Statistical Analysis Plan for VA Cooperative Studies Program (CSP) #596 (OpTION) Optimal Treatment for Recurrent Clostridium difficile Infection

NCT number: NCT02667418

Version date: July 30, 2021

Statistical Analysis Plan

for

VA Cooperative Study Program (CSP) #596

(OpTION)

Optimal Treatment for Recurrent Clostridium Difficile Infection (CDI)

(Version 07.30.2021)

OVERVIEW

This is a double-blinded, randomized, prospective pivotal trial. Veterans presenting with a first or second *C. difficile* infection (CDI) recurrence and satisfy all inclusion/exclusion criteria will be consented, screened, and randomly assigned in a double-blind manner and equal ratio (1:1:1) into one of the 3 treatment groups: 1) a 10 day course (125 mg QID per day) of oral vancomycin (VAN) , 2) a 10 day course (200 mg BID per day) of fidaxomicin (FID) or 3) a 31 day course of (125 mg QIDx10 days, QDx7 days, QODx7 days, Q3Dx7 days) vancomycin which includes a taper and pulse following daily treatment (VAN-TP). Participants will be followed up to 90 days (5 clinic visits at screening, days 10, 31, 59 and 90, and 10 phone calls at days 5, 17, 24, 38, 45, 52, 66, 73, 80, 87) for assessment of primary outcome (day 59) and various secondary outcomes (day 38, 59, 90).

The main objective of this study is to determine whether fidaxomicin treatment (FID) or vancomycin taper treatment (VAN-TP) are superior to standard vancomycin (VAN) for sustained clinical response in diarrhea composite outcome (D-COM) by day 59. If both FID and VAN-TP are found to be superior to VAN, then the non-inferiority of VAN-TP to FID will be assessed. The study plans to recruit 459 participants (includes 3% dropout by day 59) to obtain 85% global power to detect a 16% absolute difference in a composite outcome (D-COM) for at least one comparison (VAN-TP vs. VAN, FID vs. VAN) at the family-wised error rate (FWER) 0.05 level (2-sided). The marginal probability (disjunctive power) of detecting 16% absolute difference in each comparison is 60%. Two interim looks for efficacy are proposed for the study (i.e., 40% and 70% of information collected). The proposed study will be the first to systematically compare the effectiveness of three alternative therapies for the management of recurrent CDI for which there is currently inadequate comparative data.

OBJECTIVES

Primary Objective:

To determine whether fidaxomicin treatment (FID) or vancomycin taper treatment (VAN-TP) is superior to standard vancomycin (VAN) for the composite outcome (D-COM) of sustained clinical response at day 59.

Secondary objectives:

1) The comparison of sustained clinical response rates (D-COM) at 28 days post end of therapy and at day 90;
2) The comparison of sustained clinical response rates in *C Difficile* composite outcome (CDI-COM) at 28 days post end of therapy, at day 59 and at day 90;
3) The comparison of symptom resolution rate at day 10;
4) The comparison of diarrhea recurrence rate and the CDI recurrence rate following symptom resolution;
5) Compare the sustained clinical response rates (D-COM, CDI-COM) at day 59 in participants
    a) infected with the BI/NAP1/027 strain of *C. difficile* or non-BI/NAP1/027 strains
    b) 1 or 2 previous CDI episodes within 8 weeks of enrollment
    c) receiving concomitant antibiotics (during CDI treatment and subsequent follow up period till day 59) and those without concomitant antibiotics
    d) by correlating sustained clinical response rates with Horn's Severity of Illness and ATLAS severity scores
6) Compare the change in C.diff HRQOL from baseline to day 10 (end of treatment of oral Vancomycin and Fidaxomicin) and to day 59 (primary outcome assessment).
7) Evaluate safety of the three treatment groups
8) The assessment of non-inferiority of VAN-TP treatment to FID treatment (day 59), if both VAN-TP and FID are superior to standard VAN in sustained D-COM at day 59.

Goal of the pilot phase: Prior to initiation of the full study, a one-year pilot phase will be conducted in the same manner as the full study to: 1) evaluate compliance with and efficiency of the primary data collection tool, the simplified daily patient stool diary; 2) develop a patient-centered outcome questionnaire; and 3) assess the recruitment rate of the study.

OUTCOMES

Primary Outcome: Proportion of subjects who achieve sustained clinical response in diarrhea composite outcome (D-COM) that includes symptom resolution (defined as an improvement or resolution of diarrhea with $\leq$3 unformed bowel movements over 24 hours for 48 consecutive hours by day 10) during treatment without any of the following by day 59:

1. Diarrhea recurrence (defined as having diarrhea with >3 loose or semi-formed stools over 24 hours for 48 consecutive hours following initial resolution)

2. Other non-fatal clinical events including severe abdominal pain, toxic megacolon, and colectomy
3. Death

Secondary Outcomes:

- Sustained clinical response in Diarrhea Composite Outcome (D-COM) at 28 days post end of therapy (day 38 for VAN and FID, day 59 for VAN-TP) and at day 90.
- Sustained clinical response in CDI Composite Outcome (CDI-COM) at 28 days post end of therapy (day 38 for VAN and FID, day 59 for VAN-TP), at day 59 and at day 90. Sustained response in CDI-COM is defined using the same composite endpoint criteria as was used in the D-COM composite outcome but with confirmation of no CDI recurrence by a negative *C. difficile* stool assay test (i.e., proportion of subjects who achieve symptom resolution by day 10 without recurrent CDI, without non-fatal clinical events, and without death).
- Proportion of subjects with symptom resolution by day 10. Days from randomization to symptom resolution (measured from day 1-10).
- Diarrhea recurrence and diarrhea recurrence with confirmation of recurrent CDI following initial symptom resolution.
- Sustained clinical response in D-COM and CDI-COM at day 59 for subgroups (infection with the BI/NAP1/027 strain at baseline (yes, no), 1 or 2 previous CDI episode at baseline, receipt of concomitant antibiotics during study treatment (yes, no), correlated to Horn's index score and ATLAS severity score
- Change in patient reported C.*diff* Health Related Quality of Life (HRQOL) total score (CDiff32-QOL) and sub-scales (physical QOL, emotional/psychological QOL, social QOL) from baseline (day 0) to day 10 and to day 59. Total score and sub-scale are transformed to a 100-point scale with higher score indicating better QOL overall or in that function domain.
- Safety measures
    - Laboratory parameters (CBC and serum chemistry panel to include Cr, AST, ALT, alkaline phosphatase, albumin, total bilirubin) on day 0, day 10, and day 31
    - Treatment related AEs and SAEs
    - Treatment discontinuation due to any AE or SAE

SAMPLE SIZE AND POWER

EAST 6.5 allows sample size and power calculation for a multi-arm multi-stage (MAMS) design that compares multiple active treatments to a common control for a binary outcome [Chapter 38 of East 6.5 user manual] utilizing the generalization of a single-step Dunnett's test [1, 2] with unpooled variance. The study was first designed with 549 participants to obtain 91% global power to detect a 16% absolute difference in composite outcome (D-COM) for at least one comparison (VAN-TP vs. VAN, FID vs. VAN). Due to the suspension of recruitment due to the COVID-19 pandemic and its impact on the slow recruitment after the site resumed recruitment activities, the sample size of the study was reduced to 459, and the global power was reduced by 85%. The parameters used for sample size calculation are as follows.

Parameters used for sample size calculation

- Sustained D-COM rates at day 59 will be **31%** for VAN (control) and be **16**% higher in each active treatment (VAN-TP or FID) than that of VAN.
- Equal allocation to each arm.
- Two active treatment arms (VAN-TP, FID) and one active control (VAN), so there are 2 sets of hypotheses to test the efficacy of individual treatment (VAN-TP, FID) compared to VAN in sustained day 59 D-COM rate.

  Hypothesis H_1: H_{10}: $\delta_1 = P_1 - P_0 = 0$ vs. H_{11}: $\delta_1 \neq 0$ (δ_1=16%)
  Hypothesis H_2: H_{20}: $\delta_2 = P_2 - P_0 = 0$ vs. H_{21}: $\delta_2 \neq 0$ (δ_2=16%)

  P_1 and P_2 are the sustained day 59 D-COM rates for active treatment arms (VAN-TP, FID), and P_0 is the sustained day 59 D-COM rate for control arm (VAN). The family-wise error rate (FWER) for both comparisons is controlled at 0.05 level (2-sided), which is the probability of falsely claiming a treatment difference when it is actually no difference (i.e., reject any H_{i0} where δ_i=0, i=1, 2).

- Global power (i.e., the probability of rejecting at least one null hypothesis H_{i0} (i.e., δ_i=0) given H_{i1} δ_i=16%) is 85%.
- Sample size inflation due to 2 interim looks when 40% and 70% of target randomized participants have completed the 59-day follow-up to allow early stopping of active treatment arms for efficacy. Lan-DeMets O'Brien-Fleming α spending boundary function is used for efficacy boundary.
- Assumes that up to 3% of randomized subjects will not be included in the modified intent-to-treat (mITT) analysis population, which is considered as the primary analysis population

for efficacy. This population includes all randomized subjects who received at least one dose of study medication and does not violate study entry criteria.

- Subjects failed to achieve symptom resolution by day 10 or dropouts prior to day 59 due to adverse events are considered as failures for the D-COM. Subjects withdrawal prior day 59 for unknown reasons or reasons unrelated to the study drug will be handled by multiple imputation. Therefore, the sample size will not be further inflated with dropouts.

Using the software EAST (version 6.5) module MaMs, 444 participants (148 per group) will yield a 85% global power to detect a 16% absolute difference in sustained day-59 D-COM rate for at least one comparison (VAN-TP vs. VAN, or FID vs. VAN). Given ~3% subjects randomized will not be included in the modified intent to treat (mITT) analysis population, 459 subjects (153 per group) will be recruited to the study. The sample size has been adjusted for reasons listed in the preceding paragraphs. Powers at varying control response rate (31-43%) and treatment effect (16-22%) are given in **Table 1**.

Table 2 gives several powers of the study with the proposed efficacy boundary at different alternative hypotheses of treatment effect. Each result is based on 10,000 simulations with 444 subjects (148 per group) per simulation using East v6.5 and assumes two options: a) trial stops if any arm (VAN-TP and/or FID) crosses efficacy boundary, and b) trial continues with the remaining arms that do not enter efficacy region and only terminates the arm that crosses the efficacy boundary. Option b helps to detect a treatment with delayed effect.

For example, for option a, when the day 59 D-COM rates of both active treatments (VAN-TP, FID) are 16% higher than that of VAN control (Scenario 1.4), the study has an 86.5% probability (global power) of declaring any active treatment (VAN-TP, FID) to be efficacious (i.e., rejecting any null hypotheses of no treatment difference), approximately 60% probability of declaring an active treatment arm to be efficacious (P(1), P(2)), and a 34% probability of claiming both treatment arms to be efficacious (conjunctive power). When the D-COM rate in one active treatment is 16% higher than VAN control, and the other one has no difference (Scenario 1.2), the study has 73% probability claiming the treatment with 16% difference (treatment 1) to be efficacious and 1.8% chance of falsely declaring treatment 2 different from control. The global power is 74%. When both active treatments have no difference from VAN control (Scenario 1.1), the FWER is 5.1%.

For option b, when the day 59 D-COM rates of both active treatments (VAN-TP, FID) are higher than that of VAN control and in similar magnitudes (Scenarios 1.4, 1.6, 1.8), with the same parameter assumptions the probabilities of declaring an active treatment arm and both treatment

arms to be efficacious than VAN control are higher than those in option a, but the expected sample sizes are increased. The global powers remain the same (Scenarios 2.4, 2.6, 2.8). For example (scenario 2.4), the study has an 86.5% probability of declaring any active treatment (VAN-TP or FID) to be efficacious (i.e., rejecting any null hypotheses of no treatment difference), approximately 73% probability of declaring an active treatment arm to be efficacious (compared to 60% in scenario 1.4), and a 59% probability of claiming both treatment arms to be efficacious (compared to 34% in scenario 1.4). The expected sample size is increased from 360 (scenario 1.4) to 394 (scenario 2.4). When the D-COM rate in arm 1 is much higher than arm 2 (or vice versa) and are both higher than VAN control (Scenarios 1.3, 1.5, 1.7), option b (Scenarios 2.3, 2.5, 2.7) improves the probability of detecting the active treatment with lower D-COM rate. The probability of detecting the active treatment with higher D-COM rate and the global power remain vastly unchanged.

Table 3 shows the probability of stopping one or both active treatment arms for efficacy at each interim analysis for a range of alternative hypothesis. Each result is based on 10,000 simulations (East v6.5). For example, for option a, when the D-COM rates of both treatments (VAN-TP, FID) are 16% higher than that of VAN control (Scenario 1.4), the chance of declaring one or both treatment arms to be efficacious than VAN is 7.8% at the very first look, 47.8% at the 2nd look and 30.9% at the final look, which yields 86.5% total probability. The marginal probability for each treatment arm (VAN-TP, FID) across efficacy boundary is about 4% at look 1, 33% at look 2 and 23% at final look. The marginal power for each treatment arm is 60-61%. For option b, with the same parameter assumptions (Scenario 1.4), the chance of declaring one or both treatment arms to be efficacious than VAN is similar to option a (i.e., 8% at the very first look, 48.4% at the 2nd look and 30.1% at the final look, which yields 86.5% total probability). The marginal probability for each treatment arm (VAN-TP, FID) across efficacy boundary is about 4.5% at look 1, 35% at look 2 and 33% at the final look.

Table 1. Power at Varying Control Response Rate and Treatment Effect

(2 interim looks when 40% and 70% information is obtained)

FWER (α) 2-sided	Control Group Sustained D-COM Rate by Day 59 (P_0)	Treatment Group Sustained D-COM Rate by Day 59 (P_1, P_2)	$\delta_1=\delta_2$	Global Power [a]	N per Group (total for 3 groups)	
					No Inflation	Inflated with 3%
0.05	31%	47%	16%	85.2%	148 (444)	153 (459)
0.05	35%	51%	16%	84.8%	148 (444)	153 (459)
0.05	39%	55%	16%	84.8%	148 (444)	153 (459)
0.05	43%	59%	16%	85.3%	148 (444)	153 (459)
0.05	31%	49%	18%	92.5%	148 (444)	153 (459)
0.05	35%	53%	18%	92.2%	148 (444)	153 (459)
0.05	39%	57%	18%	92.3%	148 (444)	153 (459)
0.05	43%	61%	18%	92.7%	148 (444)	153 (459)
0.05	31%	51%	20%	96.7%	148 (444)	153 (459)
0.05	35%	55%	20%	96.6%	148 (444)	153 (459)
0.05	39%	59%	20%	96.7%	148 (444)	153 (459)
0.05	43%	63%	20%	96.9%	148 (444)	153 (459)
0.05	31%	53%	22%	98.8%	148 (444)	153 (459)
0.05	35%	57%	22%	98.7%	148 (444)	153 (459)
0.05	39%	61%	22%	98.8%	148 (444)	153 (459)
0.05	43%	65%	22%	98.9%	148 (444)	153 (459)

[a] Reject at least one hypothesis (H_{10} or H_{20} or both)

The highlighted information identifies the assumptions used for this study.

Table 2. Power of the Study from 10,000 Simulations

(each simulation has 444 subjects, 2 interim looks at 40% and 70%, and 1 final look, 2-sided test)

Scenario	Sustained D-COM Rate by Day 59 (P_0, P_1, P_2) [a]	P(1) [b]	P(2) [b]	P(1 only) [c]	P(2 only) [c]	P(1&2) [d]	P(none) [e]	Disjunctive Power: P(reject any H_{i0} where $\delta i \neq 0$)	Conjunctive power: P(reject both H_{i0} where $\delta i \neq 0$)	Global Power: P(reject any H_{i0})	FWER= P(reject any H_{i0} where $\delta i = 0$)	Expected Sample Size
a. Trial stops, if any treatment (VAN-TP or FID) crosses efficacy boundary												
1.1	(0.31, 0.31, 0.31)	2.81%	2.71%	2.4%	2.3%	0.41%	94.89%	-	-	5.11%	5.11%	442.6
1.2	(0.31, 0.47, 0.31)	73.05%	1.75%	72.28%	0.98%	0.77%	25.96%	73.05%	73.05%	74.03%	1.75%	386
1.3	(0.31, 0.47, 0.36)	72.32%	4.8%	68.54%	1.02%	3.78%	26.66%	73.34%	3.78%	73.34%	-	385.8
1.4	(0.31, 0.47, 0.47)	60.51%	60.01%	26.47%	25.97%	34.04%	13.52%	86.48%	34.04%	86.48%	-	360.1
1.5	(0.31, 0.51, 0.37)	89.66%	4.4%	85.79%	0.53%	3.87%	9.81%	90.19%	3.87%	90.19%	-	346.3
1.6	(0.31, 0.51, 0.51)	72.45%	72.66%	24.36%	24.57%	48.09%	2.98%	97.02%	48.09%	97.02%	-	319
1.7	(0.31, 0.55, 0.39)	97.36%	5.59%	92.07%	0.3%	5.29%	2.34%	97.66%	5.29%	97.66%	-	309.7
1.8	(0.31, 0.55, 0.55)	77.77%	77.78%	21.78%	21.79%	55.99%	0.44%	99.56%	55.99%	99.56%	-	280.3
b. Trial continues with treatments which have not crossed efficacy boundary												
2.1	(0.31, 0.31, 0.31)	2.65%	2.74%	2.25%	2.34%	0.4%	95.01%	-	-	4.99%	4.99%	443.3
2.2	(0.31, 0.47, 0.31)	73.52%	2.58%	71.97%	1.03%	1.55%	25.45%	73.52%	73.52%	74.55%	2.58%	423.9
2.3	(0.31, 0.47, 0.36)	73.28%	10%	63.91%	0.63%	9.37%	26.09%	73.91%	9.37%	73.91%	-	422
2.4	(0.31, 0.47, 0.47)	72.68%	73.02%	13.43%	13.77%	59.25%	13.55%	86.45%	59.25%	86.45%	-	394.3
2.5	(0.31, 0.51, 0.37)	90.02%	12.46%	77.61%	0.05%	12.41%	9.93%	90.07%	12.41%	90.07%	-	408.4
2.6	(0.31, 0.51, 0.51)	89.85%	89.86%	6.94%	6.95%	82.91%	3.2%	96.8%	82.91%	96.8%	-	357.8
2.7	(0.31, 0.55, 0.39)	97.62%	22.08%	75.6%	0.06%	22.02%	2.32%	97.68%	22.02%	97.68%	-	391.4
2.8	(0.31, 0.55, 0.55)	97.7%	97.58%	2.11%	1.99%	95.59%	0.31%	99.69%	95.59%	99.69%	-	318.5

[a] Sustained day 59 D-COM rate for active control group (VAN) is P_0; for treatment groups (VAN-TP or FID) are P_1 and P_2.
[b] P(*i*), *i*=1 or 2, is the probability that treatment *i* (VAN-TP or FID) is declared significantly different to control (VAN) (i.e., cross efficacy boundary)
[c] P(*i* only), *i*=1 or 2, is the probability that treatment *i* only is declared significantly different to control (VAN).
[d] P(1&2) is the probability that both treatments (VAN-TP, FID) are declared significantly different to control (VAN).
[e] P (none) is the probability that neither VAN-TP nor FID are declared significantly different to control (VAN).

Note: The disjunctive power and conjunctive power for scenario 2 are based on the comparison of P_1 vs. P_0 where $\delta_1 \neq 0$.
The comparison of P_2 vs. P_0 under null yields type I error rate.

P(1 only) + P(2 only) + P(1&2) + P(none) =100%; and P(1 only) + P(2 only) + P(1&2) = Global power.

The highlighted information identifies the assumptions used for this study.

Table 3. Probability of Trial/Arm Termination at Each Look from 10,000 Simulations

Scenario	Sustained D-COM Rate by Day 59 (P_0, P_1, P_2) [a]	Look	Info. Fraction (n/n max)	P(1) [b]	P(2) [b]	Probability (Incremental) one or two Active arms Claiming Efficacy
a. Trial stops, if any treatment (VAN-TP or FID) crosses efficacy boundary						
1.1	(0.31, 0.31,0.31)	1	0.40	0.06%	0.08%	0.14%
		2	0.70	0.78%	0.64%	1.37%
		3 (Final)	1	1.97%	1.99%	3.6%
			Total	2.81%	2.71%	5.11%
1.2	(0.31, 0.47, 0.31)	1	0.40	4.56%	0.07%	4.6%
		2	0.70	34.55%	0.82%	34.98%
		3 (Final)	1	33.94%	0.88%	34.47%
			Total	73.05%	1.77%	74.05%
1.3	(0.31, 0.47, 0.36)	1	0.40	4.43%	0.23%	4.55%
		2	0.70	34.39%	2.22%	34.88%
		3 (Final)	1	33.5%	2.35%	33.91%
			Total	72.32%	4.8%	73.34%
1.4	(0.31, 0.47, 0.47)	1	0.40	4.37%	4.5%	7.81%
		2	0.70	32.92%	32.91%	47.76%
		3 (Final)	1	23.22%	22.6%	30.91%
			Total	60.51%	60.01%	86.48%
1.5	(0.31, 0.51, 0.37)	1	0.40	10.62%	0.24%	10.73%
		2	0.70	52.02%	2.67%	52.3%
		3 (Final)	1	27.02%	1.49%	27.16%
			Total	89.66%	4.4%	90.19%
1.6	(0.31, 0.51, 0.51)	1	0.40	10.06%	9.81%	16.5%
		2	0.70	47.12%	47.39%	61.34%
		3 (Final)	1	15.27%	15.46%	19.18%
			Total	72.45%	72.66%	97.02%
1.7	(0.31, 0.55, 0.39)	1	0.40	20.03%	0.44%	20.15%
		2	0.70	60.83%	3.97%	60.98%
		3 (Final)	1	16.5%	1.18%	16.53%
			Total	97.36%	5.59%	97.66%
1.8	(0.31, 0.55, 0.55)	1	0.40	19.72%	20.32%	31.78%
		2	0.70	50.94%	50.28%	59.7%
		3 (Final)	1	7.11%	7.18%	8.08%
			Total	77.77%	77.78%	99.56%
b. Trial continues with treatments which have not crossed efficacy boundary						
2.1	(0.31, 0.31,0.31)	1	0.40	0.01%	0.06%	0.07%
		2	0.70	0.73%	0.72%	1.37%

		3 (Final)	1	1.91%	1.96%	3.55%
			Total	2.65%	2.74%	4.99%
2.2	(0.31, 0.47, 0.31)	1	0.40	4.43%	0.04%	4.46%
		2	0.70	35.64%	0.73%	36.08%
		3 (Final)	1	33.45%	1.81%	34.01%
			Total	73.52%	2.58%	74.55%
2.3	(0.31, 0.47, 0.36)	1	0.40	4.27%	0.19%	4.41%
		2	0.70	35.91%	2.76%	36.36%
		3 (Final)	1	33.1%	7.05%	33.14%
			Total	73.28%	10%	73.91%
2.4	(0.31, 0.47, 0.47)	1	0.40	4.57%	4.49%	7.97%
		2	0.70	34.81%	35.55%	48.35%
		3 (Final)	1	33.3%	32.98%	30.13%
			Total	72.68%	73.02%	86.45%
2.5	(0.31, 0.51, 0.37)	1	0.40	10.21%	0.31%	10.3%
		2	0.70	51.99%	3.62%	52.09%
		3 (Final)	1	27.82%	8.53%	27.68%
			Total	90.02%	12.46%	90.07%
2.6	(0.31, 0.51, 0.51)	1	0.40	10.42%	10.34%	17.33%
		2	0.70	51.76%	52.17%	60.92%
		3 (Final)	1	27.67%	27.35%	18.55%
			Total	89.85%	89.86%	96.8%
2.7	(0.31, 0.55, 0.39)	1	0.40	20.6%	0.53%	20.69%
		2	0.70	61.11%	7.43%	61.18%
		3 (Final)	1	15.91%	14.12%	15.81%
			Total	97.62%	22.08%	97.68%
2.8	(0.31, 0.55, 0.55)	1	0.40	19.8%	20.3%	31.53%
		2	0.70	61.95%	61.87%	60.83%
		3 (Final)	1	15.95%	15.41%	7.33%
			Total	97.7%	97.58%	99.69%

Note: The trial continues with remaining arms if any arm enters efficacy region.

The highlighted information identifies the assumptions used for this study.

INTERIM ANALYSIS FOR POTENTIAL EARLY STUDY TERMINATION FOR EFFICACY

There will be two interim looks at the primary outcome for consideration of possible early stopping one or more arms for early detected efficacy. No futility boundary is proposed for the trial. The first interim analysis already conducted when 1/3 participants (mITT population) completed at least 59 days of follow-up under the original sample size is rescaled accordingly to 40% of the reduced sample size (n=177, Table 4). With the reduced sample size, the second interim will occur when 70% of participants are randomized (n= 312 subjects in the mITT population) and completed at least 59 days of follow-up. In order to preserve Type-I error, proposed O'Brien-Fleming stopping boundary in *p*-values and Z-scales are given in **Table 4** below.

Table 4. Stopping Boundary Values at Interim and Final Look
(P_0 =0.31, P_1= P_2=0.47, global power=0.85, α=0.05, 2-sided)

Look #	Sample Size [a] (n)	Information Fraction	Cumulative α Spent [b]	Efficacy Boundary[c]			Boundary Crossing probability (Incremental)	
				Critical p-value	Z scale			
					Lower	Upper	Under H0	Under H1
1	177	0.40	0.0008	0.0004	-3.546	3.546	0.001	0.054
2	312	0.70	0.015	0.0077	-2.667	2.667	0.014	0.460
3 (Final Analysis)	444	1.00	0.05	0.0242	-2.253	2.253	0.035	0.337
Total							0.05 (α)	0.85 (Global Power)

[a] Sample size is based on 444 subjects included in the mITT population. The 1st and 2nd interim analysis will be conducted, respectively, when 177 and 312 subjects in the mITT population complete day 59 follow-up. Dropout prior day 59 will be handled using MI method except for scenarios defined in the section of "Analysis of Primary Outcome Measures" (page 14). Same efficacy boundary will be used when missing outcome is imputed.

[b] α = Family-wise error rate (2-sided).

[c] If the allocation to the three arms is unequal at one or more interim looks, boundary at each look will be recomputed using the average allocation to each arm as an approximation.

RECRUITMENT AND NUMBER OF PARTICIPATING SITES

Recruitment started with 6 pilot sites and expanded to 24 sites (26 locations. Note that given Gainesville and Hines each recruiting from another VA location, we are referencing the dual location as one unit. e.g., Gainesville and Lake City VA locations are considered one unit). The original timeline is to complete enrollment within 6 years (2 years of pilot phase with 6 sites plus transition period from pilot phase to full study, and 4 years of full study with 26 sites (24 units)) at the assumption that sites will

recruit **6** participants (on average) per site per year for a site that primarily recruits from the main hospital, and **9** participants (on average) per unit per year for sites recruiting from dual locations. Due to the suspension of recruitment by the COVID-19 pandemic and its significant impact on the slow recruitment after the sites resumed recruitment activities, as well as the termination of underperformed sites and the startup of replacement sites, the sample size of the study is reduced from 549 to 459 to yield 444 subjects in the mITT population. The study recruitment timeline is extended with an intention of funding a full 31-month extension of study recruitment till August 31, 2024, pending the 2nd interim analysis outcome in 2023. Site recruitment goal will be readjusted according to the new recruitment timeline.

RANDOMIZATION

The treatment allocation ratio for the three treatment regimens will be 1:1:1 using a permuted block randomization scheme. The randomization will be stratified by study site. We anticipate that on average 6-7 subjects will be randomized to each of the three treatment groups within each site. The number of participates recruited by sites will be more or less than the average number, depending on the site recruitment length and its ability to recruit.

STATISTICAL ANALYSIS

Analysis Population

The modified intent-to-treat population (mITT) will consist of all randomized subjects who received at least one dose of study treatment medication and did not violate study inclusion criteria. Participants will be analyzed according to their randomized study medication. This population will be considered the primary analysis population for analyses on efficacy outcomes unless otherwise specified. Per protocol analysis population (PP) will consist of subgroup of mITT participants who both (1) failed to respond by day 10 or completed day 59 follow-up (note: participants who are transferred to primary care physician prior to day 59 due to failure in achieving sustained D-COM are included since their day 59 D-COM status is known) and (2) took 80% of their assigned drug according to the pill count. This population will be mainly used for secondary analysis of the primary composite outcome as well as the non-inferiority analysis of VAN-TP vs. FID. Non-inferiority analysis will also be repeated on mITT population. For all safety outcomes subjects who took at least one dose of study drug will be analyzed based on the actual treatment they received.

All statistical tests will be 2-sided. The only exception will be the 1-sided non-inferiority test on VAN-TP vs. FID if both primary comparisons (VAN-TP vs VAN, FID vs. VAN) are significant.

Baseline Characteristic Comparisons

The distribution of baseline participant characteristics between the 3 randomization groups will be evaluated using descriptive statistics (n, mean, median, minimum, maximum for continuous variables, and n, percentage for categorical variables). For overall group comparisons parametric F test from one-way ANOVA or Kruskal-Wallis nonparametric test will be used, respectively, for normally or nonnormally distributed continuous variables. Pearson Chi-square test (or Fisher's Exact test for small cell counts) will be used for overall group comparison in nominal categorical variables (i.e., no intrinsic ordering to the categories). For ordered categorical variables, such as Horn's Severity Index (1-4) and ATLAS score (0-10), Cochran-Armitage test for trend will be used.

Analysis of Primary Outcome Measures

The primary outcome measure of sustained clinical response at day 59 will be analyzed using Z-statistic for equality of proportions for each comparison based on the mITT population.

Hypothesis 1: H_{10}: $P_1 - P_0 = 0$ vs. H_{11}: $P_1 - P_0 \neq 0$

Hypothesis 2: H_{20}: $P_2 - P_0 = 0$ vs. H_{21}: $P_2 - P_0 \neq 0$

Z-statistic comparing the *i*th (*i*=1,2) treatment arm (VAN-TP, FID) with the control (VAN) at the *j*th look (*j*=1, 2 and 3):

$Z_{ij} = \frac{\hat{P}_{ij} - \hat{P}_{0j}}{\sqrt{\frac{\hat{P}_{ij}(1-\hat{P}_{ij})}{n_{ij}} + \frac{\hat{P}_{0j}(1-\hat{P}_{0j})}{n_{0j}}}}$, where $\hat{P}_{ij}$ and $\hat{P}_{0j}$ are respectively the sample proportions for treatment *i* and control group from data collected up to the *j*th look.

In comparison of proportion of sustained D-COM in the VAN-TP group or FID group to that of the VAN group, the observed Z-statistic for each comparison will be compared to the efficacy boundary in Z scale ±3.546, ±2.667, and ±2.253, respectively, at looks 1, 2 and 3 (final look) for equal allocation. The difference in proportions for each comparison and associated repeated 95% confidence interval will be reported at each look.

Repeated 95% for δ_i at look *j* (*i*=1, 2; *j*=1, 2 and 3):

$\hat{P}_{ij} - \hat{P}_{0j} \pm c_j \sqrt{\frac{\hat{P}_{ij}(1-\hat{P}_{ij})}{n_{ij}} + \frac{\hat{P}_{0j}(1-\hat{P}_{0j})}{n_{0j}}}$ where δ_i is the treatment difference between treatment *i* (VAN-TP, FID) and control (VAN), and c_j is the efficacy boundary on the Z-scale (i.e. c_1=3.546, c_2=2.667, c_3=2.253).

Possible ways the trial or an arm may be stopped early are listed below.

- Both hypotheses (H_{10}, H_{20}) are rejected at look 1 (|Z statistic| > 3.546)

  → **Recommend stopping trial at look 1**

- Both hypotheses (H_{10}, H_{20}) are rejected at look 2 (|Z statistic| > 2.667)

  → **Recommend stopping trial at look 2**

- One hypothesis is rejected at look 1 (|Z statistic| > 3.546) or at look 2 (|Z statistic| > 2.667)

  →Based on the interim analysis results and the recruitment status, DMC will make recommendation if **trial stops or trial continues with remaining arm not crossing efficacy boundary.**

  Since VAN-TP and FID are usually more effective than VAN to prevent recurrence of CDI in established literatures, therefore it is unlikely that observed Z statistic from either hypothesis will be less than the lower boundary.

Participants who failed to achieve symptom resolution by day 10 or dropped out prior to day 59 due to study-drug related adverse events or because they felt the study drug was ineffective and their symptoms were not improved compared to baseline are considered as failures to achieve sustained day 59 D-COM. Participants who achieved symptom resolution by day 10 but subsequently had CDI symptoms to warrant clinical determination of CDI and were withdrawn from the study for re-treatment for CDI although no 2 consecutive days of ≥4 diarrhea stools are considered as failures for day 59 D-COM. Participants who terminated prior to day 59 for unknown reasons or reasons unrelated to the study treatment (about 10%) will be handled by multiple imputations. Those missing responses at day 59 will have their values imputed from an imputation model which will include baseline covariates that are known risk factors for recurrence of CDI (such as age, prior episode of CDI, baseline CDI severity measured by ATLAS score, baseline comorbid condition measured by modified Horn's index score). Covariates that causes nonconvergence in the MI model will be removed from the imputation model. Multiple imputations will be implemented in SAS PROC MI. We will create n=20 imputed datasets [3] for analysis and combine the results with PROC MIANALYZE.

In addition, a sensitivity analysis will be performed, one assuming all patients with previously imputed values are non-responders. Sensitivity analysis on completers will also be explored.

<u>Secondary Analysis of Primary Outcome Measure</u>

- Per Protocol Analysis of Primary Outcome Measure

The same analysis outlined above will be repeated for per protocol (PP) analysis population. Since the PP population includes only the subjects who took 80% of their assigned study drug and also their day 59 D-COM status (yes or no) are known, there will be no missing data involved. The α spent at the interim looks will be based on the information fractions spent at the interim looks. At the final look, the information fraction is set at 1.

- VAN-TP vs. FID Non-inferiority

If both the FID and VAN-TP vs. VAN comparisons in D-COM (day 59) are significant (above upper efficacy boundary), we will perform a non-inferiority test with minimal loss of power as per Proschan [4]. The test will be a one-side Z-test of non-inferiority of proportions of sustained clinical response at day 59 with the hypothesis that VAN-TP is non-inferior to FID within a margin of 10% at the 0.05 significance level. The analysis will be performed for both PP and mITT populations. The conclusions from mITT analyses will be considered as the primary one. PP analyses will be additionally as a secondary analysis. Based on 296 subjects whose day 59 D-COM status is known, the calculated power is about 53%, assuming both treatments have a 47% response rate, and all 3 arms continue to the final look (PASS v16). Report the test statistic, *p*-value, cell frequencies, estimated proportion difference and one-sided 95% lower confidence limit (equivalent to a 90% two-sided CI) of the difference in response rates. If the one-sided 95% lower confidence limit for $P_{VAN\text{-}TP} - P_{FID}$ lies within (-10%, + ∞), non-inferiority of VAN-TP to FID is achieved based on 5% significance level.

- Subgroup Analysis of Primary Outcome Measure

No interim analysis will be conducted for subgroup analysis on the primary outcome measure. Subgroup analyses to evaluate variation in treatment effect will be explored by performing tests of subgroup by treatment interaction using logistic regression model with sustained clinical response as the dependent variable. Additionally, treatment effects will be displayed for each of six subgroups defined by

BI/NAP1/027 strain positive (yes, no)

2 prior CDI recurrences or 1 prior recurrence

Patient received concomitant antibiotics during study (yes, no).

For each subgroup, report cell frequencies, estimated proportion difference and 95% confidence interval in primary outcome between each treatment group and VAN under Dunnett's adjustment for multiple comparisons. This can be handled, for example, using R package 'MCPAN' for multiple comparisons based on normal approximation and extensions (i.e., function 'binomRDci' provides simultaneous CI of

proportion differences based on Dunnett adjustment, function 'binomRDtest provides associated p-values).

Calculate the overall sustained clinical response rate by ATLAS score (0-10) and group-wise response by ATLAS score, along with cell frequencies. Perform a two-sided Cochran-Armitage test for trend of the 3x11 contingency table with treatment as rows and ATLAS score as columns. Report the cell frequencies, test statistic and *p*-values. If the two-sided test for trend is significant at the 0.05 level, fit a logistic regression model with sustained clinical response as the dependent variable and the following independent variables:

(1) Treatment (Three level class variable: 1=VAN (reference), 2=FID, 3=VAN-TP).

(2) ATLAS score (range 0-10) as a continuous variable.

(3) Treatment * ATLAS interaction

Additionally, if the Cochrane-Armitage test for trend is not significant, a logistic regression analysis with ATLAS treated as a categorical class variable will be performed following the procedure outlined above.

Calculate the overall sustained clinical response rate 59 days post randomization by modified Horn's Index and group-wise response by modified Horn's Index, along with cell frequencies. Perform a two-sided Cochran-Armitage test for trend of the 3x4 contingency table with treatment as rows and modified Horn's Index as columns. Report the cell frequencies, test statistic and *p*-value. If the two-sided test for trend is significant at the 0.05 level, fit a logistic regression model with sustained clinical response as the dependent variable and the following independent variables:

(1) Treatment (Three level class variable: 1=vancomycin-placebo (reference), 2=fidaxomicin-placebo 3=vancomycin-taper).

(2) Modified Horn's Index (range 1-4) as a continuous variable.

(3) Treatment * modified Horn's Index

Additionally, if the Cochrane-Armitage test for trend is not significant, fit a logistic regression model with modified Horn's Index treated as a categorical class variable will be performed following the procedure outlined above.

<u>Assessing the influence of predictors with logistic regression on Primary Outcome Measure</u>

In order to determine the effect of predictors on sustained clinical response at day 59, fit one logistic regression model with multiple independent variables. Model will include binary independent variables:

- Treatment (Three level class variable: 1=vancomycin-placebo (reference),

2=fidaxomicin-placebo 3=vancomycin-taper).

- strain (1=BI/NAP1/027 strain positive, 0=otherwise)
- treatment*strain interaction
- prior CDI (1=2 prior CDI recurrences, 0 = otherwise)
- treatment *prior CDI interaction
- concom (1=patient received concomitant antibiotics during study, 0=otherwise)
- treatment *concom

In the first step model, Wald Chi-square tests will be performed on each independent variable and if the tests of the interaction terms are non-significant at the 0.05 level those terms will be dropped from the model and the model will be refitted in the second step. In the second step, report all for individual predictors and intercept β, SE of β, Wald's χ^2, df, unadjusted *p*-value, simulation-adjusted p-values.

Analysis of Secondary Outcome Measures

No interim analysis will be conducted for secondary outcomes.

- D-COM (28 days post end of the therapy, day 90)

Dunnett's test will be performed to compare each of the FID and VAN-TP groups to VAN group at FWER=0.05 level. This is consistent with the reporting of primary outcome but without interim looks.

- CDI-COM (28 days post end of the therapy, day 59, day 90)

Response rates in CDI-COM share very similar nature as the corresponding response rates in D-COM. Therefore CDI-COM rates will be analyzed in the same way as D-COM rates.

- Proportion of symptom resolution

Using all study participants, report the *p*-value and cell frequencies, estimated proportion difference and 95% confidence interval under Dunnett's adjustment for multiplicity.

- Proportion of CDI recurrence following symptom resolution

Including only study participants who experience symptom resolution, report the *p*-value and cell frequencies, estimated proportion difference and 95% confidence interval under Dunnett's adjustment for multiplicity. Additionally, time to CDI recurrence in days will be compared among treatment groups using log-rank test. The survivor function will be estimated using Kaplan-Meier (KM) estimator.

- Proportion of diarrhea recurrence following symptom resolution

Proportion of diarrhea recurrence following symptom resolution will be analyzed in a similar way as CDI recurrence outlined above.

- Patient Reported Outcome

Repeated measures model will be used to analyze the changes in the total CDiff32-QOL score and 3 sub-scale scores (physical: P-QOL, emotional/psychological: E-QOL, social: S-QOL) from baseline (day 0) to follow up visits (days 10 and 59), including treatment groups (3-level class variable: VAN - reference group), time (3-level class variable: day 0- reference category), and interactions of treatment group and time. Interaction terms between treatment and time will be assessed to determine if there are significant differences in mean change scores between VAN-TP and VAN, and between FID and VAN. The estimates of the differences in mean changes (days 10 or 59 to day 0) between groups will be obtained by specifying appropriate contrasts for the model. Other potential baseline predictors like age, Horn's Index Score, ATLAS Score and number of prior CDI episode prior enrollment will also be tested in the repeated measures model by likelihood ratio test. Estimated difference in mean changes and 95% simultaneous CIs for comparison of VAN-TP (or FID) vs. VAN under Dunnett's adjustment for multiple comparisons will be provided at each follow-up timepoint (days 10 and 59). No further multiplicity adjustment for multiple HRQOL measures (CDiff32-QOL, P-QOL, M-QOL, S-QOL) and a HRQOL measure on multiple timepoints (days 10 and 59).

Participants who fail to meet sustained D-COM (i.e., treatment failure within first 10 days, or recurrent diarrhea/complications requiring re-treatment for CDI, or die) prior to day 59 will result in discontinuation of the treatment regimen and/or discontinuation from the study, therefore HRQOL scores are not measurable after patients' discontinuation. To account for informative censoring of longitudinal HRQOLs upon disease progression (non-ignorable dropout), a sensitivity analysis will utilize a shared parameter model [5, 6] that jointly model the longitudinal HRQOLs, and the time from randomization to fail in achieving sustained D-COM that requires re-treatment for CDI (event occurred) or dropout due to other reasons (event censored) whichever comes first. Longitudinal portion of the joint model will include treatment groups (3-level class variable: VAN - reference group), time (3-level class variable: day 0- reference category), interactions of treatment group and time, aforementioned potential baseline predictors, and a random subject intercept effect. Survival portion of the joint model will consider a semi-paramedic cox proportional hazard model. Parametric survival model (Weibull) will also be explored. Covariates included in the survival model are treatment groups, baseline risk factors of CDI recurrence, a random subject intercept effect, and a common association parameter between longitudinal and survival through the random intercept. Model comparison will use likelihood ratio test. Additional

sensitivity analysis using Pattern Mixture Model [7] that assumes Missing not at Random (MNAR) will also be explored. For example, patterns could be as simple as a group of day 59 completers vs. a group of dropouts (i.e., not measured at the final HRQOL assessment timepoint day 59). In this case, covariates in the pattern-mixture model could include treatment groups, time, interactions of treatment group and time, dropout indicator, interactions of dropout indicator and time, interactions of dropout indicator and treatment, 3-way interactions of dropout indicator and time and treatment. The 3-way interactions indicate whether any differential change across time for VAN-TP (or FID) relative to VAN varies between dropouts and completers. The estimates from each pattern are then combined using weights based on the proportion of individuals in each pattern. Aforementioned potential baseline predictors will also be tested. Model comparison will use likelihood ratio test.
Participants in the pilot phase that were not measured by CDiff32 HRQOL questionnaire will be excluded from the analysis.

Analysis of Safety Data

The p-values for safety data analyses will not be adjusted for multiple comparisons. As is typical for randomized clinical trials, the study is underpowered for events that are expected to occur at low-frequency and the p-values provided will be considered by the DMC as one component of assessing safety along with examination of actual event rates that rise to a level that the DMC would consider an elevated risk.

- SAE and AE Data

Calculate the proportion of participants experiencing an adverse event for each treatment group. Perform a Pearson's chi-square test for the difference in proportions to compare treatment groups over all adverse events by MedDRA system group. Compare by treatment groups: the number of patients experiencing an SAE, the number if treatment-related SAE, and the number of SAE by MedDRA system group. Do the same for related AEs.

- Toxicity and Kidney Function

Compare the proportion of participants who experienced an absolute increase of 0.3 or more of serum creatinine or a > 1.5 times relative increase the level recorded at baseline at any time point between treatment group using Pearson's chi-squared test of a 3 by 2 contingency table. If the *p*-value is significant at the alpha=0.05 level, perform pair-wise comparisons using Pearson's chi-squared tests (or Fisher's exact test if cell count is not sufficient), and report associated *p*-values, cell frequencies, estimated difference in proportion and associated 95% confidence intervals. The p-values will not be adjusted for multiple comparisons as the tests are exploratory and will be interpreted descriptively.

- Continuous Laboratory Values

Calculate the changes in laboratory values (CBC and serum chemistry panel to include Cr, AST, ALT, alkaline phosphatase, albumin, total bilirubin) from baseline to days 10 and 31. Report the descriptive statistics of the laboratory values at days 0, 10, and 31, as well as changes in laboratory values from baseline to days 10 and 31 for each treatment group. Compare the changes from baseline to days 10, and 31 between the three treatment groups using longitudinal analysis (covariance pattern model with unstructured covariance matrix). The dependent variable is each safety measure at days 0, 10 and 31. The covariates include 2 dummies that contrast group comparisons (VAN-TP vs. VAN and FID vs. VAN), 2 dummies that contrast time comparisons (day 10 vs. 0; day 31 vs. 0), and group by time interactions. The mean change estimate for each treatment group is derived from the model parameter estimates. The P-value for overall group comparisons in change score from day 0 to day 10 is the simultaneous testing of group (VAN-TP vs. VAN and FID vs. VAN) by time (day 10 vs. 0) interaction from likelihood ratio test with 2 degrees of freedom; the P-value for VAN-TP vs. VAN comparisons is the testing of group (VAN-TP vs. VAN) by time (day 10 vs. 0) interaction. The P-value for FID vs. VAN comparisons is the testing of group (FID vs. VAN) by time (day 10 vs. 0) interaction. P-value for overall group comparisons in change score from day 0 to day 31 is the simultaneous testing of group (VAN-TP vs. VAN and FID vs. VAN) by time (day 31 vs. 0) interaction from likelihood ratio test with 2 degrees of freedom. The P-value for VAN-TP vs. VAN comparisons is the testing of group (VAN-TP vs. VAN) by time (day 31vs. 0) interaction. The P-value for FID vs. VAN comparisons is the testing of group (FID vs. VAN) by time (day 31 vs. 0) interaction. No multiplicity adjustment will be done.

- Discontinuation of Treatment due to adverse events

Compare the proportion of patients who discontinue treatment due to lack of tolerance of the drug between treatment groups using Pearson's chi-squared test of a 3 by 2 contingency table. If the *p*-value is significant at the alpha=0.05 level, perform pair-wise comparisons using Pearson's chi-squared tests (or Fisher's exact test if cell count is not sufficient), and report associated *p*-values, cell frequencies, estimated difference in proportion and associated 95% confidence intervals. No multiplicity adjustment is considered.

- Analysis of study drug adherence data

Estimate proportion of adherent participants overall and in each group, defined as having taken more than 80% of study medication. Compare the treatment groups using a 3x2 Pearson's chi-square test of equality of proportions of adherent patients in each of the three groups: VAN-TP, VAN and FID. If the *p*-value is significant at the 0.05 level, make pairwise comparisons using Pearson's chi-squared tests, and report associated *p*-values, cell frequencies, estimated difference in proportion and associated 95% confidence intervals. No multiplicity adjustment is considered.

References

[1] Design and Monitoring of Multi-Arm Multi-Stage Clinical Trials. Ghosh P, Liu L, Senchaudhuri P, Gao P and Mehta C. Biometrics, 73: 1289-1299, 2017

[2] Adaptive Sequential Testing for Multiple Comparisons. Gao P, Liu L and Mehta C. Journal of Biopharmaceutical Statistics, 24: 1035–1058, 2014

[3] Graham, John W., Allison E. Olchowski and Tamika D. Gilreath (2007) "How many imputations are really needed? Some practical clarifications of multiple imputation theory." *Prevention Science* 8: 206–213.

[4] Proschan MA. A multiple comparison procedure for three and four-armed controlled clinical trials. Stat Med 1999; 18(7): 787-98

[5] Lawrence Gould A, Boye ME, Crowther MJ, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med. 2015; 34(14):2181-95

[6] Xu Guo and Bradley P. Carlin. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. American Statistician 2004; 58 (1):16-24

[7]. Donald Hedeker and Robert D. Gibbons.(2006). *Longitudinal Data Analysis* (pp 302-312). New York, NY: John **Wiley** & Sons.

CSP # 596 OpTION Statistical Analysis Plan (SAP) Review Sheet

Study Title: "CSP #596 Optimal Treatment for Recurrent Clostridium difficile Infection" (OpTION)

SAP Version Date: 07/30/2021
Protocol Version: 9.0

Name of the person who prepares document: Xue Li, PhD

Key changes since last reviewed SAP (dated: 03/11/2020):

1. Due to the suspension of recruitment caused by COVID-19 pandemic and its significant impact on recruitment after the site resumed recruitment activities, the sample size is reduced from 549 (91% global power) to 459 (85% global power).
2. An option has been added to allow trial termination at an interim look if any arm (VAN-TP, FID) crosses the efficacy boundary.
3. The above-mentioned changes were added to protocol version 9.0

Name of Reviewer	Reviewer Signature	Review/Approval Date
Tamara Haegerich, PhD, ***Hines Acting Center Director***		
Stuart Johnson, MD, ***Study Chair***		