Supplementary Data

Dataset 1

Dataset 1 contained colon and rectal cancer cases from the SEER Program. The histopathologic codes were limited to 8000-8152, 8154-8231, 8243-8245, 8250-8576, 8940-8950, and 8980-8981. (The same restrictions for histopathologic codes were imposed for AJCC 7th TNM staging system.) Starting from 2010, SEER included *Derived AJCC-7 T, N, M* variables [1] that recorded the matching T, N, M levels for the 7th AJCC cancer staging systems [2]. Therefore, the year of diagnosis for dataset 1 was restricted from 2010 to 2012 (minimum follow-up of 3 years). The following records were collected for dataset 1: T, N, M, survival time, and SEER cause-specific death classification variable [3]. The survival time was measured in months. The cause-specific death classification variable was used to capture deaths caused by colon cancer. 6 levels were used for T: Tis, T1, T2, T3, T4a and T4b. T0 was not included in our dataset 1 since it was not used in the AJCC system. 4 levels were used for N: N0, N1, N2a and N2b. 3 levels were used for M: M0, M1a and M1b. Definitions for the levels of T, N, and M are provided in supplementary Table S1. We excluded patients with a missing or unknown value on any of the T, N, M, survival time, and SEER cause-specific death classification variable. The remaining data for our analysis had 71382 patients.

Due to the use of statistical techniques in our algorithm, we only used combinations (in terms of T, N, M) containing at least 100 patients, which excluded 27 "rare" combinations (567 cases). The final dataset 1 consisted of 45 combinations (70815 cases with a median follow up of 50 months calculated by using the reverse Kaplan-Meier method [4]).

Dataset 2

Cases of colon and rectal cancer diagnosed between 2004 and 2010 (minimum follow-up of 5 years) were collected from SEER to create dataset 2. The same histopathologic codes as in dataset 1 applied. The following records were collected for dataset 2: T, N, M, A, C, L, survival time, and SEER cause-specific death classification variable. The levels of the T, N, and M were recorded according to *Derived AJCC-6 T, N, M* variables [1] that were initially included in SEER in 2004. T had 5 levels: Tis, T1, T2, T3, and T4. N had 3 levels: N0, N1, and N2. M had 2 levels: M0 and M1. We discretized age into 2 levels: A1 (age of diagnosis < 80), and A2 (age of diagnosis ≥ 80). The cut-off 80 was selected on the basis of the discussion in [5, 6, 7]. We used 2 levels for C: C1 (positive/elevated) and C2 (negative/normal; within normal limits) [8]. With the most commonly used definition of left- and right-sided regions of the colon and rectum [9], we further separated the rectum from the left-sided colon for differentiation. Therefore, we considered 3 levels for tumor location: Lr (right-sided), Ll (left-sided), and Lb (bottom or rectum). Definitions for the levels of T, N, M, A, C, and L are provided in supplementary Table S1. We excluded patients with a missing or unknown value on any of the T, N, M, A, C, L, survival time, and SEER cause-specific death classification variable. The remaining data had 93796 patients.

Due to statistical techniques used in our algorithm, we only retained combinations (in terms of T, N, M, A, C and L) containing a minimum of 100 patients. This excluded 223 “rare” combinations (4826 cases). The final dataset 2 consisted of 137 combinations (88970 cases with a median follow up of 88 months calculated by using the reverse Kaplan-Meier method.

1. SEER Research Data Record Description. Available from URL: https://seer.cancer.gov/data/seerstat/nov2017/TextData.FileDescription.pdf
2. Edge S, Byrd D, Compton C et al. AJCC Cancer Staging Manual. 7th edition. Springer, New York, 2010.
3. SEER Cause-specific Death Classification. Available from URL: https://seer.cancer.gov/causespecific/
4. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996; 17(4): 343-6.
5. Barrier A, Ferro L, Houry S et al. Rectal cancer surgery in patients more than 80 years of age. Am. J. Surg 2003; 185(1): 54-7.
6. Al-Refaie WB, Parsons HM, Habermann EB et al. Operative outcomes beyond 30-day mortality: colorectal cancer surgery in oldest old. Ann Surg 2011; 253(5): 947-52.
7. Neuman HB, Weiss JM, Leverson G et al. Predictors of short-term postoperative survival after elective colectomy in colon cancer patients≥ 80 years of age. Ann Surg Oncol 2013; 20(5): 1427-35.
8. CS Site-Specific Factor 1. Available from URL: https://staging.seer.cancer.gov/cs/input/02.05.50/colon/ssf1/
9. Stintzing S, Tejpar S, Gibbs P et al. Understanding the role of primary tumour localisation in colorectal cancer treatment and outcomes. Eur J Cancer 2017; 84: 69-80.