Simulating Ulcerative Colitis Clinical Trials Using Knowledge Graph–Enhanced Real-World Data Modeling

Validation Across 21 Studies

Flavio Dormont, PhD[1] · **Michael Shapiro, MD, MSc**[2] · **Amichai Perlman, PharmD, PhD**[2,5]
Amina Alaskarov, MSc[2] · **Annie J. Kruger, MD, PhD**[1] · **Bruce E. Sands, MD**[4]
Ramon Hernandez, MD, PhD[3*] · **Omri Matalon, PhD**[2*]

[1] Sanofi US, Cambridge, MA, USA

[2] QuantHealth, Tel Aviv, Israel
[3] Sanofi, Gentilly, France
[4] Dr. Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5] Department of Clinical Pharmacy, School of Pharmacy, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

Funded by Sanofi · Preprint, Research Square (not peer-reviewed)

ABSTRACT

BACKGROUND: Clinical trials are costly, time-consuming, and often yield uncertain outcomes. Predictive modeling using machine learning (ML), real-world data (RWD), and biomedical knowledge graphs offers new opportunities to improve translational efficiency and trial outcomes.

METHODS: This study validated an ML-based modeling framework, ClinBoost, which integrates de-identified RWD from insurance claims and electronic health records with a drug-centered biomedical knowledge graph. Patient severity was modeled using an electronic Mayo Score, and longitudinal data inputs were transformed into patient-drug journey embeddings. Model performance was evaluated against 21 randomized controlled trials using a robust validation framework with multiple performance metrics.

RESULTS: ClinBoost achieved high concordance with outcomes for 16 of the 21 trials, achieving an F1 score of 80%. The model demonstrated minimal bias in estimating treatment effect sizes. Additionally, the framework was able to refine trial design by identifying sub-cohorts with better response and increased statistical power.

CONCLUSIONS: These findings demonstrate the potential of the ClinBoost modeling approach, which combines RWD and knowledge graph-informed ML, to improve clinical trial outcomes, optimize study design, and accelerate drug development.

Keywords: *Machine learning; real-world data; knowledge graphs; graph neural networks; predictive modeling.*

INTRODUCTION

The development of new therapeutic agents is resource-intensive and often hindered by clinical trial challenges. Traditional randomized controlled trials, despite being considered the gold standard for evaluating drug efficacy and safety, are limited by high costs, lengthy durations, and recruitment difficulties. Predictive modeling has sought to address these challenges by shifting from traditional in vitro and in vivo studies to advanced in silico simulations, improving translational efficiency and accelerating drug discovery. Computational in silico models can simulate biological processes and help predict drug responses, providing a cost-effective and efficient exploration of new therapeutic mechanisms of action (MoA). These models help identify promising compounds based on their structure and anticipated interactions with other target biomarkers, simulate pharmacokinetic and pharmacodynamic behaviors, optimize dosing, and detect potential

adverse effects early. Such approaches can markedly reduce reliance on extensive in vivo studies and expedite drug candidate identification, yet their application to clinical-trial modeling and protocol optimization remains limited.

Recent legislative and technological advances, such as the European Health Data Space (EHDS) and the US HITECH and Cures Acts, have facilitated the integration of clinical real-world data (RWD) in drug development. Coupled with advanced machine learning (ML) methods, RWD-based models offer dynamic time-dependent predictions of patients' trajectories, crucial for management and decision-making in progressive or fluctuating diseases. These models leverage sophisticated ML techniques, enhancing predictions by parsing complex patient data and the causal frameworks influencing treatment response. Integration of knowledge graphs, encoding biological relationships, further enhances ML-driven predictive modeling. Knowledge graphs enable the representation of complex and interconnected biomedical processes in a way amenable to machine learning frameworks. This facilitates discovering novel drug-target interactions, repositioning of existing drugs, and enhancing the causal inference of clinical outcomes.

The integration of real-world data and knowledge graph–driven modeling holds particular promise for predicting clinical outcomes and optimizing trial protocols. Building on these advances, we developed ClinBoost, a modeling framework that combines RWD and knowledge graph drug embeddings to simulate clinical trial outcomes and guide trial design. The unique capabilities of ClinBoost enable scalable trial designs by modeling clinical responses using endpoints aligned with those used in prospective trials. It supports comparative simulations between investigational therapies, standard-of-care, and competitors, including scenarios absent from historical data such as first-in-class or repurposed drugs. The framework can simulate diverse treatment settings, including placebo, monotherapies, and combination therapies (regimens), while capturing drug–patient interactions to enable subgroup-level predictions. It also supports protocol optimization through flexible evaluation of cohorts, treatments, and endpoints.

Although predictive models of drug efficacy based on RWD have advanced considerably, few studies subject them to rigorous validation, and those that do typically include only one or two comparisons, limiting generalizability. To address this gap, we applied a robust validation framework to compare model predictions with outcomes from 21 clinical trials assessing 13 different drugs, through rigorous statistical and Bland–Altman analyses to ensure technical reliability and clinical relevance in line with contemporary regulatory standards. Our approach assessed not only binary trial outcomes (e.g., success vs. failure), but also continuous treatment effect estimates by precisely simulating trial conditions. This included matching clinical endpoints, eligibility criteria, time points (e.g., 12 weeks, 24 weeks), and sample sizes to replicate trial designs using real-world patient cohorts.

In this study, we demonstrated the practical applicability of our ML process by developing and validating a RWD predictive drug response model using ulcerative colitis (UC) as a use case. UC is a chronic inflammatory disease with significant unmet medical needs where patient heterogeneity hampers drug development efforts. By integrating ML, RWD, and knowledge graph methodologies, we demonstrate accurate predictions of drug efficacy against past clinical trials and demonstrate how such modeling frameworks can be applied

to optimize future trial design, thereby optimizing recruited patient populations, increasing the probability of technical success, and accelerating therapeutic development.

METHODS

Data Sources

Real-World Clinical Data

For modeling patient outcomes, we used deidentified patient-event level data from the PurpleLab® insurance open claims database and electronic health records (EHR) from Eversana's Electronic Health Record database (EVERSANA Life Sciences Inc.), which were linked via a third-party tokenization service (Datavant Inc., San Francisco, California). The PurpleLab® open claims database covers approximately 350 million patients representing 98% of healthcare claims filed in the United States between 2014 and 2023, containing anonymized information on surgeries, procedures, medications, diagnoses, and various healthcare encounters, plus comprehensive demographic and death information. The Eversana EHR dataset covers 120 million patients treated primarily in the community setting in the United States, including medical procedures, diagnoses, healthcare encounters, medications, laboratory results, vitals, and socio-demographic factors. The contents of the linked dataset were standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) prior to model development.

Bio-pharmacological Data for Knowledge Graph Construction

A comprehensive biomedical dataset compiled from a combination of open-source and proprietary data sources was used to build a biomedical knowledge graph, containing information on a wide range of small molecules and biologics and their pharmacological profiles — pharmacokinetics/-dynamics (PK/PD), absorption, metabolism, excretion, toxicity, mechanism of action, cellular targets, and known drug-drug interactions. Data for relevant biological processes, diseases, anatomical structures, genes, biological pathways, and molecular functions affected by these agents were also included. This knowledge graph consists of approximately 100,000 nodes representing biomedical entities and over 5 million edges representing relationships between these entities. To represent novel drugs lacking mechanistic information, the knowledge graph was manually enriched by adding connections to newly created compound nodes based on published preclinical studies, including mechanism of action, additional indications, and pharmacodynamic components such as metabolism and excretion. All medications indicated for UC patients were thoroughly represented in the graph, with further enrichment performed as necessary.

Framework for Model Training and Prediction

Overall Cohort Definition

The study cohort included patients with a confirmed diagnosis of UC. To emulate this criterion using RWD, eligible patients were required to have at least three documented diagnosis records of ulcerative colitis and at least one prescription for a UC-related medication. The disease start date was defined as the earliest recorded diagnosis date. For the training set, patients were permitted to enter the training process multiple times, based

on the first occurrence of mild, moderate, or severe disease classification after the disease start date — allowing up to 3 separate cohort entries per patient, with severity classified using the electronic Mayo score described below.

This overall cohort was used to train the machine learning model to predict the joint effect of patient features and treatments on outcomes. When predicting the outcomes of the validation clinical trials, a distinct cohort was used for each trial, consisting only of patients meeting that trial's eligibility criteria. During prediction, patient characteristics were paired with intervention or control arm drugs, rather than the actual drugs patients received in the RWD.

Engineering of Patient Outcomes — EHR-based Mayo Score (eMS)

The Mayo Score (MS) is a clinical measure of UC activity extensively used to evaluate treatment response, with four components: stool frequency, rectal bleeding, endoscopic findings of mucosal inflammation, and physician's global assessment. The partial MS (pMS) — composed of stool frequency, bleeding, and physician assessment on a 9-point scale — performs on par with the full MS and has been used in several clinical trials to monitor response. To model UC disease severity, we constructed a retrospective EHR-based Mayo score (eMS), summarized in Table 1. The weights were determined through expert consultation and empirical calibration: a gastroenterologist specializing in inflammatory bowel disease initialized weights reflecting clinical severity, then monthly eMS scores were compared to Mayo score distributions from the literature and curated datasets.

Element	Record Weight
Abdominal pain	1
Oral steroid usage	1
Frequent endoscopy procedures (>1 within 6 months)	1
Anemia due to blood loss	2
Hospitalization due to UC	2
Diarrhea	2
Rectal bleeding	3

Table 1. Retrospective EHR-based Mayo Score (eMS) — element weights used to derive monthly disease-severity scores from real-world clinical records. Total score range: 0–12, harmonized to 0–9 for comparability with the partial Mayo score.

The total eMS score ranges from 0 to 12, and was harmonized to a 0–9 range for consistency with the pMS framework. Validation using curated data showed high concordance between eMS and pMS. The harmonized eMS was categorized into three severity levels: mild (0–2), moderate (3–6), and severe (7–9), calculated monthly per patient. Disease progression was defined as an increase in eMS by one point in a given month relative to the previous 3-month window, a definition validated against published retrospective relapse-rate analyses. The eMS's capacity to monitor progression was validated via absolute progression rates relative to the literature, relapse rates one year after diagnosis, and progression by neutrophil/lymphocyte ratio (NLR) and C-reactive protein (CRP) strata. The predictive model was trained to predict progression in patients with UC.

Engineering of Patient Outcomes — Progression and Remission

As clinical visits in RWD often reflect episodes of heightened disease activity, they provide a granular view of progression and exacerbation, while low-activity episodes can reflect missing data rather than true remission. We therefore developed a two-stage model to predict remission in line with clinical trial benchmarks despite the lack of directly observed remission in RWD. First, the ML model was trained to predict UC progression over time using real-world clinical data. Second, an additional model inferred expected remission rates in clinical trials based on the predicted treatment effect on progression, using a reference sigmoidal remission curve constructed from reported placebo remission rates across trials of moderate-to-severe UC (5–52 weeks post-randomization). A transformation function mapped this reference curve to predicted placebo progression rates, capturing the inverse relationship between progression and remission, which was then applied to the model's predicted treatment effect on progression to estimate corresponding remission rates.

Drug-centered Knowledge Graph Embeddings

The knowledge graph was converted to graph embeddings for model training. Several embedding methods were tested — TransE, TransR, MURE, metapath2vec, and graph neural networks — evaluated using a scoring metric assessing relative cosine similarity of drugs with similar versus dissimilar mechanisms of action. The embeddings achieving the highest evaluation score were used for model training.

Knowledge Graph Enrichment for Novel Drug Representation

To enable modeling of treatment effects for novel drugs not represented in the knowledge graph (KG), a KG enrichment process was applied: relevant information (mechanisms of action, drug-drug interactions, molecular targets, side effects, approved or investigated indications) was extracted from the literature, structured to match the KG format, and used to construct a corresponding subgraph. This subgraph was integrated into the main KG, and updated drug embeddings were generated to include the novel drug alongside relevant standards of care (SoC) and comparators. These embeddings were then used to train models predicting the novel drug's effect versus SoC and alternative treatments. Drugs added via this mechanism are referred to throughout as enriched drugs.

Patient-Drug Journey Representation

The index date for the patient-drug journey was defined as the first instance the patient's disease was classified as mild, moderate, or severe following the UC start date. "Patient journey" was defined as the time series of clinical events starting six months prior to the index date, with the end date determined by the corresponding clinical trial matching date — allowing prediction over different follow-up durations. Patient demographics, current and prior treatments, hospitalization within the last year, and inflammatory marker levels (CRP, erythrocyte sedimentation rate, fecal calprotectin) were included as predictors. This patient-drug journey data was converted into a vector embedding used in the prediction model.

Temporal Prediction Framework — Model Training and Hyperparameter Optimization

The ClinBoost algorithm combines patient-drug journey embeddings and knowledge graph embeddings into the XGBoost algorithm, followed by per-time-window logistic regression, with trainable weights combining patient clinical data, demographics, and drug embeddings into a single model predicting outcome probabilities per patient per time window. The

model was trained using real-world clinical data and graph embeddings with an 80:10:10 train/validation/test split for hyperparameter tuning. Hyperparameter optimization considered a three-dimensional grid covering (1) drug-centered knowledge graph embedding, (2) clinical data definitions and filters for the overall training cohort, and (3) prediction model hyperparameters, evaluated via ROC-AUC (c-statistics) comparing predicted to observed RWD outcomes. After selecting and validating the best hyperparameters using the RWD, the best model was tested on the validation clinical trial set.

Model Validation Using Randomized Controlled Trial (RCT) Data

Selection of RCTs for Validation

Trial design and results information was sourced from ClinicalTrials.gov and PubMed. Trials were included if they were phase II or III RCTs evaluating a pharmacological treatment for moderate and/or severe ulcerative colitis versus placebo or standard of care, were registered at ClinicalTrials.gov with well-documented inclusion and exclusion criteria modelable using RWD, and reported disease remission results. Trials without published results were excluded. The 21 trials ultimately included spanned 2007 to 2022. Extracted elements included inclusion/exclusion criteria, investigational and comparator treatments, endpoint definitions, sample sizes per arm, baseline characteristics, and trial results (point estimates, confidence intervals, and reported p-values).

Prediction of RCT Outcomes

For each validation RCT, eligibility criteria were applied to select an eligible population from the study's real-world dataset, and all replicable clinical parameters were applied (full list in Appendix 1). Patient cohorts matching the RCT's age, sex distribution, and sample size were sampled to create the prediction cohort. The trained model predicted patient-level outcomes separately under treatment and control regimens, regardless of the treatment patients actually received in the RWD. Remission rate predictions were made at the same time point as the corresponding RCT to enable comparability, with subgroup-level simulations based on additional eligibility criteria across varying sample sizes.

Performance Metrics for Model Validation

Model validation used the following criteria:

- Trial success agreement — a binary analysis of accuracy in predicting trial success or failure based on direction and statistical significance, reported across four levels: Concordant Alignment (same direction and significance); Directionally Aligned with Threshold Misalignment (same direction, prediction not statistically significant); Paradoxical Agreement (both non-significant, opposite direction); Complete Divergence (both significant, opposite direction).
- Absolute outcome agreement — achieved when the predicted estimate fell within the 95% CI of the observed RCT estimate, visualized via forest plots and Bland-Altman plots.
- Cohen's d statistic — standardized difference between observed and predicted effect sizes (0.2/0.5/0.8 = small/moderate/large), averaged across all validation trials.

Trial success and failure across the 21 trials was evaluated using a confusion matrix and quantified via F1 score, sensitivity, and specificity, with performance also assessed by drug mechanism of action strata (biological versus non-biological).

Model-guided Trial Design Optimization

To evaluate the prediction engine's utility for informing trial design, simulation analyses identified patient sub-cohorts with differential treatment responses, optimized eligibility criteria, and assessed power/sample-size requirements:

- Cohort Variant Generation and Effect Prediction — starting from a base cohort derived from an existing trial population, multiple cohort variants were generated by systematically modifying inclusion/exclusion criteria (e.g., prior biologic exposure, baseline inflammatory markers, systemic steroid use), predicting expected treatment effect and estimated real-world cohort size for each variant.
- Longitudinal Outcome Projection — weekly remission rates over 52 weeks were predicted for selected cohorts using the model's time-to-event outputs.
- Power and Sample Size Estimation — the total sample size needed to reach statistical significance ($\alpha = 0.05$) was determined for each cohort, maintaining the original treatment-to-control allocation ratio.
- Monte Carlo Simulation for Probability of Success (PTS) — repeated sampling of eligible patient profiles generated a distribution of predicted treatment effects; PTS was defined as the proportion of simulations exceeding a predefined threshold (e.g., 10% remission difference).

RESULTS

Modeling Framework Overview

Figure 1. Overview of the predictive model architecture and workflow. (a) Prediction workflow. (b) Model architecture, covering patient drug-journey embeddings, the ClinBoost model, and time-to-event outcome prediction.

The model was trained using data from a bio-pharmacological knowledge graph and longitudinal real-world clinical data from a linked claims/EHR database of patients with UC. Embeddings of the knowledge graph and patient-drug profiles from these two sources were combined into a gradient-boosted tree model to predict patient-level outcomes on therapy over time.

The study cohort included 482,870 unique patients, who could be recruited up to 3 times in relation to a change in clinical status. The median age was 56 years (range 18–85), and 59.5% were women. Patients were permitted to enter the simulation multiple times, yielding 723,263 patient events used for model development (full characteristics in Supplementary Table S1).

The drug-centered knowledge graph and derived drug embeddings included both approved therapies represented in the RWD and novel therapies with known mechanisms of action manually added to the graph. Embedding quality was assessed by comparing cosine similarity of drugs sharing similar mechanisms of action and therapeutic indications. A two-dimensional t-SNE visualization showed that learned embeddings preserve the underlying

graph topology, with agents sharing mechanisms of action clustering together and demonstrating high cosine similarity, as expected.

Figure 2. QuantHealth Knowledge Graph (QHKG) and embedding evaluation — (a) example subgraph with drug, gene, disease, side-effect, and pathway nodes; (b) 2D t-SNE projection of drug embeddings colored by pharmacological subclass; (c) scatter plot showing a strong negative correlation (Pearson = -0.83) between structural distance in the KG and embedding cosine similarity; (d) heatmap of average cosine similarity within and between drug subgroups.

Model Validation Results

Table 2 provides an overview of the 21 randomized controlled trials included for validation. Of these, 7 (33.3%) were unsuccessful trials, 5 (23.8%) were phase 2 trials, and 6 (29%) included a novel agent not commonly administered in the real world at the time of the study. Sample sizes ranged from 28 to 476 patients per arm, with outcome measurement timing spanning 6 to 52 weeks post-randomization. Treatments spanned diverse mechanisms of action, including biologics (e.g., TNF inhibitors) and small molecules.

#	Trial	NCT ID	S/F	Wk	Treatment	Trt %	Control	Ctrl %	Target/MoA	N (T/C)	Ph	New
1	ELEVATE UC 12	NCT03996369	S	12	Etrasimod	24.8	Placebo	15.2	S1PR1 agonist	222/112	3	Y
2	NCT03934216	NCT03934216	F	12	Deucravacitinib	14.8	Placebo	16.3	TYK2 inhibitor	88/43	2	Y
3	U-Accomplish	NCT03653026	S	8	Upadacitinib	33.5	Placebo	4.1	JAK inhibitor	341/174	3	Y
4	NCT03482635	NCT03482635	F	12	Spesolimab	7.1	Placebo	0	Anti-IL-36	28/23	2	Y
5	SELECTION	NCT02914522	S	10	Filgotinib	26.1	Placebo	15.3	JAK inhibitor	245/137	3	N
6	NCT02819635	NCT02819635	S	8	Upadacitinib	26.1	Placebo	4.8	JAK inhibitor	319/154	3	Y
7	VARSITY	NCT02497469	S	52	Vedolizumab	31.3	Adalimumab	22.5	Integrin/Anti-TNF-α	386/383	3	N
8	NCT02435992	NCT02435992	S	10	Ozanimod	18.4	Placebo	6	S1PR1 agonist	429/216	3	Y
9	NCT02289417	NCT02289417	F	12	Apremilast	21.8	Placebo	12.1	PDE4 inhibitor	55/58	2	N
10	HIBISCUS 2	NCT02171429	F	10	Etrolizumab	18.2	Placebo	11.1	Integrin blocker	143/72	3	N
11	HIBISCUS 1(A)	NCT02163759	F	10	Etrolizumab	19.4	Adalimumab	22.5	Integrin blocker	144/142	3	N
12	HIBISCUS 1(B)	NCT02163759	S	10	Etrolizumab	19.4	Placebo	6.9	Integrin blocker	144/72	3	N
13	GARDENIA	NCT02136069	F	10	Etrolizumab	20.6	Infliximab	32.8	Integrin blocker	199/198	3	N
14	HICKORY	NCT02100696	S	14	Etrolizumab	18.5	Placebo	6.3	Integrin blocker	384/95	3	N
15	NCT01694485	NCT01694485	S	8	Abrilumab	13.4	Placebo	4.4	Integrin blocker	79/116	2	N
16	OCTAVE(A)	NCT01465763	S	8	Tofacitinib	18.5	Placebo	8.2	JAK inhibitor	476/112	3	N
17	OCTAVE(B)	NCT01458951	S	8	Tofacitinib	16.6	Placebo	3.6	JAK inhibitor	429/112	3	N
18	GEMINI 1	NCT00783718	S	6	Vedolizumab	16.9	Placebo	5.4	Integrin blocker	225/149	3	N
19	NCT00488774	NCT00488774	F	6	Golimumab	13	Placebo	11	Anti-TNF-α	77/73	2	N
20	PURSUIT-SC	NCT00487539	S	6	Golimumab	17.9	Placebo	6.4	Anti-TNF-α	257/251	3	N
21	ACT 1	NCT00096655	S	8	Infliximab	38.8	Placebo	14.9	Anti-TNF-α	121/121	3	N

Table 2. UC validation trials and experimental drug mechanism of action. S/F = trial success/failure. The 21 trials span diverse drug modalities, phase 2/3 designs, novel vs. approved drugs, and study sizes.

Prediction cohorts matched original-trial age/sex distributions and arm sizes, with remission rates predicted at the trial's actual outcome time point.

Model validation results are presented in Table 3 and Figure 3. The model correctly predicted the effect size — treatment vs. control remission difference — for 16 of 21 trials (76.2%), correctly matching direction and statistical significance. Among the five discordant trials, four (80%) showed effect sizes in the same direction as the RCTs but differed in statistical significance, altering the conclusion. Only one trial, VARSITY (adalimumab vs. vedolizumab), showed a misclassification in direction; it had the longest follow-up (52 weeks), and the discrepancy stemmed primarily from inaccurate control-arm predictions. Notably, the related HIBISCUS-1 (NCT02163759) adalimumab comparison was accurately predicted at week 10, fully consistent with the corresponding RCT. For all six trials involving novel agents, predicted effect sizes aligned with RCT outcomes, supporting the same regulatory decisions.

#	Trial	MoA	Observed Δ (95% CI)	Predicted Δ (95% CI)	Difference (95% CI)	Agree
1	ELEVATE UC 12	S1PR1 agonist	0.10 (0.04, 0.15)	0.15 (0.09, 0.20)	0.05 (-0.03, 0.13)	✓
2	NCT03934216	TYK2 inhibitor	-0.01 (-0.10, 0.07)	0.06 (-0.03, 0.14)	0.07 (-0.05, 0.19)	✓
3	U-Accomplish	JAK inhibitor	0.29 (0.25, 0.34)	0.11 (0.07, 0.15)	-0.18 (-0.24, -0.12)	✓
4	NCT03482635	Anti-IL-36	0.07 (-0.02, 0.15)	-0.04 (-0.12, 0.04)	-0.11 (-0.21, -0.01)	✓
5	SELECTION	JAK inhibitor	0.11 (0.05, 0.16)	0.19 (0.13, 0.24)	0.08 (0.00, 0.16)	✓
6	NCT02819635	JAK inhibitor	0.21 (0.17, 0.26)	0.12 (0.08, 0.16)	-0.09 (-0.15, -0.03)	✓
7	VARSITY	Integrin/Anti-TNF-α	0.09 (0.04, 0.13)	-0.15 (-0.19, -0.10)	-0.23 (-0.30, -0.17)	✗
8	NCT02435992	S1PR1 agonist	0.12 (0.09, 0.16)	0.10 (0.06, 0.14)	-0.02 (-0.07, 0.03)	✓
9	NCT02289417	PDE4 inhibitor	0.10 (0.00, 0.19)	0.25 (0.13, 0.35)	0.15 (0.00, 0.30)	✗
10	HIBISCUS 2	Integrin blocker	0.07 (0.00, 0.14)	0.05 (-0.02, 0.12)	-0.02 (-0.11, 0.07)	✓
11	HIBISCUS 1(A)	Integrin blocker	-0.03 (-0.10, 0.04)	-0.13 (-0.20, -0.07)	-0.10 (-0.20, -0.01)	✓
12	HIBISCUS 1(B)	Integrin blocker	0.12 (0.06, 0.19)	0.05 (-0.02, 0.11)	-0.08 (-0.17, 0.01)	✗
13	GARDENIA	Integrin blocker	-0.12 (-0.18, -0.06)	-0.12 (-0.18, -0.06)	0.00 (-0.08, 0.09)	✓
14	HICKORY	Integrin blocker	0.12 (0.08, 0.16)	0.01 (-0.03, 0.04)	-0.11 (-0.17, -0.06)	✗
15	NCT01694485	Integrin blocker	0.09 (0.03, 0.15)	0.04 (-0.02, 0.10)	-0.05 (-0.13, 0.03)	✗
16	OCTAVE(A)	JAK inhibitor	0.10 (0.06, 0.14)	0.08 (0.04, 0.11)	-0.03 (-0.08, 0.02)	✓
17	OCTAVE(B)	JAK inhibitor	0.13 (0.09, 0.16)	0.08 (0.04, 0.11)	-0.05 (-0.10, 0.00)	✓
18	GEMINI 1	Integrin blocker	0.11 (0.07, 0.16)	0.04 (0.00, 0.08)	-0.07 (-0.13, -0.01)	✓
19	NCT00488774	Anti-TNF-α	0.02 (-0.05, 0.08)	0.06 (0.00, 0.12)	0.04 (-0.05, 0.13)	✓
20	PURSUIT-SC	Anti-TNF-α	0.12 (0.08, 0.15)	0.06 (0.03, 0.09)	-0.06 (-0.11, 0.00)	✓
21	ACT 1	Anti-TNF-α	0.24 (0.16, 0.31)	0.11 (0.05, 0.17)	-0.13 (-0.22, -0.03)	✓

Table 3. Effect size estimates and agreement metrics. Observed and predicted treatment effect = difference in remission rates between investigational and control arms (95% CI shown). "Agree" reflects Concordant Alignment as defined in Methods.

Overall, the model achieved an F1 score of 0.80 on the binary success/failure level. On patient-level real-world data, a high-performing model was selected using ROC-AUC, which declined by only 7.2% when moving from a 10-day to a 400-day simulated follow-up window (AUC 0.83 vs. 0.77, respectively).

> **Figure 3.** Remission rates in clinical trials, observed vs. predicted outcomes. (a) Treatment vs. control outcomes across 21 trials, with 95% CIs; grayed trials denote binary-outcome misclassifications. (b) Confusion matrix of binary performance. (c) Quantitative binary performance metrics. (d) ROC-AUC curve on the patient-level RWD test set.

Visual comparison of point estimates confirmed close alignment between predicted and observed treatment effects. Absolute differences (percentage points) ranged from 0 (GARDENIA) to a maximum of -0.23 (VARSITY), with a median difference of -0.05. Bland–Altman analysis of arm-level remission (treatment and control) showed minimal systematic bias (mean difference -0.009), with most trial arms falling within the 95% limits of agreement (LoA: -0.148 to 0.129) — though prediction error increased slightly for arms with higher average remission (>0.2). Bland–Altman analysis of predicted vs. observed effect sizes showed a modest negative bias (mean difference -0.045, i.e., slight underestimation of treatment effect), with the majority of values within the LoA bounds (-0.219 to 0.129), demonstrating stable behavior across small and large effect sizes.

> **Figure 4.** Trial prediction and model validation. (a) Forest plot of absolute differences between observed and predicted treatment effect (percentage points), with an inverse-variance-weighted overall estimate. (b) Forest plot of standardized mean differences (Cohen's d) between observed and predicted effect sizes. (c) Bland–Altman plot of study-arm remission agreement. (d) Bland–Altman plot of study effect-size agreement.

These analyses confirm that ClinBoost provides accurate and unbiased estimates of both individual arm outcomes and relative treatment effects, with persistent performance across a diverse range of trial conditions.

Sub-analyses by trial strata showed the predictive model performed slightly better, on average, for non-biologic agents (F1 = 0.94 vs. 0.80 for biologics) and for phase III versus phase II trials (F1 = 0.90 vs. 0.75). By mechanism of action, JAK inhibitors (5 trials) achieved the best F1 (1.00) versus anti-integrins (5 trials, F1 = 0.75); a third group spanning other mechanisms achieved an F1 of 0.84. Finally, the six trials requiring manual knowledge-graph enrichment — agents withheld from the training RWD dataset — showed robust alignment with observed results (F1 = 1.00) compared with the standard automated pipeline (15 trials; F1 = 0.80), validating the consistency of the graph-embedding technique even when integrating pharmacological profiles absent from the primary training data.

> **Figure 5.** Stratified performance of the predictive model by clinical trial design. The model performs robustly across Phase 2 and Phase 3 trials and diverse mechanisms of action; the knowledge-graph enrichment technique enables accurate prediction of compounds absent from both RWD and the original biomedical KG. (Biologics: 10 trials; Non-biologics: 11; Phase II: 6; Phase III: 15; Anti-integrin: 5; JAK inhibitor: 6; Other mechanisms: 10; KG-enriched: 6; Non-enriched: 15.)

Model-guided Trial Design Optimization

A key application of drug-response modeling is informing clinical trial design — predicting responder populations, optimizing eligibility criteria, and improving sample-size estimation. Using the ELEVATE UC-12 base cohort — selected as a recent trial with modestly positive

efficacy findings for a novel mechanism of action not observed in RWD — treatment-effect predictions were plotted alongside cohort size for various sub-cohorts defined by additional inclusion/exclusion criteria, revealing distinct response subgroups with higher and lower effect sizes than the base cohort.

Comparing the original validation cohort to an optimized sub-cohort in a phase 3 setting, simulations showed a significant improvement in 24-week remission rates. The model also estimated the minimum sample size required for statistical significance: patient-level simulations across varying arm sizes showed the optimized sub-cohort required fewer participants to reach significance at the standard $p < 0.05$ threshold. Finally, Monte Carlo simulation of possible trial results estimated the probability of technical success (PTS) for a specific remission-difference threshold (10% treatment vs. control), showing how PTS varied by cohort type.

Figure 6. Model-guided trial design optimization — ELEVATE UC-12. (a) Predicted effect size vs. estimated U.S. cohort size across sub-cohort variants; grey points denote cohorts unlikely to reach significance at the original sample size. (b) Predicted 52-week remission for Optimized and Low Effect cohorts vs. the original cohort (blue = treatment, red = placebo). (c) Power analysis across sample sizes. (d) Monte Carlo distributions of predicted treatment effect by cohort. Low Effect cohort: prior anti-TNF exposure, no high inflammatory markers. Optimized cohort: no prior biologic therapy, no systemic steroids at baseline.

These findings emphasize the potential of model-guided patient selection to enhance clinical trial efficiency and success rates.

DISCUSSION

Our study demonstrates that a predictive model leveraging a combined knowledge graph and RWD-based machine learning framework can effectively generate out-of-sample drug response predictions that align well with clinical trial findings in UC. This is evidenced by the strong convergence between observed and model-predicted trial results, with 16 of 21 comparisons agreeing on direction, effect size, and statistical significance. This robust performance extended across a diverse set of trials spanning drug mechanisms (11 small molecules and 10 biologics), phases (II and III), and endpoint durations (6 to 52 weeks), showcasing the model's adaptability. Notably, the model showed high concordance with outcomes for all six trials involving manually enriched agents, highlighting the utility of the knowledge graph framework in estimating effects based on pharmacological attributes even when historical utilization data is unavailable.

While alignment was broadly strong, some discrepancies were observed. In four of the five cases lacking full concordance, effect sizes remained directionally consistent with the RCTs but conclusions diverged due to differences in statistical significance. Only one comparison — VARSITY — showed a contradictory result, with the original RCT showing a positive effect for vedolizumab [0.09 (0.03–0.15)] versus a model-generated negative effect versus adalimumab [-0.15 (-0.21, -0.08)], both statistically significant but in opposite directions. A plausible explanation is VARSITY's extended follow-up (52 weeks) — substantially longer than most other trials (8–14 weeks) — which can introduce increased uncertainty from informative censoring in RWD-derived estimates and potential uncontrolled concomitant interventions.

Differences in effect size between observed and predicted results were generally small, ranging from 0 to -0.23 with a median of -0.05 and a mean of -0.045 (SD 0.09).

Our efforts build upon and expand previous work using real-world evidence to emulate clinical trial findings, including target trial emulation frameworks and initiatives like RCT DUPLICATE. Unlike replication-focused studies, our approach incorporates knowledge graphs to enable outcome estimation for mechanisms not yet represented in the RWD — a significant advantage where historical utilization data is unavailable. The successful validation across 21 trials represents a more extensive comparison than typically reported in studies validating cohort-level predicted trial results. Importantly, our validation went beyond binary success/failure predictions to assess the accuracy of precise treatment-effect estimates, matching clinical endpoints, patient eligibility criteria, evaluation time points, and sample sizes with the original trials. The use of multiple performance metrics — trial success agreement, absolute outcome agreement, Cohen's d, and F1 scores — provided a comprehensive evaluation against contemporary regulatory standards.

Considering the substantial heterogeneity among patients with UC — varying clinical expressions and disease trajectories including frequent relapses versus more stable courses, driven by environmental factors, diet, life stress, psychological influences, and genetics — achieving strong concordance between predicted and observed results is noteworthy. Our analysis encompassed seven drug types across two major categories: small molecules (JAK inhibitors, S1P receptor modulators, TYK2 inhibitors, PDE4 inhibitors) and biologics (monoclonal antibodies targeting TNF-α, β7 integrin, and IL-36). This consistency across heterogeneity underscores the model's robustness, predictive reliability, and validity for UC treatment — likely driven by the comprehensive knowledge graph capturing diverse mechanisms of action, rigorous validation of individual outcome model components, and the large RWD sample (N = 723,263 patient events) facilitating well-matched cohort creation.

A critical application of a validated drug-response model lies in informing clinical trial design. Focusing predictions on specific patient subgroups allows preemptive identification of responder subgroups and treatment-effect estimation across diverse patient profiles, supporting personalized therapeutic strategies. Supplementary analyses identified factors associated with modified treatment effects for S1P modulator treatment — for instance, men under 40 without prior biologic exposure showed enhanced response. Subgroup identification also directly informs cohort optimization and increases the probability of technical success. Notably, drug response model predictions for optimized patient cohorts in ELEVATE-UC-12 aligned closely with post-hoc analyses of the trial not available at the time of the study — confirming that patients with low inflammatory markers who were naïve to biologic therapy consistently showed higher response rates to etrasimod, and that patients not receiving concomitant steroid therapy demonstrated increased S1P modulator response. Future research should elucidate whether these predictive markers are specific to S1P modulator therapy or generally predictive of induction-therapy response in UC.

It is important to emphasize that predictive models should be applied within the context in which they were validated. In our case, induction therapy trials may be more amenable to implementing these insights than maintenance therapy trials, given the patient populations and treatment paradigms used in model development. Beyond trial optimization, the frameworks presented here aim to reinforce the evolving role of real-world evidence (RWE) in

drug development and evaluation. While RCTs remain the gold standard, our findings suggest that well-designed RWE studies, enhanced by appropriate ML approaches, can confirm and extend RCT findings to broader, more heterogeneous real-world populations where adherence and concomitant care differ, and provide out-of-sample estimations for novel drug classes — offering a robust framework for comparative research where head-to-head trials are scarce and assessing efficacy in smaller subpopulations is often impractical.

The strong concordance between our real-world predictions and RCT findings, established through a comprehensive validation process, reinforces the reliability of model-based RWE approaches and carries significant implications for regulatory and clinical decision-making. As healthcare practitioners, drug developers, and regulatory agencies increasingly consider the role of AI and RWD, robust, patient-level validation frameworks offer tangible pathways to evaluate and integrate these data sources effectively. Gradual implementation of such models could support prioritizing resource allocation, optimizing designs, and assessing the feasibility of new indications or stratifications before undertaking expensive RCTs — potentially supporting adaptive approvals, label expansions, and health technology assessments where traditional trials are impractical. Given its demonstrated validity across multiple drug classes, this approach could be extended to other disease areas, broadening its applicability in real-world drug evaluation and accelerating the translation of biomedical knowledge into tangible patient benefit.

CONCLUSIONS

This study demonstrates the value of integrating RWD, biomedical knowledge graphs, and machine learning to improve the prediction of clinical trial outcomes in UC. The ClinBoost framework showed strong concordance with observed results across 21 RCTs, accurately reproducing treatment effects in diverse settings, including those involving novel agents not represented in existing datasets. By leveraging enriched knowledge graph embeddings and longitudinal patient journeys, the model provided unbiased estimates of remission outcomes and treatment effects while identifying responder subgroups that could enhance trial efficiency. Importantly, ClinBoost not only emulated past trials but also offered actionable insights to optimize trial design — refining eligibility criteria, guiding sample-size estimation, and improving probability of technical success. These findings highlight the potential of data-driven modeling to complement traditional randomized trials by informing trial strategies, reducing uncertainty, and accelerating drug development, with broader application to other therapeutic areas supporting more efficient, personalized, and cost-effective pathways for evaluating novel interventions.

LIST OF ABBREVIATIONS

Abbreviation	Definition
CDM	Common Data Model
CI	Confidence Interval
CRP	C-Reactive Protein

Abbreviation	Definition
EHDS	European Health Data Space
EHR	Electronic Health Records
eMS	Electronic Mayo Score
JAK	Janus Kinase
KG	Knowledge Graph
LoA	Limits of Agreement
ML	Machine Learning
MoA	Mechanism of Action
MS	Mayo Score
NLR	Neutrophil/Lymphocyte Ratio
OMOP	Observational Medical Outcomes Partnership
PDE4	Phosphodiesterase 4
pMS	Partial Mayo Score
PTS	Probability of Technical Success
RCT	Randomized Controlled Trial
RWD	Real-World Data
RWE	Real-World Evidence
SD	Standard Deviation
SoC	Standard of Care
S1P	Sphingosine-1-Phosphate
TNF	Tumor Necrosis Factor
t-SNE	t-distributed Stochastic Neighbor Embedding
TYK2	Tyrosine Kinase 2
UC	Ulcerative Colitis

STUDY INFORMATION

Funding

This study was funded by Sanofi.

Acknowledgements

The authors thank Jean-Paul Collet and Maninder Anand of Evidinno Outcomes Research Inc. (Vancouver, BC, Canada) for medical writing assistance.

Competing Interests

Flavio Dormont, Annie Kruger, and Ramon Hernandez are employees and shareholders of Sanofi (Morristown, NJ, USA). Amina Alaskarov, Amichai Perlman, Omri Matalon, and Michael Shapiro are employees of QuantHealth (Tel Aviv, Israel), which was contracted by Sanofi to conduct this study. Dr. Sands reports personal fees and/or non-financial support from

numerous pharmaceutical and biotechnology companies, outside the submitted work — full disclosure available in the original manuscript.

Data Availability

The datasets used in this study, including RWD derived from EHRs and insurance claims, as well as components of the biomedical knowledge graph, were obtained under license from proprietary vendors and are not publicly available. Researchers interested in accessing the datasets may contact the corresponding author to discuss the possibility of data access, subject to licensing agreements and confidentiality obligations.

Ethics Approval and Consent to Participate

All methods were carried out in accordance with relevant guidelines and regulations. This study utilized retrospective, de-identified patient-level data from the PurpleLab® insurance open claims database and EHR data from the EVERSANA EHR database, linked via a third-party tokenization service (Datavant Inc.). As a retrospective analysis of de-identified data with no direct patient identifiers, this study was exempt from IRB review under 45 CFR 46.104(d)(4), and the requirement for informed consent was waived.

Supplementary Files

- Supplementary File 1 (.pptx) — Supplementary Tables & Figures
- Appendix 1 (.docx) — List of features recreated in the QuantHealth platform

REFERENCES

1. Hwang TJ, Carpenter D, Lauffenburger JC, et al. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. JAMA Intern Med. 2016;176(12):1826-33.

2. Farid SS, Baron M, Stamatis C, Nie W, Coffman J. Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. mAbs. 2020;12(1):1754999.

3. Mullard A. Parsing clinical success rates. Nat Rev Drug Discov. 2016;15(7):447.

4. Schlander M, Hernandez-Villafuerte K, Cheng CY, Mestre-Ferrandiz J, Baumann M. How Much Does It Cost to Research and Develop a New Drug? PharmacoEconomics. 2021;39(11):1243-69.

5. Augustine EF, Adams HR, Mink JW. Clinical trials in rare disease: challenges and opportunities. J Child Neurol. 2013;28(9):1142-50.

6. Rehman AU, Li M, Wu B, et al. Role of Artificial Intelligence in Revolutionizing Drug Discovery. Fundamental Research. 2024.

7. Fu T, Huang K, Xiao C, Glass LM, Sun J. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. Patterns. 2022;3(4):100445.

8. Wang Z, Xiao C, Sun J. SPOT: Sequential Predictive Modeling of Clinical Trial Outcome with Meta-Learning. ACM-BCB 2023.

9. Al-Mohaya M, Demiralp B, Kurt A, Çelik Y. In silico approaches used in pharmacy. J Appl Pharm Sci. 2024.

10. Duruisseaux M, Masson P, Nativel A, et al. Prospective validation of in silico clinical trials prediction using an EGFR-mutated NSCLC mechanistic computational disease model. J Clin Oncol. 2024;42(16_suppl):8614.

11. Angoulvant D, Granjeon-Noriot S, Amarenco P, et al. In-silico trial emulation to predict the cardiovascular protection of new lipid-lowering drugs. Eur J Prev Cardiol. 2024;31(15):1820-30.

12. Brbić M, Yasunaga M, Agarwal P, Leskovec J. Predicting drug outcome of population via clinical knowledge graph. medRxiv. 2024.

13. Joshi P, V M, Mukherjee A. A knowledge graph embedding based approach to predict adverse drug reactions using a deep neural network. J Biomed Inform. 2022;132:104122.

14. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021;19:4538-58.

15. Franklin J, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther. 2017;102.

16. Raman K, Kumar R, Musante CJ, Madhavan S. Integrating Model-Informed Drug Development With AI. Clin Transl Sci. 2025;18(1):e70124.

17. Sharma A, Lysenko A, Jia S, Boroevich KA, Tsunoda T. Advances in AI and machine learning for predictive medicine. J Hum Genet. 2024;69(10):487-97.

18. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414-28.

19. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67.

20. Gao Z, Ding P, Xu R. KG-Predict: A knowledge graph computational framework for drug repurposing. J Biomed Inform. 2022;132:104133.

21. FDA. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. 2025.

22. Li P, Wu Y, Xiong W, et al. Association between the immune-inflammation index and severity/clinical outcomes of IBD: a systematic review and meta-analysis. BMC Gastroenterol. 2025;25(1):414.

23. Sands BE, Peyrin-Biroulet L, Loftus EV, et al. Vedolizumab versus Adalimumab for Moderate-to-Severe Ulcerative Colitis. N Engl J Med. 2019;381(13):1215-26.

24. Naegeli AN, Hunter T, Dong Y, et al. Full, Partial, and Modified Permutations of the Mayo Score. Crohns Colitis 360. 2021;3(1):otab007.

25. Lewis JD, Chuai S, Nessel L, et al. Use of the noninvasive components of the Mayo score to assess clinical response in UC. Inflamm Bowel Dis. 2008;14(12):1660-6.

26. Louis E, Schreiber S, Panaccione R, et al. Risankizumab for Ulcerative Colitis: Two Randomized Clinical Trials. JAMA. 2024;332(11):881-97.

27. Sandborn WJ, Feagan BG, D'Haens G, et al. Ozanimod as Induction and Maintenance Therapy for Ulcerative Colitis. N Engl J Med. 2021;385(14):1280-91.

28. Feagan BG, Rutgeerts P, Sands BE, et al. Vedolizumab as Induction and Maintenance Therapy for Ulcerative Colitis. N Engl J Med. 2013;369(8):699-710.

29. Rutgeerts P, Sandborn WJ, Feagan BG, et al. Infliximab for Induction and Maintenance Therapy for Ulcerative Colitis. N Engl J Med. 2005;353(23):2462-76.

30. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New York: Routledge; 1988.

31. Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation. 2021;143(10):1002-13.

32. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther. 2017;102(6):924-33.

33. Franklin JM, Pawar A, Martin D, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making. Clin Pharmacol Ther. 2020;107(4):817-26.

34. Ungaro R, Colombel JF, Lissoos T, Peyrin-Biroulet L. A Treat-to-Target Update in Ulcerative Colitis: A Systematic Review. Am J Gastroenterol. 2019;114(6):874-83.

35. Ferretti F, Cannatelli R, Monico MC, Maconi G, Ardizzone S. An Update on Current Pharmacotherapeutic Options for the Treatment of Ulcerative Colitis. J Clin Med. 2022;11(9).

36. Sands BE, Dubinsky MC, Kotze PG, et al. Efficacy and Safety of Etrasimod in Patients With Moderately to Severely Active Ulcerative Colitis Stratified by Baseline Modified Mayo Score. Inflamm Bowel Dis. 2025.

37. Sands BE, Leung Y, Rubin DT, et al. Etrasimod Corticosteroid-Free Efficacy, Impact of Concomitant Corticosteroids on Efficacy and Safety, and Corticosteroid-Sparing Effect in Ulcerative Colitis. J Crohns Colitis. 2024;19(3).