

Research Article

Survival–Toxicity Trade-Offs Between Trifluridine/Tipiracil and Regorafenib in Refractory Metastatic Colorectal Cancer: A Simulation-Based Reinforcement Learning Study

 Perihan Perkin,  Fahriye Tuğba Köş

Department of Medical Oncology, Ankara Bilkent City Hospital, Ankara, Türkiye

Abstract

Objectives: We examined whether survival and grade 3–4 toxicity could be jointly modeled for trifluridine/tipiracil (FTD-TPI) and regorafenib in refractory metastatic colorectal cancer using reinforcement learning (RL).

Methods: A synthetic cohort of 1,039 virtual patients was generated from later-line trials and allocated to FTD-TPI (n=534) or regorafenib (n=505). Survival was simulated with Weibull distributions incorporating age, ECOG performance status, and KRAS status. Grade 3–4 adverse events were modeled using published probabilities. A one-step Q-learning model integrated survival and toxicity through a prespecified reward function. Cox models descriptively compared survival distributions.

Results: Median simulated survival was 6.91 months with FTD-TPI and 6.59 months with regorafenib. The unadjusted hazard ratio was 0.893 (95% CI, 0.790–1.008; p=0.068), and the adjusted hazard ratio was 0.905 (95% CI, 0.801–1.023; p=0.110). Mean modeled rewards were –18.38 and –25.24, respectively. The learned policy mapped 68.5% of profiles to FTD-TPI.

Conclusion: FTD-TPI showed a more favorable modeled survival–toxicity profile, but the survival difference was not statistically significant and clinical superiority was not established. The framework is hypothesis-generating rather than a validated decision-support tool.

Keywords: Colorectal neoplasms, computer simulation, regorafenib, reinforcement machine learning, trifluridine tipiracil drug combination

Cite This Article: Perkin P, Köş FT. Survival–Toxicity Trade-Offs Between Trifluridine/Tipiracil and Regorafenib in Refractory Metastatic Colorectal Cancer: A Simulation-Based Reinforcement Learning Study. EJMI 2026;10(4):601–609.

Metastatic colorectal cancer (mCRC) remains a major cause of cancer-related mortality despite continued advances in systemic therapy.^[1] For patients whose disease progresses after fluoropyrimidine-, oxaliplatin-, and irinotecan-based therapy and appropriate biologic or molecularly targeted treatment, the therapeutic goal shifts toward prolonging survival while preserving function and quality of life. In this refractory setting, treatment benefit is often modest, and the clinical consequences of toxicity become increasingly important.

Trifluridine/tipiracil (FTD-TPI; formerly known as TAS-102) and regorafenib are established later-line treatment options in previously treated mCRC and demonstrated overall survival (OS) benefit over best supportive care (BSC) in the RECURSE and CORRECT trials, respectively.^[2,3] No head-to-head randomized trial has directly compared the two monotherapies, and treatment selection has therefore relied on indirect evidence, patient characteristics, prior therapy, performance status, comorbidities, and the different toxicity profiles of the drugs.^[4,5] The later-line treat-

Address for correspondence: Perihan Perkin, MD. Department of Medical Oncology, Ankara Bilkent City Hospital, Ankara, Türkiye

Phone: +90 505 884 48 04 **E-mail:** perihanperkin@gmail.com

Submitted Date: August 4, 2026 **Revised Date:** September 15, 2026 **Accepted Date:** September 25, 2026

©Copyright 2026 by Eurasian Journal of Medical Investigation - Available online at www.ejmi.org

OPEN ACCESS This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.



ment landscape has since evolved further, including the addition of bevacizumab to FTD-TPI and the introduction of fruquintinib.^[6,7] Nevertheless, FTD-TPI and regorafenib remain clinically relevant benchmark agents for examining how efficacy and tolerability can be weighed when direct comparative evidence is unavailable.

Existing indirect comparisons and clinical reviews may inform this choice, but their interpretation is limited by differences in eligibility criteria, baseline risk, treatment exposure, follow-up, and adverse-event reporting.^[4,5] This is particularly important when between-treatment differences in survival are small, because toxicity, dose intensity, treatment continuity, and patient preference may materially influence the perceived net benefit of therapy. A modeling framework that considers efficacy and toxicity together may therefore provide a useful way to explore the assumptions underlying treatment prioritization, provided that its outputs are clearly distinguished from clinical evidence.

Reinforcement learning (RL) is a computational approach in which alternative actions are evaluated according to a reward function, allowing the model to learn action values or policies within a defined environment.^[8–10] In healthcare, RL has been explored for both sequential and context-specific treatment decision modeling under uncertainty.^[9–12] When applied to synthetic data, however, RL does not discover new clinical effects; rather, it explores the consequences of the assumptions, outcome distributions, and reward structure encoded in the simulated environment. This distinction is essential when interpreting simulation-based treatment models.

In the present study, we used refractory mCRC as a clinical use case to examine whether survival and treatment-related toxicity could be integrated within a single simulation-based RL framework. A common synthetic patient cohort was generated from published aggregate trial characteristics, and treatment-specific survival and toxicity outcomes were modeled for FTD-TPI and regorafenib. A one-step Q-learning approach was then used to explore how prespecified survival–toxicity trade-offs influenced modeled treatment preferences. The objective was not to determine which drug is clinically superior but to provide a transparent, hypothesis-generating framework for evaluating survival–toxicity trade-offs in a setting without direct randomized comparative evidence.

Materials and Methods

Study Design and Overview

This was a fully simulation-based methodological study using FTD-TPI and regorafenib as a clinical use case in refractory mCRC. The objective was to explore how simulated OS

and severe treatment-related toxicity could be integrated within an RL framework rather than to estimate real-world comparative effectiveness or establish clinical superiority between the two treatments.

The analytical workflow included generation of a common synthetic patient cohort, simulation of treatment-specific survival and toxicity outcomes, construction of a prespecified reward function, descriptive survival analyses, and implementation of a one-step Q-learning model. All outputs therefore represent the behavior of the specified virtual environment and remain conditional on the input distributions and modeling assumptions.

Computational Environment and Reproducibility

The final analysis was implemented in Python. NumPy and pandas were used for random number generation and data structuring, SciPy for probability-distribution functions, statsmodels for survival analyses, and Matplotlib for visualization. The Q-learning algorithm was implemented using custom Python routines. Fixed random seeds were used to ensure reproducibility, and the complete simulated patient-level dataset, model parameters, and analysis code were retained for auditability.

Synthetic Patient Cohort

A common synthetic cohort of 1,039 virtual patients was generated. Baseline distributions were derived by sample-size-weighted pooling of aggregate characteristics reported for the active-treatment populations of the pivotal RECURSE and CORRECT trials.^[2,3]

Age was generated from a truncated normal distribution with a mean of approximately 62.5 years, a standard deviation of 9 years, and prespecified bounds of 22–82 years. The pooled cohort distributions were approximately 45.5% for Eastern Cooperative Oncology Group (ECOG) performance status 1, 61.0% for male sex, 53.4% for KRAS-mutant disease, and 65.4% for a colon primary tumor. The proportions of patients who had received two or fewer, three, and four or more previous treatment lines were approximately 22.4%, 23.5%, and 54.2%, respectively.

The variables included age, sex, ECOG performance status, previous treatment line, KRAS mutation status, and primary tumor site. Sex, treatment line, and primary tumor site were included to characterize the synthetic cohort but were not used as modifiers of survival or as components of the RL state.

All synthetic variables were stored in a pandas DataFrame structure for subsequent analyses.

In practical terms, 1,039 represents the sum of the two active-treatment sample sizes (534 in RECURSE and 505 in

CORRECT), not a pooled individual-patient dataset. For each reported categorical characteristic, the corresponding trial-level proportions were combined using these sample sizes as weights, and virtual characteristics were generated to approximate the resulting pooled marginal distributions. Ages were generated separately from the prespecified truncated normal distribution described above. Each resulting record represented a newly generated virtual patient rather than an actual participant from either trial. Because only aggregate information was used, the construction should not be assumed to reproduce the original within-patient correlations among baseline characteristics. The common virtual cohort was subsequently permuted and divided into the two treatment arms, while potential outcomes for both treatments were generated for the RL environment.

Treatment Allocation

After generation of the common synthetic cohort, patients were randomly allocated to the two virtual treatment arms using an exact allocation ratio corresponding to the active-treatment sample sizes of the pivotal trials. A total of 534 patients were assigned to FTD-TPI and 505 to regorafenib.

This allocation was performed using a random permutation of the common cohort and was not stratified according to baseline characteristics. The procedure was intended to construct a virtual comparison and should not be interpreted as clinical trial randomization.

Overall Survival Modeling

Treatment-specific OS was simulated using Weibull distributions. A common Weibull shape parameter of 1.5 was used as a prespecified modeling assumption.

The value of 1.5 was specified as an illustrative survival-shape assumption: in a Weibull model, a shape parameter greater than 1 permits an increasing hazard over time, whereas a value of 1 corresponds to a constant hazard. It was not estimated by fitting patient-level data or reconstructed Kaplan–Meier curves from RECURSE or CORRECT and should not be regarded as an empirically validated description of either treatment arm.

Treatment-specific median OS calibration targets were derived from the treatment and placebo median OS values reported in RECURSE and CORRECT.^[2,3] The mean of the two placebo medians, 5.15 months, was used as a common reference. The calibration targets were calculated as follows:

$$M_j^* = \frac{M_{\text{treatment},j}}{M_{\text{placebo},j}} \times 5.15$$

This ratio-based approach informed the prespecified simulation targets of 6.91 months for FTD-TPI and 6.59 months

for regorafenib. Because the published trial medians were rounded, these numerical targets should be regarded as approximate calibration inputs, not as estimates from a formal adjusted indirect treatment comparison.

For each virtual patient and each treatment option, an action-specific OS time was generated according to:

$$T_{ij} = \lambda_{ij} [-\ln(U_{ij})]^{1/k}$$

where T_{ij} represents the simulated OS for patient i under treatment j , U_{ij} is a random value drawn from a uniform distribution, k is the Weibull shape parameter, and λ_{ij} is the patient- and treatment-specific scale parameter.

To introduce prognostic heterogeneity without altering the marginal treatment calibration targets, the Weibull scale parameter was modified using centered age, ECOG performance status, and KRAS mutation status:

$$\lambda_{ij} = \lambda_{0j} \exp\{\beta_{\text{age}}[\text{Age}_i - \overline{\text{Age}}] + \beta_{\text{ECOG}}[\text{ECOG}_i - \overline{\text{ECOG}}] + \beta_{\text{KRAS}}[\text{KRAS}_i - \overline{\text{KRAS}}]\}$$

The prespecified coefficients were:

- age: $\beta_{\text{age}} = -0.015$ per year;
- ECOG performance status 1: $\beta_{\text{ECOG}} = -0.20$;
- KRAS-mutant disease: $\beta_{\text{KRAS}} = -0.10$.

These coefficients were used as simplified modeling assumptions to introduce clinically plausible variation and should not be interpreted as validated prognostic effect estimates.

The coefficients were fixed on the log-scale of the Weibull time parameter to introduce directional variation around the centered cohort characteristics, rather than to reproduce trial-derived covariate effects. Under the specified equation, an additional year of age multiplies the time scale by $\exp(-0.015) \approx 0.985$, ECOG performance status 1 by $\exp(-0.20) \approx 0.819$, and KRAS-mutant status by $\exp(-0.10) \approx 0.905$, with other variables held constant. These numerical choices are illustrative assumptions; their magnitudes were not estimated or validated using individual patient data and should not be interpreted as observed hazard ratios or clinically established prognostic effects.

Following application of the covariate modifiers, a treatment-specific multiplicative calibration factor was applied so that the observed median OS in each randomly assigned virtual treatment arm matched its prespecified target. No censoring or treatment crossover was modeled, and all simulated OS times were treated as observed events.

Adverse Event Modeling

For every virtual patient, potential grade 3–4 treatment-related adverse-event outcomes were generated separately under FTD-TPI and regorafenib using Bernoulli sampling.

The prespecified probabilities were 0.49 for FTD-TPI, as reported for treatment-related grade 3–4 adverse events in the FDA clinical review of RECOURSE^[13], and 0.54 for regorafenib, as reported in CORRECT.^[3]

Severe toxicity was represented as a single binary variable. The model did not reproduce the type, timing, duration, reversibility, cumulative burden, or management of individual adverse events. Toxicity occurrence was simulated independently of baseline characteristics and simulated survival.

For the virtual-trial survival analyses, only the adverse-event and survival outcomes corresponding to the randomly allocated treatment were considered observed. The action-specific potential outcomes under both treatments were retained for construction of the RL environment.

Reward Function

A prespecified reward function was used to integrate simulated survival and severe toxicity. The common OS reference was 6.75 months, corresponding to the mean of the two treatment-specific target medians.

The survival component of the reward was defined as follows:

- OS at least 15% above the common reference: +100;
- OS within $\pm 15\%$ of the common reference: +20;
- OS at least 15% below the common reference: -100.

A grade 3–4 treatment-related adverse event resulted in an additional penalty of -50. No treatment-discontinuation penalty was included because a reproducible discontinuation model was not available.

The resulting reward for each patient-treatment combination was therefore:

$$\text{Reward}_{ij} = \text{SurvivalReward}_{ij} - 50 \times \text{SevereAE}_{ij}$$

The reward values were predefined for modeling purposes and were not intended to represent validated clinical utility weights, quality-adjusted survival, or patient preferences. The magnitude and direction of the model outputs were consequently dependent on the selected reward structure.

Reinforcement Learning Framework

A one-step episodic tabular Q-learning model was implemented. The model was formulated as a contextual treatment-choice problem rather than as a longitudinal dynamic treatment regime.

The pretreatment state was defined using:

- age younger than 65 years versus 65 years or older;
- ECOG performance status 0 versus 1;
- KRAS wild-type versus mutant status.

These variables generated eight possible pretreatment states. The two available actions were FTD-TPI and regorafenib. Simulated OS and adverse-event outcomes were not included in the pretreatment state. No external clinical validation dataset was used.

The model was trained over 1,000 episodes, with each episode representing one complete pass through the synthetic patient profiles. An epsilon-greedy strategy with an exploration probability of 0.10 was used. Because each episode consisted of a single treatment decision without a subsequent clinical state, the discount factor was set to zero.

Q-values were updated using a sample-average learning rate:

$$\alpha_t = \frac{1}{N_t(s, a)}$$

where $N_t(s, a)$ represents the cumulative number of visits to a specific state-action pair. Under this one-step structure, the learned Q-value approximates the mean modeled reward associated with each treatment within each pretreatment state.

The learned policy assigned the treatment with the higher Q-value to each of the eight states. The proportion of synthetic patient profiles mapped to each treatment was then calculated according to the prevalence of these states in the simulated cohort.

Outcome Measures and Statistical Analysis

The virtual-trial survival outputs included treatment-specific median OS, simulated Kaplan-Meier curves, and a log-rank comparison. Cox proportional hazards models were fitted to summarize separation between the simulated treatment distributions. An unadjusted model included treatment alone, while the adjusted model additionally included age and ECOG performance status.

Hazard ratios with 95% confidence intervals and two-sided p values were reported. Because the treatment-specific survival distributions were prespecified and calibrated during simulation, the Cox estimates were interpreted as descriptive model outputs rather than independent validation or causal comparative treatment-effect estimates.

The RL outputs included:

- mean potential reward under FTD-TPI and regorafenib;
- state-specific Q-values;
- the treatment preferred within each modeled pretreatment state;
- the proportion of synthetic profiles mapped to each treatment by the learned policy.

No permutation test or inferential statistical comparison of reward scores was performed.

Sensitivity analyses examined the dependence of the model outputs on the reward specification. The severe-toxicity penalty was varied between -25, -50, and -75, while the survival threshold was varied between $\pm 10\%$, $\pm 15\%$, and $\pm 20\%$ of the common OS reference.

Ethical Considerations

Since this study used only synthetic data generated from published aggregate statistics and did not involve human participants, identifiable information, or individual patient-level data, ethics committee approval and informed consent were not applicable.

Results

Baseline Characteristics of the Synthetic Treatment Cohorts

A common synthetic cohort of 1,039 virtual patients was generated and randomly allocated to FTD-TPI (n=534) or regorafenib (n=505). The two simulated treatment cohorts were broadly comparable with respect to age, ECOG performance status, sex, previous treatment exposure, KRAS

mutation status, and primary tumor site (Table 1). Median age was 63 years (range, 38–82) in the FTD-TPI cohort and 62 years (range, 29–82) in the regorafenib cohort. ECOG performance status 1 was present in 44.0% and 46.9% of patients, respectively, while KRAS-mutant disease was present in 54.3% and 52.5%. These values describe the realized composition of the synthetic cohorts after random allocation and were not subjected to inferential baseline-comparison testing.

Simulated Overall Survival and Severe Toxicity

By design, the treatment-specific survival models reproduced the prespecified calibration targets. Median simulated overall survival was 6.91 months for FTD-TPI and 6.59 months for regorafenib. The Kaplan-Meier curves showed substantial overlap between the treatment cohorts (Fig. 1). The log-rank comparison did not reach statistical significance ($p=0.068$). In the descriptive Cox model, the unadjusted hazard ratio for FTD-TPI relative to regorafenib was 0.893 (95% CI, 0.790–1.008; $p=0.068$). After adjustment for age and ECOG performance status, the hazard ratio was 0.905 (95% CI, 0.801–1.023; $p=0.110$) (Table 2). Because the treatment-specific survival distributions were prespecified and calibrated within the simulation, these estimates summarize separation within the modeled environment and do not represent independent comparative-effectiveness estimates.

The observed simulated rate of grade 3–4 treatment-related adverse events was 46.6% in the FTD-TPI arm and 53.9%

Table 1. Baseline characteristics of the simulated treatment cohorts

Characteristic	FTD-TPI (n=534)	Regorafenib (n=505)
Age, median (range), years	63 (38–82)	62 (29–82)
ECOG performance status, n (%)		
0	299 (56.0)	268 (53.1)
1	235 (44.0)	237 (46.9)
Sex, n (%)		
Male	329 (61.6)	305 (60.4)
Female	205 (38.4)	200 (39.6)
Previous treatment lines, n (%)		
≤2	120 (22.5)	112 (22.2)
3	129 (24.2)	115 (22.8)
≥4	285 (53.4)	278 (55.0)
KRAS status, n (%)		
Mutant	290 (54.3)	265 (52.5)
Wild-type	244 (45.7)	240 (47.5)
Primary tumor site, n (%)		
Colon	353 (66.1)	327 (64.8)
Rectum	181 (33.9)	178 (35.2)

ECOG: Eastern Cooperative Oncology Group; KRAS: Kirsten rat sarcoma viral oncogene homolog.

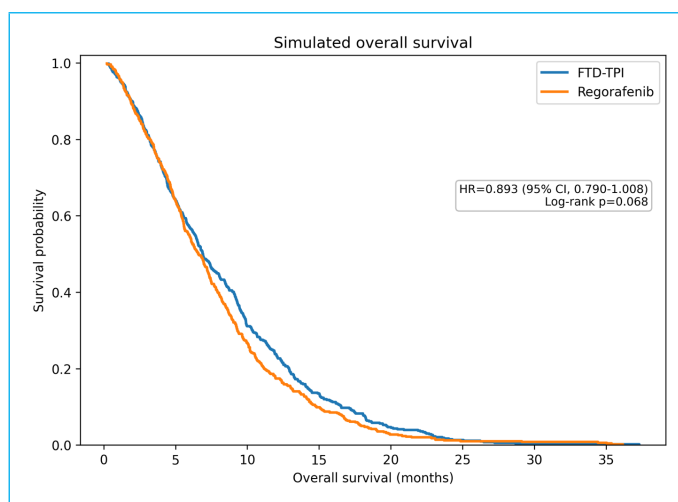


Figure 1. Simulated Kaplan-Meier curves for overall survival in the FTD-TPI and regorafenib cohorts.

The curves represent modeled overall survival in the randomly allocated virtual treatment arms. The hazard ratio and log-rank p value summarize separation within the simulated environment and should not be interpreted as independent comparative clinical-effect estimates.

Table 2. Primary simulated survival, toxicity, and reinforcement-learning outputs

Outcome	FTD-TPI	Regorafenib	Comparative/model output
Median simulated OS, months	6.91	6.59	-
Grade 3-4 treatment-related AE, n (%)	249 (46.6)	272 (53.9)	-
Unadjusted Cox model	-	-	HR 0.893 (95% CI, 0.790–1.008); p=0.068
Adjusted Cox model*	-	-	HR 0.905 (95% CI, 0.801–1.023); p=0.110
Log-rank comparison	-	-	chi-square 3.335; p=0.068
Mean potential reward	-18.38	-25.24	-
Profiles mapped by learned policy, %	68.5	31.5	FTD-TPI preferred in 5 of 8 states

*Adjusted for age and ECOG performance status. AE: Adverse event; CI: Confidence interval; HR: Hazard ratio; OS: Overall survival. All estimates are outputs of the prespecified simulation model; FTD-TPI: Trifluridine/tipiracil.

in the regorafenib arm (Table 2). These realized rates were close to the treatment-specific probabilities used to generate the toxicity outcomes.

Reinforcement Learning Outcomes

Under the primary reward specification, the mean modeled reward was –18.38 for FTD-TPI and –25.24 for regorafenib (Table 2). The negative mean values reflect the numerical scale of the prespecified reward function, which assigned penalties for short survival and severe toxicity; they should not be interpreted as measures of net clinical harm. The distribution of individual modeled rewards is shown in Figure 2. The learned policy preferred FTD-TPI in five of the eight modeled pretreatment states and regorafenib in three.

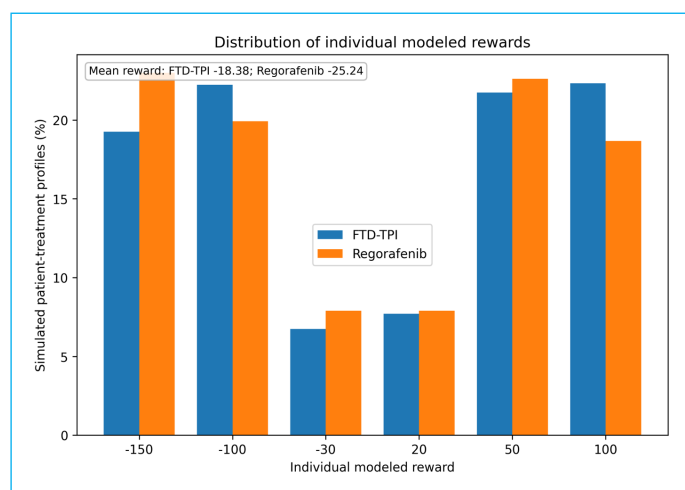


Figure 2. Distribution of individual modeled rewards under the primary reward specification.

Bars show the percentage of simulated patient-treatment profiles assigned to each discrete reward value for FTD-TPI and regorafenib. Reward values combine the prespecified survival categories and the penalty for a simulated grade 3–4 treatment-related adverse event.

When weighted according to the prevalence of these states in the synthetic cohort, 68.5% of simulated patient profiles were mapped to FTD-TPI and 31.5% to regorafenib. State-specific Q-values and treatment preferences are provided in Supplementary Table S1. These proportions characterize policy behavior within the virtual environment and should not be interpreted as the proportion of real patients expected to benefit from either treatment.

In particular, the 68.5% allocation is a frequency of assignments produced by the learned policy under the specified synthetic cohort, outcome distributions, and reward function; it is neither a treatment-effect estimate nor evidence that 68.5% of patients in clinical practice would benefit from, or should receive, FTD-TPI.

Sensitivity Analyses

Across all nine combinations of survival thresholds ($\pm 10\%$, $\pm 15\%$, and $\pm 20\%$ of the common OS reference) and severe-toxicity penalties (-25, -50, and -75), the mean modeled reward remained higher for FTD-TPI than for regorafenib. Depending on the survival threshold, the learned policy mapped 68.5%–77.8% of simulated profiles to FTD-TPI and 22.2%–31.5% to regorafenib. Varying the toxicity penalty changed the magnitude of the mean rewards but did not alter the number of states assigned to each treatment at a given survival threshold (Supplementary Table S2).

Discussion

This study used refractory metastatic colorectal cancer as a clinical use case to examine how simulated overall survival and severe treatment-related toxicity could be integrated within a one-step RL framework. Under the primary reward specification, FTD-TPI had a higher mean modeled reward than regorafenib, and the learned policy mapped 68.5% of the simulated patient profiles to FTD-TPI. How-

ever, the simulated survival curves showed substantial overlap, and neither the unadjusted nor the adjusted Cox analysis demonstrated statistically significant separation between the treatment distributions. The principal finding is therefore not that FTD-TPI is clinically superior, but that a modest difference in modeled survival combined with a lower prespecified probability of severe toxicity produced a model-based preference for FTD-TPI within the defined virtual environment.

The clinical question remains relevant because later-line mCRC is characterized by limited expected survival, cumulative treatment burden, and considerable heterogeneity in treatment tolerance. The pivotal RECURSE and CORRECT trials established FTD-TPI and regorafenib as active later-line therapies, respectively, but did not directly compare the two agents.^[2,3] Indirect comparisons, clinical reviews, and real-world data have generally suggested broadly similar efficacy, with clinically distinct toxicity profiles.^[4,14–17] In such a setting, small differences in expected survival may be amplified or offset by differences in tolerability. The present model illustrates this principle by assigning both survival outcomes and grade 3–4 toxicity to a common reward structure.

The simulated median OS values were 6.91 months for FTD-TPI and 6.59 months for regorafenib, reflecting the prespecified calibration targets. The corresponding hazard ratios numerically favored FTD-TPI but did not reach statistical significance. These estimates should not be interpreted as independent comparative-effectiveness findings because the treatment-specific survival distributions were calibrated before the Cox analyses were performed. Rather, the Cox models provide descriptive summaries of the separation generated within the simulation. The substantial overlap between the Kaplan–Meier curves is consistent with the limited absolute difference between the modeled median survival values.

The RL component extended this descriptive survival comparison by incorporating severe toxicity into the treatment evaluation. The observed simulated grade 3–4 adverse-event rates were 46.6% for FTD-TPI and 53.9% for regorafenib, closely reflecting the treatment-specific probabilities used to generate these outcomes. Under the primary reward specification, the mean modeled reward was –18.38 for FTD-TPI and –25.24 for regorafenib. These negative values arise from the numerical structure of the reward function, which applies penalties for short survival and severe toxicity, and should not be interpreted as measures of net clinical harm. The difference between the mean rewards instead indicates that FTD-TPI generated a relatively more favorable balance of the modeled outcomes under

the selected assumptions.

The learned policy preferred FTD-TPI in five of the eight modeled pretreatment states and regorafenib in three. After weighting these states according to their prevalence in the synthetic cohort, 68.5% of the simulated profiles were mapped to FTD-TPI and 31.5% to regorafenib. These percentages do not represent the proportions of real patients expected to benefit from either treatment. They summarize state-specific action values learned from synthetic potential outcomes generated by the model. Nevertheless, they demonstrate how relatively small differences in survival and toxicity can produce different treatment preferences across clinically defined patient profiles.

An important feature of the final framework is that the treatment-selection state included only variables available before treatment: age group, ECOG performance status, and KRAS mutation status. Simulated OS and adverse-event outcomes were used to calculate rewards but were not included in the pretreatment state. This separation makes the framework conceptually clearer than a model that uses post-treatment outcomes as treatment-selection variables. However, the present analysis remains a one-step contextual treatment-choice model rather than a longitudinal dynamic treatment regime. It does not simulate repeated treatment decisions, dose modifications, disease progression, or changes in patient condition over time.

The sensitivity analyses provided additional information regarding the internal stability of the modeled preference. Across all nine combinations of survival thresholds and severe-toxicity penalties, the mean modeled reward remained higher for FTD-TPI than for regorafenib. Depending on the survival threshold, the learned policy mapped between 68.5% and 77.8% of the simulated profiles to FTD-TPI. Variation in the toxicity penalty changed the magnitude of the reward values but did not alter the number of states assigned to each treatment at a given survival threshold. These findings suggest that the direction of the modeled preference was stable across the examined reward specifications. However, this internal stability should not be regarded as external validation, because all sensitivity scenarios were evaluated within the same synthetic environment and depended on the same underlying outcome distributions.

RL offers a structured way to formalize multi-outcome treatment trade-offs because survival and toxicity can contribute simultaneously to a common reward signal.^[9–12] Its usefulness, however, depends on the validity of the data, assumptions, and reward definitions used to construct the environment. In the present study, the algorithm learned from outcomes generated by the simulator itself and there-

fore could not produce new comparative clinical evidence beyond the information encoded in the model. The value of the framework lies instead in making those assumptions explicit and in allowing investigators to examine how different representations of survival and toxicity influence modeled treatment prioritization.

The study has several methodological strengths. First, it uses a common synthetic cohort and explicitly separates pretreatment state variables from simulated post-treatment outcomes. Second, the complete analytical workflow was reconstructed within a single reproducible codebase, with fixed random seeds, retained patient-level synthetic data, and clearly specified model parameters. Third, the framework combines clinically interpretable outcomes within a transparent reward structure and examines the dependence of its conclusions on alternative reward specifications. Finally, the model is modular and could be extended to incorporate additional outcomes such as quality of life, treatment duration, dose intensity, time without symptoms or toxicity, cost, and patient preferences.

The clinical landscape of refractory mCRC has evolved since the original monotherapy trials. The SUNLIGHT trial demonstrated improved survival with FTD-TPI plus bevacizumab compared with FTD-TPI alone, while FRESCO-2 established fruquintinib as an additional later-line treatment option.^[6,7] Accordingly, the present comparison of FTD-TPI monotherapy and regorafenib should be regarded as a benchmark modeling scenario rather than a comprehensive representation of contemporary treatment sequencing. Future analyses could extend the framework to include newer treatment combinations and agents, provided that sufficiently detailed and comparable data are available.

The model therefore provides a restricted historical-monotherapy benchmark for studying reward-dependent allocation, not a comparison against FTD-TPI plus bevacizumab or fruquintinib and not a recommendation for contemporary treatment sequencing.

Several limitations warrant emphasis. The analysis relied entirely on synthetic data derived from aggregate trial-level information, and no individual patient-level clinical trial or real-world dataset was available for model development or external validation. Cross-trial calibration is vulnerable to differences in eligibility criteria, supportive care, assessment schedules, geographic populations, and baseline risk. The Weibull shape parameter and the prognostic coefficients for age, ECOG performance status, and KRAS mutation status were predefined modeling assumptions rather than estimates derived from the source trials. No censoring was modeled, and all simulated OS times were treated as observed events.

Sensitivity analyses varied the reward specification but did not vary the Weibull shape parameter or the prognostic coefficients; changes to these survival-model assumptions could also alter state-specific rewards and policy allocations. Toxicity was represented as a single binary grade 3–4 adverse-event outcome and did not capture the type, timing, duration, reversibility, cumulative burden, or clinical management of individual toxicities. The model also assumed that toxicity was independent of baseline patient characteristics and simulated survival. The reward values were predefined for modeling purposes and did not represent validated clinical utility weights, quality-adjusted survival, or patient preferences. Although sensitivity analyses examined several alternative reward specifications, other reasonable definitions could yield different state-specific preferences. Furthermore, the treatment-specific potential outcomes were generated from calibrated distributions rather than observed counterfactual outcomes, and the resulting Q-values therefore remain model-dependent.

Future work should prioritize the use of individual patient-level trials or real-world data, with independent development and validation cohorts. More clinically meaningful reward functions could incorporate quality-adjusted survival, time without symptoms or toxicity, treatment discontinuation, dose modification, financial burden, and patient-reported preferences. Longitudinal models could additionally represent repeated treatment decisions and changing clinical states. External validation and calibration assessment would be essential before any related framework could be considered for clinical decision support.

Conclusion

This simulation-based RL study provides a transparent and reproducible framework for jointly modeling survival and severe toxicity trade-offs in refractory mCRC. Under the prespecified assumptions, the model assigned a relatively more favorable reward profile to FTD-TPI and mapped a greater proportion of simulated pretreatment profiles to FTD-TPI than to regorafenib. However, the simulated survival difference was not statistically significant, and the model outputs do not establish clinical superiority between the treatments. The framework should therefore be viewed as a hypothesis-generating methodological tool that may inform the design of future analyses using validated patient-level data.

Disclosures

Ethics Committee Approval: Ethics approval was not required because this study used fully simulated patient cohorts derived from published aggregate trial data and did not involve real patient information.

Informed Consent: No individual patient-level clinical data were used.

Conflicts of Interest: The authors declared they have no competing interests.

Funding: The authors declared that they received no specific funding for this work.

Use of AI for Writing Assistance: The authors declared that during the preparation of this manuscript, ChatGPT (OpenAI) was used to assist with language editing, manuscript organization, and development of the simulation code. The authors reviewed the model assumptions, code, analyses, generated outputs, and manuscript and take full responsibility for the accuracy and integrity of the work.

Authorship Contributions: Concept – FTK; Design – FTK, PP; Supervision – FTK, PP; Data Collection or Processing: FTK, PP; Analysis or Interpretation: FTK, PP; Literature Search: PP; Writing: PP; Critical Review: FTK.

Peer-review: Externally peer-reviewed.

Acknowledgements: The authors thank their colleagues for constructive discussions during the development of this work.

References

1. Cremolini C, Chalabi M, Elez E, Fassin M, Gelli M, Goéré D, et al; ESMO Guidelines Committee. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol* 2026;37(6):759–76. [\[CrossRef\]](#)
2. Mayer RJ, Van Cutsem E, Falcone A, Yoshino T, Garcia-Carbonero R, Mizunuma N, et al; RECURSE Study Group. Randomized trial of TAS-102 for refractory metastatic colorectal cancer. *N Engl J Med* 2015;372(20):1909–19. [\[CrossRef\]](#)
3. Grothey A, Van Cutsem E, Sobrero A, Siena S, Falcone A, Ychou M, et al; CORRECT Study Group. Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet* 2013;381(9863):303–12. [\[CrossRef\]](#)
4. Abrahao ABK, Ko YJ, Berry S, Chan KKW. A comparison of regorafenib and TAS-102 for metastatic colorectal cancer: a systematic review and network meta-analysis. *Clin Colorectal Cancer* 2018;17(2):113–20. [\[CrossRef\]](#)
5. Bekaii-Saab T, Kim R, Kim TW, O'Connor JM, Strickler JH, Malika D, et al. Third- or later-line therapy for metastatic colorectal cancer: reviewing best practice. *Clin Colorectal Cancer* 2019;18(1):e117–29. [\[CrossRef\]](#)
6. Prager GW, Taieb J, Fakih M, Ciardiello F, Van Cutsem E, Elez E, et al; SUNLIGHT Investigators. Trifluridine-tipiracil and bevacizumab in refractory metastatic colorectal cancer. *N Engl J Med* 2023;388(18):1657–67. [\[CrossRef\]](#)
7. Dasari A, Lonardi S, Garcia-Carbonero R, Elez E, Yoshino T, Sobrero A, Yao J, et al; FRESCO-2 Study Investigators. Fruquintinib versus placebo in patients with refractory metastatic colorectal cancer (FRESCO-2): an international, multicentre, randomised, double-blind, phase 3 study. *Lancet* 2023;402(10395):41–53. [\[CrossRef\]](#)
8. Watkins CJCH, Dayan P. Q-learning. *Mach Learn* 1992;8(3–4):279–92. [\[CrossRef\]](#)
9. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018;24(11):1716–20. [\[CrossRef\]](#)
10. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. *Nat Med* 2019;25(1):16–8. [\[CrossRef\]](#)
11. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: comprehensive review. *J Med Internet Res* 2020;22(7):e18477. [\[CrossRef\]](#)
12. Coronato A, Naeem M, De Pietro G, Paragliola G. Reinforcement learning for intelligent healthcare applications: A survey. *Artif Intell Med* 2020;109:101964. [\[CrossRef\]](#)
13. U.S. Food and Drug Administration. Clinical Review: TAS-102 (Lonsurf), NDA 207981. 2015; p. 10. Reference ID: 3809894. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2015/207981Orig1s000MedR.pdf. Accessed September 17, 2026.
14. Sonbol MB, Benkhadra R, Wang Z, Firwana B, Walden DJ, Mody K, et al. A systematic review and network meta-analysis of regorafenib and TAS-102 in refractory metastatic colorectal cancer. *Oncologist* 2019;24(9):1174–9. [\[CrossRef\]](#)
15. Parmar A, Chan KKW, Ko YJ. Metastatic colorectal cancer: therapeutic options for treating refractory disease. *Curr Oncol* 2019;26(Suppl 1):S24–32. [\[CrossRef\]](#)
16. Byrne M, Saif MW. Selecting treatment options in refractory metastatic colorectal cancer. *Onco Targets Ther* 2019;12:2271–8. [\[CrossRef\]](#)
17. Moriwaki T, Fukuoka S, Taniguchi H, Takashima A, Kumekawa Y, Kajiwarra T, et al. Propensity score analysis of regorafenib versus trifluridine/tipiracil in patients with metastatic colorectal cancer refractory to standard chemotherapy (REGO-TAS): A Japanese society for cancer of the colon and rectum multicenter observational study. *Oncologist* 2018;23(1):7–15. [\[CrossRef\]](#)

SUPPLEMENTARY TABLES

Supplementary Table S1. State-specific Q-values and preferred treatment

Pretreatment state	n profiles	Q-value, FTD-TPI	Q-value, regorafenib	Preferred treatment	Profiles, %
Age <65, ECOG 0, KRAS wild-type	145	1.30	-10.62	FTD-TPI	14.0
Age <65, ECOG 0, KRAS mutant	173	-6.16	-15.14	FTD-TPI	16.7
Age <65, ECOG 1, KRAS wild-type	132	-11.63	-19.16	FTD-TPI	12.7
Age <65, ECOG 1, KRAS mutant	151	-17.26	-39.74	FTD-TPI	14.5
Age ≥65, ECOG 0, KRAS wild-type	111	-7.84	-21.25	FTD-TPI	10.7
Age ≥65, ECOG 0, KRAS mutant	138	-27.43	-22.70	Regorafenib	13.3
Age ≥65, ECOG 1, KRAS wild-type	96	-42.98	-39.94	Regorafenib	9.2
Age ≥65, ECOG 1, KRAS mutant	93	-55.45	-43.63	Regorafenib	9.0

Q-values are model-based mean reward estimates within each pretreatment state.

Supplementary Table S2. Sensitivity analyses across reward specifications

Toxicity penalty	Survival threshold	Mean reward, FTD-TPI	Mean reward, regorafenib	Profiles to FTD-TPI, %	Profiles to regorafenib, %	States to FTD-TPI	States to regorafenib
-25	±10%	-8.26	-13.01	77.8	22.2	6	2
-25	±15%	-6.45	-11.86	68.5	31.5	5	3
-25	±20%	-5.22	-9.16	68.5	31.5	5	3
-50	±10%	-20.19	-26.39	77.8	22.2	6	2
-50	±15%	-18.38	-25.24	68.5	31.5	5	3
-50	±20%	-17.15	-22.54	68.5	31.5	5	3
-75	±10%	-32.13	-39.77	77.8	22.2	6	2
-75	±15%	-30.32	-38.61	68.5	31.5	5	3
-75	±20%	-29.09	-35.92	68.5	31.5	5	3

The primary analysis used a toxicity penalty of -50 and a survival threshold of ±15%.