated post-hoc given the observed sample. As an example, with $n = 400$and a 1:1 split between early and late stage, the study would have >80% power to detect an absolute difference of approximately 10 percentage points in a binary predictor with baseline prevalence 30% at $\alpha = 0.05$. Exact detectable differences will be reported with final sample counts.

Statistical Analysis Plan

General Principles

- **Analysis population**: all eligible subjects with sufficient data for the outcome (analysis will note denominators for each variable).
- **Software**: analyses will be performed using standard statistical software (e.g., R or SAS).
- **Significance**: two-sided tests with $\alpha = 0.05$. Report 95% confidence intervals.

Descriptive Analyses

- Summarize continuous variables with **mean ± SD** or **median (IQR)** depending on distribution.
- Summarize categorical variables with **counts and percentages**.
- Compare EO-CRC vs HGIN and early vs late stage using **t test or Wilcoxon rank-sum** for continuous variables and **chi-square or Fisher's exact test** for categorical variables.

Primary Analysis

- **Univariable analysis**: evaluate association between each candidate predictor (demographics, symptoms, tumor location, laboratory markers, endoscopic features, pathology features) and late stage (T2+) using logistic regression.
- **Multivariable analysis**: build a logistic regression model for late stage (dependent variable) including clinically relevant covariates and those with $p < 0.10$in univariable screening. Use **stepwise selection** or **penalized regression (e.g., LASSO)** as sensitivity. Report adjusted odds ratios (aOR) with 95% CIs.
- **Model diagnostics**: assess multicollinearity, calibration (Hosmer-Lemeshow), discrimination (AUC), and internal validation via bootstrap or cross-validation.

Secondary Analyses

- Compare distributions of tumor location and pathologic high-risk features between EO-CRC and HGIN using logistic or multinomial regression as appropriate.
- Time-to-event outcomes (e.g., recurrence) will be analyzed with **Kaplan-Meier** estimates and **Cox proportional hazards** models if follow-up data permit; proportional hazards assumption will be tested.

Handling of Missing Data

- Report extent and patterns of missingness for each variable.
- If missingness is <5% for a variable, perform complete-case analysis for that variable.
- For variables with substantial missingness, apply **multiple imputation by chained equations (MICE)** under missing at random assumption; perform sensitivity analyses comparing imputed and complete-case results.

Subgroup and Sensitivity Analyses

- Predefined subgroups: age strata (≤40 vs 41–50), sex, tumor location (right vs left vs rectum).
- Sensitivity analyses: exclude patients with prior malignancy; restrict to cases with molecular marker data; alternative stage dichotomization.

Multiplicity Secondary and exploratory analyses will be interpreted cautiously; no formal multiplicity adjustment for exploratory comparisons, but key secondary

outcomes may be adjusted using the Benjamini-Hochberg procedure where appropriate.

Data Management Ethical Considerations and Timeline

Data Management

- **De-identification**: all records will be coded; direct identifiers removed prior to analysis.
- **Quality control**: double data extraction for a random sample; logic checks and range checks; audit trail for edits.
- **Storage**: secure institutional servers with access limited to authorized study personnel.

Regulatory and Ethics

- Study will proceed under institutional review board approval or waiver of consent as required by local regulations. Data use will comply with applicable privacy laws and institutional policies.

Timeline

- **Case ascertainment and extraction**: 1–3 months after IRB approval.
- **Data cleaning and coding**: 1–2 months.
- **Statistical analysis**: 1–2 months.
- **Manuscript preparation and registry updates**: subsequent 1–2 months.