Supplementary Online Content

Hou J, Zhao R, Cai T, et al. Temporal trends in clinical evidence of 5-year survival within electronic health records among patients with early-stage colon cancer managed with laparoscopy-assisted colectomy vs open colectomy. *JAMA Netw Open*. 2022;5(6):e2218371. doi:10.1001/jamanetworkopen.2022.18371

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods 1. Details of the Data Curation

The construction of study cohorts is illustrated by the first row in the diagram **eFigure 1**, whose results were displayed in **Figure 1**. A total of 65,968 patients from the Mass General Brigham (MGB) electronic health records (EHR) who had ⩾1 International Classification of Disease (ICD) codes of colorectal cancer were included in colorectal cancer data mart (**eTable 2**), representing patients who may have colorectal cancer. The multimodal automated phenotyping (MAP) phenotyping algorithm (main text citation 11) classified n=28,859 subjects with colorectal cancer with a positive predictive value (PPV) of 0.90. Of patients identified with colorectal cancer, 3,655 underwent colectomy, 2,070 laparoscopic-assisted colectomies (LAC) and 1,585 open colectomies (OC), between 2006 and 2018. After excluding participants who did not meet the key eligibility criteria, we obtained a cohort of 943 patients (518 LAC and 425 OC) of subjects who would have been eligible for the COST Study Group Trial to form the EHR emulation cohort.

1. Cohort construction with MAP
   For each patient, we extracted the numbers of colorectal cancer ICD codes (mapped to PheWAS code 153) and mention of colorectal cancer in medical notes according to Concept Unique Identifier (CUI) C0009402 of the Unified Medical Language System (ULMS). We used the total number of days with any ICD code as the measure for hospital utilization. Feeding the ICD counts, CUI counts and hospital utilization into the MAP program, we obtain the MAP scores. An abstractor in the team manually reviewed the 171 randomly selected patients' medical history to annotate their colorectal cancer diseases status. The MAP scores achieve 0.945 area-under-curve (AUC) of receptor-operating-characteristics (ROC) over all 171 annotated patients, and 0.885 AUC of ROC over 67 patients with at least one colorectal cancer ICD codes (see **eFigure** 2). We constructed the colorectal cancer cohort with 28,859 patients of MAP > .371 (0.95 specificity, 0.70 sensitivity, 0.90 positive predictive value).

2. Outcome extraction: overall survival
   We obtained the mortality information by linking MGB healthcare data to death registry of Massachusetts. The overall survival was defined as the number of year(s) from the date of the identified colectomy to the death date. Patients without death record were marked as censored at the last date of their EHR records.

3. Feature extraction: eligibility criteria
   We extracted the EHR features that can be mapped to the target RCT eligibility criteria and features according to **eTable** 2. For prior cancers, adhesions, Crohn's disease, familial polyposis, and chronic ulcerative colitis, we searched for ICD codes between 5 years prior to the colectomy date and the colectomy date. For colon obstruction and perforation, we searched for ICD codes between 30 days prior to the colectomy date and the colectomy date. For cancer stage, we extracted from natural language processing (NLP) of the medical notes between 1 years prior to the colectomy date and the colectomy date.
   We extracted cancer clinical stage using the NLP Interpreter for Cancer Extraction (NICE) tool (main text citation 12). In the MGB healthcare EHR, clinical stages (e.g., "stage 1" or "stage IV") or more detailed tumor-node-metastasis (TMN) stages (e.g., T1N2M0) are commonly documented in medical notes for cancer patients. Among the colorectal cancer patients identified by MAP, there are 65.2% with at least one mention of either clinical or TMN stages. We validated the extraction accuracy by 140 notes from the colorectal cancer data mart with annotated cancer stage mentions (79 with cancer stage information, 61 without cancer stage information). NICE achieved very high capture rate (no false negative extraction, 1 false positive extraction). The extractions of numerical stage for all 79 notes with cancer stage information were accurate.

4. Feature extraction: obesity from structured data

For LAC, obesity is identified as a determinant besides typical cancer risk factors. We extracted the body-mass-index (BMI) from EHR during the year prior to the colectomy date, and BMI>30 was defined as obese. We extracted BMI from both structured EHR data and natural language processing (NLP) of the medical notes. In medical notes, BMI is sometimes documented as numeric value indicated by the key term “body-mass-index” or its abbreviation “BMI”. Using the Extraction of EMR Numerical Data (EXTEND, main text citation 14) tool, we were able to identify the occurrence of the indicating term and extract the following numeric value. In **eFigure 3**, we illustrated the concordance between BMI from structured EHR data and BMI from EXTEND over patients with both types of BMI extractions. BMI from two sources generally agree with each other. For the EHR emulation cohort, identification of obesity (BMI > 30) was generally consistent between two sources, except for 5 patients near the overweight range (25<BMI<30).

5. Feature extraction: hospital utilization and broad range of other comorbidities
We used the total number of ICD codes as the measure of hospital utilization. We used the Phenome Wide Association Studies (PheWas) catalog groups (the Phecodes) to derive the broad range of other comorbidities from ICD codes. For each integer level Phecode, we defined its count as the occurrence of all ICD codes mapped to the Phecode. For each healthcare feature, we included the total count until the date of colectomy and the count within one year before colectomy. We also included a healthcare utilization rate variable defined as the overall healthcare utilization divided by years observed in EHR before colectomy.

6. Follow-up durations:
For the 943 patients in the EHR emulation cohort, we presented the distribution of their total EHR follow-up (from first EHR record to last EHR record) and pre-colectomy follow-up (from first EHR record to the date of identified colectomy) in **eFigure 4**. Most patients had more than one year of total EHR follow-up for MAP and more than one-year of pre-colectomy follow-up.

eMethods 2. Details of the Statistical Analysis

The statistical analysis is illustrated by the second row in the diagram **eFigure 1.** For t up to 5, we measured the treatment effect in terms of difference in mortality rate at t-year after the identified colectomy. We accounted for confounding from 1) clinically relevant variables: age, gender, cancer stage, tumor location, colon adhesion, procedure subtypes, obesity, 2) a broad range of co-morbidities according to the PheWAS catalogs 3) calendar year and its interaction with other variables. We adopted a doubly robust causal modeling strategy that combines 1) the propensity score approach 2) regression adjustment approach.

We adopted a doubly robust causal modeling strategy under which we trained 1) the outcome regression (OR) model of overall survival by a Cox model; 2) the propensity score (PS) model by logistic regression models. To account for temporal changes for ATE estimation, we allowed the covariate effects in both the OR and PS models to vary across the temporal periods but adopted a co-training strategy allowing the data to determine the degree of similarity between the three sets of models across time periods. We used group adaptive least absolute shrinkage and selection operator (LASSO) estimation (main text citations 19-21) for both the OR and PS training to incorporate feature selection and high dimensionality of the confounders. We assessed the effect of temporal trends on OR by testing whether the coefficients of the confounders on the mortality risk differ across temporal periods based on the Chi-square test. Similarly, we tested whether the odds ratios of the confounders on treatment assignment in the PS model differed by temporal periods based on another Chi-square test. To further illustrate the impact of temporal trends as a confounding factor, we additionally performed causal modeling using the entire EHR emulation cohort without adjusting for temporal periods as a benchmark analysis.

1. <u>Notation</u>

We denote the EHR data as $\{(X_i, \delta_i, A_i, Z_i, t_i), i = 1, \ldots, n\}$, where X_i is the observation time (from colectomy to death or loss-to-follow-up), $\delta_i$is the death indicator, $A_i$is the treatment indicator (1=LAC, 0=open colectomy), $Z_i$is the vector of variables adjusted in the analysis, $t_i = 1,2,3$ indicates the temporal periods (1=2006-2009, 2=2010-2013, 3=2014-2017). For Analysis I, we denote the RCT as $\{(X_i, \delta_i, A_i, Z_i), i = n + 1, \ldots, n + m\}$.

2. <u>OR model</u>

We train the Cox regression model using (X_i, δ_i)as the response. The covariates are A_i, Z_i and $A_i Z_i$ with temporal-period-dependent coefficients. Let the fitted relative risks be

$$\hat{r}_i = exp\left(\hat{\beta}_{1,t_i} A_i + \hat{\beta}_{2,t_i} Z_i + \hat{\beta}_{3,t_i} A_i Z_i\right)$$

and the estimated baseline cumulative hazard function as $\hat{\Lambda}_0(t)$. The OR predictions for OS for two arms are

$$\hat{S}_i(1, t) = exp\left(\hat{\beta}_{1,t_i} A_i + \hat{\beta}_{2,t_i} Z_i + \hat{\beta}_{3,t_i} A_i Z_i\right) \hat{\Lambda}_0(t), \hat{S}_i(0, t) = exp\left(\hat{\beta}_{2,t_i} Z_i\right) \hat{\Lambda}_0(t).$$

3. <u>PS model</u>

We train the logistic regression model using $A_i$as the response. The covariates are Z_i with temporal-period-dependent coefficients. The predicted PS is

$$\hat{\pi}_i = exp\left(\hat{\alpha}_{0,t_i} + \hat{\alpha}_{1,t_i} Z_i\right).$$

4. <u>Variable selection</u>

The variables in OR and PS models were selected by the following rules:

1) Marginal association with either overall survival or LAC assignment is significant at 0.05 level.

2) Binary or counting variables must occur for more than 20 patients. Variables with less than 20 occurrences were excluded from Z. Variables with less than 20 occurrences in LAC arm were set to have no interaction in the OR model $\beta_{3,t,j} = 0$. Variables with less than 20 occurrences in temporal period t for $t = 2,3$ were set to have no temporal shift in the temporal period $\beta_{2,t,j} = \beta_{2,t-1,j}, \alpha_{1,t,j} = \alpha_{1,t-1,j}$. Variables with less than 20 occurrences in LAC arm and temporal period t for $t = 2,3$ were set to have no temporal shift for interaction in OR model in the temporal period, $\beta_{3,t,j} = \beta_{3,t-1,j}$.
3) Remaining variables were selected by the co-trained adaptive group lasso with each variable group formed by temporal-period-dependent components for the same coefficients, $\{\beta_{i,t,j}: t = 1,2,3\}$ in OR model $\{\alpha_{i,t,j}: t = 1,2,3\}$ in PS model.

5. Matching key patient characteristics

We train the exponential tilt model using $Z_i$as the variables to be balanced. We estimate the model by solving the equations

$$\frac{1}{n_t}\sum_{i:t_i=t} exp\ (\gamma_{0,t} + \gamma_{1,t}Z_i)\ Z_i\ = \frac{1}{m}\sum_{i=n+1}^{n+m} Z_i\,, \frac{1}{n_t}\sum_{i:t_i=t} exp\ (\gamma_{0,t} + \gamma_{1,t}Z_i) = 1,$$

where $n_t = \#\{i: t_i = t\}, t = 1,2,3$. The estimated balancing weight is

$$\widehat{w}_i = exp\ (\hat{\gamma}_0 + \hat{\gamma}_i Z_i).$$

6. Censoring distribution

We estimate the censoring distribution by the Kaplan-Meier estimator stratified by treatment arm and year of colectomy. The predicted survival function is $\hat{G}_i(t)$.

7. Estimation of average OS and ATE: pooled EHR cohort

We estimate the average OS for two arms in the following ways:

$$\hat{\mu}_{AIPW}(1,t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i[\hat{S}_i(1,t) + \frac{A_i}{\hat{\pi}_i}\{1 - \frac{\delta_i}{\hat{G}_i(X_i)} I(X_i \le t) - \hat{S}_i(1,t)\}],$$

$$\hat{\mu}_{AIPW}(0,t) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_i[\hat{S}_i(0,t) + \frac{1-A_i}{1-\hat{\pi}_i}\{1 - \frac{\delta_i}{\hat{G}_i(X_i)} I(X_i \le t) - \hat{S}_i(0,t)\}].$$

The three estimators for average OS are used to construct three ATE estimators:

$$\hat{\Delta}_{AIPW}(t) = \hat{\mu}_{AIPW}(1,t) - \hat{\mu}_{AIPW}(0,t).$$

8. Estimation of average OS and ATE: time stratified analysis.

Suppose Ω be a subset of $\{1, \ldots, n\}$of size k, which applies to the 4-year strata. We estimate the average OS for two arms over the subpopulation:

$$\hat{\mu}_{AIPW,\Omega}(1,t) = \frac{1}{k}\sum_{i\in\Omega}[\hat{S}_i(1,t) + \frac{A_i}{\hat{\pi}_i}\{1 - \frac{\delta_i}{\hat{G}_i(X_i)} I(X_i \le t) - \hat{S}_i(1,t)\}],$$

$$\hat{\mu}_{AIPW,\Omega}(0,t) = \frac{1}{k}\sum_{i\in\Omega}[\hat{S}_i(0,t) + \frac{1-A_i}{1-\hat{\pi}_i}\{1 - \frac{\delta_i}{\hat{G}_i(X_i)} I(X_i \le t) - \hat{S}_i(0,t)\}].$$

The three estimators for average OS are used to construct three ATE estimators:

$$\hat{\Delta}_{AIPW,\Omega}(t) = \hat{\mu}_{AIPW,\Omega}(1,t) - \hat{\mu}_{AIPW,\Omega}(0,t).$$

9. Analysis procedure

Our analysis takes 3 steps: 1) Estimate the initial PS model and exclude patients with extreme PS (<0.1 or >0.9); 2) Estimate OR, density ratio and censoring models and re-estimate the PS model; 3) Use the estimated models to calculate the ATE estimators.

10. Benchmark analysis without temporal adjustments

For the OR model without temporal adjustment, the covariates are A_i, Z_i and A_iZ_i. For the PS model without temporal adjustment, the covariates are Z_i. The definition of OR and PS predictions is the same. The rest of analysis is the same as described in **analysis procedure** above.

11. Test for confounding from temporal trends

The confounding from temporal trends is identified by their simultaneous association with treatment assignment and overall survival outcome. The goodness-of-fit test for OR model is testing $\beta_{j,1} = \beta_{j,2} = \beta_{j,3}, j = 1,2,3$. The goodness-of-fit test for PS model is testing $\alpha_{j,1} = \alpha_{j,2} = \alpha_{j,3}, j = 1,2,3$.

eResults. Temporal Trends in Identified Confounding Factors and Comparison With Benchmark Analyses

1. Temporal trends in identified confounding factors

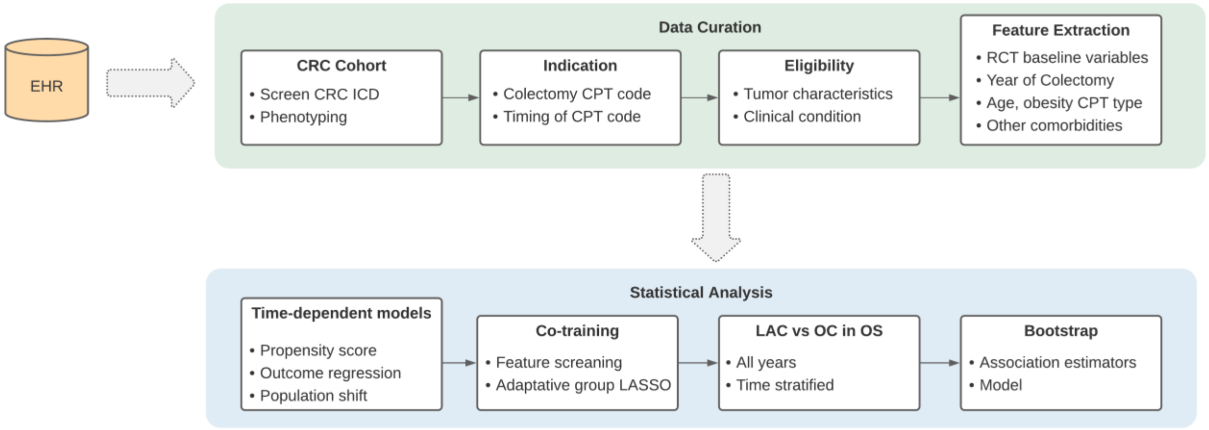
   The descriptions of temporal trends were based on **Figure 2**.

   Patients with a code for “recent other symptoms involving abdomen and pelvis” before colectomy were more likely to have had an OC and had poorer survival, and the prevalence of the code decreased over time. Advanced cancer stages (II/III) or no colon adhesion were associated with OC, but the association decreased over time. Patients with sigmoid colon cancer had better survival and this survival advantage increased over time. Colon adhesion was associated with poorer survival across all years, and its association with LAC for 2006-2009 and 2014-2017.

   Interestingly, the data-driven analysis identified temporal trends in variables seemingly not associated with colectomy. Patients with a code for “sciatica” before colectomy were more likely to have had a LAC and had better survival, and the prevalence of the code increased over time. Patients with a code for “superficial cellulitis and abscess” before colectomy had better survival if they underwent OC, and the prevalence of the code decreased over time.
2. Comparison with benchmark analyses

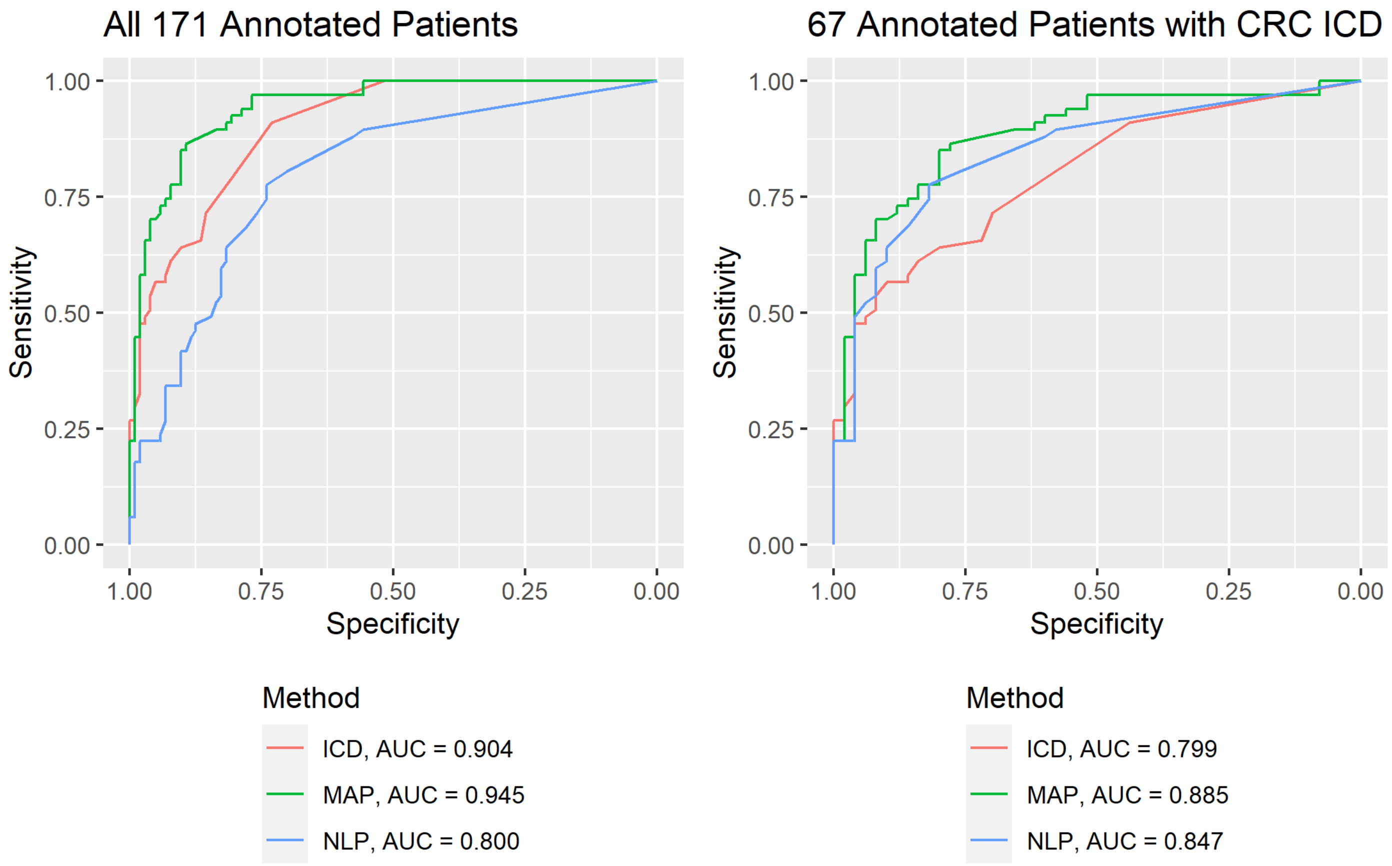
   In **eFigure 5**, we compared the predicted survival rates and point wise 95% confidence bands from crude analysis without adjusting for any confounders, benchmark analysis ignoring temporal trends, temporal effect adjusted analysis and the randomized controlled trial (RCT), the COST Study Group Trial. The benchmark analysis and the crude analysis both showed larger and significant treatment differences between LAC and OC compared to the temporal effect adjusted analysis and RCT. These differences suggested that temporal trends are an important confounding factor in real-world evidence analysis. Compared to the benchmark analysis, temporal effect adjusted analysis has a similar width in confidence band, which demonstrated that the increase of estimation variance from stratification with co-training was moderate.

eFigure 1. Study Schematics

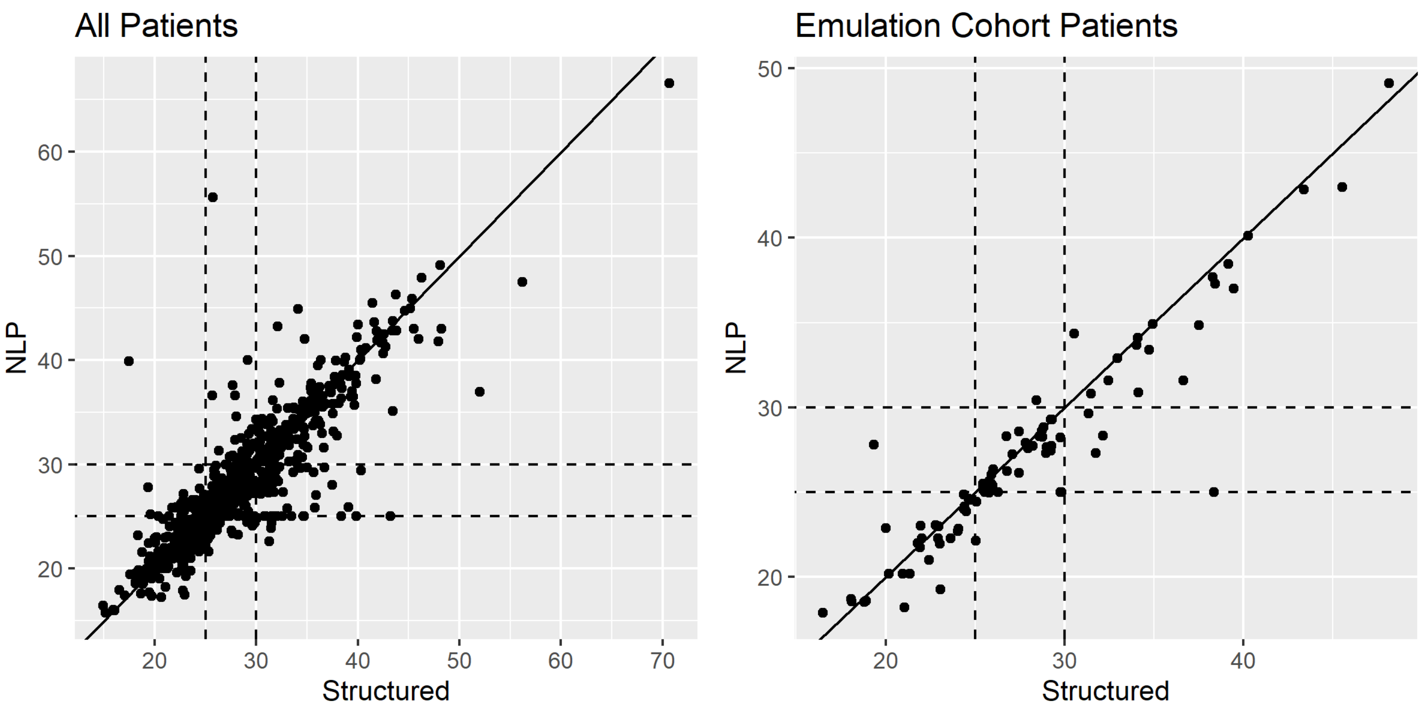


eFigure 2. Receptor Operating Characteristics (ROC) for MAP Prediction, Number of Colorectal Cancer Diagnosis Code and Number of Mentions of Colorectal Cancer in Medical Notes

The sensitivity and specificity are validated over 171 patients with annotated colorectal cancer diagnosis status from manual chart review of medical history.

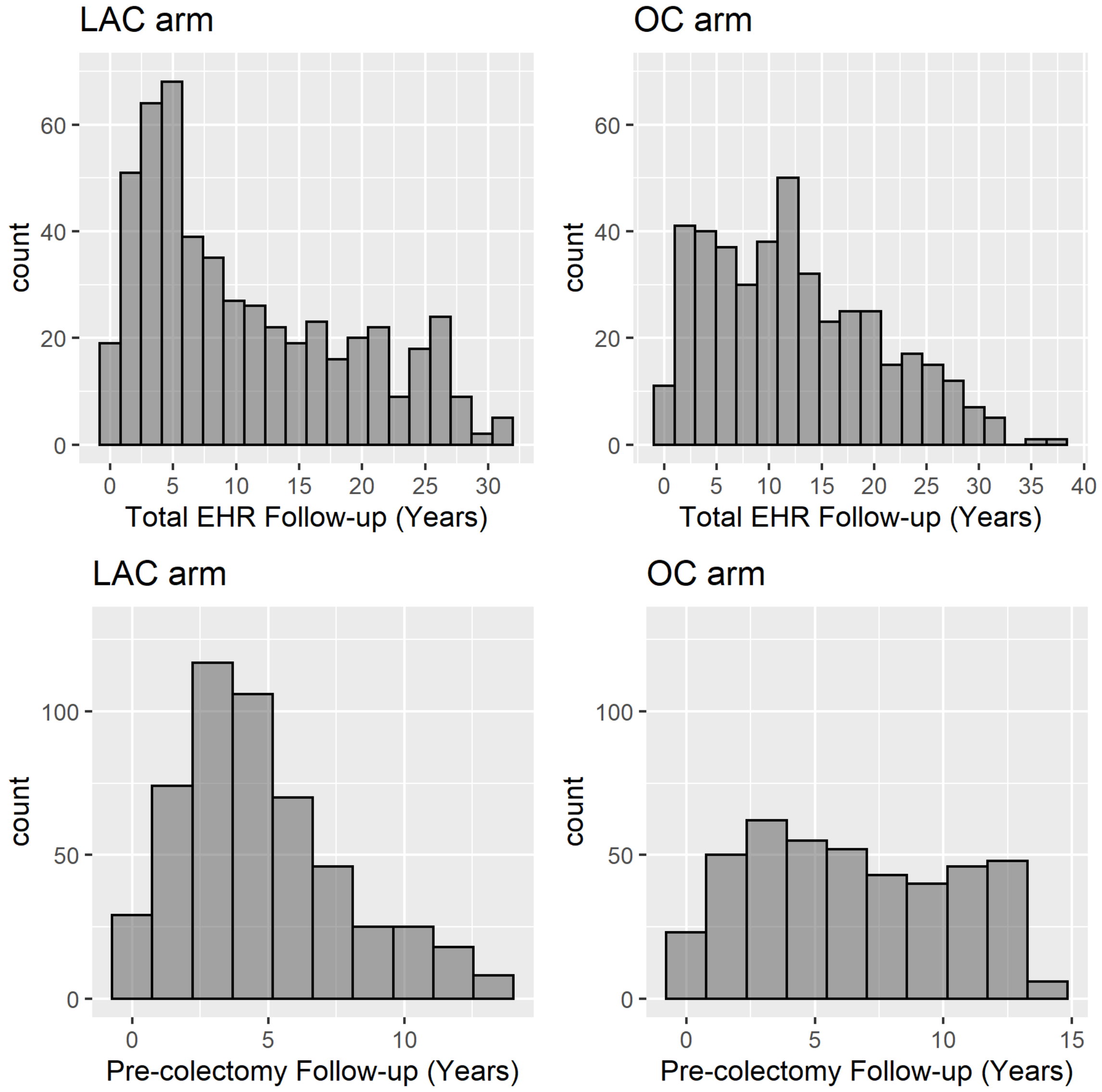


eFigure 3. Concordance Between NLP BMI Extraction and Structured BMI Over Patients With Both Extractions



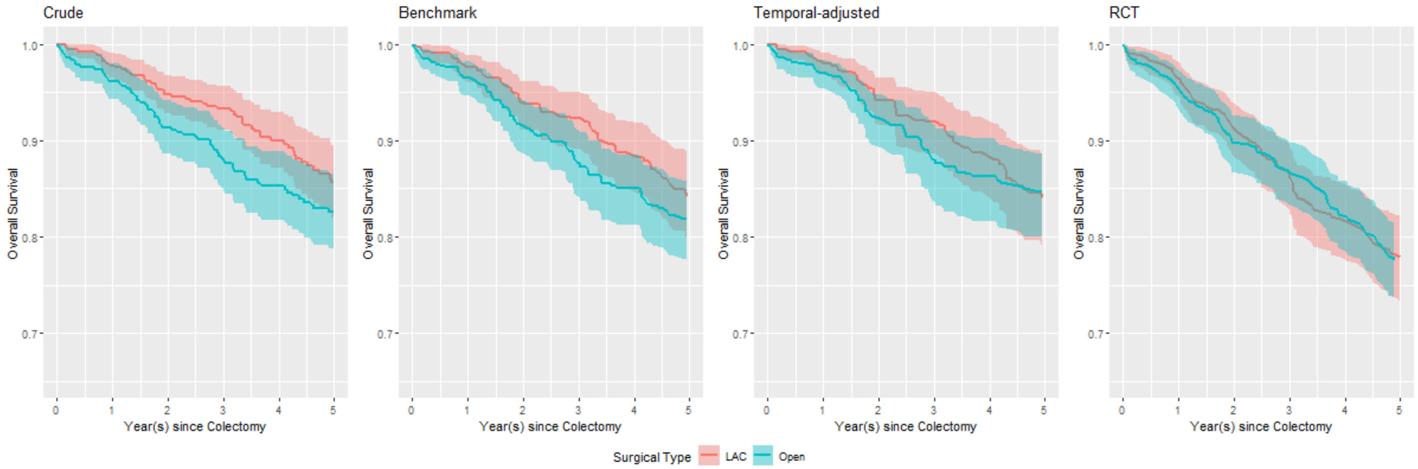
eFigure 4. Summaries on Follow-up Durations for 943 Patients in EHR Emulation Cohort

Most patients have sufficiently long follow-up in the MGB healthcare system and medical history before colectomy.



eFigure 5. Estimated Overall Survival With 95% Confidence Band From Crude Analysis, Benchmark Analysis, Temporal Effect Adjusted Analysis, and RCT

The benchmark analysis adjusted for other covariates except for the temporal effect. The lines depicted the survival curves for time-to-death, obtained from (1) Kaplan-Meier for crude analysis and RCT, (2) mean predicted overall survival probability based on doubly robust estimation for benchmark and temporal effect adjusted analysis.



At risks, all years

Yr	0	1	2	3	4	5
LAC	518	478	420	333	263	190
OC	425	393	355	319	269	231

At risks, 2006-2009

Yr	0	1	2	3	4	5
LAC	518	478	420	333	263	190
OC	425	393	355	319	269	231

At risks, 2010-2013

Yr	0	1	2	3	4	5
LAC	518	478	420	333	263	190
OC	425	393	355	319	269	231

At risks, 2014-2017

Yr	0	1	2	3	4	5
LAC	425	408	384	361	337	311
OC	414	395	368	353	327	296

eTable 1. Key Eligibility Criteria of the Target RCT and Their Correspondence in EHR Data

Type	**Description**	**EHR variable**
Treatment	Open colectomy	CPT: C44140 (general), C44145 (sigmoid), C44146 (sigmoid), C44160 (right)
Treatment	Laparoscopy-assisted colectomy	CPT: C44204 (general), C44205 (right), C44207 (sigmoid), C44208 (sigmoid)
Indication	Adenocarcinoma involving a single colon segment	ICD10: C18.x; ICD9: 153.x
Exclusion	Transverse colon cancer	ICD10: C18.4; ICD10: 153.1
Exclusion	Rectal cancer	ICD10: C20; ICD9: 154
Feature	Tumor in right segment of colon	ICD10: C18.0-3; ICD9: 153.0, 4-6
Feature	Tumor in left segment of colon	ICD10: C18.5-6; ICD9: 153.2,7
Feature	Tumor in sigmoid colon	ICD10: C18.7, C19; ICD9: 153.3, 154.0
Exclusion	Concurrent or previous malignant tumor within 5 years	PheWAS codes 145-230
	Exception: squamous cell carcinoma of skin	PheWAS code 172.22
	Exception: basal cell carcinoma of skin	PheWAS code 172.21
	Exception: cervical cancer	PheWAS code 180.3
Exclusion	Stage IV disease	NLP: NICE
Exclusion	Prohibitive adhesions	PheWAS code 568.1
Exclusion	Crohn’s disease	PheWAS code 555.1
Exclusion	Chronic ulcerative colitis	PheWAS code 555.2
Feature	Obesity	Height, weight, BMI and PheWAS code 278
Exclusion	Acutely obstructed colon	ICD10: K56.52,609,699; ICD9: 560.81,9
Exclusion	Acutely perforated colon	ICD10: K63.1; ICD9: 569.83

eTable 2. EHR Variables Forming the Filter for Colorectal Cancer Data Mart

EHR variable	**Description**
ICD9: 153.x	Malignant neoplasm of colon
ICD9: 154.x	Malignant neoplasm of rectum rectosigmoid junction and anus
ICD9: 159	Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum
ICD9: 197.5	Secondary malignant neoplasm of large intestine and rectum
ICD9: 209.1x	Malignant carcinoid tumors of the appendix, large intestine, and rectum
ICD9: 230.x	Carcinoma in situ of digestive organs
ICD10: C18.x	Malignant neoplasm of colon
ICD10: C19	Malignant neoplasm of rectosigmoid junction
ICD10: C20.x	Malignant neoplasm of rectum
ICD10: C78.5	Secondary malignant neoplasm of large intestine and rectum
ICD10:C7A.02x	Malignant carcinoid tumors of the appendix, large intestine, and rectum
ICD10: D01.x	Carcinoma in situ of other and unspecified digestive organs
	Other internal codes used by MGB equivalent to the ICD codes above