1. Title Page

Title	TARGET-EU: Comparative effectiveness and safety studies using the target trial emulation and estimand frameworks: Clinical Benefit of Bevacizumab in Metastatic Colorectal Cancer
Research question & Objectives	The objective is to assess progression-free survival (PFS) in patients with mCRC when treated with CapOx-B compared with CapOx alone.
Protocol version	V1.0
Last update date	04 March 2026
Contributors	**Primary investigator contact information:** Julia E.M. van Dommelen (MSc); j.e.m.vandommelen@uu.nl **Contributor names:** Dr. Thijs J. Giezen Prof. Helga Gardarsdottir Prof. Toine C.G. Egberts Dr. Frederieke H. van der Baan Dr. Robert Jan Kwakman Prof. Ian Douglas Prof. Olaf H. Klungel Dr. Daniala L. Weir
Study registration	**Site:** https://catalogues.ema.europa.eu/node/4440/administrative-details **Identifier:** EUPAS1000000539
Sponsor	**Organisation:** EU PE&PV research network **Contact:** eupepv@uu.nl
Conflict of interest	None

Table of contents

2. Abstract

Background: Colorectal cancer (CRC) is the second most commonly diagnosed type of cancer in women and third most commonly diagnosed type in men, with an incidence of 73.5 per 100,000 inhabitants-year in 2022 in the European Union. [1] The mortality in that year was 32.3 per 100,000 inhabitants. In the Netherlands, 23% of the patients with colon cancer and 19% of the patients with rectal cancer presented with metastatic colorectal cancer (mCRC) at diagnosis [2]. Guidelines recommend fluoropyrimidine-based chemotherapy in combination with oxaliplatin or irinotecan for first-line initially unresectable mCRC [3,4]. The addition of bevacizumab, an anti-Vascular Endothelial Growth Factor (VEGF) antibody, improved prognosis compared to chemotherapy alone [5]. A randomised clinical trial (RCT) found that addition of bevacizumab had a median overall survival (OS) of 9.4 months compared to 8.0 months for placebo (HR, 0.83; 97.5% CI, 0.72 to 0.95), in combination with oxaliplatin-based chemotherapy [6].

Objectives: The acceptance of bevacizumab in mCRC treatment landscape varies largely throughout the Netherlands, presumably due to differences in policy and attitude [7]. This study will evaluate the comparative effectiveness of first-line addition of bevacizumab to the capecitabine-oxaliplatin (CapOx) regimen versus the CapOx regimen alone, in initially unresectable mCRC patients.

Methods: We will use the Netherlands Cancer Registry (NCR), a nationwide, population-based, Dutch cancer registry. [8] The selected data source is widely used for research, and the data holders provide thorough documentation of data contents, assumptions and limitations. Data is extracted from the individual patients' electronic health record by data managers from the NCR. The NCR contains longitudinal, date stamped information on patient enrolment (marked by a new cancer diagnosis), demographics, diagnoses, procedures, medication prescriptions, outcomes and covariates. The primary analysis uses a Cox proportional hazards model, with supplemental analyses using an accelerated failure time model to estimate restricted mean survival time (RMST) at 52 weeks and 104 weeks. Sensitivity analyses will be conducted to assess the impact of censoring assumptions, using inverse probability of censoring weighting (IPCW) and tipping point analysis, as well as outcome misclassification using probabilistic bias analysis.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
04 March 2026	V1.0			

4. Milestones

Table 1. Milestones

Milestone	Date
Study protocol for RWD study	8 August 2025
Preliminary results RWD study	April 2026
Final Study report	10 June 2026

5. Rationale and background

What is known about the condition: The third most common type of cancer worldwide is colorectal cancer (CRC) [3,4]. In 2022, the incidence of colorectal cancer in the European Union was 73.5 per 100,000 inhabitants [1]. In 15-30%, metastases were already present at the time of diagnosis. [4] Although the mortality rate has declined since 2012 due to screening and advanced treatments, the mortality rate for colorectal cancer was 32.3 per 100,000 inhabitants in 2022 [1]. There are various treatment options due to the heterogeneous patient population and specific tumour characteristics [3]. As a foundation for initially unresectable metastatic colorectal cancer (mCRC), guidelines recommend first-line fluoropyrimidine-based chemotherapy in combination with oxaliplatin (e.g., the CapOx regimen, consisting of capecitabine and oxaliplatin) [4,5].

What is known about the exposure of interest: Bevacizumab can be added to this regimen [9]. Bevacizumab is an anti-Vascular Endothelial Growth Factor (VEGF) antibody. By binding VEGF, it limits the vascularisation of the tumour and subsequently inhibits tumour growth [10]. A meta-analysis of 7 randomised controlled trials (RCTs) compared bevacizumab plus chemotherapy to chemotherapy alone and found that the hazard ratio (HR) for progression-free survival (PFS) was 0.71 (95% confidence interval (CI), 0.65 to 0.77) and the HR for overall survival (OS) was 0.85 (05% CI, 0.78 to 0.94) [11]. One of the RCTs found that addition of bevacizumab to oxaliplatin-based chemotherapy improves prognosis, reflected by a median PFS of 9.4 months for bevacizumab versus 8.0 months for placebo (HR, 0.83; 97.5% CI, 0.72 to 0.95). [6] CI, 0.76 to 1.03), when combined with an oxaliplatin-based chemotherapy. An observational study found that adding bevacizumab to first-line palliative chemotherapy improved the OS from 14 months (95% CI, 11 to 16) to 22 months (95% CI, 19 to 24). [7] The multivariable analysis showed a HR of 0.6 (95% CI, 0.45 to 0.73) for death.

Gaps in knowledge: In the Netherlands, the evidence for adding bevacizumab to standard chemotherapy regimen was re-evaluated in 2008 and appeared to be moderate [12]. There is uncertainty about the comparative effectiveness of bevacizumab + CapOx (CapOx-B) regimen vs CapOx regimen alone. Even the guideline from the European Society of Medical Oncology (ESMO) does not strongly advice on whether to add bevacizumab to first-line therapy [4].

What is the expected contribution of this study? A head-to-head comparison of first-line bevacizumab + CapOx regimen versus CapOx regimen alone in initially unresectable mCRC patients. This could provide additional evidence on the clinical benefit of including bevacizumab in first-line treatment of mCRC.

6. Research questions and objectives

The overall aim is to assess progression-free survival (PFS) in patients with mCRC when treated with CapOx-B compared with CapOx alone.

6.1 Primary Estimand 1

Research question targeted by the estimand: What is the HR of disease progression or death for CapOx-B vs CapOx alone in patients with mCRC, regardless of treatment discontinuation, partial discontinuation, treatment switch or local treatment?

Table 2. Core Emulation Table - Estimand 1

Attribute	Target Trial	Target Trial Emulation	Comment
Population:	Patients with mCRC	Patients with mCRC,	-
Treatment Conditions:	CapOx-B regimen vs CapOx regimen alone	CapOx-B regimen vs CapOx regimen alone, identified using administration data	Exposure defined based on first observed administration within first-line treatment regimen (new-user design)
Endpoint:	Time to first occurrence of disease progression or death.	Time to first occurrence of disease progression or death. Disease progression is marked by the end of an "episode"[1] in the NCR.	Slight deviation from target trial due to differences in PFS measurement. The impact of the difference in PFS measurements between the target trial and the target trial emulation is expected to be limited. PFS is generally longer in observational studies than in trials due to less frequent disease progression assessment. More heterogeneity can also be expected as PFS is not assessed centrally as in clinical trials. However, since disease progression is established in the same manner for both treatment groups, the different PFS measurements are not expected to bias the results. The measurement applied in the target trial

			emulation is validated and regularly used in observational studies using NCR data [13].
Summary Measure:	Hazard ratio	Hazard ratio	Target trial and emulation identical. Since PFS will be assessed in the same manner in both treatment groups, the HR should be valid.
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation (bevacizumab): treatment policy Partial discontinuation: treatment policy Treatment switch i.e. use of new anticancer therapy: treatment policy Local treatment: treatment policy	Same: intercurrent events handled according to prespecified strategies; implemented primarily via administration data Intercurrent events will be measured in NCR as follows: Treatment discontinuation (bevacizumab): stop date of individual drugs is registered Partial discontinuation: stop date of individual drugs is registered Treatment switch: stop date and start date of individual drugs is registered. If the treatment is switched due to toxicity, this is registered within the same "episode". Treatment switching due to disease progression is not an intercurrent event. Local treatment: - Surgery: the performance of surgery for metastases is registered, alongside the specific codes and dates - Radiotherapy: the performance of radiotherapy for metastases is	Target trial and emulation identical Treatment policy reflects real-world effectiveness There may be some limitations regarding the measurement of intercurrent events. Especially partial discontinuation of capecitabine may be difficult to assess because capecitabine is administered orally. (In)adherence might be a problem here. Treatment policy implies using all available data regardless of the occurrence of the intercurrent events. Hence, issues with the identification of the corresponding intercurrent events in real-world data will bear no consequences for the estimand as long as relevant post-intercurrent event data are available in the data source.

		registered, alongside the specific codes and dates - Ablation: the performance of ablation for (liver) metastases is registered, alongside the specific codes and dates	

[1]An "episode" is defined as the clinical trajectory of a patient up until disease progression. Disease progression marks the end of the "episode".

Rationale for handling of intercurrent events: The strategy to deal with all intercurrent events for estimand 1 is treatment policy. This strategy best reflects the evaluation of the intervention under study within clinical practice [14]. This estimand reflects the effectiveness of the intervention regardless of treatment discontinuation (of bevacizumab), partial discontinuation (of oxaliplatin and/ or capecitabine before the protocolised discontinuation), use of new anti-cancer therapy (treatment switch) or local treatment.

6.2 Supplementary Estimand 2

Research question targeted by the estimand: What is the HR of disease progression or death for CapOx-B vs CapOx alone in patients with mCRC, regardless of treatment discontinuation and in the hypothetical absence of partial discontinuation or treatment switch?

Table 3. Core Emulation Table - Estimand 2

Attribute	Target Trial	Target Trial Emulation	Comment
Population:	Patients with mCRC	Patients with mCRC	-
Treatment Conditions:	CapOx-B regimen vs CapOx regimen alone	CapOx-B regimen vs CapOx regimen alone, identified using administration data	Exposure defined based on first observed administration within first-line treatment regimen (new-user design)
Endpoint:	Time to first occurrence of disease progression or death	Time to first occurrence of disease progression or death. Disease progression is marked by the end of an "episode" in the NCR.	Slight deviation from target trial due to differences in PFS measurement. The impact of the difference in PFS measurements between the target trial and the target trial emulation is expected to be limited. PFS is generally longer in

			observational studies than in trials due to less frequent disease progression assessment. More heterogeneity can also be expected as PFS is not assessed centrally as in clinical trials. However, since disease progression is established in the same manner for both treatment groups, the different PFS measurements are not expected to bias the results. The measurement used in the target trial emulation is validated and regularly used in observational studies using NCR data [13].
Summary Measure:	Hazard ratio	Hazard ratio	Target trial and emulation identical Since PFS will be assessed in the same manner in both treatment groups, the HR should be valid.
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation (bevacizumab): treatment policy Partial discontinuation: hypothetical Treatment switch i.e. use of new anticancer therapy: hypothetical Local treatment: composite	Same strategies implemented primarily via administration data. Using censoring for the hypothetical strategy and creating a composite outcome that includes local treatment. The occurrence of local treatment leads to patients begin assumed to attain the final administrative censoring date. Intercurrent events will be measured in NCR as follows: Treatment discontinuation (bevacizumab): stop date of individual drugs is registered Partial discontinuation: stop date of individual drugs is registered	Target trial and emulation identical Treatment policy reflects real-world effectiveness; hypothetical strategy reflects a hypothetical scenario in which the intercurrent event would not occur. Composite outcome created including local treatment alongside disease progression and death. Tumour shrinkage would have needed to occur before local treatment can be applied, which is deemed a good health outcome. This will be implemented by assuming that the patient has not died or

		Treatment switch: stop date and start date of individual drugs is registered. If the treatment is switched due to toxicity, this is registered within the same "episode". Treatment switching due to disease progression is not an intercurrent event. Local treatment: - Surgery: the performance of surgery for metastases is registered, alongside the specific codes and dates - Radiotherapy: the performance of radiotherapy for metastases is registered, alongside the specific codes and dates - Ablation: the performance of ablation for (liver) metastases is registered, alongside the specific codes and dates	experienced disease progression by the end of the study. There may be some limitations regarding the measurement of intercurrent events. Especially partial discontinuation of capecitabine may be difficult to assess because capecitabine is administered orally. (In)adherence might be a problem here. No issues are expected for the identification of treatment switch or local treatment in the NCR, since those are registered with administration or procedure codes. Treatment policy implies using all available data regardless of the occurrence of the intercurrent events. Hence, issues with the identification of the corresponding intercurrent events in real-world data will bear no consequences for the estimand as long as relevant post-intercurrent event data are available in the data source.

Rationale for handling of intercurrent events: For estimand 2, multiple strategies are applied. The treatment policy strategy is applied when treatment discontinuation (of bevacizumab) occurs. The hypothetical strategy is applied when partial discontinuation (of oxaliplatin and/ or capecitabine before the protocolised discontinuation) or treatment switch occurs. A composite outcome is created that includes local treatment. The occurrence of local treatment then leads to patients being assumed to attain the final administrative censoring date. This estimand reflects the effects of the intervention on PFS, regardless of treatment discontinuation, in the hypothetical absence of partial discontinuation or treatment switch. Final administrative

censoring date. This estimand reflects the effects of the intervention on PFS, regardless of treatment discontinuation, in the hypothetical absence of partial discontinuation or treatment switch.

6.3 Supplementary Estimand 3

Research question targeted by the estimand: What is the difference in RMST to disease progression or death for CapOx-B vs CapOx alone in patients with mCRC, regardless of treatment discontinuation, partial discontinuation, treatment switch or local treatment?

Table 4. Core Emulation Table - Estimand 3

Attribute	Target Trial	Target Trial Emulation	Comment
Population:	Patients with mCRC	Patients with mCRC	-
Treatment Conditions:	CapOx-B regimen vs CapOx regimen alone	CapOx-B regimen vs CapOx regimen alone, identified using administration data.	Exposure defined based on first observed administration within first-line treatment regimen (new-user design).
Endpoint:	Time to first occurrence of disease progression or death	Time to first occurrence of disease progression or death. Disease progression is marked by the end of an "episode" in the NCR.	Slight deviation from target trial due to differences in PFS measurement. The impact of the difference in PFS measurements between the target trial and the target trial emulation is expected to be limited. PFS is generally longer in observational studies than in trials due to less frequent disease progression assessment. More heterogeneity can also be expected as PFS is not assessed centrally as in clinical trials. However, since disease progression is established in the same manner for both treatment groups, the different PFS measurements are not expected to bias the results. The

			measurement used in the target trial emulation is validated and regularly used in observational studies using NCR data [13].
Summary Measure:	Difference in restricted mean survival time	Difference in restricted mean survival time.	Target trial and emulation identical.
Intercurrent events and strategies to handle them	Same for both treatment conditions Treatment discontinuation (bevacizumab): treatment policy Partial discontinuation: treatment policy Treatment switch i.e. use of new anticancer therapy: treatment policy Local treatment: treatment policy	Same: intercurrent events handled according to prespecified strategies; implemented primarily via administration data. Intercurrent events will be measured in NCR as follows: Treatment discontinuation (bevacizumab): stop date of individual drugs is registered. Partial discontinuation: stop date of individual drugs is registered. Treatment switch: stop date and start date of individual drugs is registered. If the treatment is switched due to toxicity, this is registered within the same "episode". Treatment switching due to disease progression is not an intercurrent event. Local treatment: - Surgery: the performance of surgery for metastases is registered, alongside the specific codes and dates - Radiotherapy: the performance of radiotherapy for metastases is registered, alongside the specific codes and dates	Target trial and emulation identical. Treatment policy reflects real-world effectiveness. There may be some limitations regarding the measurement of intercurrent events. Especially partial discontinuation of capecitabine may be difficult to assess because capecitabine is administered orally. (In)adherence might be a problem here. Treatment policy implies using all available data regardless of the occurrence of the intercurrent events. Hence, issues with the identification of the corresponding intercurrent events in real-world data will bear no consequences for the estimand as long as relevant post-intercurrent event data are available in the data source.

		- Ablation: the performance of ablation for (liver) metastases is registered, alongside the specific codes and dates	

Rationale for handling of intercurrent events: The strategy to deal with all intercurrent events for estimand 3 is treatment policy. This strategy best reflects the evaluation of the intervention under study within clinical practice [14]. This estimand reflects the effectiveness of the intervention regardless of treatment discontinuation (of bevacizumab), partial discontinuation (of oxaliplatin and/ or capecitabine before the protocolised discontinuation), use of new anti-cancer therapy (treatment switch) or local treatment.

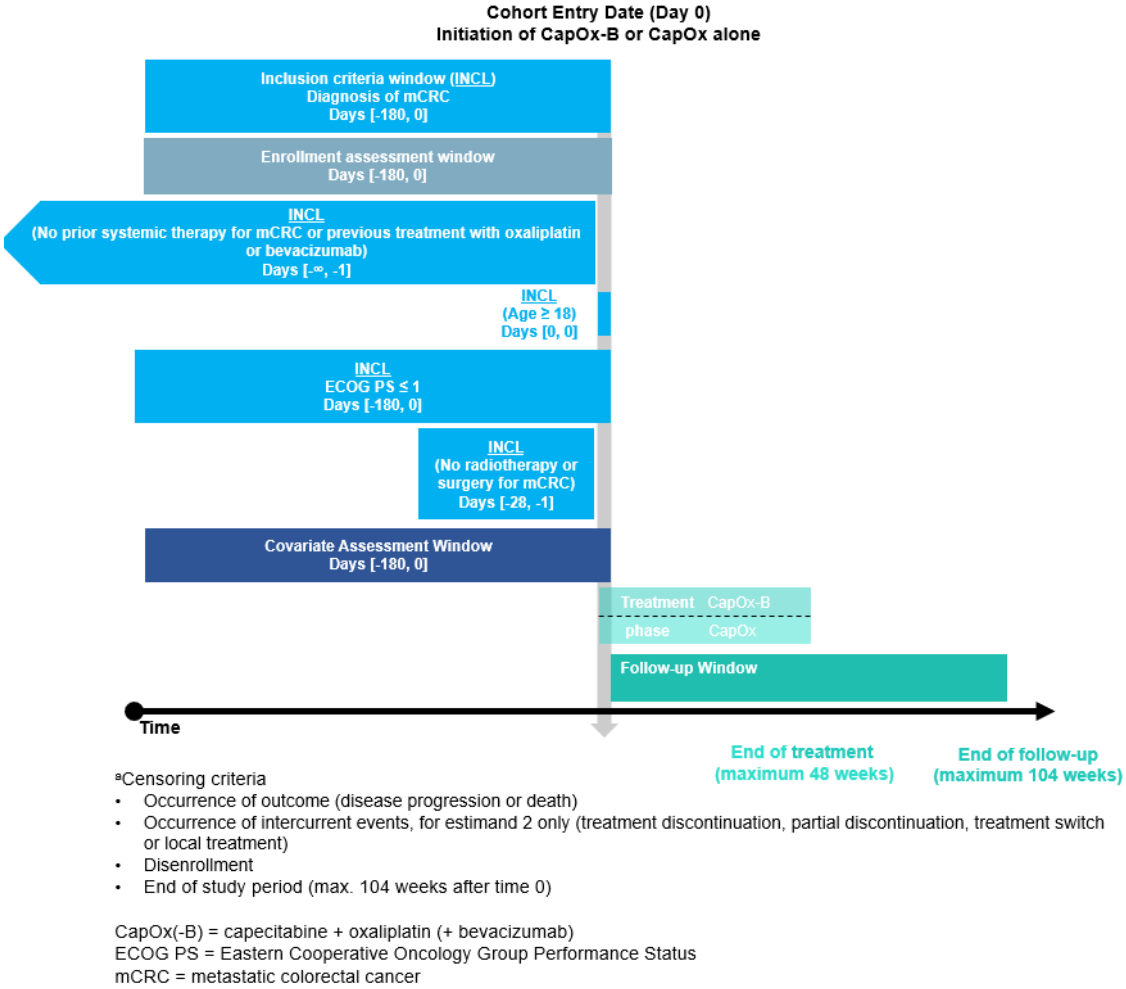
7. Research methods

7.1 Study design

Research design (e.g. cohort, case-control, etc.): New user active comparator cohort study

Rationale for study design choice: This study design reduces risk of bias from unmeasured confounding by indication and fits neatly into the target trial framework. In this target trial emulation, there are different strategies to handle the intercurrent events in the different estimands, following the hypothetical target trial. The treatment policy that is applied for all intercurrent events in estimands 1 and 3 can be seen as the former intention-to-treat approach of the RCTs. In estimand 2, the hypothetical strategy is considered to handle the intercurrent events of partial discontinuation and treatment switch, and a composite strategy is used to handle local treatment.

7.2 Study design diagram



[a]Censoring criteria

- Occurrence of outcome (disease progression or death)
- Occurrence of intercurrent events, for estimand 2 only (treatment discontinuation, partial discontinuation, treatment switch or local treatment)
- Disenrollment
- End of study period (max. 104 weeks after time 0)

CapOx(-B) = capecitabine + oxaliplatin (+ bevacizumab)
ECOG PS = Eastern Cooperative Oncology Group Performance Status
mCRC = metastatic colorectal cancer

Figure 1. Study design diagram for the CapOx-B vs CapOx trial replication.

7.3 Setting

This study is conducted using routinely collected electronic health records from 2021 to 2024, reflecting the period of bevacizumab use in routine clinical practice. The study is set in secondary care, drawing on longitudinal cancer registry data from the NCR. Data are sourced from the Netherlands, providing nation-wide, population-based and representative coverage of real-world clinical care.

7.3.1 Definition of time 0 (and other primary time anchors) for entry to the study population

The eligible cohort entry period is set from January 1st 2021 until December 31st 2022. The actual study period is defined from January 1st 2021 until December 31st 2024. The study period takes into account the maximum follow-up period of 104 weeks.

Time 0 is the date of initiation of CapOx-B regimen or CapOx regimen alone. This is when patients enter the study population and mimics the initiation of therapy at randomisation in the target trial framework. The inclusion criterium of a mCRC diagnosis is assessed within the 180 days prior to time 0. The absence of prior systemic therapy for mCRC or previous treatment with oxaliplatin and/ or bevacizumab is assessed using all available information prior to time 0. Other inclusion criteria such as ECOG PS ≤ 1 is assessed within the 180 days prior to time 0 and age ≥ 18 is assessed at time 0. The absence of radiotherapy or surgery is assessed within the 28 days prior to time 0. The covariate assessment window (see section 7.4) ranges from the 180 days prior to time 0 until time 0. Patients are followed up until the occurrence of the outcome (i.e., disease progression or death), loss to follow-up (e.g., disenrollment) or until a maximum follow-up of 104 weeks after treatment initiation. Under the hypothetical strategy in estimand 2, the occurrence of partial discontinuation and treatment switch will inform the end of follow-up.

Table 5. Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Exposure: CapOx-B	Date of first-time use of CapOx-B (time 0)	Single	Incident	[-∞, -1]	OP	ATC	n/a	CapOx-B or CapOx alone (intravenous and oral formulations)	No validation study	n/a
Comparator: CapOx alone	Date of first-time use of CapOx (time 0)	Single	Incident	[-∞, -1]	OP	ATC	n/a	CapOx-B or CapOx alone (intravenous and oral formulations)	No validation study	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

[2]See appendix for listing of clinical codes for each study parameter

7.3.2 Study inclusion criteria:

We require information from the medical history prior to time 0 in order to capture the clinical codes to measure inclusion and exclusion criteria and covariates. We restrict the population to adult patients with mCRC, an ECOG PS ≤ 1, no prior systemic therapy for mCRC or previous treatment with bevacizumab and/ or oxaliplatin, and no radiotherapy or surgery for mCRC within 4 weeks prior to time 0. Finally, the study population will be restricted to study drug initiators because of the new user active comparator study design.

Table 6. Operational Definitions of Inclusion Criteria

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Histologically confirmed mCRC	Based on registration in the Dutch pathology database PALGA or a cancer diagnosis in the national registration of hospital care (LBZ)	[-180, 0]	OP	ICD-O	Any	Exposure: CapOx-B, Comparator: CapOx alone	No validation study	n/a
Age ≥ 18	Defined as: (time 0 – date of birth)	[0, 0]	n/a	n/a	n/a	Exposure: CapOx-B, Comparator: CapOx alone	n/a	n/a
ECOG PS ≤ 1	At the time of recorded diagnosis	[-180, 0]	OP	n/a	n/a	Exposure: CapOx-B, Comparator: CapOx alone	No validation study	n/a
No prior systemic therapy for mCRC or previous	n/a	[-∞, -1]	OP	ATC	Any	Exposure: CapOx-B, Comparator: CapOx alone	No validation study	Only the first "episode" is recorded, so by definition there is no prior

treatment with oxaliplatin and/ or bevacizumab								systemic therapy for mCRC. We will look into previous cancer diagnoses and assess previous treatment with oxaliplatin and/ or bevacizumab.
Radiotherapy or surgery for mCRC was permitted if completed ≥ 4 weeks before treatment assignment	n/a	[-28, -1]	OP & IP	n/a	Any	Exposure: CapOx-B, Comparator: CapOx alone	No validation study	n/a
Treatment initiation with either CapOx-B or CapOx	Time 0 is the date of initiation of CapOx-B or CapOx	[0, 0]	OP & IP	ATC	Any	Exposure: CapOx-B, Comparator: CapOx alone	No validation study	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

[2] See appendix for listing of clinical codes for each study parameter

[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

The following inclusion criteria were applicable to the target trial but not to the target trial emulation:

- Not felt to be amenable to curative resection. This inclusion criterion is not implemented in the target trial emulation because we are unable to identify the patients with the intention to undergo curative resection before initiation of systemic treatment.
- Life expectancy ≥ 3 months. This inclusion criterion is not implemented in the target trial emulation because life expectancy at baseline is not recorded in the NCR.

Adequate hematologic/clotting, hepatic and renal function. This inclusion criterion is not implemented in the target trial emulation because there is no information available on hematologic/clotting, hepatic and renal function.

7.3.3 Study exclusion criteria

No exclusion criteria will be applied. The target trial aimed to apply the following exclusion criteria: pregnant or breastfeeding women and serious nonhealing wounds or ulcers. Unfortunately, information on pregnancy is not completely available for the whole of the Netherlands. In addition, we do not have information on breastfeeding, serious nonhealing wounds or ulcers.

The fact that these potential exclusion criteria cannot be applied, is not expected to have a large impact. The use of capecitabine, oxaliplatin and bevacizumab is contra-indicated in pregnant women and/or breastfeeding women. [10,15,16] The Summary of Product Characteristics of bevacizumab also advises precaution in complicated wound-healing cases [10]. Moreover, guidelines recommend leaving at least 5 weeks in between chemotherapy/bevacizumab administration and resection [5]. Hence, it is assumed that these oncological drugs are not applied to patients with these characteristics. Therefore, the lacking information on these potential exclusion criteria is not expected to be problematic.

Records with missing values for key demographic variables (e.g., age or sex) will be excluded from the analysis.

Table 7. Operational Definitions of Exclusion Criteria

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

7.4 Variables

7.4.1 Exposure(s) of interest

We chose an active comparator design because it is the most appropriate for this observational study. As mentioned earlier, guidelines recommend fluoropyrimidine-based chemotherapy in combination with oxaliplatin or irinotecan for first-line initially unresectable mCRC [4,5]. An example of such a regimen is the CapOx regimen. Bevacizumab can be added to the CapOx regimen because it is hypothesised to improve prognosis compared to chemotherapy alone [9].

Algorithm to define duration of exposure effect:

The assessment of exposure to the CapOx-B regimen and CapOx regimen requires some aspects to be defined. A regimen is composed of a number of treatment cycles. In each cycle, one or more drugs are administered over the course of one or multiple days. Often, the cycle also contains a period without drug administration. Depending on the regimen, patients may be treated with a predetermined number of cycles or until disease progression or intolerable toxicity occurs. In some regimens, the combination of drugs is administered for a predetermined number of cycles, after which a

modified regimen (in which one or more of the anticancer drugs, often oxaliplatin, are ceased) may be continued until disease progression or intolerable toxicity.

The CapOx(-B) regimen consists of 21-day cycles that are repetitively administered. Within a cycle, oxaliplatin (and bevacizumab) are administered intravenously on day 1. Capecitabine is administered orally, twice daily, on day 1-14. The remaining 7 days in the cycle are rest days, in which no drug will be administered. Discontinuation of oxaliplatin after 6 cycles is a part of the standard treatment protocol, as are adjustments to the dose and dosing interval.

The exposure assessment will be based on administration data. In the NCR, drug exposure is captured within so-called episodes. An "episode" captures the clinical trajectory of a patient up until disease progression. Disease progression marks the end of the "episode". These "episodes" also take cycle prolongations and potential treatment holidays into account. During a treatment holiday, the systemic anticancer therapy (SACT) can be temporarily paused while the disease is stable. The regimen may be continued as soon as disease progression occurs. This is still regarded as the first-line therapy but not captured in the first "episode" anymore. The dataset from the NCR contains information on the start and stop date of SACT, including capecitabine, oxaliplatin and bevacizumab. The stop date is defined as the termination of the final cycle. If this information is lacking, which is the case for capecitabine, the start date of the final cycle is registered as the stop date. We have to add the cycle length of 21 days to the start date of the final cycle to obtain the stop date. The NCR dataset also contains information on the number of cycles that were administered. Exposure is based on administration data, not on prescription or dispensation data.

Throughout the study period, there may be some exposure-related intercurrent events, such as (partial) treatment discontinuation and treatment switching. Treatment discontinuation is defined as the discontinuation of bevacizumab, while continuing oxaliplatin and/ or capecitabine. Partial discontinuation is defined as the premature discontinuation of oxaliplatin and/ or capecitabine. A treatment switch primarily reflects the discontinuation of the allocated treatment and initiation of a different (chemotherapy-backbone) regimen. The incidental addition of another component to the allocated regimen is also considered a treatment switch. Switching to another treatment regimen due to toxicity is a part of the same "episode" and can easily be recognised as such. Switching to another treatment regimen due to disease progression will not be perceived as an intercurrent event because it marks the occurrence of the outcome.

Table 8. Operational Definitions of Exposure

Exposure group name(s)	Details	Washout window	Assessment Window	Care Setting[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Incident with respect to...	Measurement characteristics/ validation	Source of algorithm
Exposure: CapOx-B	CapOx-B; oral capecitabine, intravenous oxaliplatin and	[-∞, -1]	[0, censor]	OP	ATC	n/a	Patients with histologically confirmed mCRC	Administrations measured in the entire registered	No validation study	n/a

	intravenous bevacizumab							period prior to first use		
Comparator: CapOx	CapOx; oral capecitabine and intravenous oxaliplatin	[-∞, -1]	[0, censor]	OP	ATC	n/a	Patients with histologically confirmed mCRC	Administrations measured in the entire registered period prior to first use	No validation study	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.2 Outcome(s) of interest

Progression-free survival is a common effectiveness outcome within oncology. The end of progression-free survival is marked by the occurrence of either disease progression or death. This outcome parallels the outcome used in the RCT by Saltz et al, comparing bevacizumab to placebo, when added to the CapOx regimen [6]. In estimand 2, local treatment will be added as a composite outcome, its occurrence leading to patients being assumed to attain the final administrative censoring date.

Table 9. Operational Definitions of Outcome

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Progression of disease	No imaging available. Disease progression is marked by the end of an "episode", defined as the discontinuation of a regimen due to disease progression.	Yes	Time-to-event	n/a	IP, OP, ED	n/a	n/a	Patients with histologically confirmed mCRC, both in the exposure and comparator group.	Similar method to assess disease progression have been used in Zwart et al 2023 Cancer Med [13]	n/a

Death	n/a	Yes	Time-to-event	n/a	n/a	n/a	n/a	Patients with histologically confirmed mCRC, both in the exposure and comparator group.	No validation study	n/a
Local treatment	This entails surgery, radiotherapy or ablation.	No for estimands 1 and 3; Yes for estimand 2	Time-to-event	[-28, 0]	IP & OP	n/a	n/a	Patients with histologically confirmed mCRC, both in the exposure and comparator group.	n/a	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.3 Intercurrent events

The following intercurrent events have been defined:

- Treatment discontinuation: the discontinuation of bevacizumab, while continuing oxaliplatin and/ or capecitabine. This intercurrent event can only occur in the CapOx-B group.
- Partial discontinuation: the premature discontinuation of oxaliplatin and/ or capecitabine. Discontinuation of oxaliplatin according to standard protocol is not considered an intercurrent event.
- Treatment switch: the discontinuation of the allocated treatment (CapOx-B or CapOx) and the initiation of a different (chemotherapy-backbone) regimen. The incidental addition of another component to the allocated regimen is also considered within this intercurrent events.
- Local treatments: surgery, radiotherapy or ablation. Local treatment may be applied if the disease is sufficiently downstaged by chemotherapy.

Table 10. Operational Definitions of Intercurrent events

Intercurrent events	Details	Washout window	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Treatment discontinuation	Discontinuation of bevacizumab.	[-∞, -1]	[0, end of follow-up]	IP	ATC	n/a	Patients with histologically confirmed mCRC, only in the exposure group (CapOx-B).	n/a	Stop date of individual drugs is registered.
Partial discontinuation	Discontinuation of oxaliplatin and/ or capecitabine before the protocolised discontinuation	[-∞, -1]	[0, end of follow-up]	IP & OP	ATC	n/a	Patients with histologically confirmed mCRC, both in the exposure and comparator group.	n/a	Stop date of individual drugs is registered.
Treatment switch	Use of new anticancer therapy	[-∞, -1]	[0, end of follow-up]	IP & OP	ATC	n/a	Patients with histologically confirmed mCRC, both in the exposure and comparator group.	n/a	Stop date and start date of individual drugs is registered. If the treatment is switched due to toxicity, this is registered within the same treatment "episode". Treatment switching due to disease progression is not an intercurrent event.
Local treatment	The occurrence of surgery, radiotherapy or ablation.	[-∞, -1]	[0, end of follow-up]	IP		n/a	Patients with histologically confirmed mCRC, both in the exposure and	n/a	Surgery: the performance of surgery for metastases is registered, alongside

							comparator group.		the specific codes and dates Radiotherapy: the performance of radiotherapy for metastases is registered, alongside the specific codes and dates Ablation: the performance of ablation for (liver) metastases is registered, alongside the specific codes and dates.

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.4 Follow up

Follow-up will start at time 0, which is the initiation of the CapOx-B regimen for the intervention group and initiation of the CapOx regimen for the control group.

Table 11. Operational Definitions of Follow Up

Follow up start	Day 0	
Follow up end[1]	**Select all that apply**	**Specify**
Date of outcome	Yes	See table 9
Date of death	Yes	According to Centraal Bureau voor de Statistiek (CBS), which NCR is linked to

End of observation in data	Yes	Maximum of 104 weeks after time 0 or until censoring (due to loss to follow-up)
Day X following index date *(specify day)*	Yes	End of data availability of database Data until the end of the first “episode” will be complete. Further follow-up won't be complete, except for death
End of study period (specify date)	Yes	December 31st 2024
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	Yes	Maximum treatment duration following time 0 The intercurrent event treatment discontinuation is defined by the discontinuation of bevacizumab. Follow-up is continued regardless. The intercurrent event partial discontinuation is defined by the discontinuation of capecitabine and/ or oxaliplatin before protocolised discontinuation. In estimand 1 and 3, follow-up is continued regardless. In estimand 2, follow-up is censored
Date of add to/switch from exposure *(specify algorithm)*	Yes	Defined by the discontinuation of the allocated regimen and incident administration of a different (chemotherapy-backbone) regimen or addition of another component to the regimen In estimand 1 and 3, follow-up is continued regardless of add on or switch. In estimand 2, follow-up is censored at add on or switch
Other date *(local treatment)*	Yes	Defined by the occurrence of local treatment (i.e. radiotherapy, surgery or ablation) In estimand 1 and 3, follow-up is continued regardless of local treatment. In estimand 2, local treatment is part of the composite outcome and its occurrence leads to patients being assumed to attain the final administrative censoring date.

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

7.4.5 Covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications)

Table 12. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Age	(time 0 – date of birth)	Continuous	[0, 0]	n/a	n/a	n/a	Exposure and control group	n/a	n/a
Sex	Male, female	Binary	[0, 0]	n/a	n/a	n/a	Exposure and control group	n/a	n/a
ECOG PS	Eastern Cooperative Oncology Group Performance Score, at the time of diagnosis 0, 1	Binary	[-180, 0]	OP	n/a	n/a	Exposure and control group	n/a	n/a
(History of) cardiovascular disease	Yes, no	Binary	[-∞, 0]	OP	ICD-10	n/a	Exposure and control group	n/a	n/a
(History of) hypertension	Yes, no	Binary	[- ∞, 0]	OP	ICD-10	n/a	Exposure and control group	n/a	n/a
(History of) hypercholesterolemia	Yes, no	Binary	[- ∞, 0]	OP	ICD-10	n/a	Exposure and control group	n/a	n/a
Diabetes	Yes, no	Binary	[- ∞, 0]	OP	ICD-10	n/a	Exposure and control group	n/a	n/a
Site of primary tumour	Colon Rectosigmoid Rectum	Categorical	[-180, 0]	OP	ICD-O	n/a	Exposure and control group	n/a	n/a
Tumour sidedness	Left-sided Right-sided	Categorical	[-180, 0]	OP	n/a	n/a	Exposure and control group	n/a	n/a

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Topography of metastases	Liver Lung Lymph nodes Peritoneum Other Multiple	Categorical	[-180, 0]	OP	n/a	n/a	Exposure and control group	n/a	n/a
RAS mutation status	Absent Present, type unknown Present, RAS other Present, KRAS G12C Unknown	Categorical	[-180, 0]	OP	n/a	n/a	Exposure and control group	n/a	n/a
BRAF mutation status	Absent Present, type unknown Present, BRAF-V600E Present, BRAF-nonV600E Unknown	Categorical	[-180, 0]	OP	n/a	n/a	Exposure and control group	n/a	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.5 Core Emulation Table - Design Summary

Table 13. Core Emulation Table - Design

	Target Trial	Target Trial Emulation	Comment
Eligibility	Inclusion criteria		
	- Histologically confirmed mCRC	- Have a record of histologically confirmed mCRC based on the registration in the Dutch pathology database PALGA or a cancer diagnosis in the national registration of hospital care (LBZ)	Eligibility applied using structured EHR data.
	- Age 18 and over at the time of eligibility screening	- Age 18 and over at time 0	n/a
	- Not felt to be amenable to curative resection	-	Probably no information is available on amenability for curative resection. It should be possible to identify which patients underwent (curative) resection, but we cannot identify the patients who intended to undergo curative resection before initiation of systemic treatment.
	- ECOG PS ≤ 1	- ECOG PS ≤ 1 at the time of recorded diagnosis	n/a
	- Life expectancy ≥ 3 months	-	Life expectancy at baseline is not recorded.
	- No prior systemic therapy for mCRC or previous treatment with oxaliplatin and/ or bevacizumab	- No prior systemic therapy for mCRC or previous treatment with oxaliplatin and/ or bevacizumab	Emulation restricts to new users in routine care to align with eligibility and mitigate immortal time bias.

	- Radiotherapy or surgery for mCRC was permitted if completed ≥ 4 weeks before random assignment	- Radiotherapy or surgery for mCRC was permitted if completed ≥ 4 weeks before treatment assignment	n/a
	- Adequate hematologic/clotting, hepatic and renal function	-	We do not have any information on adequate hematologic/clotting, hepatic and renal function.
	-	- Treatment initiation with either CapOx-B or CapOx using administration records	Because of the new user active comparator study design, the study population will be restricted to study drug initiators. Patients who receive treatment or comparator in clinical practice reflect the target population. Hence, no relevant differences are expected between patients eligible to initiate treatments approved within the same indication. Moreover, using the first administration record can be interpreted as an emulation of intention-to-treat, which is based on eligibility in RCTs.
Eligibility	Exclusion criteria		
	- Pregnant or breastfeeding women	-	Information on pregnancy, breastfeeding or comorbidities present at the time of diagnosis is registered in some regions of the Netherlands only. We do not have information on breastfeeding, serious nonhealing wounds or ulcers. The use of capecitabine, oxaliplatin and/or bevacizumab is contra-indicated in
	- Serious nonhealing wound or ulcer	-	

			patients with these conditions. Moreover, guidelines recommend leaving at least 5 weeks in between chemotherapy/bevacizumab administration and resection. The lack of information on these exclusion criteria is not expected to have a large impact because the drugs are unlikely to be used in such patients in clinical practice.
Settings	Multicentre, parallel-group, randomised controlled open-label trial	Routine care data from NCR database; capturing patient and disease characteristics, prescriptions, outcomes and covariates from the in- and outpatient hospital setting	Real-world data (RWD) captures care as delivered; pragmatic design supports effectiveness focus.
Treatment Conditions	CapOx-B regimen vs. CapOx regimen alone Discontinuation of oxaliplatin is a part of the standard treatment protocol as are adjustments to the dose and dosing interval	Initiation of first-line CapOx-B regimen or CapOx regimen alone (no prior use), identified using administration data The CapOx(-B) regimen consists of repetitively administered 21-day cycles. Within a cycle, oxaliplatin (and bevacizumab) are administered intravenously on day 1. Capecitabine is administered orally, twice daily, on day 1-14 Discontinuation of oxaliplatin (after 6 cycles) is part of the standard treatment protocol Real-world use without restriction	Reflects new-user, active comparator design.

Treatment assignment	1:1 randomisation, stratified by ECOG PS and liver as a metastatic site	Randomisation cannot be directly emulated. At the estimation stage, randomisation will preferably be emulated through 1:1 matching by ECOG PS and liver as a metastatic site. Alternatively, propensity-score based weighting (inverse probability of treatment weighting (IPTW)), including ECOG PS and liver as a metastatic site, will be used to account for baseline differences between treatment groups.	1:1 matching or PS methods used to balance confounders in absence of randomisation; design emulates stratified, randomised treatment assignment.
Follow-up	Begins at the time of randomisation; ends at the first occurrence of the outcome (disease progression or death), loss to follow-up, study withdrawal or after a maximum follow-up period of 104 weeks	Begins at treatment initiation (index date) and ends at outcome (disease progression or death), loss to follow-up or after a maximum follow-up time of 104 weeks. In estimand 2, the occurrence of some intercurrent events will inform the end of follow-up. This is the case for the intercurrent events partial discontinuation and treatment switch, under the hypothetical strategy. For local treatment, under the composite strategy, patients are assumed to have completed the intended follow up and will be administratively censored at 104 weeks. Follow-up is continued regardless of the occurrence of intercurrent events when the treatment	Aligns start of follow-up with treatment initiation to mimic randomisation. From the NCR, disease progression will be identified through the end of an "episode". This identification from the NCR is validated and regularly used in observational studies using NCR data [13]. With little expected impact, it differs from the target trial measurement, in which disease progression is identified through medical imaging. The hypothetically handled intercurrent events from estimand 2 can be identified from the NCR reliably due to the availability of start and stop dates of SACT.

		policy strategy is applied, which is the case in estimand 1 and 3.	
Outcome	Time to first occurrence of disease progression, based on abdominal CT scans and/or MRI, or death In estimand 2, local treatment (surgery, radiotherapy or ablation) will be added as a composite outcome	Time to first occurrence of disease progression, based on end of "episodes" in NCR database, or mortality records (in linked databases) In estimand 2, local treatment (surgery, radiotherapy or ablation) will be added as a composite outcome	PFS measurement differs between target trial and emulation. A new "episode" in the NCR database is only made if the reason for treatment switch is disease progression. A new "episode" indicates disease progression and thus the end of PFS. PFS is assessed locally instead of centrally, which may introduce some heterogeneity in the outcome assessment. If treatment is switched due to toxicity, this is added to the same "episode". Outcome measurement is validated and regularly used in observational studies using NCR data. Date of death is always available through linkage with the municipal basic administration (GBA), also if the patient is transferred to another (Dutch) hospital.
Intercurrent events	Estimand 1 and 3: treatment policy for all intercurrent events Estimand 2: treatment policy for the treatment discontinuation intercurrent	Same strategies implemented using censoring and creating a composite outcome including local treatment	Estimand-aligned censoring rules and use of a composite outcome.

	event, hypothetical for partial discontinuation or treatment switch, composite for local treatment	Intercurrent events will be measured in NCR as follows: Treatment discontinuation (bevacizumab): stop date of individual drugs is registered Partial discontinuation: stop date of individual drugs is registered Treatment switch: stop date and start date of individual drugs is registered. A treatment switch due to toxicity is registered within the same "episode". Local treatment: - Surgery: the performance of surgery for metastases is registered, alongside the specific codes and dates - Radiotherapy: the performance of radiotherapy for metastases is registered, alongside the specific codes and dates - Ablation: the performance of ablation for (liver) metastases is registered, alongside the specific codes and dates	A treatment switch due to toxicity is registered within the same "episode" and can be identified correctly.
Loss to follow-up	A participant will be considered lost to follow-up if they repeatedly fail to return for scheduled visits, they cannot	Moving to another country.	Loss to follow-up in the target trial emulation could happen if the patient moved to another country or de-

	be contacted by the study staff and their health condition and vital status remains unknown despite attempts to contact them.	De-registration from the data source upon request from the patient.	registered from the data source. Both are assumed to be very rare in this population and therefore inconsequential.

7.6 Data analysis

7.6.1 Analysis plan

Overview

The analyses are conducted within a target trila emulation framework to estimate the effect of the CapOx-B regimen compared with the CapOx regimen on progression-free survival (PFS).

For **Estimand 1**, the main estimand supporting decision making, the primary causal effect summary measure is the hazard ratio for time to first disease progression or death, estimated using an inverse probability of treatment weighted (IPTW) Cox proportional hazards model.

Sensitivity analyses will assess robustness of the primary findings to key assumptions, including inverse probability of censoring weighting (IPCW), tipping point analysis, and probabilistic bias analysis for non-differential outcome misclassification (details in section 7.6.5).

Two supplemental estimands are also defined: **Estimand 2**, applying a hypothetical and composite strategy for intercurrent events, and **Estimand 3**, estimating treatment effects using restricted mean survival time (RMST) derived from an IPTW-weighted Weibull accelerated failure time (AFT) model. In addition, supplemental analyses (e.g., crude and IPTW-adjusted Kaplan-Meier curves, crude Cox models, event counts and incidence rates, propensity score and weight distributions, covariate balance before and after weighting, censoring and intercurrent event patterns, proportional hazads diagnostics, positivity checks, and multiple-imputation diagnostics) will be conducted to support interpretation of the main analysis.

7.6.2 Primary Estimand (1) Analysis

i. Objective

No hypothesis testing will be performed. This study will investigate whether the risk of disease progression or death is higher in mCRC patients treated with CapOx compared to mCRC patients treated with CapOx-B.

ii. Exposure contrast

CapOx-B regimen vs.CapOx regimen

iii. Outcome

Time to disease progression or death

iv. Analytic software

R

v. Handling of intercurrent events

Do not exclude follow-up data after the occurrence of intercurrent events
Handle intercurrent events according to the following strategies:

- Treatment discontinuation (bevacizumab): Apply a treatment policy strategy
- Partial discontinuation: Apply a treatment policy strategy
- Treatment switch: Apply a treatment policy strategy
- Local treatment: Apply a treatment policy strategy

vi. Outcome Modelling

A Cox proportional hazards model, weighted by inverse probability of treatment (IPTW), will be used to estimate the effect of initiating CapOx-B versus CapOx on time to disease progression or death.

- Start of follow-up: time will be measured from the day after treatment initiation of either CapOx-B or CapOx
- Endpoint: time from treatment initiation to the first occurrence of disease progression or death
- Censoring:
    - Non-administrative censoring: loss to follow-up
    - Administrative censoring: end of study follow-up in the absence of disease progression or death (maximum 104 weeks).
    - Model covariates: treatment group (CapOx-B vs CapOx))

Assumptions of Cox Model

- Proportional hazards:
    - The effect of treatment is assumed to be constant over time.
- Non-informative censoring:
    - Censoring is assumed to be independent of the outcome, conditional on model covariates and no disease progression of death up to the

time of censoring.

Diagnostics for Cox Model

- Proportional hazards assessed using log(-log) survival plots or Schoenfeld residuals

Note: The assumption of non-informative censoring cannot be verified with observed data; it will be addressed through sensitivity analyses.

vii. Confounding Adjustment

Matching to obtain stratified randomisation

The preferable method to emulate the stratified randomisation would be 1:1 matching patients from the CapOx-B group and CapOx group by ECOG PS (0 or 1) and liver as a metastatic site (yes or no). This ensures comparability at treatment initiation. There is a risk that the obtained stratification subgroups become too small. If there are less than 220 patients in either one of the treatment arms (see sample size estimation in section 7.10) after stratified randomisation, matching is deemed unfeasible, and we may resort to an IPTW model, as described in the following section.

Inverse Probability of Treatment Weighting (IPTW)

In case 1:1 matching on ECOG PS and liver as a metastatic site is not feasible, IPTW will be used to adjust for baseline confounding. Propensity scores, defined as the probability of initiating CapOx-B versus CapOx, will be estimated using logistic regression. The model will include ECOG PS, liver as a metastatic site and the number of comorbidities (a covariate deviated from the presence of prespecified comorbidities, i.e., cardiovascular disease, hypertension, hypercholesterolemia and diabetes). Razenberg et al. identified that the covariate "number of comorbidities" influences the presence of bevacizumab within the first-line palliative treatment in metachronous mCRC patients [7].

Stabilised weights will be calculated by dividing the marginal probability of receiving the treatment actually received (i.e., the overall proportion treated in the study population) by the individual's estimated propensity score (i.e., the conditional probability of receiving their observed treatment). Weights will be truncated at the 1st and 99th percentiles to limit the influence of extreme values.

Weight truncation reduces the influence of individuals with highly improbable treatment assignments but does not resolve non-overlap. Therefore, if regions of the propensity score distribution show insufficient overlap, we plan to restrict analyses to the overlapping region (trimming) or apply overlap weights.
Truncated stabilised IPTW weights will then be applied in the Cox proportional hazards model (weighted likelihood) to estimate the marginal treatment effect (CapOx-B vs CapOx) on time to first disease progression or death.

Assumptions underlying IPTW

- **No unmeasured confounding** (all relevant baseline confounders are included in the propensity score model)
- **Positivity** (each individual has a non-zero probability of receiving either treatment, given their covariates)
- **Correct model specification** (the propensity score model is correctly specified [functional form, covariate inclusion])
- **Consistency** (each individual's potential outcome under the observed treatment equals their actual outcome)

Diagnostics for IPTW

- **Covariate balance:** check that baseline characteristics are balanced across treatment groups after weighting.
    - Evaluate standardised mean differences (SMDs): SMDs < 0.1 will be considered acceptable.
- **Positivity check:** ensure adequate overlap in propensity score distributions between treatment groups to support estimation (graphically).

viii. Missing Data Handling

Missing exposure data

We assume that missing administration data for CapOx(-B) reflect true treatment discontinuation. These treatments can only be provided in the hospital setting or from the outpatient pharmacy and this is captured on a national level in the NCR.

Missing outcome data

The Cox proportional hazards model implicitly assumes non-informative censoring given model covariates and PFS up to the time of censoring, meaning that censored participants contribute partial information (i.e. time at risk up to the time of censoring) and their censoring is unrelated to the outcome, conditional on model covariates and PFS up to the censoring time (i.e., outcome data is missing at random (MAR) under these assumptions).

Missing covariate data

The absence of a diagnosis code is assumed to indicate the absence of the corresponding condition. Missing values for disease-related variables will be addressed using multiple imputation with chained equations (MICE) under a MAR assumption, using the MICE package in R.

Assessment of missingness

Before performing imputation, we will examine the extent and patterns of missingness to evaluate whether imputation is appropriate. Specifically, we will:

- Quantify the percentage of missing data for each covariate
- Compare the proportion of missing values across treatment groups to assess differential missingness
- If a covariate has more than 40% missing data, we will consider alternative approaches (e.g., exclusion of the variable, sensitivity analyses) and justify the decision. Thresholds of 40% have been cited because effect estimates begin to be less reliable as the level of missingness increases beyond this threshold [17].

Imputation model

The MICE procedure will include all covariates used in the outcome and treatment models, as well as predictors of missingness (if there are any additional factors not covered by the covariates in the treatment and outcome models). The treatment and outcomes of interest will also be included.

Key covariates included in the imputation model will be:

- Demographics (age, sex)

Full conditional distributions
MICE will use variable-specific conditional models:
- Logistic regression for binary variables
- Multinomial logistic regression for categorical variables with >2 categories
- Predictive mean matching for continuous variables

Number of imputations and diagnostics
We will generate at least 10 imputed datasets (to ensure stable estimates given the level of missingness) and pool results across imputations using Rubin's rules. Diagnostics will include:
- Checking whether imputed values are plausible and consistent with observed distributions.
- Evaluating convergence of the chained equations.
- Assessing stability and consistency of results across imputed datasets.

Effect estimation under Multiple imputation
The imputation model will be applied prior to effect estimation. IPTW and outcome models will then be fitted in each imputed dataset, and treatment effect estimates (e.g., hazard ratios) will be pooled across datasets using Rubin's rules.

ix. Subgroup Analyses

Not applicable

7.6.3 Supplemental Estimand (2) Analysis

i. Objective

No hypothesis testing will be performed. This study will investigate whether the risk of disease progression or death is higher in mCRC patients treated with CapOx compared to mCRC patients treated with CapOx-B.

ii. Exposure contrast

CapOx-B regimen vs.CapOx regimen

iii. Outcome

Time to disease progression or death

iv. Analytic Software

R

v. Handling of intercurrent events

Assume complete follow up was attained after the occurrence of local treatment, do not exclude follow-up data after the occurrence of intercurrent events
Handle intercurrent events according to the following strategies:

- Treatment discontinuation (bevacizumab): Apply a treatment policy strategy
- Partial discontinuation: Apply a hypothetical strategy
- Treatment switch: Apply a hypothetical strategy
- Local treatment: Apply a composite strategy

vi. Outcome Model

Same as primary estimand, including censoring for the intercurrent events that are handled using the hypothetical strategy.

Cox proportional hazards model

vii. Confounding Adjustment

Same as primary estimand

viii. Missing Data Handling

Same as primary estimand

ix. Subgroup Analyses

Not applicable

7.6.4 Supplemental Estimand (3) Analysis

i. Objective

No hypothesis testing will be performed. This study will investigate whether the risk of disease progression or death is higher in mCRC patients treated with CapOx compared to mCRC patients treated with CapOx-B.

ii. Exposure contrast

CapOx-B regimen vs.CapOx regimen

iii. Outcome

Time to disease progression or death

iv. Analytic software

R

v. Handling of intercurrent events

Same as primary estimand.

vi. Outcome Model

Accelerated failure time (AFT) model with Weibull distribution, followed by estimation of the Restricted Mean Survival Time (RMST) at fixed time points (52 weeks and 104 weeks).

- Covariates in the model will include treatment group
- Start of follow-up: date of treatment initiation for either CapOx-B or CapOx
- Endpoint: time from treatment initiation to the first occurrence of disease progression or death
- Censoring:
    - Non-administrative censoring: loss to follow-up
    - Administrative censoring: end of study follow-up in the absence of disease progression or death (maximum 104 weeks).

Assumptions of AFT model

- Survival times follow a Weibull distribution
- Non-informative censoring (conditional on included covariates and PFS up to the censoring time)
- Log-linear relationship between covariates and log survival time

Diagnostics for AFT model
Log(-log(S(t))) vs log(t) should be linear Q-Q plot of residuals

To estimate the RMST at 52 weeks and 104 weeks from the Weibull AFT model, we first use the model to obtain the predicted cumulative incidence curves for each treatment group. The RMST is then calculated as the average survival time up to a fixed time point, which corresponds to the area under the cumulative incidence curve between time zero and the chosen time horizon (52 or 104 weeks).

- Fit the Weibull AFT model, which gives the shape and scale of the cumulative incidence curve for each group.
- From this model, generate the predicted survival probability at each time.
- Integrate (i.e., add up) the survival probabilities from time 0 to 52 weeks and separately from time 0 to 104 weeks. The result is the expected survival time lived within those windows.

Compare the RMST values between treatment groups to obtain the difference in average survival time over 52 and 104 weeks.

We will quantify uncertainty using a nonparametric bootstrap of individuals (resampling with replacement) with B = 1,000 replicates. Within each bootstrap sample, we will perform multiple imputation (M = 10) using the same imputation model described in the missing data section, fit the Weibull AFT model, compute RMST at 3 and 5 years, and average the M imputation-specific estimates to obtain one bootstrap estimate. 95% confidence intervals will be derived from the percentile distribution of the B bootstrap estimates. For reporting point estimates, we will additionally pool across M = 40 imputations on the original dataset.

vii. Confounding Adjustment

Same as primary estimand. The AFT model will preferably be 1:1 matched, but alternatively IPT-weighted, to obtain estimates of RMST at 52 weeks and 104 weeks.

viii. Missing Data Handling

Same as primary estimand

ix. Subgroup Analyses

Not applicable

7.6.5 Sensitivity Analyses

Table 14. Sensitivity analyses – rationale, strengths and limitations

The following sensitivity analyses will be conducted for the primary estimand only.

	Sensitivity analysis 1 Inverse Probability of Censoring Weighting (IPCW)	Sensitivity analysis 2 Outcome misclassification
Analysis method	This analysis will examine the impact of varying assumptions about the censoring-at-random condition on the estimated treatment effect. In the primary analysis, we assumed censoring independent of the outcome, *conditional* on treatment group and survival up to the time of censoring. An additional assumption is that IPTW indirectly balances covariates between the censored and uncensored (i.e., outcome data is MAR conditional on observed exposure and outcome). In the IPCW analysis, we use inverse probability of treatment weights and inverse probability of censoring weights. This analysis assumes censoring is independent of the outcome, with all common causes of both the outcome and censoring being accounted for. Follow-up will be divided into equal 30-day intervals. At the start of each interval, we will update the information available on each patient and assess whether they remain followed or have been censored. If they remain under observation, they contribute to the risk set for that interval. • The weight for each participant at each interval is calculated as: 1/(the estimated probability of remaining uncensored, given a set of baseline and time-updated covariates that could affect both censoring and the outcome.Characteristics that could affect both censoring and the outcome include: treatment group (CapOx-B vs CapOx) and demographics (age, sex). The weight is calculated separately for each interval, and then multiplied together across all intervals of follow-up to give each participant's cumulative weight. The denominator probabilities will be estimated using pooled logistic regression model fit to the person-interval dataset. In this model, the outcome is whether the participant was censored in that interval. We will truncate weights at prespecified percentiles (1st and 99th). The IPCW will be applied as time-varying weights in the Cox model for time to disease progression or death. Because inverse probability of treatment weights	Probabilistic Bias Analysis (PBA) using Monte Carlo Simulation at the summary level measure • This analysis will be conducted after pooling the hazard ratios across imputed datasets. In other words, multiple imputation will first address uncertainty due to missing data, producing a pooled hazard ratio and variance that account for the imputation process. The pooled estimate (and its variance) will then be used as the input for the Monte Carlo simulation in the probabilistic bias analysis, which will quantify the additional uncertainty due to outcome misclassification. • Plausible probability distributions for the sensitivity and specificity of outcome classification are specified based on available evidence. • Within each Monte Carlo iteration, a hazard ratio will be sampled from a probability distribution informed by the pooled hazard ratio and its variance from the main analysis. This step propagates uncertainty due to sampling variability and multiple imputation into the bias analysis. • Within each iteration a new pair of sensitivity and specificity values is sampled, and these are applied to correct the observed effect estimate for outcome misclassification using bias-adjustment formulas. • This process is repeated across many iterations (e.g., 10,000 times), resulting in a distribution of bias-adjusted effect estimate. • The risk estimate of the originally imputed analysis will be used as the input for the Monte Carlo Simulation. Hence, it is not

	(IPTW) is also used, the final analysis weights will be the product of IPTW and IPCW.	necessary to re-run the original analysis (and imputation).
	Tipping Point Analysis under the censoring not at random (CNAR) assumption: • Tipping point analysis is applied to explore how the estimated treatment effect varies depending on CNAR assumptions departing from the assumption made in the primary analysis (CAR given treatment group and survival up to the time of censoring). To implement the tipping point analysis, we will vary assumptions on the hazard rate after the censoring time for subjects who are non-administratively censored and impute a time-to-event to those participants under the assumed hazard rates. Assumptions post-censoring should take the form of relative increases/decreases with respect to the baseline hazard and should cover a wide range from best to worst case scenarios, i.e. from a scenario with very low assumed post-censoring hazard rates where almost surely none of those non-administratively censored subjects will experience the event to a scenario where almost surely all of them will experience the event at the time of censoring. Assumptions should be varied independently in each treatment group. For each combination of assumed post-censoring hazard rates in the two arms, multiple imputation of the time-to-event will be performed. The primary analysis will then be performed for each imputed dataset with complete data (observed or imputed) for all subjects, and the resulting treatment effects should be combined across imputations. For each grid point we will then obtain a treatment effect estimate and associated uncertainty (95% confidence interval) which will allow us to determine what pattern of assumed hazard rates post-censoring would have led to different conclusions. For example, the point at which a treatment effect would no longer be distinguishable from a chance finding, or vice versa. The plausibility of the assumed post-censoring changes in hazard underpinning the tipping point will then be evaluated. The results of the tipping point analysis (estimated treatment effect and 95% confidence interval) for each combination of assumed post-censoring hazard rates will be presented both numerically (in a table) and graphically (e.g. using a heatmap). We plan to perform analyses to populate a 10x10 grid.	

Assumptions	For IPCW: • Censoring is conditionally independent of the outcome given covariates (i.e., non-informative censoring/censoring at random (CAR), conditional on other covariates beyond those in the analysis model). • Correct model specification and positivity. • Outcome does not directly influence its' own missingness. For Tipping Point Analysis: • Under the assumption that the changes in hazard follow after censoring, time-to-event data for non-administratively censored subjects is imputed in the Cox regression model. • Assumes CNAR. • Does not require formal modelling of the censoring mechanism	• Misclassification is either non-differential or differential (depending on the scenario). • Sensitivity and specificity of outcome classification are known or reasonably estimated from external data or expert opinion. • Misclassification affects only the outcome, not exposure or covariates. • The misclassification process can be simulated accurately.
What is Being Varied?	For IPCW: • The condition of the CAR assumption • Tests a different MAR assumption. For Tipping Point Analysis: • Tipping Point Analysis explores what assumptions about censored outcomes would be needed to alter conclusions with respect to the treatment effect.	• The assumption that the outcome is measured without error is relaxed. • Varying values of sensitivity and specificity are drawn from defined distributions in each simulation iteration. • For differential misclassification, separate Se/Sp values are specified for individuals with and without exposure. Plausible parameter ranges for outcome measurement in real-world data: Sensitivity: 0.70-0.95 • (lower bound reflects poorer capturing of the outcome; upper bound reflects good capturing of the outcome) Specificity: 0.90-0.95 • (high specificity expected, as those without disease progression or death are unlikely to be classified as progressed or dead
Why (Objective)	For IPCW: • To evaluate whether the treatment effect estimate is sensitive to changes in the condition of the MAR assumption. For Tipping Point Analysis: • To evaluate the robustness of findings to a range of	• To assess the robustness of the estimated treatment effect to plausible levels of outcome misclassification. • To determine whether conclusions change under realistic measurement error for outcome.

	assumptions underpinning a MNAR scenario	
Strengths Compared to Primary Analysis	For IPCW: • IPCW adjusts for measured predictors of censoring, providing effect estimates under an alternative MAR condition to the primary analysis. For Tipping Point Analysis: • Tipping Point Analysis provides effect estimates under a range of MNAR assumptions	• Explicitly accounts for uncertainty in outcome measurement. • Provides a distribution of adjusted estimates rather than a single corrected value. • Can reflect differential or non-differential misclassification. • Enhances transparency around the impact of measurement error.
Limitations Compared to Primary Analysis	• IPCW is sensitive to model misspecification. • Cannot account for unmeasured factors affecting censoring. • If censoring is determined by the outcome value itself or if it shares an unmeasured common cause with the outcome, then IPCW estimates remain biased. • Weighting can increase variance, especially if weights are unstable. • Sensitivity analyses rely on varying the assumptions of the primary analysis, but for censoring these assumptions cannot be verified from the observed data; their plausibility can be discussed yet ultimately remains unknown.	• Requires external data or expert assumptions to specify sensitivity and specificity. • Results are only as reliable as the plausibility of input parameters. • May be computationally intensive. • Does not account for other sources of bias (e.g., unmeasured confounding, exposure misclassification) unless jointly modelled.

7.6.6 Other Supplemental Analyses

Baseline characteristics will be presented overall and stratified by treatment group. Categorical ad binary variables will be summarised as counts (n) and percentages, while continuous variables will be reported using means and standard deviations or medians and interquartile ranges (IQR), as appropriate.

Kaplan-Meier methods will be used to compare the time-to-event distribution of the outcome 'time until disease progression or death' between patients treated with the CapOx-B regimen and those treated with the CapOx regimen. Crude Kaplan-Meier cumulative incidence curves will be estimated separately for patients initiating the CapOx-B regimen and for those initiating the CapOx regimen. Time will be measured from the day after treatment initiation (index date) until the first occurrence of disease progression or death.

The cumulative incidence (absolute risk) of disease progression or death will be estimated from the Kaplan-Meier curves at pre-specified time points of 52 and 104 weeks for each treatment group, together with 95% confidence intervals.

IPTW Kaplan-Meier curves will also be estimated

We will also conduct descriptive analyses to characterise censoring patterns overall and across treatment groups. This will include median (IQR) time to censoring overall and according to the reason for censoring. This will be estimated separately for the overall study population and by treatment arm (CapOx-B vs. CapOx).

Reasons for censoring will include:

- Administrative censoring: reaching the maximum follow-up period 104 weeks or the end of the study period (31 December 2024).
- End of data availability: last recorded healthcare encounter, database end date, or practice withdrawal.
- Loss to follow-up: de-registration from the contributing practice or migration out of the healthcare system.

We will also compare baseline characteristics between eligible population and resulting study population after PS based trimming.

7.6.7 Core Emulation Table – Estimation Summary

Table 15. Core Emulation Table - Estimation Summary Estimand 1

		Target Trial	Target Trial Emulation	Comment
Analysis Method		Cox Proportional Hazards model to estimate the hazard ratio for time to first disease progression or death. Randomisation ensures balance in measured and unmeasured confounders.	Weighted Cox model (e.g. IPTW) to estimate marginal HR Preferably 1:1 matching by ECOG PS and liver as a metastatic site ensures comparability at treatment initiation. Alternatively, IPTW based on propensity score estimation may be applied.	Cox model selected to reflect treatment policy estimand. Preferably 1:1 matching, but alternatively IPTW, used to emulate randomisation in observational data. In case of the latter, trimming of PS represents a departure from the original target trial but is considered best practice when using propensity score methods in emulation. By removing patients in regions of non-overlap, the analysis is restricted to a population where treatment assignment is comparable across groups. As a result, the estimated effect no longer applies to the entire population but to this more comparable subset.

				If some patients have an extremely low probability of receiving one of the treatments, valid causal contrasts cannot be identified for them. Without trimming, effect estimates in these regions rely on unsupported extrapolation, making the results unstable and potentially biased.
Missing Data Assumptions and Methods to Handle		Outcome: assumes non-informative censoring conditional on treatment, and survival time; censored participants contribute partial information. Exposure: N/A (trial monitoring ensures exposure data completeness) Covariates: minimized through trial data collection.	Outcome: same assumption, covariates included in condition are different (1:1 matching or alternatively PS matching, included). Exposure: missing administration data for CapOx(-B) reflect true treatment discontinuation. Covariates: absence of a diagnosis code will be interpreted as absence of the condition. Missings for disease-related variables will be imputed using multiple imputation by chained equations (MICE) under the MAR assumption.	Mechanisms of missing exposure, covariate and outcome data differ between target trial and target trial emulation. Multiple imputation would not occur for missing covariate data in target trial.
Statistical Model Assumptions		Proportional hazards assumption for Cox model. Censoring is non-informative (given assumption re: missing outcome data)	Same: proportional hazards assumption assessed with Schoenfeld residuals and log(-log) plots. IPTW assumptions: no unmeasured confounding, positivity, correct model specification, consistency.	Diagnostics confirm appropriateness of Cox model; violations addressed in supplemental estimands and analyses (RMST). Some assumptions for IPTW are difficult to verify (e.g., unmeasured confounding).
Sensitivity Analyses		None	IPCW: varies conditions of the CAR assumption.	Potential for outcome misclassification only present in emulation since disease

			Tipping Point Analysis: conducted under the missing not at random (MNAR) assumption. Probabilistic Bias Analysis: Monte Carlo simulation to assess impact of non-differential exposure misclassification. See table 14 for more detail	progression is assessed differently.

Table 16. Core Emulation Table – Estimation Summary Estimand 2

	Target Trial	Target Trial Emulation	Comment
Analysis Method	Cox Proportional Hazards model to estimate the hazard ratio for time to first disease progression or death. Randomisation ensures balance in measured and unmeasured confounders.	Weighted Cox model (e.g. IPTW) to estimate marginal HR Preferably 1:1 matching by ECOG PS and liver as a metastatic site ensures comparability at treatment initiation. Alternatively, IPTW based on propensity score estimation may be applied.	Cox model selected to reflect treatment policy estimand. Preferably 1:1 matching, but alternatively IPTW, used to emulate randomisation in observational data.
Missing Data Assumptions and Methods to Handle	Outcome: assumes non-informative censoring conditional on treatment, and survival time; censored participants contribute partial information. Exposure: N/A (trial monitoring ensures exposure data completeness)	Outcome: same assumption, covariates included in condition are different (1:1 matching or alternatively PS matching, included). Exposure: missing administration data for CapOx(-B) reflect true treatment discontinuation. Covariates: absence of a diagnosis code will be interpreted as absence of	Mechanisms of missing exposure, covariate and outcome data differ between target trial and target trial emulation. Multiple imputation would not occur for missing covariate data in target trial.

	Covariates: minimized through trial data collection.	the condition. Missings for disease-related variables will be imputed using multiple imputation by chained equations (MICE) under the MAR assumption.	
Statistical Model Assumptions	Proportional hazards assumption for Cox model. Censoring is non-informative (given assumption re: missing outcome data)	Same: proportional hazards assumption assessed with Schoenfeld residuals and log(-log) plots. IPTW assumptions: no unmeasured confounding, positivity, correct model specification, consistency.	Diagnostics confirm appropriateness of Cox model; violations addressed in supplemental estimands and analyses (RMST). Some assumptions for IPTW are difficult to verify (e.g., unmeasured confounding).
Sensitivity Analyses	Not applicable	Only applied in primary analysis (Estimand 1)	Sensitivity analyses (e.g., IPCW, tipping point, probabilistic bias analysis) only used for primary Cox model analysis, not for Estimand 2.

Table 17. Core Emulation Table – Estimation Summary Estimand 3

	Target Trial	Target Trial Emulation	Comment
Analysis Method	Accelerated failure time (AFT) model with Weibull distribution	1:1 matched (or alternatively IPT-weighted) AFT model with Weibull distribution, followed by estimation of RMST at 52 and 104 weeks.	Preferably 1:1 matching, but alternatively IPTW, used to emulate randomisation in observational data.
Missing Data Assumptions and Methods to Handle	Outcome: assumes non-informative censoring conditional on treatment, and survival time; censored	Outcome: same assumption, covariates included in condition are different (1:1 matching or alternatively PS matching, included).	Mechanisms of missing exposure, covariate and outcome data differ between target trial and target trial emulation. Multiple imputation would

	participants contribute partial information. Exposure: N/A (trial monitoring ensures exposure data completeness) Covariates: minimized through trial data collection.	Exposure: missing administration data for CapOx(-B) reflect true treatment discontinuation. Covariates: absence of a diagnosis code will be interpreted as absence of the condition. Missings for disease-related variables will be imputed using multiple imputation by chained equations (MICE) under the MAR assumption.	not occur for missing covariate data in target trial.
Statistical Model Assumptions	Weibull survival distribution; log-linear relationship between covariates and log survival time.	Same; assessed using diagnostics such as log(-log(S(t))) vs log(t) for Weibull assumption and Q-Q plot for residuals.	-
Sensitivity Analyses	Not applicable	Only applied in primary analysis (Estimand 1)	Sensitivity analyses (e.g., IPCW, tipping point, probabilistic bias analysis) only used for primary Cox model analysis, not for AFT-based Estimand 3.

7.7 Data sources

7.7.1 Data sources

Rationale for selection and feasibility:

The NCR collects data with the primary aim of enabling several stakeholders to reflect on and improve oncological and palliative care [18]. The NCR compiles clinical data (i.e., hospital inpatient and outpatient data) of all individuals newly diagnosed with cancer in the Netherlands. A record in the NCR is triggered by biopsies from the national pathology database (PALGA) or the registration of a cancer diagnosis in the national registration of hospital care (Landelijke Basisregistratie Ziekenhuiszorg; LBZ).

A group of data managers daily screen for new information of the patients registered in the NCR. The relevant data for each tumour (and the corresponding patient) is registered after a set amount of time, typically 6-12 months after diagnosis. Registration of patients is typically done a year

after the incidence date. The vital status of patients is once a year. The data managers are often specialised in a specific type of cancer: colorectal cancer in this case study. The data managers strictly follow data extraction instructions.

NCR will be used to conduct this target trial emulation because critical data elements are readily available and fairly reliable, with reservations regarding a design element endpoint. Data is expected to be received within 2 months after the data request. A data recency of approximately 12 months is reasonable for the research question. Moreover, the expected sample size of 440 patients is achievable. The NCR recorded 22,192 patients aged ≥ 70 years with metastatic colon cancer between 2005 and 2020, of whom 23% received targeted therapy [19].

Strengths of data source(s):

The strength of the NCR lies in the coverage of the whole of the Netherlands and of critical variables. For instance, information on variables such as mCRC diagnosis, age, date of death is available for 100% of the patients. Administration data is available for the first-line treatment. With an estimated sample size of 440 patients, the NCR should easily suffice, since it recorded over 5,000 mCRC patients (aged ≥ 70 years) treated with targeted therapy between 2005 and 2020 [19].

Limitations with potential impact in the study results:

Potentially major limitations of using the NCR as a data source are the following. First of all, the primary endpoint, PFS, is not directly provided. Instead, an algorithm of prognostic markers is used to predict PFS. Secondly, the ECOG PS seems to be missing in 15% of the data. Lastly, the median length of follow-up per patient (with any type of cancer) is approximately 9 months. This variation is likely non-differential, meaning it is not expected to bias the results for any particular cancer group. If the patients included in this study have a longer survival time, the registry will allow for the required follow-up of 104 weeks. A minor limitation is the fact that data is registered with a lag of 6-12 months after diagnosis. Another minor quality issue is the fact that some variables are only registered in certain regions, which is the case for pregnancy for instance.

Data Quality:

The NCR is maintained by the Netherlands Comprehensive Cancer Organisation (IKNL). [20] It is the only nationwide oncological hospital registry, with data availability from 1989 onwards. Data is recorded by registration employees of IKNL, resulting in reliable and objective data.

Databases' suitability and case-study feasibility assessments followed three key steps: (I) characterisation of data source systems and processes, using the EMA data quality checklist to evaluate foundational aspects and their maturity; (II) assessment of data quality metrics for each data source (data reliability), based on published research and open-access catalogues; and (III) fitness-for-use evaluation (data relevance), assessing database suitability for each case study based on question-specific determinants. Steps 1 and 2 were database-specific, while step 3 was both database- and case-specific, i.e., it could only be assessed in view of the specific research question to be addressed. From these steps, two tables containing qualitative information (I and III) and one with quantitative metrics (II) were created. The overall feasibility of the case studies using the candidate data

sources was determined by critically analysing the collected information. Additional insights were gathered from DEAPs. All of the information was compiled into a report accompanying the generated tables, with our narrative assessment (appendix).

The overall feasibility is summarised in table 18. The NCR was deemed a feasible data source for studying CapOx(-B) and PFS, with achievable sample sizes and reasonably up-to-date data. The estimated sample size of 440 participants is supported by the NCR's coverage of 22,912 mCRC patients (aged ≥ 70 years) between 2005 and 2020, of whom 23% (~ 5000 patients) treated with targeted therapy. The database provides reliable in- and outpatient hospital data and timely updates, but some limitations were identified. For instance, PFS is not captured in the NCR through medical imaging directly, but through a validated algorithm that marks the end of "episodes", as previously discussed. The missingness in ECOG PS will not have much impact, since ECOG PS is one of the inclusion criteria.

Overall, these minor limitations are manageable within the study design and do not prevent the study from being feasible.

Table 18. Overall feasibility assessment summary for CS10 using NCR.

Case study	RWD source	Sample size estimation from the hypothetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessment	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
10 (Capecitabine with Oxaliplatin (CapOx) plus Bevacizumab versus CapOx in patients with Metastatic Colorectal Cancer)	NCR	With an approximate estimated sample size of 440 individuals (based on a 1:1 ratio between treatment arms, comparing CAPOX plus bevacizumab versus CAPOX alone), and considering that the Netherlands Cancer Registry (NCR) recorded 22,192 patients aged ≥70 years with metastatic colon cancer between 2005 and 2020—of whom 23% received targeted therapy—the target	Yes, with limitations on a design element	Elements with high criticality are available and fairly reliable, **with reservations regarding a design element endpoint**. The time elapsed from when a user requests the data to when they actually receive it is 2 months. Data recency is ~12 months before extraction, reasonably enough for the research question. Sample size is achievable.	·Potentially major: Progression-free survival (key endpoint) is not directly provided, although an algorithm using prognostic markers has been used in this database to predict PFS. ·Potentially major: ECOG is missing in 15% of the patients included in the DAP. ·Potentially major: The median length of	Although PFS is not directly available, a previously developed algorithm using prognostic markers has been applied in this database to estimate PFS. Missing ECOG data may prevent us from including certain subjects. Although the median follow-up time in the NCR is 9 months, this includes patients with all types of cancer with different

		sample size is anticipated to be reached. [19]			follow-up per patient is approximately 9 months. ·Minor: Some cancer patients do not have a biopsy and pathology, but might be picked by diagnostic code. ·Minor: Only prescription of first line of treatment is available, but cancer stage changes mean a new first treatment line is started; so, we will be able to identify previous treatments. ·Minor: Data is registered 6-12 months after diagnosis so there is a lag. ·Minor: Imaging information to assess progression-free survival is not available, only death is captured ·Minor: Procedure codes are available, but cancer-related surgery might only be picked if a specific code is available.	survival durations. However, this variation is likely non-differential, meaning it is not expected to bias the results in favour of or against any particular cancer group. If the patients included in the study have a longer survival time, the registry will allow for the follow-up required by protocol.

Table 19. Metadata about data sources and software

	Data 1
Data Source(s):	Netherlands Cancer Registry (NCR)
Study Period:	January 1st 2021 - December 31st 2024
Eligible Cohort Entry Period:	January 1st 2021 – December 31st 2022
Data Version (or date of last update):	NCR updates are daily. However, data is registered 6-12 months after diagnosis so there is a lag. Vital status is checked once per year.
Data sampling/extraction criteria:	Patients ⩾ 18 with mCRC initiating CapOx-B or CapOx; ⩾ 1 year lookback and 2 years follow-up
Type(s) of data:	Hospital data, administration data from individual patients' electronic health records and death records
Data linkage:	NCR is linked to the CBS for data regarding deaths. The last linkage was performed in early 2025
Conversion to CDM*:	Data is registered in OMOP CDM
Software for data management:	RANK, developed and maintained by the in-house Software Development Department. The changes in the database are loaded into a datawarehouse (DWH) every night.

*CDM = Common Data Model

7.8 Data management

The study will be conducted in a distributed manner using the UMCU, ARS Toscana and VAC4EU tools, procedures, and pipeline. Figure 2 specifies the data sets (D) and transformation processes (T), programming follows this pipeline, with involvement of different types of experts.

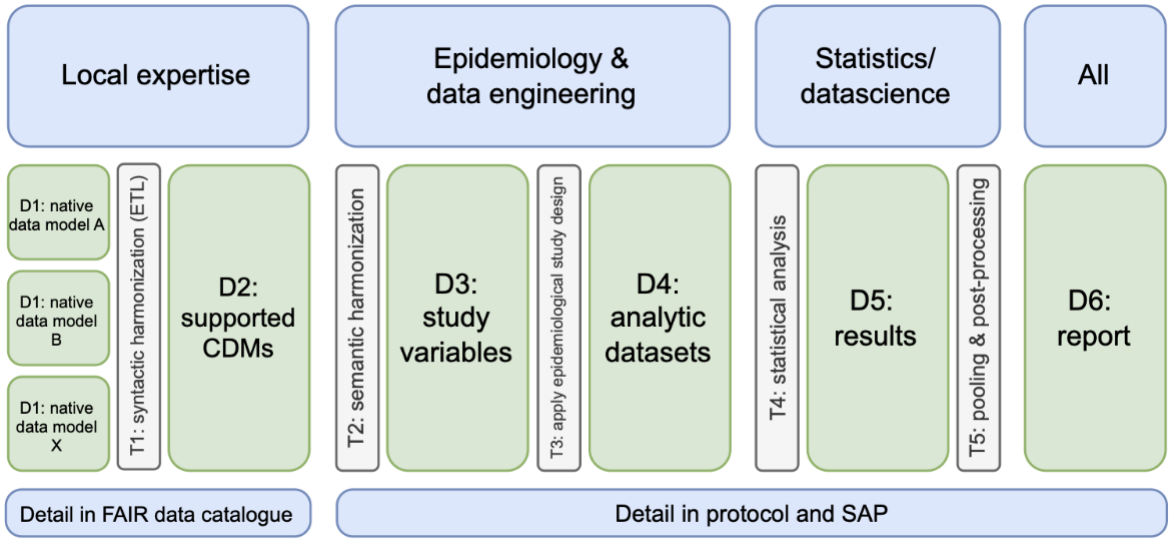


Figure 2. Data Management from the data transformation perspective

D1: Original data can be in any native format

The RWD-RWE pipeline used by VAC4EU starts with data banks that are controlled by the Data Expert and Access Partner (DEAP) and can be in any format. Data always stays local and never leaves the secure environments of the DEAPs. The ETL (extract, transform, load, see below for more details under 'T1') design is shared in a searchable FAIR VAC4EU catalogue. The VAC4EU FAIR Molgenis data catalogue is a meta-data management tool designed to contain searchable meta-data describing organisations that can provide access to specific data sources.d**T1: Syntactic harmonisation (ETL)**dT1: Syntactic harmonisation is conducted through an extraction, transformation, and loading (ETL) process of native data into the ConcePTION common data model (CDM) (see section 'D2: Common data model'). To harmonise the structure of the data sets stored and maintained by each data partner, a shared syntactic foundation is used. The ETL process has various structured steps as described by Thurin et al [21]:dDEAPs are asked to share the data dictionaries of their data banks (selected tables and variable names/structure)

- Metadata (descriptive data about the data sources and databanks) & data dictionaries, are uploaded in FAIR data catalogue (Molgenis).

D2: Common data model

For this project, the Common Data Model (CDM) (D2) is the Observational Medical Outcomes Partnership (OMOP) common data model. The CDM version that is used is v5.4, which is available as an open-source CDM [22]

T2: Semantic harmonisation

During the T2 step, many data transformations occur related to the completion of missing features in the data. Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g., medications), one or more phenotype algorithms are constructed (typically one sensitive, or broad, algorithm and one specific, or narrow, algorithm) to operationalise the identification and measurement of each event. In this step we conduct time anchoring (observation periods, look back periods), clean the data such as the dose of vaccines, sort on record level, aggregate across multiple records, and combine concepts for implantation of algorithms, and rule-based creation of study variables.,In this phase of the creation of study variables, semantic mapping is conducted. This semantic mapping across different vocabularies is conducted as part of the R-study script using different functionalities. To reconcile differences between different terminologies and native data availability, machine-readable code lists are used that comprise the terminologies that are used in the network (e.g. ICD-9, ICD10, SNOMED, ICPC and DEAP specific adaptations). This is combined with the BRIDGE metadata file that defines risk windows, look-back periods, and algorithms for each study variable [23].,**D3: Study variables** ,D3 datasets are interim data sets with information on study variables for each study participant, the unit may be a person, a medicine, or episode of time. The design of these datasets is described in codebooks. Examples of D3 datasets are the outputs of the ConcePTION pregnancy algorithm (https://github.com/IMI-ConcePTION/ConceptionTools/wiki#conception-pregnancy-algorithm), and outputs of functions that define smoking. Multiple functions/packages exist within the VAC4EU, for different study variables.

T3: Application of epidemiological design

In the T3 step epidemiological designs are applied such as sampling, matching (on specific variables and/or propensity scores), and selection based on inclusion and exclusion criteria using the study variables in the D3 datasets. The designs will be implemented for the various study objectives using R-scripts, and these may use the existing functions (R-cran) or functions that have been developed in the VAC4EU community (e.g. matching).

D4: Analytical data set

D4 is an analytical dataset, and multiple D4 data sets may be produced based on the objectives of the study. The format is described initially in a code book for communication between programmers and statisticians.

T4: Statistical analysis

This step in the data transformation pipeline will produce statistical estimates such as descriptives (counts, percentages), distributions (mean, percentiles), rates (prevalence, incidence), regression coefficients, or other relevant estimates. This will be conducted using R.

D5: Results

D5 is the set of estimates, tables or aggregate data that is transferred from the DEAPs to the Digital Research Environment (DRE). The aggregated results produced by these scripts at the DEAP's site will be uploaded to the UMCU DRE for post-processing, pooling and visualisation (Figure 2). The DRE is a cloud-based, globally available research environment where data are stored and organised securely and where researchers can collaborate. The DRE is made available through UMCU. The DRE applies double authentication where researchers can collaborate using data that are stored and organised securely [24]. UMCU is responsible for data processing and data security.

All researchers who need access to the DRE will be granted access to study-specific secure workspaces by UMCU. Access to the workspaces will be possible only after double authentication using an identification code and password together with the user's mobile phone for authentication.

Uploading files will be possible for all researchers with access to the workspace within the DRE. Downloading of files will be possible only after requesting and receiving permission from a workspace member with an "owner" role, who will be a UMCU team member.

T5: Post-processing/pooling.

In this step, the result from different DEAPs is pooled and converted into tables and figures for reporting.

7.9 Quality control

All key study documents such as the hypothetical trial protocol, target trial emulation protocol and study reports will undergo senior scientific and editorial review.

Data quality

For quality control of the data instance from the Netherlands Cancer Registry, we will use the Observational Health Data Sciences and Informatics (OHDSI) DataQualityDashboard [25]. This tool performs more than 3,000 standardized checks on a populated OMOP CDM instance. Its objective is to evaluate the quality of observational data in a systematic, reproducible, and transparent manner.

The quality checks are structured according to the Kahn Framework, which defines categories and contexts representing different strategies for assessing data quality. The DataQualityDashboard implements 24 core check types within this framework, which are systematically applied across all relevant tables and fields of the OMOP CDM [26].

Code Quality

These coding practices define how the TARGET programming team collaborates to write clean, reliable, and reproducible code for the VAC4EU Real-World Evidence (RWE) Analytical Pipeline. They aim to ensure clarity, consistency, and maintainability across all case studies within the project.

Coding conventions

To ensure clarity, consistency, and maintainability across the project, the following conventions will be applied to all codebases within the project:

- Consistent style: Code follows a consistent and readable style (see the tidyverse style guide for R).
- Meaningful names: Use clear, descriptive names for variables, functions, and files to convey their purpose.
- Modular code: Break down code into small, reusable functions where possible.
- No hardcoded paths: Use configuration files or relative paths to ensure portability.

Following these conventions makes the code easier to understand, test, and reuse across case studies and teams.

Documenting Code

Code documentation is used to promote good coding practices and ensure our work is understandable, maintainable, and reproducible. To achieve this, we will:

- Use descriptive comments that explain the purpose and rationale behind code sections, focusing on why something is done, not just what.
- Clearly document function inputs, outputs, and side effects, using standardised formats (e.g., roxygen2 in R) where appropriate and supported.
- Write meaningful variable and function names to make the code as self-explanatory as possible.

Version Control

We use Git and GitHub to manage version control. These tools support good coding practices by enabling collaboration, tracking changes, accessing a project's history, and ensuring code quality through review and documentation.

A dedicated GitHub organisation has been created for the project (https://github.com/target-roc19). Each case study is managed in its own repository within this organisation. Repositories are structured consistently across case studies, to reinforce modularity. Access to repositories is controlled through teams.

During development, all repositories remain private to ensure confidentiality. Once the project is finalised, relevant repositories will be made public and assigned a digital object identifier (DOI) via Zenodo to support transparency, reproducibility, and reuse by the wider research community.

To maintain code quality and clarity, we follow the git and GitHub guidelines below.

- Always use pull requests (PRs): never push directly to the main branch.
- Open an issue before creating a new branch. Ideally, one PR resolves one issue to keep changes focused and reviewable.
- Every PR must be reviewed by at least one other person before merging.
- The PR author merges the PR after it has been reviewed and approved.
- Write clear, descriptive commit messages.
- Write informative PR descriptions, including:
    - A concise title
    - Links to related issues
    - A summary of the changes

Continuous Integration

Continuous Integration (CI) is set up to automatically check code quality and run tests whenever changes are pushed to the repository or submitted through a pull request (PR). The CI workflow ensures that the package adheres to predefined style guidelines and that all automated tests pass before changes are merged.

Coding Template

Every case study follows the general coding template used across all code in the TARGET project. The folder structure is organised as follows:

```
case-study-template
|____data
| |____D2_cdm
| |____D3_study_variables
| |____D4_analytic_datasets
| |____D5_results
| |____D6_report
|____docs
|____logs
```

|____run
|____tests
|____transformations
| |____T2_semantic_harmonisation
| |____T3_study_design
| |____T4_statistical_analysis
| |____T5_processing_results
|____CHANGELOG.md
|____LICENSE
|____README.md

Project Data Structure and Storage

The data folder follows the Real-World Evidence pipeline structure. Data conforming to the common data model is stored in the D2_cdm folder.

Results from transformations T2, T3, T4, and T5 are saved in the respective folders:

- D3_study_variables
- D4_analytic_datasets
- D5_results
- D6_report

Each dataset is associated with a codebook, explained in more detail below.

All data remain securely stored on the Data Expert and Access Partners (DEAPs) servers and are never transferred externally. For testing purposes, dummy datasets are created. These fall into two categories:

- Unit test data: Small, predefined input and output pairs used to test individual transformation steps. These are stored in the tests folder, not in data, and can support automated testing.
- Pipeline test data: Larger, more complex dummy datasets used to test whether the full pipeline runs as expected. These may be included in the repository only if they remain below GitHub's 100 MiB file size limit and will otherwise be shared via SharePoint.

Logging System

When the pipeline is executed, log files are saved in the logs folder. These logs are especially helpful when running the code in the DEAPs environment, as they help trace and diagnose potential errors. We recommend using the logger R package to handle logging throughout the pipeline. A sample logging setup can be found in the logger.R script located at the root of the project directory.

Executing the Analytical Pipeline

The run folder contains scripts used to execute each transformation step in the pipeline.

- A central script, run_pipeline.R, orchestrates the full pipeline from start to finish.
- Subscripts (e.g., run_T2.R or similar) are available to run individual transformation steps separately.

Typically, the run_pipeline.R script is the main entry point used by a DEAP to execute the full pipeline. Before running it in the DEAP environment, the pipeline may need to be adapted to local settings. This can be done using a configuration file that defines variables required to tailor the pipeline to a specific DEAP. Please note that configuration files should not include sensitive information.

Such a file might include variables like:

- The name of the DEAP
- The path to the local data instance
- The path to any required external resources

Testing and Quality Assurance

The tests folder contains scripts to test the analytical pipeline. Tests will be used to ensure code behaves as expected and remains stable over time. By systematically checking inputs, outputs, and edge cases, tests help catch errors early and make future changes safer. We use the testthat R package to structure and run unit tests.

Continuous integration (CI) is used to automate testing. With CI, tests are automatically run each time code is pushed to the repository (e.g., via GitHub Actions). This helps identify issues immediately, ensures that new changes do not break existing functionality, and supports better collaboration by enforcing consistent code quality across contributors.

Modular Data Transformation Workflow

The transformations folder follows the Real-World Evidence pipeline structure. It contains the source code for all transformation steps, which is typically written in R. Each subfolder corresponds to a specific step in the pipeline (e.g., T2_semantic_harmonisation, T3_study_design, T4_statistical_analysis, T5_processing_results) and includes the relevant scripts and helper functions for that step.

During the T2 step, a database is usually created (e.g., using DuckDB). This database can be queried using SQL, and it is recommended that all SQL queries be saved as clearly named, standalone SQL script files to ensure readability and reusability.

The purpose of the transformations folder is to structure and modularise the processing logic, making it easier to maintain, test, and reuse across different case studies. By organising code by transformation step, teams can work in parallel, increasing efficiency.

Changelog

A changelog will be kept for all notable changes in the project. Changelogs help track the evolution of the project over time, making it easier for collaborators to understand what has changed between versions. We follow the structure and best practices outlined in Keep a Changelog.

Codebooks

Before developing code, codebooks are created to describe each dataset (D) within the pipeline. A codebook is a comprehensive document that outlines the structure, contents, and metadata of a dataset. It serves as a detailed reference guide for anyone working with the data and plays a crucial role in guiding the development of the analytical pipeline by clearly defining both the inputs and expected outputs.

All codebooks are summarised in a central index file, which provides a high-level overview of the pipeline's structure. For each codebook, the index file includes:

- A brief description of its purpose,
- A list of the scripts used to generate the corresponding dataset,
- A description of the input datasets and input parameters required.

The datasets D2, D3, D4, and D5 are typically subdivided into multiple smaller transformation steps, each detailed within their respective codebooks. These smaller transformation steps ensure that each part of the pipeline is clearly scoped and well-documented.

In addition to supporting development, codebooks help ensure quality control by making transformation logic transparent and verifiable, and they enhance reproducibility by documenting exactly how data is structured and used throughout the analytical pipeline.

Deployment

The analytical pipeline is delivered to DEAPs as a GitHub release, tagged with a version number. Versioning follows the format: YYYYYMMDD.XX, where the date indicates the release date and XX denotes the sub-version or revision number.

Any deployment issues can be reported via the GitHub repository using the issues feature, where the programming team responsible for the R code will collaborate with the local DEAP to resolve them as needed.

Reproducibility

It is recommended to locally use the renv R package to maintain the R version and version of packages for reproducibility purposes.

At this time, however, using renv reliably across different systems and environments remains challenging. For this reason, we currently recommend its use only in local development setups.

We are actively monitoring developments in the R ecosystem related to cross-platform reproducibility. As soon as a more stable and portable solution becomes available, we will revisit this guidance and promote broader adoption.

Open Source Licensing

The code will be made available under an open source license.

README Guidelines

Each case study repository includes a README that covers the following points:

- Project Overview: brief summary of the study goals and key research questions.
- Background: context and rationale for the study.
- Repository Structure: Outline of main folders and their contents.
- Data Overview: Description of data sources, formats, and data privacy considerations.
- How to Run: Instructions for running the pipeline and key scripts, plus where outputs are saved.
- Testing: How to run tests to verify code functionality.
- Contributing: Guidelines for code contributions and issue tracking.
- License: Information about the code license.
- Contact: Who to reach out to for help or questions.

7.10 Study precision

Sample size estimation from the hypothetical trial protocol:

In the target trial, the sample size was calculated for a hypothesis test as follows: Assuming a two-sided alpha of 0.05, 90% power, a 1% loss to follow up and median PFS of 8 months in the CapOx arm, the number of patients required to detect a HR of 0.83 is (assuming 70% experience event during follow up), is 440 (i.e. 220) per randomisation arm [6]. The target sample size is anticipated to be reached, considering that the NCR recorded 22,192 patients aged ≥ 70 years with metastatic colon cancer between 2005 and 2020, of which approximately 5,104 received targeted therapy [19].

Sample size estimation in this NIS protocol:

In this non-interventional study, no hypothesis test will be performed. The focus is on the precision of the estimated treatment effect. Assuming the study size in each RWD source will be similar to the sample size of the target trial, the precision is estimated as described below.

To estimate the level of precision that is achievable with a fixed sample size, we can estimate the expected width of the confidence interval (CI) for the effect estimate.

Estimation of the precision of the HR

To estimate the expected 95% CI for a hazard ratio (HR) from a Cox proportional hazards model, the standard error (SE) of the log(HR) is derived from the total number of events.

Assumptions

- Equal allocation to treatment groups
- Large-sample normal approximation for log(HR)
- Symmetric CI on the log scale

The confidence interval width for the HR can be calculated using the following formula:

$CI\_width\_HR = exp(\hat{\beta} + 1.96 \times SE) - exp(\hat{\beta} - 1.96 \times SE)$

Where:

- $\hat{\beta}$ is the log hazard ratio (log(HR))
- SE is the standard error of the log(HR)

– 1.96 is the z-score for a 95% confidence interval

Calculation of 95% CI

1. Assume equal allocation:
Number of events per group: d1 = d2 = d / 2

2. Calculate SE of log(HR):
SE[log(HR)] = sqrt(1/d1 + 1/d2) = sqrt(2/d)

3. Construct the 95% CI on log scale:
log(HR) ± 1.96 × SE[log(HR)]

4. Convert back to HR scale:
CI_HR = exp(log(HR) ± margin)

Scenario 1
Calculation based on 308 events, HR = 0.83
SE = sqrt(2 / 308) = 0.08058

Log(HR) = log(0.83) = -0.1863

Margin = 1.96 x SE = 1.96 x 0.08058 = 0,15794
Lower bound = -0.1863 - 0.15794 = -0.34427
Upper bound = -0.1863 + 0.15794 = -0.53060

Lower CI = exp(-0.34427) = 0.709
Upper CI = exp(-0.53060) = 0.972

% Precision= (0.97/0.83) -1 = 16.87%

Scenario 2
Calculation under the assumption that the overall event rate for disease progression/death is 10% lower than expected (277), HR = 0.83(277), HR = 0.83
SE = sqrt(2 / 277) = 0.08497

Log(HR) = log(0.83) = -0.1863

Margin = 1.96 x SE = 1.96 x 0.08497 = 0,166545
Lower bound = -0.1863 - 0.166545 = -0.35287
Upper bound = -0.1863 + 0.166545 = -0.01978

Lower CI = exp(-0.35287) = 0.703
Upper CI = exp(-0.01978) = 0.980

% Precision= (0.98/0.83) -1 = 18.07%

Scenario 3
Calculation under the assumption that the overall event rate for disease progression/death is 30% lower than expected (215), HR = 0.83
SE = sqrt(2 / 215) = 0.09645

Log(HR) = log(0.83) = -0.1863

Margin = 1.96 x SE = 1.96 x 0.09645 = 0,18904
Lower bound = -0.1863 - 0.18904 = -0.37537
Upper bound = -0.1863 + 0.18904 = 0.00271

Lower CI = exp(-0.37537) = 0.687
Upper CI = exp(0.00271) = 1.003
% Precision= (1.00/0.83) -1 = 20.48%

Table 20. Power and sample size

Scenario	Number of Events	Hazard Ratio (HR)	Log(HR)	Standard Error (SE)	Margin of Error	Lower CI (HR)	Upper CI (HR)	Precision
Scenario 1	308	0.83	-0.1863	0.08058	0.15794	0.709	0.972	16.87%
Scenario 2	277	0.83	-0.1863	0.08497	0.16655	0.703	0.980	18.07%
Scenario 3	215	0.83	-0.1863	0.09645	0.18904	0.687	1.003	20.48%

8. Limitation of the methods

There are certain limitations to this protocol that should be noted, mainly regarding emulation departures.

First, the PFS measurement is based on the occurrence of disease progression or death. In this target trial emulation, the occurrence of disease progression is assessed differently from the disease progression assessment often used in clinical trials and clinical practice. As also done in Saltz et al, disease progression is assessed through imaging. Imaging data is not available from the NCR. In the NCR, disease progression is measured through the end of the first "episode". An "episode" can consist of one or multiple regimens. If the first regimen is discontinued and switched to a second regimen due to toxicity, this is still part of the first "episode". The end of an "episode" is only marked if the reason for discontinuation or switching is disease progression. This difference in measurement may result in differences in the observed PFS; the time till disease progression or death is expected to be longer in this non-interventional study (NIS) compared to the RCT by Saltz et al [6]. But assuming a high degree of specificity, the HR would not be expected to differ.

Second, since capecitabine is administered orally on day 1-14 of each cycle in the outpatient setting, we cannot be sure about the patients' (in)adherence. This means that we cannot be entirely sure about the patients' exposure to capecitabine. For capecitabine, the first day of the last cycle is registered as the stop date. Hence, we know that the patient initiated the last cycle, but we don't know if the last cycle was also finished completely. Hence, unknown (in)adherence of capecitabine is expected to have limited impact on the identification of the intercurrent event partial discontinuation.

Third, there are some deviations in the in- and exclusion criteria as well. The target trial inclusion criterion that a patient should not be amenable for curative resection is not applied in the emulation because we cannot identify patients who had the intention to undergo curative resection. The inclusion criterion of a life expectancy $\geq$ 3 months is not applied in the emulation because this information is not recorded at baseline. Besides, including patients who do not survive past 3 months will take away immortal time bias and represent the real-world effectiveness better. Information on hematologic/clotting, hepatic and renal function is not available so this inclusion criterion is not applied in the emulation. Because of the new user active comparator study design, the initiation of CapOx(-B) treatment is an inclusion criterion that is applied in the emulation but not in the target trial. In the target trial, pregnant/breastfeeding patients and patients with serious nonhealing wounds and ulcers are excluded. These exclusion criteria are not applied in the emulation because this information is not available. This is not expected to be problematic because the use of CapOx(-B) is contra-indicated if one of these conditions is present. Hence, the drugs are unlikely to be used in such patients in clinical practice.

Finally, for estimand 2, several strategies to handle local treatment as an intercurrent event have been explored but all seemed to have some limitations. The composite strategy seemed to be the most suitable. Tumour shrinkage, due to systemic treatment, must occur before local treatment can be applied. When local treatment occurs, the patient is assumed to not have died or experienced disease progression up to the date of administrative censoring. Hence, the patient is followed up for the whole hypothetical study period. It should be taken into account that the occurrence

of disease progression or death are regarded as clinically negative, whereas the occurrence of local treatment can be regarded as clinically positive. As in every observational study, there is a risk of residual confounding.

9. Protection of human subjects

This is a non-interventional study using secondary data collection and does not pose any risks for individuals. Each data source research partner will apply for an independent ethics committee review according to local regulations. Data protection and privacy regulations will be observed in collecting, forwarding, processing, and storing data from study participants.

Patient information

This study involves data that exists in an anonymised structured format and contains no patient personal information. All parties will comply with all applicable laws, including laws regarding the implementation of organisational and technical measures to ensure the protection of patient personal data. Such measures will include omitting patient names or other directly identifiable data in any reports, publications, or other disclosures, except where required by applicable laws. Patient personal data will be stored at DAPs in encrypted electronic form and will be password protected to ensure that only authorised study staff have access. DAPs will implement appropriate technical and organisational measures to ensure that personal data can be recovered in the event of a disaster. In the event of a potential personal data breach, DAPs shall be responsible for determining whether a personal data breach has in fact occurred and, if so, providing breach notifications as required by law.

Patient consent

As this study does not involve data subject to privacy laws according to applicable legal requirements, obtaining informed consent from individuals is not required.

10. Reporting of adverse events

For studies in which the research team uses only data from automated healthcare databases, according to the International Society for Pharmacoepidemiology Guidelines for GPP, “Aggregate analysis of database studies can identify an unexpected increase in risk associated with a particular exposure. Such studies may be reportable as study reports, but typically do not require reporting of individual cases. Moreover, access to automated databases does not confer a special obligation to assess and/or report any individual events contained in the databases. Formal studies conducted using these databases should adhere to these guidelines.” For non-interventional study designs that are based on secondary use of data, such as studies based on medical chart reviews or electronic health records, systematic reviews, or meta-analyses, reporting of adverse events/adverse drug reactions is not required. Reports of adverse events/adverse drug reactions should only be summarised in the study report, where applicable. According to the EMA Guideline on GVP, Module VI – Management and Reporting of Adverse Reactions to Medicinal Products, “All adverse events/reactions collected as part of [non-interventional post-authorisation studies with a design based on secondary use of data], the submission of suspected adverse reactions in the form of [individual case safety reports] is not required. All adverse events/reactions collected for the study should be recorded and summarised in the interim safety analysis and in the final study report.” Module VIII – Post-Authorisation Safety Studies

echoes this approach. Legislation in the EU further states that for certain study designs such as retrospective cohort studies, particularly those involving electronic health records, it may not be feasible to make a causality assessment at the individual case level.

11. References

[1] European Cancer Information System. Estimates of cancer incidence and mortality in 2022, for all countries n.d. https://ecis.jrc.ec.europa.eu/explorer.php?$0-0$1-All$4-1,2$3-16$6-0,85$5-2022,2022$7-7,8$2-All$CEstByCountry$X0_8-3$X0_19-AE27$X0_20-No$CEstBySexByCountry$X1_8-3$X1_19-AE27$X1_-1-1$CEstByIndiByCountry$X2_8-3$X2_19-AE27$X2_20-No$CEstRelative$X3_8-3$X3_9-AE27$X3_19-AE27$CEstByCountryTable$X4_19-AE27 (accessed September 12, 2025).

[2] Integraal Kankercentrum Nederland (IKNL). Cijfers darmkanker n.d. https://iknl.nl/kankersoorten/darmkanker/registratie (accessed December 1, 2023).

[3] Morris VK, Kennedy EB, Baxter NN, Benson AB 3rd, Cercek A, Cho M, et al. Treatment of Metastatic Colorectal Cancer: ASCO Guideline. J Clin Oncol 2023;41:678–700.

[4] Cervantes A, Adam R, Roselló S, Arnold D, Normanno N, Taïeb J, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2023;34:10–32. https://doi.org/10.1016/j.annonc.2022.10.003.

[5] Colorectaal carcinoom (CRC). 2021.

[6] Saltz LB, Clarke S, Díaz-Rubio E, Scheithauer W, Figer A, Wong R, et al. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: A randomized phase III study. Journal of Clinical Oncology 2008;26:2013–9. https://doi.org/10.1200/JCO.2007.14.9930.

[7] Razenberg LGEM, van Gestel YRBM, de Hingh IHJT, Loosveld OJL, Vreugdenhil G, Beerepoot L V., et al. Bevacizumab for metachronous metastatic colorectal cancer: A reflection of community based practice. BMC Cancer 2016;16:11. https://doi.org/10.1186/s12885-016-2158-8.

[8] Integraal Kankercentrum Nederland. Registratie n.d. https://iknl.nl/nkr/registratie (accessed September 12, 2025).

[9] Nederlandse Vereniging voor Medische Oncologie. Adviezen commissie BOM - Bevacizumab als eerstelijns behandeling van het colorectaal carcinoom. vol. 23. 2005. https://doi.org/10.1200/JCO.2005.05.112.

[10] European Medicines Agency (EMA). Avastin n.d. https://www.ema.europa.eu/en/medicines/human/EPAR/avastin (accessed November 9, 2023).

[11] Baraniskin A, Buchberger B, Pox C, Graeven U, Holch JW, Schmiegel W, et al. Efficacy of bevacizumab in first-line treatment of metastatic colorectal cancer: A systematic review and meta-analysis. Eur J Cancer 2019;106:37–44. https://doi.org/10.1016/j.ejca.2018.10.009.

[12] Nederlandse Vereniging voor Medische Oncologie. Adviezen commissie BOM -Herbeoordeling: bevacizumab bij gemetastaseerd colorectaalcarcinoom. 2008.

[13] Zwart K, van der Baan FH, Cohen R, Aparicio T, de la Fouchardiére C, Lecomte T, et al. Prognostic value of Lynch syndrome, BRAFV600E, and RAS mutational status in dMMR/MSI-H metastatic colorectal cancer in a pooled analysis of Dutch and French cohorts. Cancer Med 2023;12:15841–53. https://doi.org/10.1002/CAM4.6223.

[14] Kahan BC, Hindley J, Edwards M, Cro S, Morris TP. The estimands framework: A primer on the ICH E9(R1) addendum. BMJ 2024. https://doi.org/10.1136/bmj-2023-076316.

[15] European Medicines Agency. Summary of Product Characteristics - Oxaliplatin. n.d.

[16] European Medicines Agency. Summary of Product Characteristics - Xeloda. n.d.

[17] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. BMC Med Res Methodol 2017;17. https://doi.org/10.1186/s12874-017-0442-1.
[18] Integraal Kankercentrum Nederland. About IKNL n.d.
[19] Baltussen JC, de Glas NA, Liefers GJ, Slingerland M, Speetjens FM, van den Bos F, et al. Time trends in treatment patterns and survival of older patients with synchronous metastatic colorectal cancer in the Netherlands: A population-based study. Int J Cancer 2023;152:2043–51. https://doi.org/10.1002/ijc.34422.
[20] Integraal Kankercentrum Nederland. Netherlands Cancer Registry (NCR) n.d.
[21] Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther 2022;111:321–31. https://doi.org/10.1002/cpt.2476.
[22] OMOP CDM v5.4 n.d. https://ohdsi.github.io/CommonDataModel/cdm54.html (accessed February 26, 2026).
[23] Royo AC, Elbers JHJ R, Weibel D, Hoxhaj V, Kurkcuoglu Z, Sturkenboom MCJ, et al. Real-World Evidence BRIDGE: A Tool to Connect Protocol With Code Programming. Pharmacoepidemiol Drug Saf 2024;33. https://doi.org/10.1002/pds.70062.
[24] anDREa. Bridging the Gap n.d. https://andrea-cloud.com/our-story/ (accessed September 12, 2025).
[25] Observational Health Data Sciences and Informatics. DataQualityDashboard n.d. https://github.com/ohdsi/DataQualityDashboard (accessed February 26, 2026).
[26] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association 2021;28:2251–7. https://doi.org/10.1093/jamia/ocab132.