Supplementary Material and Methods

Data Source

GEMINI 1 trial data were collected as part of the phase 3 clinical trial (NCT00783718) with corresponding ethics/institutional review board approval. The VICTORY consortium dataset was collected after ethics/institutional review board approval at all participating sites.

Data Source: VICTORY Consortium

The VICTORY Consortium is a multicenter collaborative research group in which patient demographics, disease characteristics, and treatment outcomes are pooled for inflammatory bowel disease (IBD) patients treated with biologics. Institutional review board approval was obtained from each site for ongoing retrospective data collection and transfer. Data were collected individually by sites using a standardized data collection form and transferred (after de-identification) to the coordinating site (University of California, San Diego) for data compilation and analysis.

Model and Clinical Decision Support Tool Sensitivity Analyses in GEMINI 1 Trial for VDZ Exposure-Efficacy Relationships

Albumin has previously been shown to be the strongest predictor of exposure-efficacy relationships for vedolizumab (VDZ) in ulcerative colitis (UC); however, albumin does not explain the entirety of VDZ exposure variability.[8,9] Before finalizing the model for external validation, a sensitivity analysis was performed by replacing albumin with calculated VDZ clearance profiles for GEMINI 1 trial participants.[8,9] Performance of this model within the GEMINI 1 trial derivation cohort was slightly better than the original model that included albumin (area under the receiver-operating characteristic curve [AUC] 0.69 vs 0.65). A baseline prediction model is not currently available for predicting VDZ exposure over time, and therefore clearance could not be readily substituted for albumin in the final prediction model. Albumin was therefore retained for external validation of the model in the VICTORY Consortium dataset.

GEMINI 1 Trial VDZ Concentration Assessments

VDZ concentrations were assessed in the GEMINI 1 trial using serum samples with a direct VDZ capture pharmacokinetic assay. A sandwich enzyme-linked immunosorbent assay was used for quantifying VDZ in human serum. Serum concentrations of VDZ were determined by QPS in accordance with Good Laboratory Practice. The lower limit of detection was 0.125 μg/mL. Time points for trough concentration assessments taken 30 minutes before VDZ infusions were weeks 0, 2, 6, 22, and 46. Additional concentration assessments were taken at weeks 4, 14, 38, and 52. Time points for peak concentration assessments taken 2 hours postinfusion were weeks 0, 2, 6, 22, and 46. Differences in median concentrations at each time point across the 3 probability groups were first assessed using non-parametric testing (Kruskal-Wallis), and then pairwise comparisons were done for each group at each time point.

GEMINI 1 Trial Fecal Calprotectin Assessments

Fecal calprotectin was assessed in the GEMINI 1 trial using the CAL0100 test kit. Time points for assessments were weeks 0, 6, 30, and 52. Differences in median concentrations at each time point across the three probability groups were first assessed using nonparametric testing (Kruskal-Wallis), and then pairwise comparisons were done for each group at each time point.

GEMINI 1 Intention-to-Treat Sensitivity Analyses

Intention-to-Treat Induction Analysis. In accordance with an intent-to-treat approach, the intention-to-treat (ITT) population for the induction study consisted of all randomized patients in the GEMINI 1 trial cohort who received any amount of blinded study drug during induction treatment.

ITT Maintenance Analysis. In accordance with an intent-to-treat approach, the ITT population for the maintenance treatment consisted of all rerandomized patients randomized as week 6 responders who received VDZ during induction treatment and then received any amount of study drug in the maintenance phase.

Model Validation VICTORY Consortium

External validation of the model was conducted in the VICTORY Consortium cohort. Discriminative ability was assessed by receiver-operating characteristic curve analysis and is presented as AUC. An AUC value of 0.5 denotes that the model does not discriminate any better than random guessing and 1 denotes perfect discrimination. Calibration of the model applied to the external validation cohort was evaluated in multiple ways. The observed event rates and predicted risk were graphically assessed by a calibration curve. The intercept (calibration-in-the-large) assessed whether the overall predicted event rate matches the observed rate, and should ideally be zero, indicating no systematic difference in observed predicted rates. The calibration slope gives an indication of how over- or under-fit the model is, and should ideally be 1, indicating no overfitting. A joint hypothesis test of perfect calibration using a likelihood ratio test was evaluated as an overall test of calibration. This test is

more sensitive to potential miscalibration than the Hosmer-Lemeshow goodness-of-fit test, which is also shown for comparison after splitting the sample into quintiles. This test assesses whether or not the observed event rates match expected event rates in subgroups of the model population, with P values $<.05$ indicating evidence of poor fit for both calibration tests. The overall performance of the models was evaluated with the Nagelkerke R^2 and the Brier score. Nagelkerke R^2 is a measure between 0 and 1, with 0 denoting that the model does not explain any variation and 1 denoting that it perfectly explains the observed variation in outcomes. The Brier score is a measure of prediction error with the mean squared difference between the predicted probability and the actual outcome, and values range from 0 (perfect prediction) to 0.25 (a noninformative model).[10]

Results

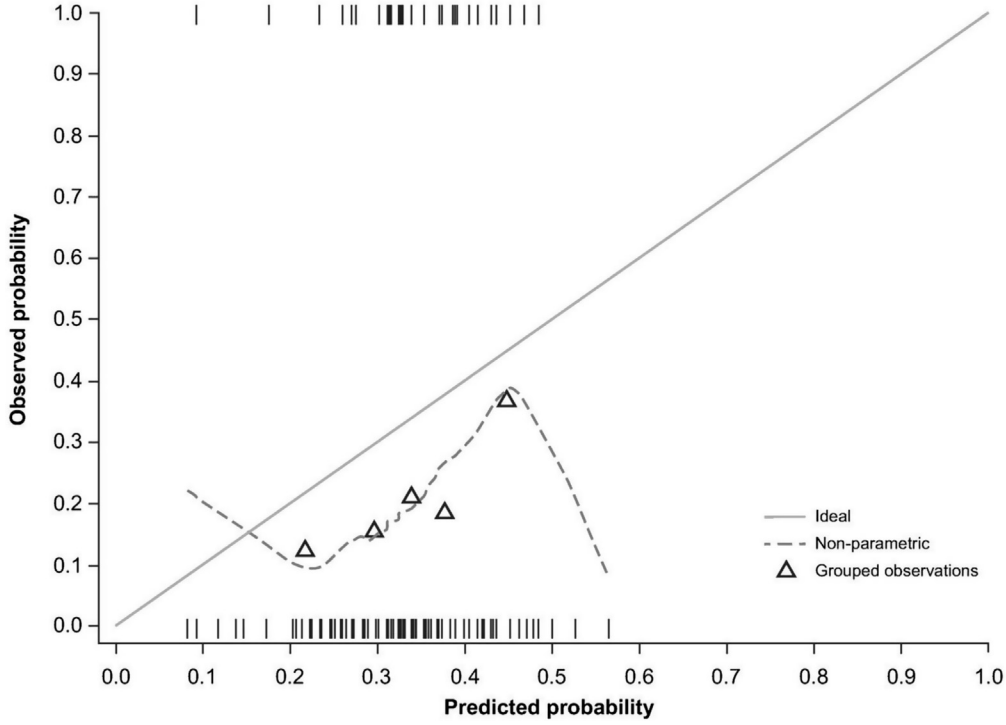
Variable Selection

A binary categorization was chosen for disease duration ($\geq$2 years vs $<$2 years) because nonlinearity was observed for the association between disease duration and corticosteroid-free remission, and a binary categorization was thought to be easier to interpret and apply clinically. Previous tumor necrosis factor (TNF) antagonist exposure was used instead of previous TNF antagonist failure, given the inability to accurately classify failure subtypes in routine practice and similarity in significance between these variables. Baseline endoscopy was used as a metric for disease activity instead of baseline stool frequency, given the subjectivity in stool frequency assessment in routine practice, the observation that the significance for stool frequency was being driven by the comparison of severe (Mayo score stool frequency 3) vs inactive (Mayo score stool frequency 0) disease, the inclusion of endoscopic disease activity as an endpoint for the model, and observed collinearity between baseline endoscopy and baseline stool frequency ($P < .01$). Current or prior smoking was further investigated as individual variables instead of as a composite variable, when separating this into current smokers vs nonsmokers and prior smokers vs nonsmokers, it was found to not meet the thresholds for inclusion and was therefore excluded before the second step of model building.

Model Equation Example

A 35-year-old man with UC is being considered for VDZ therapy. His UC was diagnosed 15 months prior, and he had no prior TNF antagonist exposure. Baseline endoscopy is performed and is notable for the absence of vascular pattern, marked erythema, and friability, without ulcers or spontaneous bleeding (Mayo endoscopic subscore 2). Baseline lab results are notable for an albumin of 40 g/L (4 g/dL).

$$\begin{aligned}&\text{This patient's model calculation} = \\ &-3.7038 + [0.2820] + [0.1847] \\ &+ [0.0647 \times 40] = -0.7464\end{aligned}$$



Supplementary Figure 1. The calibration plot compares the predicted risk for all patients to their expected risk. The ideal line (gray) shows perfect calibration (ie, predicted risk is equal to observed risk). The nonparametric line (dashed, gray) shows the risk for patients over the entire range of risks predicted by the model. The grouped observations (triangles) are the average risks within each quintile group. The calibration intercept (calibration-in-the-large) and slope were estimated from the calibration curve. The calibration-in-the-large compares the systematic difference between the average observed risk and predicted risk and should ideally be equal to 0, indicating no systematic difference. The calibration-in-the-large estimate is –0.674 (95% confidence interval, –1.379 to 0.031) for the current model. The calibration slope gives an indication of overfitting and should ideally be equal to 1, indicating no overfitting or underfitting. Its estimate is 0.999 (95% confidence interval, 0.029 to 1.968) for the current model. A format test of calibration can be done by a likelihood ratio test so that the intercept is 0 and slope is 1. This test confirms that the model predicts risks approximately twice as large as observed on the odds scale (exp(0.674) = 1.96) (likelihood ratio $\chi^2 = 16.18$, $df = 2$, $P = .00031$), confirming model miscalibration.

Supplementary Table 1. Comparison of Demographics Between the VICTORY Cohorts Included in Validation

	VICTORY Consortium		
	Entire Vedolizumab-Treated Cohort (N = 352)[a]	Vedolizumab-Treated Validation Cohort (n = 199)	*P* value
Female	184 (52)	104 (52)	1.00
Smoker (never)	256 (73)	144 (72)	.92
Age, *y*	41.6 ± 17.3	41.5 ± 17.3	.95
BMI, kg/m^2	25.6 ± 5.77	25.3 ± 5.83	.56
Disease duration, *y*	6.0 (2–12)	6.0 (2–12)	1.00
Disease duration <2 *y*	54 (16)	31 (16)	1.00
No prior hospitalization	93 (26)	55 (28)	.77
Prior TNF antagonist exposure	231 (66)	135 (68)	.64
Prior TNF antagonist failure	194 (55)	117 (59)	.42
Extensive baseline disease	210 (60)	112 (56)	.47
Baseline moderate endoscopic disease	237 (67)	126 (63)	.35
Baseline albumin, *g/L*	39.1 ± 5.54	39.4 ± 5.41	.54
Concomitant CS only	118 (34)	69 (35)	.78
Concomitant IMMs only	50 (14)	36 (18)	.27
Concomitant CS and IMMs	73 (21)	49 (25)	.37

Values are n (%), mean ± SD, or median (interquartile range).
BMI, body mass index; CS, corticosteroid; IMM, immunomodulator; TNF, tumor necrosis factor; VICTORY, Vedolizumab for Health Outcomes in Inflammatory Bowel Diseases.
[a]Includes patients who were ultimately excluded for lack of endoscopic follow-up.

Supplementary Table 2. Univariable Analyses for Corticosteroid-Free Remission After 52 Weeks of Vedolizumab in the GEMINI 1 Trial Cohort

	Univariable	
Baseline variable	Odds ratio	*P* value
Age	1.00	.958
Sex (female vs male)	1.34	.119
Ethnicity (other vs non-Hispanic/Latino)	0.78	.549
Race (non-white vs white)	0.82	.450
Smoker (current/previous vs nonsmoker)	0.72	.091
Smoker (current vs nonsmoker)	0.78	.854
Smoker (previous vs nonsmoker)	0.71	.413
BMI	1.03	.046
Disease duration (continuous)	1.04	.017
Disease duration (≥5 y vs <5 y)	1.28	.192
Disease duration (≥2 y vs <2 y)	1.49	.107
EIM (no vs yes)	1.03	.890
Prior hospitalization (no vs yes)	1.21	.352
Previous TNF antagonist exposure (no vs yes)	1.84	.001
Previous TNF antagonist failure (no vs yes)	1.88	.001
Concomitant corticosteroids (no vs yes)	0.99	.952
Concomitant immunomodulator (no vs yes)	0.72	.085
Baseline severe disease, full Mayo score (<10 vs ≥10)	1.62	.023
Baseline full Mayo score	0.91	.079
Baseline partial Mayo score	0.94	.246
Baseline rectal bleeding Mayo score (1 vs 0)	1.37	.717
Baseline rectal bleeding Mayo score (2 vs 0)	1.70	.292
Baseline rectal bleeding Mayo score (3 vs 0)	1.92	.209
Baseline stool frequency Mayo score (0–2 vs 3)	1.70	.005
Baseline stool frequency Mayo score (1 vs 0)	1.52	.010
Baseline stool frequency Mayo score (2 vs 0)	1.25	.409
Baseline stool frequency Mayo score (3 vs 0)	0.75	.015
Baseline endoscopy (moderate vs severe)	1.57	.016
Baseline albumin	1.08	<.001
Baseline albumin (>35 g/L vs ≤35 g/L)	1.88	.002
Baseline fecal calprotectin	1.00	.939

BMI, body mass index; EIM, extraintestinal manifestation; TNF, tumor necrosis factor.

Supplementary Table 3. Baseline Differences in Demographics Between Longer Disease Duration (≥2 *y*) and Shorter Disease Duration (<2 *y*) Patients in the GEMINI 1 Cohort

Variable	Short disease duration (<2 y) (n = 126)	Long disease duration (≥2 y) (n = 492)
Age, *y*	33 (25–51)	40 (31–49)
Female	50 (40)	205 (42)
BMI, kg/m^2	23 (21–27)	24 (22–28)
Disease duration, *y*[a]	1.2 (0.7–1.5)	6.2 (3.7–10.5)
Prior hospitalization	65 (52)	145 (29)
Baseline EIM	37 (29)	177 (36)
Previous TNF antagonist exposure	48 (38)	263 (53)
Previous TNF antagonist failure	44 (35)	222 (45)
Concomitant CS only	73 (58)	252 (51)
Concomitant IMMs only	41 (33)	171 (35)
Baseline endoscopy moderate	63 (50)	214 (44)
Baseline partial Mayo score	6 (5–7)	6 (5–7)
Baseline calprotectin, mg/kg	952 (305–1800)	832 (355–1727)
Baseline albumin, *g/L*	36 (33–41)	37 (34–40)
VDZ maintenance q4	105 (83)	391 (79)

BMI, body mass index; CS, corticosteroid; EIM, extraintestinal manifestation; IMM, immunomodulator; q, quartile; TNF, tumor necrosis factor.
[a]2 patients had missing data for disease duration.

Supplementary Table 4. Differences in Median Measured VDZ Concentrations Over 52 Weeks in GEMINI 1 Trial Derivation Cohort

Time point	Low probability (μg/mL)	Intermediate probability (μg/mL)	High probability (μg/mL)	*P* value (overall)	*P* value (low vs intermediate)	*P* Value (Low vs High)	*P* Value (Intermediate vs High)
Baseline predose	0	0	0	—	—	—	—
Baseline postdose	93.8 (78.7–108)	92.6 (78–108)	98.3 (80.4–117)	.095	.842	.103	.034[a]
Week 2 predose	22.9 (17.6–29.2)	27.4 (23.3–35.7)	32 (26.6–39.5)	<.001[a]	<.001[a]	<.001[a]	<.001[a]
Week 2 postdose	112 (94.9–132)	115.5 (96.2–139)	129.5 (108–149.5)	<.001[a]	.593	<.001[a]	<.001[a]
Week 4	33.8 (25.5–45.4)	42.25 (33.1–53.6)	53.8 (41.4–64.5)	<.001[a]	<.001[a]	<.001[a]	<.001[a]
Week 6 predose	17.2 (10.5–25.2)	23.5 (16.85–33.3)	34.85 (25.9–43.6)	<.001[a]	<.001[a]	<.001[a]	<.001[a]
Week 6 postdose	108 (89.8–128)	113 (93.9–136)	126 (102–150)	<.001[a]	.062	<.001[a]	<.001[a]
Week 14	15.7 (7.79–24.1)	21.25 (12.9–32.7)	29.45 (18–43.5)	<.001[a]	<.001[a]	<.001[a]	<.001[a]
Week 22 predose	17.95 (9.27–29.8)	23.8 (13.5–37.55)	32.45 (17.55–48.1)	<.001[a]	.007[a]	<.001[a]	.004[a]
Week 22 postdose	109 (91.55–129.5)	112.5 (94.6–142)	125 (103–145)	.032[a]	.285	.009[a]	.059
Week 38	16.9 (7.59–32.7)	26.3 (13.7–42.6)	29.5 (15.9–50.2)	.002[a]	.010[a]	<.001[a]	.154
Week 46 predose	22.5 (7.97–36.4)	27.8 (13.6–42.8)	31.5 (14.1–52.3)	.015[a]	.069	.005[a]	.104
Week 46 postdose	108 (91.3–129)	115 (90–143)	127 (104–153)	.016[a]	.368	.008[a]	.023[a]
Week 52	35.25 (19.6–60.65)	43.1 (20.5–63.1)	52.25 (21.1–77.6)	.053	.498	.028[a]	.050

Values are median (interquartile range). All postdose assessments were done 2 hours postdose.
VDZ, vedolizumab.
[a]Closed test procedure used to control for type I error.

Supplementary Table 5. Differences in Partial Mayo Score Over 52 Weeks in GEMINI 1 Trial Derivation Cohort

Time point	Low probability; mean change from baseline (interquartile range)	Intermediate probability; mean change from baseline (interquartile range)	High probability; mean change from baseline (interquartile range)	*P* value (overall)	*P* value (low vs intermediate)	*P* value (low vs high)	*P* value (intermediate vs high)
Baseline	0	0	0	—	—	—	—
Week 2	–0.67 (–1.0 to 0.0)	–0.98 (–2.0 to 0.0)	–1.11 (–2.0 to 0.0)	.003[a]	.007[a]	.001[a]	.230
Week 4	–1.18 (–2.0 to 0.0)	–1.71 (–3.0 to 0.0)	–1.91 (–3.0 to 0.0)	<.001[a]	.001[a]	<.001[a]	.134
Week 6	–1.38 (–2.0 to 0.0)	–1.88 (–3.0 to 0.0)	–2.08 (–3.0 to –1.0)	<.001[a]	.002[a]	<.001[a]	.135
Week 10	–1.81 (–3.0 to 0.0)	–2.31 (–4.0 to 0.0)	–2.73 (–4.0 to –1.0)	<.001[a]	.007[a]	<.001[a]	.021[a]
Week 14	–2.19 (–4.0 to –1.0)	–2.51 (–4.0 to –1.0)	–3.06 (–4.0 to –2.0)	<.001[a]	.027[a]	<.001[a]	.003[a]
Week 18	–2.44 (–4.0 to –1.0)	–2.74 (–4.0 to –1.0)	–3.23 (–5.0 to –2.0)	<.001[a]	.050	<.001[a]	.005[a]
Week 22	–3.04 (–4.0 to –1.0)	–3.16 (–5.0 to –1.0)	–3.55 (–5.0 to –2.0)	.003[a]	.064	<.001[a]	.028[a]
Week 26	–3.15 (–5.0 to –2.0)	–3.45 (–5.0 to –2.0)	–3.77 (–5.0 to –2.0)	.002[a]	.030[a]	<.001[a]	.055
Week 30	–3.29 (–5.0 to –1.0)	–3.75 (–6.0 to –2.0)	–3.75 (–5.0 to –3.0)	.004[a]	.004[a]	.002[a]	.509
Week 34	–3.26 (–5.0 to –1.0)	–4.06 (–6.0 to –2.0)	–3.77 (–5.0 to –2.0)	<.001[a]	<.001[a]	<.001[a]	.805
Week 38	–3.52 (–5.0 to –2.0)	–4.22 (–6.0 to –3.0)	–3.95 (–6.0 to –3.0)	.002[a]	<.001[a]	.004[a]	.732
Week 42	–3.74 (–5.0 to –2.0)	–4.28 (–6.0 to –3.0)	–4.16 (–6.0 to –3.0)	.015[a]	.006[a]	.010[a]	.937
Week 46	–3.92 (–5.0 to –2.0)	–4.30 (–6.0 to –3.0)	–4.14 (–6.0 to –3.0)	.074	.037[a]	.035[a]	.812
Week 50	–3.94 (–5.0 to –2.0)	–4.45 (–6.0 to –3.0)	–4.18 (–6.0 to –3.0)	.145	.056	.098	.853
Week 52	–3.88 (–6.0 to –2.0)	–4.44 (–6.0 to –3.0)	–4.17 (–6.0 to –3.0)	.029[a]	.009[a]	.031[a]	.702

[a]Closed test procedure used to control for type I error.

Supplementary Table 6. Model Performance in GEMINI 1 Trial Derivation Cohort

Cutoff	Sensitivity (95% CI) (%)	Specificity (95% CI) (%)	PPV (95% CI)	NPV (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)
26 points	87.5 (81.8–91.9)	28.1 (23.9–32.6)	34.0 (29.8–38.5)	84.1 (77.2–89.7)	1.22 (1.12–1.32)	0.44 (0.29–0.67)
32 points	34.2 (27.4–41.6)	82.5 (78.6–85.9)	45.3 (36.9–54.0)	74.7 (70.6–78.6)	1.96 (1.47–2.60)	0.80 (0.71–0.89)

Area under the receiver-operating characteristic curve 0.65, Brier score 0.18, Nagelkerke R-square 0.07, Hosmer-Lemeshow goodness-of-fit *P* value = .46.
CI, confidence interval; LR, likelihood ratio; NPV, negative predictive value; PPV, positive predictive value.

Supplementary Table 7. Anti-TNF–Treated UC Patients

	Anti-TNF Cohort (n = 123)
Female	57 (46)
Smoker (never)	85 (69)
Age, *y*	37.7 ± 15.7
Body mass index, kg/m^2	25.4 ± 7.8
Disease duration, y	3 (1–9)
Disease duration <2 y	84 (68)
Prior hospitalization	72 (58)
Prior TNF antagonist exposure	41 (33)
Prior TNF antagonist failure	27 (22)
Extensive baseline disease	83 (68)
Baseline moderate endoscopic disease	54 (44)
Baseline albumin, g/L	3.6 ± 0.6

Values are n (%), mean ± SD, or median (interquartile range).
TNF, tumor necrosis factor; UC, ulcerative colitis.