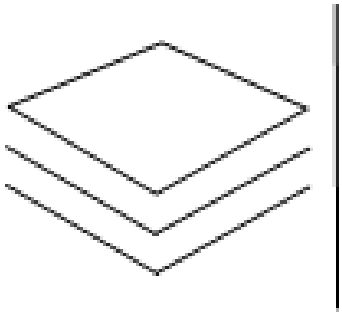


VEIN

Venous leg ulcErs: management and eradIcatioN - VEIN Platform Study

Venous Leg Ulcers: Management & Eradication:

Randomised, controlled, multi-centre platform trial assessing the clinical and cost-effectiveness of the most clinically and cost-effective treatments to aid healing and reduce recurrence in people with active and/or healed Venous leg ulcers

---

Sample Size and Optimisation Report

Version: v1.0

Imperial Clinical Trials Unit

27th March 2024

VEIN Statistics Working Group:

Nicholas Johnson – Imperial College London

Dr Rachel Phillips – Imperial College London

Dr Ed Waddingham – Imperial College London

Zohra Zenasni – Imperial College London

Dr Shaun Barber – University of Leicester

Prof. John Norrie – University of Edinburgh

Prof. Mahesh (Max) Parmar – MRC Clinical Trials Unit at University College London

Dr Babak Choodari-Oskooei – MRC Clinical Trials Unit at University College London

Contents

1. Purpose of Document

This document was created following the VEIN funding submission for NIHR HTA call 23/95 “Platform studies to efficiently evaluate the clinical effectiveness of multiple interventions in areas of strategic importance”. The text below details platform design considerations, assumptions, and the rationale behind key decisions that make up the final sample size figures for the VEIN Platform Trial. This is presented alongside coding syntax and relevant output.

2. Background

VEIN is a proposed platform trial that aims to answer the research question “What are the most clinical and cost-effective interventions to aid healing and reduce recurrence of venous leg ulcers (VLU)?“ and subsequently aims to reach a conclusion by investigating the following:

1. In acute (< 6 months old) VLU, can medication and / or antibacterial wound wash improve healing rates?

2. In chronic (≥ 6 months old) VLU, can medication and / or treatment of superficial venous reflux improve healing rates?

3. In healed VLU, can medication and / or a duplex ultrasound surveillance programme reduce recurrence rates?

The VEIN platform has been designed to comprise of three individual, four-arm, multi-arm-multi-stage (MAMS) domains running concurrently. Participants will be eligible to enrol in only one clinical domain at a time determined by disease stage as described by the Clinical Etiology Anatomy Pathophysiology (CEAP) classification system (13). This is an international staging system describing the severity of venous disease; C6 defines active (open) ulcers and C5 defines healed ulcers. Interim analyses will be performed across each domain to establish if there is evidence to stop recruitment to intervention arms based on evidence for futility or efficacy. As well as dropping intervention arms, the design allows for new interventions to be added over time.

Setting: Participants will be recruited from over 400 community / primary care, and over 30 secondary / tertiary care centres. Depending on disease stage (C5 or C6) and, in C6 disease, ulcer age (< or ≥ 6 months old), participants will be randomised to different arms.

Recruitment will continue while the interim analysis and decisions around adding / stopping arms are made. Additional interventions will be considered in a formal review process and will be added if suitable for delivery via the VEIN platform.

3. Trial Design

Initial Challenge

With the Platform call it soon became apparent that there were 3 specific cohorts of participants: those with acute ulceration, those with chronic ulceration, and those at-risk of ulceration. Each cohort of patients has their own specific needs and as such it was clear that treatment would need to be tailored to each group. As a result, the platform naturally split into three domains. The challenge now was how we run these and how they interact together within a platform setting.

The cleanest solution to stop 'contamination' across domains was to have a formal split and treat each domain individually. As such each domain would have its own control arm and would have its own comparator (though the same treatment could be tested in more than one domain).

This naturally led to the proposed creation of three individual multi-arm-multi-stage (MAMS) domains to serve each cohort.

Multi-Arm-Multi-Stage (MAMS) Design

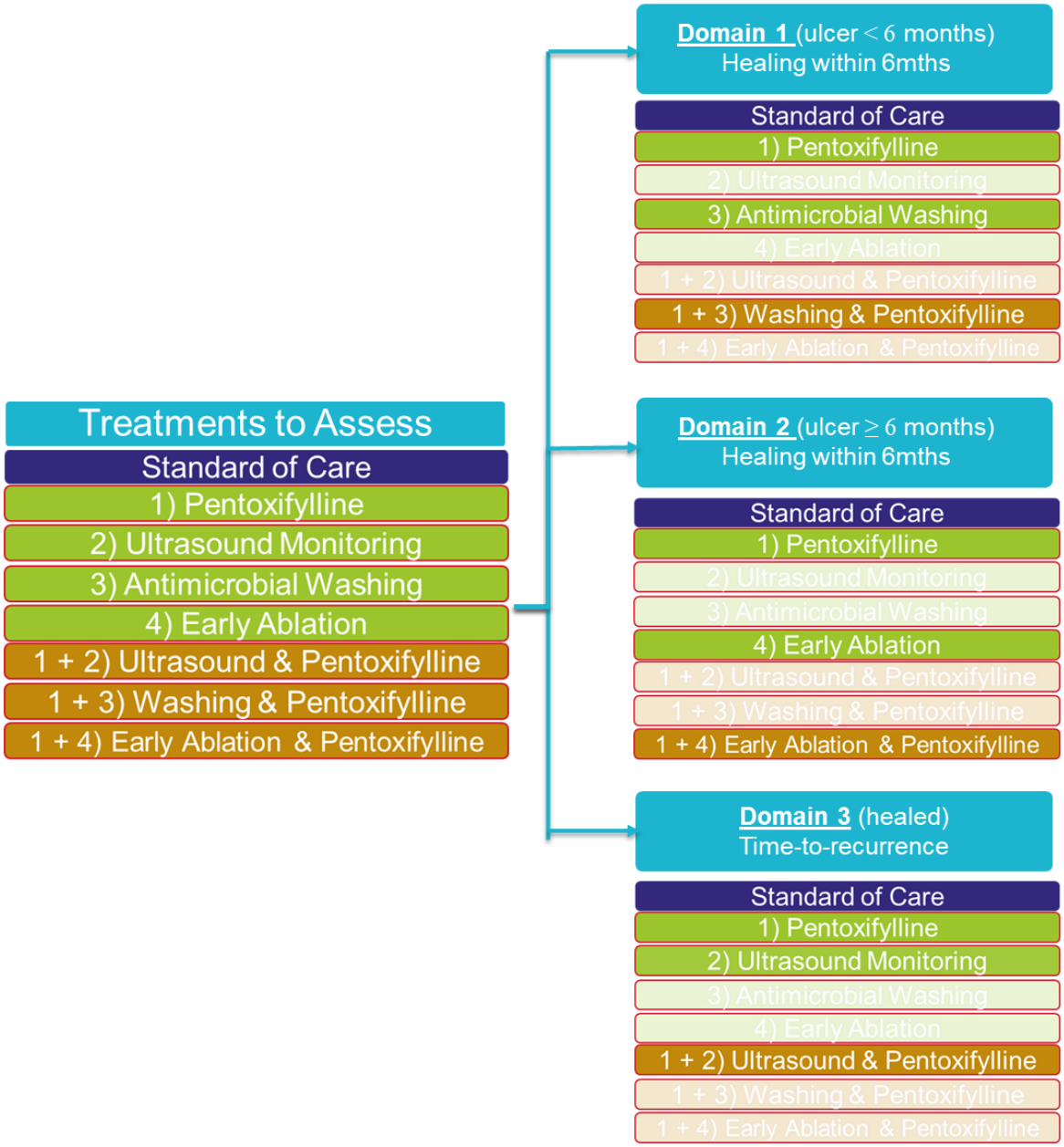
The multi-arm-multi-stage (MAMS) approach would subsequently be ideal for the Platform funding call as this would allow for interim assessments to remove arms early under overwhelming evidence for efficacy and futility (based on pre-defined boundaries).

The key question, based on the name, is a) how many arms and b) how many stages do we incorporate into each domain? To consider these we also have to ask additional questions in parallel How we consider these parameters may restrict/increase the number of arms/stages we wish to investigate. These include:

- Potential duration of grant / initial phase
- Type of endpoint for primary assessment
- Length of follow-up required.
- Extent of recruitment
    - Overall size of patient population to draw from
    - Feasible number of patients to recruit each month
    - Number of sites available

It was agreed early on during design discussions that we utilise a standard of care comparator within each domain. Subsequent therapies for initial assessment were decided on by the clinical team with Figure 1 below demonstrating how the series of investigative therapies are assigned to each domain.

Figure 1: Breakdown of treatments by domain



Statistical Assumptions

Power to be fixed at 90%.

One sided alpha to be set at 0.025 (two sided at 0.05) but with subsequent controlling for family-wise error-rate (FWER) once multiple testing is considered. This is necessary to control for Type I errors that may arise through multiple testing.

Note: We subsequently elected to fix FWER to 0.0375 as this allows us to compensate for the multiple testing of each therapy twice in a domain (i.e. pentoxifylline is assessed both individually and part of a combination therapy). The reason 0.0375 is used is that STATA programs such as NSTAGE will automatically allocate FWER control across all (three) treatment arms such that setting at an expected value of 0.025 would represent an over-correction. To think of it mathematically, using 0.0375 to control for two tests out of three would result in 0.0375/3)*2=0.0025, and as such would represent an adequate correction.

Haybittle-Peto efficacy boundaries at each interim are set at (one-sided) $p<0.0005$.

Futility boundaries are non-binding. This will allow for clinical input at assessments to establish whether a potentially futile arm may still serve purpose in continuing.

Effect size is fixed as a minimum across all 3 interventions per domain. As combination therapy is not expected to have a reductive interaction effect, and Pentoxifylline vs compression literature suggest healing effect sizes of at least 15% are feasible, domain specific therapy was used to determine minimum effect size across each Domain.

Maximum duration of initial phase is 5-years, and a minimum of 6-months is required for domain set-up.

Recruitment figures were agreed with the clinical team and represent and expected average across the recruitment period.

4. Domain 1

Introduction and PICOT

Domain 1 includes VLU patients with presence of active VLU (C6) present for < 6 months

Initial design: Four arm, four stage MAMS design

Aim: Compare Standard-of-Care (SoC) – Compression and reference to ablative therapy against 3 treatment arms:

- Prontosan Antimicrobial wound washing
- Pentoxifylline medical therapy
- Antimicrobial wound washing combined with medical therapy

Primary outcome is ulcer healing rate (yes/no) at 6-month follow-up.

Assumptions

Key assumptions underpinning this sample size calculation are:

- Anticipated event (healing) rate in SoC arm at 6-months is 70%
- Minimum treatment effect to detect +10% (absolute)
- Combination therapy is being treated as its own individual arm and not explicitly as a factorial design.
- Aim to obtain a feasible design based:
    - 90% power
    - a (two-sided) alpha of 0.05
    - 10% loss-to-follow-up rate
    - 36-month recruitment period

Design Considerations for Domain 1

- What is our maximum sample size and corresponding recruitment rate?
- How many interims is considered optimal?
- What stopping rules should be applied?

The Stata package NSTAGEBINOPT allows us to assess potential feasible designs based on how power is assigned at different stages. A 2-stage design was dismissed after investigations on 3 & 4-stage designs found the 4-stage to be preferable. A 5-stage design was also not considered after initial review based on the overall burden of assessment placed on both the study team and the DMEC. We investigated 3 and 4 stage MAMS designs using the following syntax:

3-stage design*: nstagebinopt, alpha(0.0135) power(0.90) nstage(3) arms(4) theta0(0) theta1(0.1) ctrlp(0.7) ltfu(0.10) fu(4) accrate(59 59 59) aratio(1) plot*

4-stage design*: nstagebinopt, alpha(0.0135) power(0.90) nstage(4) arms(4) theta0(0) theta1(0.1) ctrlp(0.7) ltfu(0.10) fu(4) accrate(59 59 59 59) aratio(1) plot*

3-stage NSTAGEBINOPT Output

nstagebinopt, alpha(0.0135) power(0.90) nstage(3) arms(4) theta0(0) theta1(0.1) ctrlp(0.7) ltfu(0.10) fu(4) accrate(59 59 59) aratio(1) plot

Table 1: 3-stage NSTAGEBINOPT Designs

n-stage (binary) trial design version 1.0.2, 09 June 2023

Admissible designs for a **4**-arm **3**-stage trial with binary outcome based on Choodari-Oskooei, Bratton, and Parmar (2023) Stata Journal 23(3).

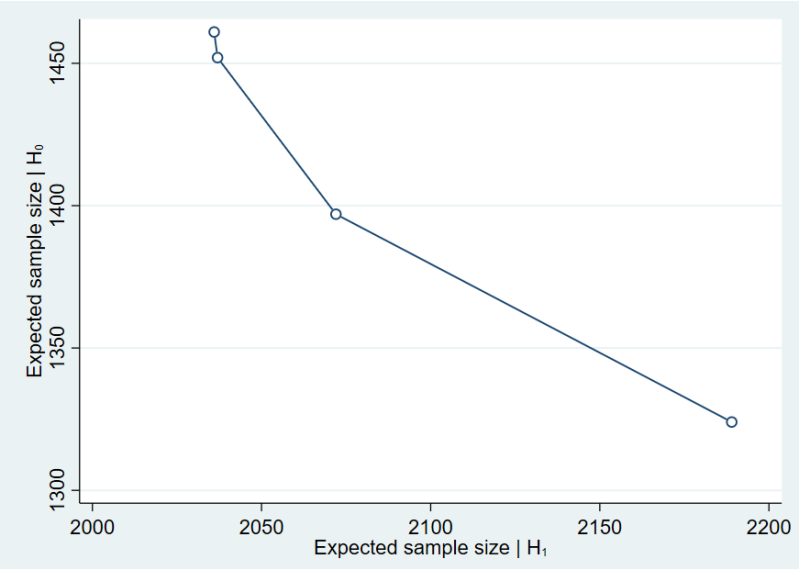
Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)
1	**[0.00,0.38]**	**1**	**0.43**	**0.96**	**1.00**	**1324**	**2189**
		2	**0.16**	**0.96**			
		3	**0.016**	**0.94**			
2	**[0.39,0.61]**	**1**	**0.36**	**0.97**	**1.00**	**1397**	**2072**
		2	**0.19**	**0.97**			
		3	**0.015**	**0.92**			
3	**[0.62,0.90]**	**1**	**0.42**	**0.98**	**1.00**	**1452**	**2037**
		2	**0.22**	**0.98**			
		3	**0.014**	**0.91**			
4	**[0.91,1.00]**	**1**	**0.44**	**0.98**	**1.00**	**1461**	**2036**
		2	**0.19**	**0.98**			
		3	**0.014**	**0.91**			

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H1) for values of q specified in q_range. H1 is the hypothesis that all of the experimental arms are effective.

Table 1 lists all (in this case) four, potential feasible designs based on the assigned parameters.

The designs are based on minimising a loss function (detailed within the table) for an arbitrary value q | q ϵ [0,1]. Stage represents the analysis number with the last stage representing a final analysis. Significance level and power represents alpha (or futility boundaries) and power at each interim stage. Finally, E(N|H0) represents the expected sample size under a scenario where all arms are futile and E(N|H1) where all arms are effective.

Figure 2: 3-stage NSTAGEBINOPT Designs comparing expected sample sizes under H_0 & H_1



4-stage NSTAGEBINOPT Output

nstagebinopt, alpha(0.0135) power(0.90) nstage(4) arms(4) theta0(0) theta1(0.1) ctrlp(0.7) ltfu(0.10) fu(4) accrate(59 59 59 59) aratio(1) plot

Table 2: 4-stage NSTAGEBINOPT Designs

n-stage (binary) trial design version 1.0.2, 09 June 2023

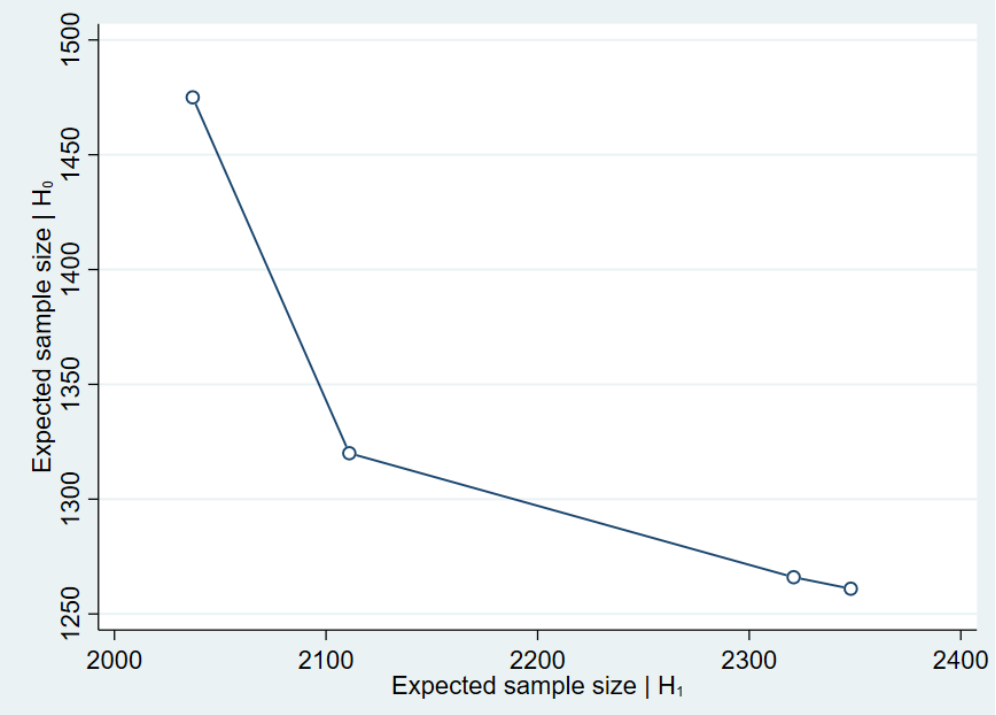
Admissible designs for a **4**-arm **4**-stage trial with binary outcome based on Choodari-Oskooei, Bratton, and Parmar (2023) Stata Journal 23(3).

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)
1	**[0.00,0.15]**	**1**	**0.50**	**0.96**	**1.00**	**1261**	**2348**
		2	**0.29**	**0.96**			
		3	**0.14**	**0.96**			
		4	**0.018**	**0.96**			
2	**[0.16,0.20]**	**1**	**0.50**	**0.96**	**1.00**	**1266**	**2321**
		2	**0.24**	**0.96**			
		3	**0.11**	**0.96**			
		4	**0.019**	**0.96**			
3	**[0.21,0.67]**	**1**	**0.46**	**0.97**	**1.00**	**1320**	**2111**
		2	**0.26**	**0.97**			
		3	**0.13**	**0.97**			
		4	**0.016**	**0.93**			
4	**[0.68,1.00]**	**1**	**0.40**	**0.98**	**1.00**	**1475**	**2037**
		2	**0.23**	**0.98**			
		3	**0.11**	**0.98**			
		4	**0.014**	**0.91**			

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H1) for values of q specified in q_range. H1 is the hypothesis that all of the experimental arms are effective.

Table 2 lists all (in this case) four, potential feasible designs based on the assigned parameters and follows the same structure as Table 1.

Figure 3: 3-stage NSTAGEBINOPT Designs comparing expected sample sizes under H_0 & H_1



Assessment of Output:

Tables 1 & 2 display potential designs for a 4 and 3 stage trial.

The choice of the final design parameters are dependent on whether we elect to prioritise a hypothetical minimum (under null hypothesis being satisfied under all arms) but at the expense of a higher maximum under the opposite scenario (alternate hypothesis satisfied). At the other end of the scale, a design exists that minimises the hypothetical maximum (alternate hypothesis satisfied) but at an expense of a higher minimum under the null scenario. In-between, alternative designs are presented within these 'minimax' scenarios which offer alternative payoffs between compensating for potential sample sizes under H_0 and H_1 being satisfied for all arms. Figures 2 & 3 demonstrate this pay-off for a 3 and 4 stage design.

A number of considerations are taken into account as to which design is preferred. This is essentially based on two design choices 1) Which design is optimal/preferred based on a 3-or 4-stage approach, 2) looking at the optimal designs, is the 3 or 4-stage option better:

- The size of the 'q-range' boundary for each design. A larger range indicates that the underlying loss-function is minimized under a wider range of values and is a safer choice
- Preference on (pairwise) power at each stage/analysis, for example a design that produces the lowest minimum under H_0 is based on equal power across all stages. Having a lower power at the 'final' analysis to ensure higher power at interim stages may be preferable.

- Trial Considerations (#1) - do we have any desire to run a minimax approach under a hypothetical scenario where H_0 and H_1 is satisfied under all arms? Would a more balanced scenario where the range between minimum and maximum is lower make more sense?
- Trial Considerations (#2) – can we rule out any designs based on the premise that a hypothetical maximum may not be feasible to recruit to within funding window based on predicted recruitment rates.

From the considerations listed, based on a) the range of q-values and b) the desire to steer away from the minimax designs due to their extreme nature (a design which assigned power equally across both interim and final assessments may have too low a power at the first interim and too high a power at final analysis), the following design was picked:

4-stage design

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)
3	[0.21,0.67]	1	0.46	0.97	1.00	1320	2111
		2	0.26	0.97			
		3	0.13	0.97			
		4	0.016	0.93			

3-stage design

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)
2	[0.39,0.61]	1	0.36	0.97	1.00	1397	2072
		2	0.19	0.97			
		3	0.015	0.92			

Whilst it was anticipated that the 3-stage design was to have lower maximum sample size, a 4-stage approach would allow for an additional look at the data at the expense of 39 patients and as such was selected on this potential net-benefit.

Adjustments to base sample size:

Once the optimal design was chosen we had to adjust the sample size to account for applying efficacy rules (and whether O'Brien-Fleming or Haybittle-Peto) and non-binding rules for futility. To apply these rules we need to be careful to ensure FWER and power is controlled and is supported by work by Blenkinsop A-et-al (1).

To account for this additional simulation work was carried out by Babak Choodari-Oskooei (MRC @ UCL) to make an appropriate inflation to the sample size for adequate control. As a manuscript is being prepared based on the simulation code utilized this will not be made available within the VEIN website. The simulation work minimised the Monte Carlo standard error for the familywise type I error rate (FWER) as is based on 10 million replications. The multiplicity-adjusted significance level for the final stage primary analysis is also slightly different from the non-adjusted initial figure of 0.016 above and is set at 0.0133.

The subsequent inflation was just over 4% and resulted in our final sample size of 2244.

5. Domain 2

Introduction and PICOT

Domain 2 includes VLU patients with presence of active VLU (C6) present for ≥ 6 months

Initial design: Four arm, four stage MAMS design

Aim: Compare Standard-of-Care (SoC) – Compression therapy

- Superficial endovenous ablation + SoC
- Pentoxifylline medical therapy + SoC
- Endovenous ablation combined with medical therapy + SoC

Primary outcome is ulcer healing rate (yes/no) at 6-month follow-up.

Assumptions

Key assumptions underpinning this sample size calculation are:

- Anticipated event (healing) rate in SoC arm at 6-months is 40%
- Minimum treatment effect to detect +15% (absolute)
- Combination therapy is being treated as its own individual arm and not explicitly as a factorial design.
- Aim to obtain a feasible design based:
    - 90% power
    - a (two-sided) alpha of 0.05
    - 10% loss-to-follow-up rate
    - 36-month recruitment period

Design Considerations for Domain 2

- What is our maximum sample size and corresponding recruitment rate?
- How many interims is considered optimal?
- What stopping rules should be applied?

The Stata package NSTAGEBINOPT allows us to assess potential feasible designs based on how power is assigned at different stages. A 2-stage design was dismissed after investigations on 3 & 4-stage designs found the 4-stage to be preferable. A 5-stage design was also not considered after initial review based on the overall burden of assessment placed on both the study team and the DMEC. We investigated 3 and 4 stage MAMS designs using the following syntax:

3-stage design*: nstagebinopt, alpha(0.0375) power(0.90) nstage(3) arms(4) theta0(0) theta1(0.15) ctrlp(0.4) ltfu(0.10) fu(4) accrate(37 37 37) aratio(1) plot fwer*

4-stage design*: nstagebinopt, alpha(0.0375) power(0.90) nstage(4) arms(4) theta0(0) theta1(0.15) ctrlp(0.4) ltfu(0.10) fu(4) accrate(37 37 37 37) aratio(1) plot fwer*

3-stage NSTAGEBINOPT Output

nstagebinopt, alpha(0.0375) power(0.90) nstage(3) arms(4) theta0(0) theta1(0.15) ctrlp(0.4) ltfu(0.10) fu(4) accrate(37 37 37) aratio(1) plot fwer

Table 3: 3-stage NSTAGEBINOPT Designs

n-stage (binary) trial design — version 1.0.2, 09 June 2023

Admissible designs for a **4**-arm **3**-stage trial with binary outcome based on Choodari-Oskooei, Bratton, and Parmar (2023) Stata Journal 23(3).

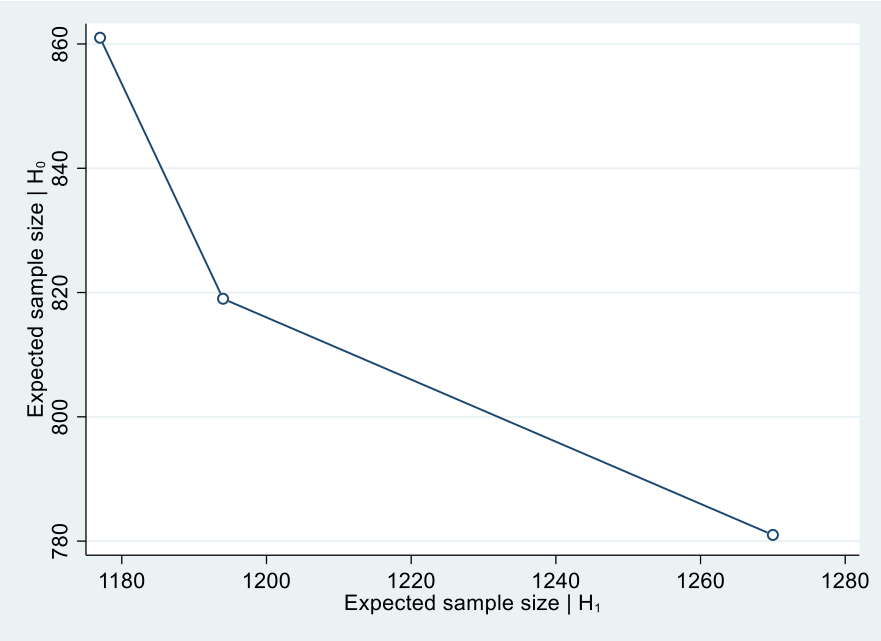
Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)	FWER (SE)
1	**[0.00,0.33]**	**1**	**0.42**	**0.96**	**1.00**	**781**	**1270**	**0.0372**
		2	**0.22**	**0.96**				**(0.0004)**
		3	**0.017**	**0.94**				
2	**[0.34,0.71]**	**1**	**0.37**	**0.97**	**1.00**	**819**	**1194**	**0.0377**
		2	**0.19**	**0.97**				**(0.0004)**
		3	**0.016**	**0.92**				
3	**[0.72,1.00]**	**1**	**0.44**	**0.98**	**1.00**	**861**	**1177**	**0.0372**
		2	**0.19**	**0.98**				**(0.0004)**
		3	**0.015**	**0.91**				

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H1) for values of q specified in q_range. H1 is the hypothesis that all of the experimental arms are effective.

Table 2 lists all (in this case) three, potential feasible designs based on the assigned parameters.

The designs are based on minimising a loss function (detailed within the table) for an arbitrary value q | q ϵ [0,1]. Stage represents the analysis number with the last stage representing a final analysis. Significance level and power represents alpha (or futility boundaries) and power at each interim stage. Finally E(N|H0) represents the expected sample size under a scenario where all arms are futile and E(N|H1) where all arms are effective.

Figure 4: 3-stage NSTAGEBINOPT Designs comparing expected sample sizes under H_0 & H_1



4-stage NSTAGEBINOPT Output

nstagebinopt, alpha(0.0375) power(0.90) nstage(4) arms(4) theta0(0) theta1(0.15) ctrlp(0.4) ltfu(0.10) fu(4) accrate(37 37 37 37) aratio(1) plot fwer

Table 4: 4-stage NSTAGEBINOPT Designs

n-stage (binary) trial design version 1.0.2, 09 June 2023

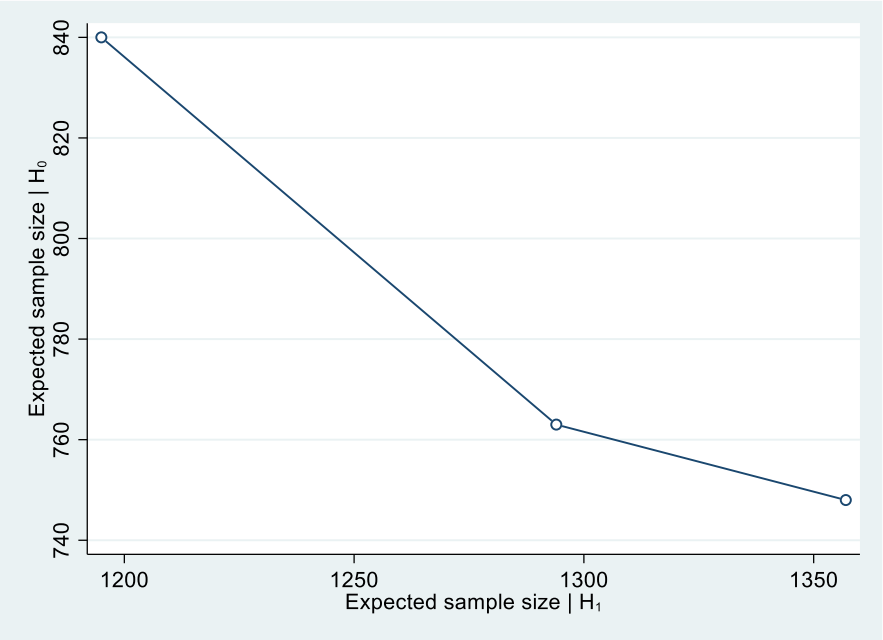
Admissible designs for a **4**-arm **4**-stage trial with binary outcome based on Choodari-Oskooei, Bratton, and Parmar (2023) Stata Journal 23(3).

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)	FWER (SE)
1	**[0.00,0.19]**	**1**	**0.50**	**0.96**	**1.00**	**748**	**1357**	**0.0376**
		2	**0.34**	**0.96**				**(0.0004)**
		3	**0.18**	**0.96**				
		4	**0.019**	**0.96**				
2	**[0.20,0.43]**	**1**	**0.40**	**0.96**	**1.00**	**763**	**1294**	**0.0380**
		2	**0.23**	**0.96**				**(0.0004)**
		3	**0.11**	**0.96**				
		4	**0.019**	**0.95**				
3	**[0.44,1.00]**	**1**	**0.33**	**0.97**	**1.00**	**840**	**1195**	**0.0371**
		2	**0.19**	**0.97**				**(0.0004)**
		3	**0.09**	**0.97**				
		4	**0.016**	**0.92**				

Note: each design minimises the loss function (1-q)E(N|H0)+qE(N|H1) for values of q specified in q_range. H1 is the hypothesis that all of the experimental arms are effective.

Table 4 lists all (in this case) four, potential feasible designs based on the assigned parameters and follows the same structure as Table 1.

Figure 5: 3-stage NSTAGEBINOPT Designs comparing expected sample sizes under H_0 & H_1



Assessment of Output:

Tables 3 & 4 display potential designs for a 4 and 3 stage trial.

The choice of the final design parameters are dependent on whether we elect to prioritise a hypothetical minimum (under null hypothesis being satisfied under all arms) but at the expense of a higher maximum under the opposite scenario (alternate hypothesis satisfied). At the other end of the scale, a design exists that minimises the hypothetical maximum (alternate hypothesis satisfied) but at an expense of a higher minimum under the null scenario. In-between, alternative designs are presented within these 'minimax' scenarios which offer alternative payoffs between compensating for potential sample sizes under H_0 and H_1 being satisfied for all arms. Figures 4 & 5 demonstrate this pay-off for a 3 and 4 stage design.

A number of considerations are taken into account as to which design is preferred. This is essentially based on two design choices 1) Which design is optimal/preferred based on a 3-or 4-stage approach, 2) looking at the optimal designs, is the 3 or 4-stage option better:

- The size of the 'q-range' boundary for each design. A larger range indicates that the underlying loss-function is minimized under a wider range of values and is a safer choice
- Preference on (pairwise) power at each stage/analysis, for example a design that produces the lowest minimum under H_0 is based on equal power across all stages. Having a lower power at the 'final' analysis to ensure higher power at interim stages may be preferable.
- Trial Considerations (#1) - do we have any desire to run a minimax approach under a hypothetical scenario where H_0 and H_1 is satisfied under all arms? Would a more balanced scenario where the range between minimum and maximum is lower make more sense?

- Trial Considerations (#2) – can we rule out any designs based on the premise that a hypothetical maximum may not be feasible to recruit to within funding window based on predicted recruitment rates.

From the considerations listed, based on a) the range of q-values and b) the desire to steer away from the minimax designs due to their extreme nature (a design which assigned power equally across both interim and final assessments may have too low a power at the first interim and too high a power at final analysis), the following design was picked:

4-stage design

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)	FWER (SE)
3	**[0.44,1.00]**	**1**	**0.33**	**0.97**	**1.00**	**840**	**1195**	**0.0371**
		2	**0.19**	**0.97**				**(0.0004)**
		3	**0.09**	**0.97**				
		4	**0.016**	**0.92**				

3-stage design

Design number	q-range	Stage	Sig. level	Power	Alloc. ratio	E(N|H0)	E(N|H1)	FWER (SE)
2	**[0.34,0.71]**	**1**	**0.37**	**0.97**	**1.00**	**819**	**1194**	**0.0377**
		2	**0.19**	**0.97**				**(0.0004)**
		3	**0.016**	**0.92**				

Whilst it was anticipated that the 3-stage design was to have lower maximum sample size, a 4-stage approach would allow for an additional look at the data at the expense of 39 patients and as such was selected on this potential net-benefit.

Adjustments to base sample size:

Once the optimal design was chosen we had to adjust the sample size to account for applying efficacy rules (and whether O'Brien-Fleming or Haybittle-Peto) and non-binding rules for futility. To apply these rules we need to be careful to ensure FWER and power is controlled and is supported by work by Blenkinsop A-et-al (1).

To account for this additional simulation work was carried out by Babak Choodari-Oskooei (MRC @ UCL) to make an appropriate inflation to the sample size for adequate control. As a manuscript is being prepared based on the simulation code utilized this will not be made available within the VEIN website. The simulation work minimised the Monte Carlo standard error for the familywise type I error rate (FWER) as is based on 10 million replications. The multiplicity-adjusted significance level for the final stage primary analysis is also slightly different from the non-adjusted initial figure of 0.016 above and is set at 0.0133.

The subsequent inflation was again just over 4% and resulted in our final sample size of 1320.

6. Domain 3

Introduction and PICOT

Domain 3 includes patients with healed VLUs.

Initial design: Four arm, four stage MAMS design

Aim: Compare Standard-of-Care (SoC) – Compression therapy

- Duplex Ultrasound Surveillance + SoC
- Pentoxifylline medical therapy + SoC
- Surveillance combined with medical therapy + SoC

Primary outcome is time to ulcer recurrence.

Assumptions

Key assumptions underpinning this sample size calculation are:

- Anticipated event (ulcer recurrence) rate in SoC arm at 12-months is 19%
- Minimum treatment effect: Hazard Ratio (HR) of 0.75
    - Equivalent to a 1-year absolute reduction of 4.6%
- Combination therapy is being treated as its own individual arm and not explicitly as a factorial design.
- Aim to obtain a feasible design based:
    - 90% power
    - a (two-sided) alpha of 0.05
    - 10% loss-to-follow-up rate
    - 36-month recruitment period
    - Minimum 12-month follow-up
    - Maximum 48-month follow-up

Design Considerations for Domain 3

- What is our maximum sample size and corresponding recruitment rate?
- How many interims is considered optimal?
- What stopping rules should be applied?

The Stata package NSTAGE allows us to assess potential feasible designs for time-to-event outcomes (based on a survival rate). Unlike NSTAGEBIN & NSTAGEBINOPT, we can incorporate efficacy stopping rules such that additional work to account for inflation is not required.

3-stage design*: nstage, accrue(87 87 87) arms(4 4 4) hr0(1 1) hr1(0.75 0.75) alpha(0.50 0.20 0.01875) fwercontrol(0.0375) omega(0.95 0.95 0.9) nonbinding t(12 12) s(0.81 0.81) aratio(1) probs tstop(36) nstage(3) tunit(4) esb(hp=0.0005)*

In terms of stage selection, 3-stages was our defined maximum. This is because incorporating 4-stages (and above) is infeasible as recruitment will have closed during the 3rd stage, making further

assessments redundant. This is driven by the low event rate resulting in a longer time to collect enough events for each stage.

Whilst feasible, a 2-stage design with just 1-interim would only give us 2 opportunities to investigate therapies across a 48-month follow-up period.

3-stage NSTAGE Output

nstage, accrue(87 87 87) arms(4 4 4) hr0(1 1) hr1(0.75 0.75) alpha(0.50 0.20 0.01875) fwercontrol(0.0375) omega(0.95 0.95 0.9) nonbinding t(12 12) s(0.81 0.81) aratio(1) probs tstop(36) nstage(3) tunit(4) esb(hp=0.0005)

n-stage trial design version 5.0.0, 04 Oct 2021

Sample size for a **4**-arm **3**-stage trial with time-to-event outcome based on Royston et al. (2011) Trials 12:81 and Blenkinsop et al. (2019) Clinical Trials 16(2)

Note: I outcome and D outcome are identical
Median survival time: **39.5** time units

Operating characteristics

Stage	Alpha (LOB)*	Alpha (ESB)*	Power	HR|H0	HR|H1	Crit.HR (LOB)	Crit.HR (ESB)	Length**	Time**
1	0.5000	0.0005	0.950	1.000	0.750	**1.000**	**0.438**	**21.037**	**21.037**
2	0.2000	0.0005	0.950	1.000	0.750	**0.907**	**0.525**	**11.649**	**32.686**
3	0.0140	.	0.900	1.000	0.750	**0.834**	.	**17.173**	**49.858**

Pairwise Error Rate **0.0145** Pairwise Power **0.9001**
Familywise Error Rate (SE) **0.0378** (**0.0002**)

LOB = lack of benefit; ESB = efficacy stopping boundary
Note: patient accrual stopped at time 36.000
 * All alphas are one-sided
 ** Length (duration of each stage) is expressed in periods and assumes survival times are exponentially distributed.
 Time is expressed in cumulative periods.

Note: the value of alpha specified at the final stage has been replaced with the value required to control the FWER at 0.0375.

Sample size and number of events

	Stage 1		
	Overall	Control	Exper.
Arms	4	1	3
Acc. rate	87	22	65
Patients*	**1831**	**458**	**1373**
Events**	**249**	**75**	**174**
	Stage 2		
	Overall	Control	Exper.
Arms	4	1	3
Acc. rate	87	22	65
Patients*	**2844**	**711**	**2133**
Events**	**572**	**170**	**402**
	Stage 3		
	Overall	Control	Exper.
Arms	4	1	3
Acc. rate	87	22	65
Patients*	**3132**	**783**	**2349**
Events**	**1120**	**328**	**792**

* Patients are cumulative across stages
** Events are cumulative across stages, but are only displayed for those arms to which patients are still being recruited
** Events are for the same outcome at all 3 stages

It should be noted that the time of 'final' analysis is at 50 months, which is greater than our proposed time of 48 months. This is because NSTAGE does not allow for loss to follow up. As a result we reduced the accrual rate from 97 to 87. For a time-to-event approach, this is an over-conservative adjustment as patients will withdraw throughout the study period and may well register their primary outcome before withdrawing.

To formally incorporate loss-to-follow-up we utilise the ARTPEP command which is detailed below:

Incorporating Attrition/Loss-to-Follow-Up and ARTPEP:

The use of the ARTPEP command in STATA allows us to incoporporate patient attrition in the form of a loss function – in other words, we can model patients withdrawing under the assumption that they will withdraw at different times within their follow-up and that some patients may register events. In the syntax below, this is represented by "ldf(0.1, 48; 0.1, 48; 0.1, 48; 0.1, 48)" which simply means that throughout the study, we anticipate 10% of patients to withdraw at a constant rate and is the same for each arm.

ARTPEP Syntax and Output

artpep, pts(0) epts(97) eperiods(48) startperiod(0) stoprecruit(36) median(39.5) method(l) ngroups(4) fp(0) edf0(0.81, 12) hratio(1, 0.75, 0.75, 0.75) alpha(0.01875) aratios(1 1 1 1) lg(1 2 3 4) ldf(0.1, 48; 0.1, 48; 0.1, 48; 0.1, 48) distant(0) detail(1) onesided(0) ni(0) tunit(4) trend(0)

month	#pats	#C-events	#events	Power
1	(no patients)			
2	97	1	1	0.01910
3	194	1	3	0.02015
4	291	2	7	0.02193
31	2910	159	531	0.58475
32	3007	169	564	0.61859
33	3104	179	598	0.65143
34	3201	189	633	0.68306
35	3298	200	668	0.71331
36	3395	210	705	0.74201
37	3395	221	741	0.76855
38	3395	231	777	0.79252
39	3395	242	812	0.81411
40	3395	252	847	0.83352
41	3395	262	881	0.85093
42	3395	271	914	0.86653
43	3395	281	947	0.88048
44	3395	290	980	0.89295
45	3395	299	1011	0.90409
46	3395	308	1043	0.91403
47	3395	317	1074	0.92290
48	3395	325	1104	0.93081
49	3395	334	1134	0.93787

The above ARTPEP output simply displays the number of patients recruited per month and the expected number of events – allowing for our 10% withdrawal rate. We see at close of recruitment that our overall sample size is at 3395 but we also see that at conclusion of our 48th month (noted as month 49 in the table) we have 1134 events for analysis. As this exceeds the value of 1120 from our NSTAGE output above we can be reassured that incorporating the loss function and using the ARTPEP value of 3395 is appropriate.

Assessment of Output:

Domain 3 had a very different challenge in comparison to the binary outcome designs of Domains 1 & 2. Using NSTAGE, the challenge was to adequately represent a 10% loss-to-follow-up and this was achieved using ARTPEP.

There are still a couple of items we need to define. Whilst we can use ARTPEP to design the overall sample size. We need to state when our two interims are expected to take place. This was achieved by taking the required events for each interim as defined within our NSTAGE output (Stage 1 requires 249 events, stage 2 requires 572) and using the corresponding ARTPEP output to assess the expected time for these benchmarks to be triggered. Subsequently 21 and 33 months were defined as our expected timings for Domain 3 interim assessments.

As our futility and efficacy bounds have already been set with subsequent considerations for inflation already considered we have our final figures in place for Domain 3 with is detailed within section 7 below.

7. Final Sample Size and Schematic (as submission)

Figure 6 – Overview Schematic of VEIN Platform Trial

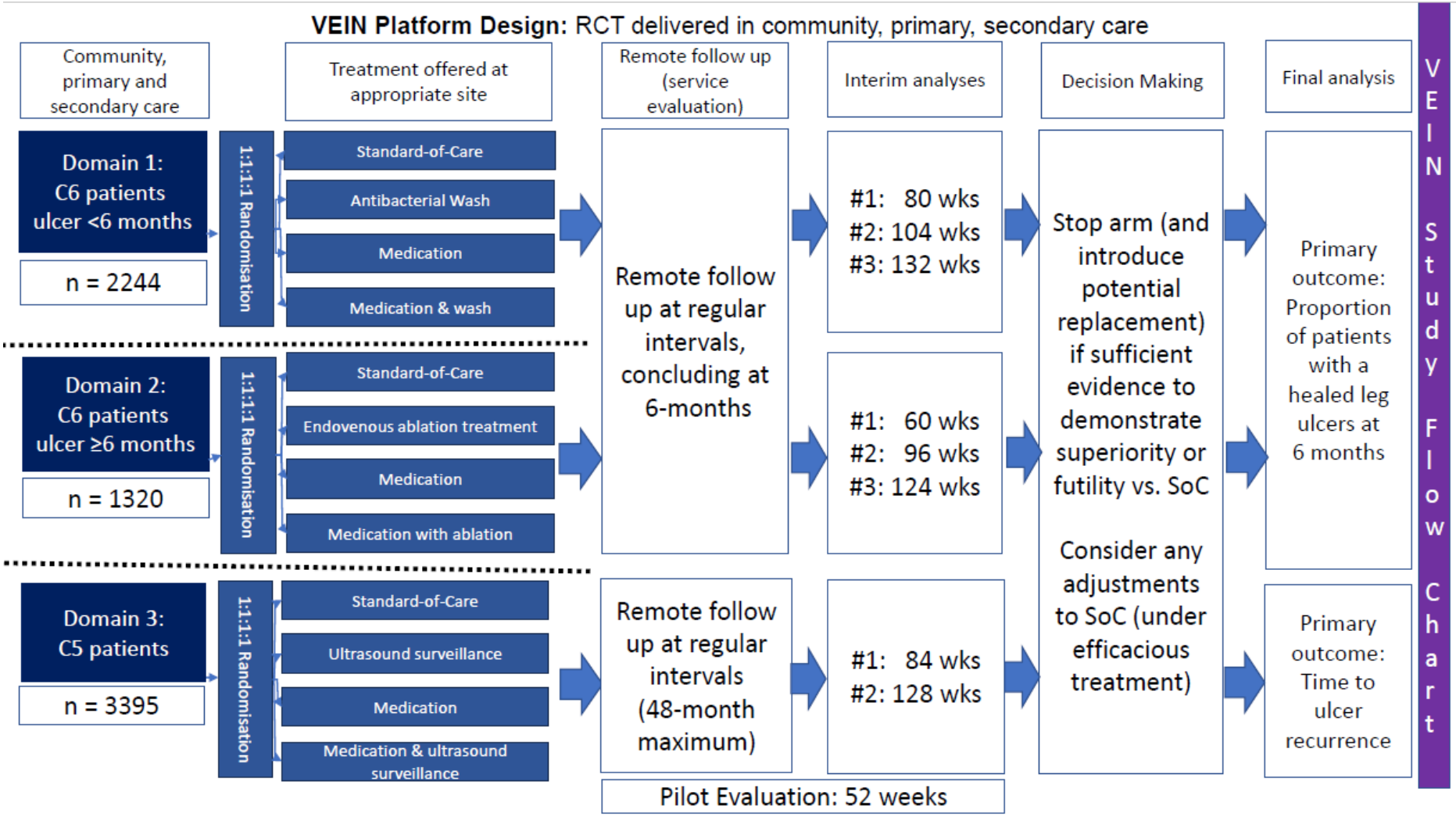


Table 5 – Final Sample Size per Domain with corresponding futility/efficacy boundaries.

	Stage (Analysis) Number			
Domain 1 (No. Healed at 6-months)	**1**	**2**	**3**	**4**
n	996	1404	1844	2244
Time (weeks)	73	103	135	182
Power (1 – β)	0.97	0.97	0.97	0.92
Futility Boundary (α)	p>0.46	p>0.26	p>0.13	p>0.0133
Efficacy Boundary	p<0.0005	p<0.0005	p<0.0005	p<0.0133
Control event rate and effect size	SoC 6-month healing rate: 70%. Effect size: +10% (absolute difference)			
Domain 2 (No. Healed at 6-months)	**1**	**2**	**3**	**4**
n	544	768	1008	1320
Time (weeks)	67	94	124	182
Power (1 – β)	0.97	0.97	0.97	0.92
Futility Boundary (α)	p>0.51	p>0.21	p>0.11	p>0.133
Efficacy Boundary	p<0.0005	p<0.0005	p<0.0005	p<0.0133
Control event rate and effect size	SoC 6-month healing rate: 40%. Effect size: +15% (absolute difference)			
Domain 3 (time to recurrence)	**1**	**2**	**3**	
n	2034	3088	3395	
Time (weeks)	91	142	182	
Power (1 – β)	0.95	0.95	0.9	
Futility Boundary (α)	p>0.5	p>0.2	p>0.0138	
Efficacy Boundary	p<0.0005	p<0.0005	p<0.0138	
Control event rate and effect size	SoC 1-year recurrence rate: 19%. Effect size: Hazard Ratio 0.75			

8. References

1. Blenkinsop A, Parmar MK, Choodari-Oskooei B. Assessing the impact of efficacy stopping rules on the error rates under the multi-arm multi-stage framework. Clin Trials. 2019 Apr;16(2):132-141. doi: 10.1177/1740774518823551. Epub 2019 Jan 16. PMID: 30648428; PMCID: PMC6442021.