

ACTIVATE Statistical Analysis Plan (SAP)

Table of Contents

ACTIVATE Statistical Analysis Plan (SAP)	1
Sec 1: Administrative Information	3
Overview	3
Sec 2: Background & aims	4
Background	4
Aims.....	6
Sec 3: Trial method	6
Sec 4: Objectives & Hypotheses	7
Primary objective	7
Secondary objectives	7
Main hypothesis.....	7
Sec 5: Estimand(s)	7
Primary estimand	7
Secondary estimands.....	8
Supplementary estimand	9
Sec 6: Sample Size Estimation	9
Sec 7: Randomisation, Concealment & Blinding	10
Randomisation & concealment.....	10
Blinding	11
Sec 8: Analysis Populations & Protocol Deviations	11
Analysis sets.....	11
Compliance definition (relevant to PP/CACE only).....	11
Protocol deviations	12
Timing rules	12
Sec 9: Outcomes	13
Overview	13
Scoring and item level missingness	13
Scale reliability	13
Sec 10: Statistical Reporting	15
Primary inference & analysis scale	15
Effect measures and standardisation	15
Model diagnostics/robustness	15
Reporting, baseline imbalance & multiplicity	15
Scale reliability	16

Sec 11: Missing data	16
Sec 12: Primary Analysis.....	17
Sec 13: Additional Analyses.....	19
Overview	19
Per protocol and CACE	19
Sec 14: SAP Reporting Checklist	21
Sec 15: Example R code snippets.....	27
Sec 16: Version history.....	34
Sec 17: Background references	37
Sec 18: SAP references	38

Sec 1: Administrative Information

Version (date) 1.0 (13/08/26)

Project title A randomised controlled evaluation of Thumos: a cognitive-behavioural intervention for medical students

Acronym ACTIVATE

SAP author Dr Luke Budworth

ISRCTN 12744098

REC reference 2024-21589-38791

Funded by The Medical Protection Society (MPS) Foundation

Chief Investigator Dr Judith Johnson

Sponsor University of Manchester

Approvals	Name	Signature	Date
Chief Investigator	Dr Judith Johnson		
Trial Statistician	Dr Luke Budworth		13/08/26
TSC Independent Statistician	Prof Mona Kanaan		

Overview

This Statistical Analysis Plan sets out how the primary results paper from the ACTIVATE trial will be analysed and reported. Any related papers, including further exploratory work, are not bound by it but will follow the same principles.

We have tried to pre-specify as much as we can, including a justified set of sensitivity analyses to test the robustness of the primary result. Not everything can be anticipated, so the plan does not rule out exploratory analysis or sensible judgement in analysis and reporting. Where it does specify something, we will follow it as closely as we can.

The SAP follows Sheffield CTRU SOP ST001, "The Statistical Analysis Plan" (02/07/2025), and the Gamble et al. (2017) guidelines on SAP content. A completed reporting checklist is in Section 14.

Sec 2: Background & aims

Background

Healthcare systems internationally are experiencing a workforce crisis, with a current global shortage of six million physicians [1]. While poor workforce planning has contributed to this problem, high turnover has been identified as the main contributing factor [2]. Inadequate staffing is linked with negative patient outcomes including extended hospital stays and mortality [3]. It also creates a vicious cycle, whereby existing staff become overworked and burnt-out, a syndrome characterised by exhaustion and disengagement [4]. Higher burnout then increases the risk of sickness absence and turnover, further exacerbating staffing shortages [5, 6] and directly leading to the delivery of poorer patient care [7].

In a review of 170 studies in doctors, higher burnout was associated with twice the risk of patient safety incidents and a two-fold reduction in patient satisfaction [7]. Within this vicious cycle, provider burnout and poor wellbeing must be recognised as an active component, serving to drive the workforce crisis. Equally, addressing this component by better supporting healthcare professionals could help to break this cycle, leading to reduced sickness absence, better retention and better patient care [8].

Most burnout research has focused on qualified professional groups, but there is a growing recognition that it could be beneficial to intervene during training, prior to professional registration, due to high rates of mental health distress in these cohorts [9]. Studies in medical students indicate half report burnout, one in three report depression and one in ten experience suicidal thoughts [10-12]. Furthermore, the immediate post-qualification years have been recognised as a high-risk time for physicians leaving the profession [13]. Poor mental health and burnout are linked with higher attrition [14], so improving their mental wellbeing could be one way to enhance retention and reduce physician shortages.

Several causal factors for poor mental health and burnout have been identified which are specific to this group, including a high academic workload, a tendency towards maladaptive perfectionism [15, 16] and exposure to stressful clinical events on placement [17]. Indeed, rates of burnout in medical students increase significantly during the years involving clinical placements [18]. However, there has been a lack of interventions tailored for medical students and none which prepare medical students for stressful workplace events [19]. The mental health interventions which have been tested in this group have been largely generic and 'off-the-shelf', including mindfulness, relaxation training and yoga [19]. Evidence for these interventions is weak and mixed, and they have been criticised for lacking relevance [20]. There is an urgent need for effective, preventative interventions which help students cope

with the specific occupational stressors they will encounter during their placements, training, and beyond.

The current trial will evaluate Thumos, an intervention that employs cognitive-behavioural therapy (CBT) techniques to prepare healthcare professionals and students to manage stressful work or training events. Thumos has been developed and refined over time in response to pilot testing and evaluations. Thumos differs from previously evaluated psychological interventions for medical students in three ways. Firstly, it has been designed with healthcare professionals in mind, based upon a combination of 1) research into the effects of patient safety incidents and other work stressors, 2) resilience theory, and 3) CBT techniques. Secondly, Thumos includes both one-to-one and group modalities, synergising the benefits of peer engagement through group workshops with professional, confidential support in a subsequent phone call. Thirdly, Thumos addresses the stressful situations that are intrinsic to healthcare work, providing support for those unavoidable events that cannot be improved through organisational interventions.

Characteristics of Thumos may make it particularly well suited for medical students. For example, it is a preventative rather than ameliorative intervention which could help to reduce stigma in participation; a known barrier for medical student uptake of psychological interventions [11, 21, 22]. It also uses case studies and materials which are tailored towards the stressful experiences medical students may face on placements. This approach increases the sense of relevance of the intervention for participants which may increase engagement [9]. Furthermore, it aims to reduce maladaptive perfectionism, a factor linked with mental health distress in medical students [16].

There have been three uncontrolled evaluations of previous versions of Thumos which indicate positive findings. The first was in a group of 66 multi-disciplinary healthcare professionals and students; results indicated it was associated with improvements in participants' confidence, coping knowledge and psychological resilience [23, 24]. The second was in 80 Critical Care Nurses (CCNs) [25, 26]. Participation was associated with increased confidence and resilience and reduced depression and burnout. Nurses also reported a lower intention to leave clinical practice following the intervention [25, 26]. The third evaluation was in 115 medical students [9]. Findings indicated the intervention was feasible and potentially effective in this group, with participants reporting significant increases in resilience and confidence in coping with adverse events, and significant decreases in burnout and depression [9]. Qualitative findings from these evaluations indicate that participants appreciate the intervention's tailored format and opportunity for peer support alongside professional input. They describe increased self-awareness and reported the acquisition of psychological skills which they could apply to cope with work stressors [9, 23-26].

Feedback from these evaluations was used to make improvements to further tailor the intervention towards medical students, including the addition of new exercises, reduction of didactic information and changes to the phone call. There is now a need to deliver the finalised version of the intervention, Thumos, and to ascertain whether observed improvements in outcome indicators can be attributed to the intervention. This study addresses this gap, through a definitive randomised controlled trial (RCT). We also extend the final follow-up time point in previous studies of 4 months to 9 months and create the possibility of conducting a subsequent 5-year follow-up study, should initial results be positive.

Aims

The main purpose of the trial is to establish whether Thumos increases resilience and improves mental health in medical students. The primary objective is to determine whether Thumos is superior to the control condition (services as usual; SAU) at 4 month follow-up in increasing psychological resilience in medical students. Secondary objectives are:

- To determine whether Thumos is superior to SAU at 9-month follow-up in increasing psychological resilience.
- To determine whether Thumos is superior to SAU at 4-month and 9-month follow-ups in the endpoints of confidence, burnout, depression, mental health and self-reported sickness absence.
- To assess acceptability and how Thumos is delivered in practice through a process evaluation with users, medical educators and providers.

Sec 3: Trial method

Table 1. Trial method summary

Element	Specification
Design	Two-arm, parallel group superiority RCT with individual randomisation 1:1 to Thumos or services as usual (SAU)
Assessments	T1 baseline; T2 post-intervention, or 6 weeks post-baseline for controls; T3 4 months (primary endpoint); T4 9 months. All questionnaires completed online via Qualtrics
Eligibility	UK medical students in a year containing a clinical placement of one week or longer, recruited nationally (typically Y4 or Y5)
Sample size	220 randomised, 110 per arm
Primary outcome	Self-reported Brief Resilience Scale (BRS) score at T3
Secondary outcomes (quantitative)	BRS at T4. At T3 and T4: Confidence in Coping with Adverse Events (CCAEC), Oldenburg Burnout Inventory (OLBI), Patient Health Questionnaire 9 (PHQ-9), Recovering Quality of Life Utility Index (ReQoL-UI), self-reported sickness absence

	(novel measure). QALYs derived from ReQoL-UI utilities, baseline to T4
Primary analysis	Mixed effects model for repeated measures with covariate adjustment, estimating the adjusted mean difference in BRS at T3
Interim analyses	None planned
Multiplicity	Primary endpoint tested two-sided at $\alpha = .05$. Secondary endpoints reported with 95% CIs and unadjusted p-values, with false discovery rate adjusted p-values alongside for transparency
Statistical software	R version 4.5.1, or the latest available at the time of analysis. All code version controlled

Sec 4: Objectives & Hypotheses

Primary objective

Determine whether Thumos increases resilience at 4 months versus SAU.

Secondary objectives

- Examine effects at T3 and T4 on confidence, burnout, depression, mental health utility and QALYs, and sickness absence
- Assess acceptability and delivery via mixed methods process evaluation (not outlined in this SAP)

Main hypothesis

We expect mean BRS at T3 to be higher in the Thumos group than in the SAU group. The trial was powered to detect a conventionally medium sized standardised mean difference (Cohen's $d = 0.5$). Assuming a baseline SD of around 0.7, this is equivalent to a difference of about 0.35 BRS points.

Sec 5: Estimand(s)

Estimands are specified following the ICH E9(R1) framework (Kahan et al. 2024) [1]. Analysis sets, the compliance definition, protocol deviations and assessment windows referred to below are defined in Section 8.

Primary estimand

Table 2. Primary estimand

Attribute	Specification
Estimand	'Policy' effect of offering Thumos on self-reported resilience at 4 months

Treatment	Assignment to Thumos versus assignment to services as usual
Population	UK medical students in a year containing a clinical placement of one week or longer, recruited nationally (typically Y4 or Y5)
Variable	Adjusted between-arm difference in mean BRS at T3 (approximately 4 months post-baseline), estimated by mixed models accounting for correlation between time points
Summary measure	Covariate adjusted unstandardised and standardised (d ; spread = pooled baseline SD) mean difference in BRS at T3, Thumos minus control, with 95% CIs and p -values
Intercurrent events and strategies	Non-adherence to Thumos components. Treatment policy: analyse as allocated.
	Use of external psychological support. Treatment policy: recorded at each time point and described
	Missing T3 outcome. Not an intercurrent event as such but an estimation assumption: MAR given observed data, with planned sensitivity analyses relaxing MAR.

Secondary estimands

Treatment, population and intercurrent event strategies are as for the primary estimand. Variables and summary measures differ by outcome. Secondary results are supportive rather than definitive (Section 10).

Table 3. Secondary estimands

Outcome	Time	Variable definition	Effect measure
BRS	T4	Mean of six items, range 1 to 5	Adjusted mean difference and Cohen's d , 95% CI
CCAE	T3, T4	Mean of three items	Adjusted mean difference and Cohen's d , 95% CI
OLBI	T3, T4	Subscale means (exhaustion, disengagement) and total mean	Adjusted mean difference and Cohen's d , 95% CI
PHQ-9	T3, T4	Sum of nine items, range 0 to 27	Adjusted mean difference and Cohen's d , 95% CI
ReQoL-UI	T3, T4	Utility index via the ReQoL-UI algorithm	Adjusted mean difference and Cohen's d , 95% CI
ReQoL-UI QALYs	Baseline to T4	QALYs, area under the utility time curve with linear interpolation	Adjusted mean difference, 95% CI

Sickness absence	T3, T4	Days lost in the past 4 weeks (two items, analysed separately)	Adjusted incidence rate ratio, 95% CI
BRS: Brief Resilience Scale; CCAE: Confidence in Coping with Adverse Events; OLB: Oldenburg Burnout Inventory; PHQ-9: Patient Health Questionnaire 9 item depression scale; ReQoL-UI: Recovering Quality of Life Utility Index; QALYs: quality adjusted life years; T1 to T4: time points 1 to 4 (T = time).			

Supplementary estimand

Table 4. Supplementary estimand

Attribute	Specification
Estimand	Effect of receiving Thumos on resilience at 4 months among students who would take it up if offered
Treatment	Assignment to Thumos versus assignment to services as usual
Population	Students who would take up Thumos if it were offered to them. Take up means completing both group workshops, the minimum adherence definition in protocol section 9.5; the one-to-one call is not required. For analysis this counts as attending both workshops within 6 weeks of randomisation with at least 75% of each session attended (Section 8). Take up is a characteristic of the student rather than of the arm they were allocated to, so it is seen directly in the Thumos arm and estimated statistically for the control arm, where nobody was offered the workshops.
Variable	Mean BRS at T3, defined as for the primary estimand, among students who would take up Thumos
Summary measure	Complier average causal effect: mean difference with 95% CI, standardised by the pooled baseline SD, reported alongside the primary effect
Intercurrent events and strategies	Non-adherence. Handled by narrowing who the estimate applies to (students who would take up Thumos), not by excluding or adjusting for anyone. All randomised participants remain in the analysis
	External psychological support and missing T3 outcome. As for the primary estimand.

Sec 6: Sample Size Estimation

We aim to recruit and randomise 220 participants, with 110 in each arm. Simulations using the R package *powerImm* showed that 200 participants would provide approximately 80% (or above) power to detect the target treatment effect at T3 under various combinations of assumptions (see Table 5). This target was further

increased by 10% (i.e. to 220) as a conservative measure to allow for uncertainty in our simulations' assumptions.

The target effect was informed by our earlier uncontrolled study; mean BRS increased from 2.72 at baseline to 3.17 (SD = 0.72) at four months, with a paired pre-post $d = 0.89$ (Johnson et al. (2023) [2]). Because this trial had no control group, and because a paired pre-post effect is not directly comparable with a between-group effect, we assumed a more conservative target: $d = 0.50$.

Table 5. Power simulation tested assumptions

Assumption	Value(s)
Primary comparison	Adjusted difference in mean BRS between Thumos and services as usual at T3
Target effect	Standardised between-group difference of $d = 0.50$. With a baseline SD of 0.70-0.90, $d = 0.50$ corresponds to approximately 0.35-0.45 BRS points, or 9-11% of the four point span of the 1-5 scale. As a rough convention, d values of 0.2, 0.5 and 0.8 are considered as small, medium and large (Cohen (1992) [3])
Power and significance level	80% power (beta = 20%) at a two-sided alpha of 0.05.
Repeated measurements/within-participant correlation	Up to three post-baseline BRS measurements per participant assumed. A range of participant ICC values were explored (0.30, 0.40, 0.50). The original BRS study reported test-retest correlations of 0.62-0.69, although these are not identical to model-based ICCs (Smith et al. (2008) [4])
Variation in change over time	A range of variance-ratio values were examined (.01, .03, .05) (in <i>powerlmm</i> , this represents random-slope variance relative to within-participant residual variance)
Attrition	Monotone attrition by T4 was conservatively assumed to be 40% in the control and 30% in the intervention arm. These were deliberately cautious based mainly on pilot data (Johnson et al. (2023) [2]) and relevant published evidence (de Miquel et al. (2024) [5])

Sec 7: Randomisation, Concealment & Blinding

Randomisation & concealment

Simple 1:1 individual randomisation via Sealed Envelope, a platform that generates and holds the allocation sequence. Allocation is revealed only after consent and baseline completion within Sealed Envelope; the sequence is not accessible to the trial team.

Blinding

Because the intervention is cognitive-behavioural and outcomes are self-reported, participants and the intervention team cannot be blinded.

The primary analysis will be conducted with the treatment groups labelled anonymously as A and B. The trial manager will record which arm is which, prepare the analysis dataset and remove or mask variables that could reveal allocation, such as intervention attendance and therapist details.

The allocation key will not be released until the primary analysis dataset has been locked and the masked primary results have been checked for completeness and accuracy.

Analyses requiring information about intervention receipt (e.g. per-protocol and CACE analyses) will be conducted after the allocation key is released. Any accidental unblinding will be documented and reported transparently.

Sec 8: Analysis Populations & Protocol Deviations

Analysis sets

- **Full analysis set (FAS aka. ITT: intention to treat, primary analysis)**

All randomised participants, analysed as allocated, regardless of intervention receipt or post-randomisation events. Participants with no post-baseline outcomes appear in flow and baseline tables but will not contribute to outcome models. This set underpins the treatment policy estimands (as outlined above).

- **Per protocol set (PP, sensitivity analysis)**

FAS excluding major protocol deviations and excluding Thumos participants who do not meet the compliance definition (see below). Requires a valid T3 BRS within the T3 window.

- **CACE (supplementary estimand)**

Binary indicator of 'full receipt' (aka take up) used in the instrumental variables (IV) analysis (see Section 13). All randomised participants will remain in the IV analysis.

Compliance definition (relevant to PP/CACE only)

For descriptive reporting only, we may present a stricter completion metric that includes the one-to-one call, but PP and CACE will use the minimum adherence definition below to avoid ambiguity:

- **Thumos arm**

Attended both group workshops within 6 weeks of randomisation, with at least 75% attendance of each session. The one-to-one call is not required for minimum adherence.

- **Control arm**

No receipt of any Thumos component. Accessing usual external psychological support is allowed in both arms and is not a deviation (as this is not specified within the trial protocol).

Protocol deviations

All deviations will be logged and classified.

- **Minor**

Unlikely to affect the primary outcome or its interpretation, for example small lateness to sessions, questionnaire completed within the window but not on the target day. Minor deviations remain in all analysis sets.

- **Major**

May affect the primary outcome. These include:

- Invalid consent or ineligibility
- Allocation error/breach of concealment
- Any receipt of Thumos intervention in the SAU/control arm
- No valid T3 BRS within the prespecified T3 window
- Data integrity issues such as very unusual or impossible values recorded, or identified falsification

Major deviations remain in the FAS but will be excluded from PP. Non-completion of Thumos is not a deviation for the FAS, but it excludes participants from the PP sample and defines non-receipt for the CACE indicator.

Timing rules

Assessment windows are: T2 = 4–10 weeks, T3 = 12–24 weeks, and T4 = 30–44 weeks post-baseline. Where multiple responses fall within a window, the response closest to the nominal time point will be used.

Any assessments outside these windows will be documented, with the number and timing reported in the main paper. If the proportion of out-of-window assessments is substantial, a sensitivity analysis will assess the impact on results.

Sec 9: Outcomes

Overview

All measures are administered online via Qualtrics at T1 (baseline), T2, T3 and T4. T3 and T4 are the pre-specified endpoints for efficacy analyses; T1 provides the baseline covariate and T2 contributes to the repeated measures models in Sections 12 and 13.

Table 6 sets out how each outcome is scored, including what happens when individual items are missing from within a measure. The rules and thresholds in the final column are justified below.

Scoring and item level missingness

Most key items are set as required in Qualtrics, so missing items within an otherwise completed measure should be rare and mainly where a participant exits midway through. Measures completed up to that point are scored normally; the interrupted measure follows the rules in Table 6.

Where published guidance on within scale missingness exists we follow it (NHS Digital for the PHQ-9; ReQoL-UI scoring documentation; Keetharuth et al. (2018) [6]). Otherwise we use person mean substitution, standard practice for patient reported outcome scales, allowing at most one missing item per scale or subscale (effective thresholds of 10-17%, supported by simulation work; Huisman (2000) [7]; Parent (2013) [8]; Jacoby et al. (2022) [9]). Beyond the threshold, scores are set to missing at that time point and handled at the analysis stage by the mixed models and multiple imputation (Section 11).

Scale reliability

Multi-item scales are useful only if their items measure the intended construct reliability in the population being studied. Several measures here have limited UK validation data in medical students. Poor reliability reduces treatment effects and complicates interpretation, so internal consistency will be reported at each time point to document how the scales perform in this sample and to support the validity of our efficacy conclusions. The estimates and thresholds are specified in Section 10.

Table 6. Outcome scoring and handling of missing items

Outcome	Items	Construct/direction	Endpoints	Derived variable	If items are missing
BRS (primary)	6	Resilience; higher is better	T3 [12 to 24 wks]	Mean of six items after required reverse scoring (range 1 to 5)	Score if at least 5 items present, imputing the single missing item by person mean; otherwise missing (max = 16.7%)
BRS	6	Resilience; higher is better	T4 [30 to 44 wks]	As above	As above
CCAE	3	Confidence in coping with adverse events; higher is better	T3, T4	Mean of three items	All three items required; no imputation (max 0%)
OLBI (total and subscales)	16, as two 8 item subscales	Burnout: exhaustion and disengagement; higher is worse	T3, T4	Subscale means and total mean, after required reverse scoring	Subscale scored if at least 7 of 8 items present, imputing the single missing item by subscale mean; if either subscale cannot be scored, the total is missing (max = 12.5%)
PHQ-9	9	Depressive symptoms; higher is worse	T3, T4	Sum of nine items (range 0 to 27)	Score if at least 8 items present, imputing the single missing item by person mean and rounding the total to the nearest integer; otherwise missing (max = 11.1%)
ReQoL-UI (utility)	10	Quality of life utility; higher is better	T3, T4	Utility index via the ReQoL-UI algorithm	Score if at least 9 items present, imputing by person mean; otherwise missing (max = 10%)
ReQoL-UI QALYs	10	Quality adjusted life years; higher is better	Baseline to T4	Area under the utility time curve (Section 13)	Derived from the utility scores above; a missing utility at any time point is handled at the analysis stage (Sections 11 and 13)
Sickness absence	2	Days lost in past 4 weeks; higher is worse	T3, T4	Two counts, analysed separately: total days lost, and working or studying days lost	Standalone items; no item level imputation

BRS: Brief Resilience Scale; CCAE: Confidence in Coping with Adverse Events questionnaire; OLBI: Oldenburg Burnout Inventory; PHQ-9: Patient Health Questionnaire 9 item depression scale; ReQoL-UI: Recovering Quality of Life Utility Index; QALYs: quality adjusted life years. Person mean substitution means replacing a missing item with the mean of that participant's completed items on the same scale or subscale at the same time point. Support access is collected at all four time points but is only summarised descriptively by arm and time point as context for the treatment policy strategy in Section 5, and is not used as a covariate in any model.

Sec 10: Statistical Reporting

See Section 15 for example R code snippets.

Primary inference & analysis scale

For the primary analysis, the BRS at T3 will be analysed with a two-sided significance level of $\alpha = 0.05$. Although our hypothesis is that Thumos is beneficial, effects could plausibly occur in either direction. All continuous outcomes will be analysed on their original scales to retain clinical interpretability (e.g. points on the BRS).

Effect measures and standardisation

For continuous outcomes, Cohen's d is reported alongside the adjusted mean difference to give a scale-free measure comparable across outcomes and to link back to the target effect size calculation (Section 6). The standardiser is the pooled baseline (T1) SD of the relevant outcome. The ratio of the observed T3 SD to the baseline SD will be reported once, allowing readers to convert if they want as the choice can meaningfully alter d (e.g. a 0.35 point BRS difference is $d = 0.50$ against a baseline SD of 0.70 but $d = 0.58$ where SD = 0.60).

Count outcomes (the two sickness absence items) are reported as adjusted incidence rate ratios. These are already scale-free, so no standardised effect size is calculated for them. A rate ratio on its own can mislead when the underlying rate is low, so each is reported with the adjusted marginal mean days lost in both arms. The absolute difference in days lost, with its 95% CI, will be given alongside the ratio.

QALYs are reported as an adjusted mean difference only (Section 13).

Adverse and serious adverse events are reported as risk differences with 95% CIs, not as ratios (Section 13).

Model diagnostics/robustness

Model assumptions will be examined using residual and influence diagnostics. If there is clear evidence of severe non-normality or heteroskedasticity, we will repeat the analysis with *clubSandwich* (R package) robust standard errors as a sensitivity (clustered by participant for repeated measures). In all cases, the target estimand remains the adjusted mean difference on the original scale.

Reporting, baseline imbalance & multiplicity

Every endpoint will be reported as a point estimate with a 95% confidence interval and an exact p -value to three decimal places (for example $p = .054$), with values below .001 reported as $p < .001$. The number and percentage of participants

contributing to each model will be given by arm and time point alongside every result.

Baseline characteristics will be described, not tested. Because allocation was random, any differences between the arms at baseline arose by chance, and testing whether chance occurred is not a useful question. We will instead report standardised differences, a measure of how far apart the arms are relative to the spread of the variable, and flag anything above 0.20 (just for the sake of interpretation).

We will not apply a formal correction for multiple testing when drawing our main conclusions. The primary endpoint is a single test at $\alpha = 0.05$. Secondary and exploratory findings will be treated as supportive rather than definitive whatever their p -values show. For transparency about the risk of chance positives, we will additionally report false discovery rate adjusted p -values for the secondary outcomes, calculated by the Benjamini-Hochberg procedure within two families: all secondary outcomes at T3 and all at T4.

Scale reliability

Following the rationale set out in Section 9, we will report internal consistency for each multi-item scale at every time point. Cronbach's α and McDonald's ω will be estimated with 95% bootstrap CIs (1000 resamples), pooled across arms per time point and by arm at baseline. Thresholds of 0.70 will be considered acceptable, 0.80 good and 0.90 excellent. For the three-item CCAE we will also report the mean inter-item correlation, with 0.20 to 0.50 indicating adequate coherence, since α is unreliable on three items and will not be interpreted alone.

Sec 11: Missing data

See Section 15 for example R code snippets.

Primary analyses use all available data. Each participant can contribute up to three post-baseline BRS observations (T2-T4). Scoring follows Section 9; where a scale cannot be scored, it is set to missing for that outcome and time point.

The primary analysis assumes outcome data are missing at random (MAR) given observed outcomes and model covariates. We will describe the pattern and predictors of missingness to judge how plausible that is: response rates by arm and time; stated reasons for non-completion where available; and comparisons of baseline characteristics and earlier outcomes (for example BRS at T1 and T2) between participants who do and do not provide BRS at T3 and T4, including logistic regression models with missingness as the outcome.

We will also conduct three types of sensitivity analysis for missing outcome data (Table 7).

Table 7. Sensitivity analyses for missing outcome data

Analysis	Specification
Multiple imputation under MAR	Imputation models include outcomes at all time points, randomised group, and key baseline covariates (baseline BRS, age, gender, medical school, and baseline values of secondary outcomes where relevant). Carried out separately by arm, reflecting each outcome's type. The analysis mirrors the primary model and pools estimates by Rubin's rules
Pattern mixture delta adjustment	Fixed negative offsets of 0.10, 0.20 and 0.30 BRS points applied to imputed values for (intervention) participants missing at T3 or T4. Given an expected BRS SD of 0.7 to 0.9, these correspond to roughly 0.15 to 0.45 SDs, keeping the range within clinically plausible bounds
Last post-baseline carried forward	The most recent observed post-baseline value replaces a missing endpoint: T3 carried to T4, T2 carried to T3 or T4. Where no post-baseline value exists, baseline is carried forward

Results will be considered robust if the direction and magnitude of the primary T3 BRS effect are broadly consistent across the primary analysis and these sensitivity analyses. We will then consider which assumptions are most plausible given the observed missingness patterns and predictors of dropout, and discuss the range of plausible treatment effects rather than relying on a single estimate.

Sec 12: Primary Analysis

See Section 15 for example R code snippets.

- **Model & outcome**

The primary outcome is BRS at T3, scored as the mean of the six BRS items (range 1–5) at the T3 assessment window (12–24 weeks). Each participant can contribute up to three post-baseline BRS measurements at T2, T3 and T4.

The primary analysis will use a linear mixed model for repeated measures (MMRM) with BRS at T2, T3 and T4 as the repeated outcome and baseline BRS (T1) included as a covariate. The primary contrast is the adjusted between-group difference in mean BRS at T3 (Thumos minus services as usual) estimated from this model.

T4 occurs after T3, but including BRS at T2 and T4 in the model does not change the definition of the primary outcome. The MMRM jointly models the mean trajectory over time and the within-participant correlation structure. It allows all available post-

baseline BRS data to contribute to estimation and gives an efficient and unbiased estimate of the T3 treatment contrast under the missing-at-random assumptions described in Section 11.

- **Specification**

Fixed effects will include:

- treatment group (Thumos vs services as usual)
- categorical time (T2, T3, T4)
- treatment-by-time interaction
- baseline BRS (T1)
- baseline covariates: age, gender, and medical school

Baseline BRS is expected to be strongly predictive of BRS at T3 and later, so adjusting for it will increase the precision (power) of the treatment effect estimate. Age, gender and medical school are pre-randomisation characteristics that may also be associated with BRS and with missingness. Including them also improves precision and supports a more plausible missing-at-random mechanism by conditioning on key baseline predictors. Randomisation already protects against confounding, so not adjusting for these covariates would not bias the treatment effect, but it would generally mean less precise estimates.

The model will include a random intercept for participant to account for clustering of repeated BRS scores within students. The within-subject covariance for the residuals at T2–T4 will be specified as unstructured in the primary analysis, allowing each time-specific variance and each covariance to be estimated freely. If this model fails to converge, we will use a more parsimonious structure, such as first-order autoregressive. Degrees of freedom for tests and confidence intervals for fixed effects will be estimated using Kenward-Roger approximation.

- **Diagnostics**

Model assumptions will be checked using:

- residual plots (fitted values vs residuals) and Q-Q plots for normality
- assessment of influential observations and fit indices (for example AIC/BIC)
- inspection of estimated covariance parameters (including whether simpler structures such as AR(1) provide an adequate fit)

If serious departures from assumptions are detected, we will report their impact through pre-specified sensitivity analyses (for example, models with robust standard errors or alternative covariance structures).

Sec 13: Additional Analyses

See Section 15 for example R code snippets.

Overview

Unless Table 8 states otherwise, secondary outcomes use the same repeated measures framework as the primary analysis: fixed effects for treatment, categorical time, treatment by time, the baseline value of that outcome, and age, gender and medical school; a participant random intercept; unstructured within-participant covariance with AR(1) as a fallback if that fails to converge.

The focus for each outcome is the adjusted between-arm contrast at T3, with T4 reported alongside. Every outcome is reported as adjusted marginal means by arm, a contrast with 95% CI and exact p -value, and an FDR adjusted p -value within its time point family (Section 10).

Thumos is delivered in therapist led groups of roughly 6 to 8, but randomisation is at the individual level and the control arm has no equivalent grouping, so group and therapist effects are not modelled.

Per protocol and CACE

The per protocol and CACE analyses need information on who received the intervention, which cannot be used under the masked A and B labelling. They follow lock of the primary analysis dataset, checking of the masked primary results, and release of the allocation key (Section 7).

- **Per protocol (PP)**

We will repeat the primary analysis in the per protocol set, defined in line with Section 8.

- **Complier average causal effect (CACE)**

We will estimate the CACE for BRS at T3 by two stage least squares, using randomised allocation as an instrument for a binary take up indicator (Section 5; Angrist et al. (1996) [10]; Dunn et al. (2005) [11]; Boatman et al. (2018) [12]). Baseline BRS, age, gender and medical school will be included in both stages. Estimation will use a single instrumental variables fit rather than two regressions run in sequence, since the manual two-step version produces the same point estimate but standard errors that are too small.

We will report the effect among students who would take up Thumos with its 95% CI, the estimated proportion of the Thumos arm meeting the take up definition, and the first stage partial F statistic as a measure of instrument strength (<10 = weak).

Table 8. Specification of additional analyses

Analysis	Model	Effect measure	Notes
Continuous secondary outcomes: BRS (T4), CCAE, OLB total and subscales, PHQ-9, ReQoL-UI utility	Linear MMRM as above	Adjusted mean difference and Cohen's d , 95% CI	Standardiser for d as in Section 10. OLB subscales interpreted alongside the total as one finding
Sickness absence: total days lost, and working or studying days lost, in past 4 weeks (analysed separately)	Poisson GLMM, log link	Adjusted incidence rate ratio, 95% CI, with adjusted mean days lost per arm	Negative binomial if overdispersed
QALYs: baseline to T4	Linear regression, one value per participant, adjusted for baseline utility, age, gender and medical school	Adjusted mean difference, 95% CI	Area under the utility time curve with linear interpolation (Manca et al. (2005) [13]). Run within the multiple imputation framework (Section 11)
Support access (four items)	None	None	Descriptive only, by arm and time point. Context for the treatment policy strategy in Section 5; not used as a covariate
Per protocol	Primary model refitted in the per protocol set (Section 8)	Adjusted mean difference in BRS at T3, 95% CI	Reported with the number and baseline characteristics of those excluded
Complier average causal effect	Two stage least squares, allocation instrumenting take up	Mean difference among students who would take up Thumos, 95% CI, raw and standardised	Specification above
Safety	None	Risk difference, 95% CI	Participants with at least one AE and at least one SAE by arm, with narrative summaries

Sec 14: SAP Reporting Checklist

In accordance with Gamble et al. (2017) [14]

Table 9. SAP reporting checklist

Section/Item	Index	Checklist description as written	How addressed
Section 1: administrative information			
Trial and trial registration	1a	Descriptive title that matches the protocol, with SAP either as a forerunner or subtitle, and trial acronym (if applicable)	A randomised controlled evaluation of Thumos: a cognitive-behavioural intervention for medical students. Acronym: ACTIVATE
	1b	Trial registration number	ISRCTN 12744098; REC reference 2024-21589-38791
SAP version	2	SAP version number with dates	Version 1, dated 13/08/2026 (seventh iteration; Section 16)
Protocol version	3	Reference to version of protocol being used	Protocol version 2.3, dated 15/01/2026
SAP revisions	4a	SAP revision history	Section 16
	4b	Justification for each SAP revision	Section 16
	4c	Timing of SAP revisions in relation to interim analyses, etc.	N/A, no interim analyses planned
Roles and responsibility	5	Names, affiliations, and roles of SAP contributors	SAP author and trial statistician: Dr Luke Budworth. TSC independent statistician: Prof Mona Kanaan. Chief Investigator: Dr Judith Johnson. Trial Co-Investigator: Dr Chris Taylor. Sponsor: University of Manchester. Funder: MPS Foundation
Signatures	6a	Person writing the SAP	Dr Luke Budworth (Section 1)
	6b	Senior statistician responsible	Prof Mona Kanaan (Section 1)
	6c	Chief investigator/clinical lead	Dr Judith Johnson (Section 1)
Section 2: introduction			
Background and rationale	7	Synopsis of trial background and rationale including a brief description of research question and brief justification for undertaking the trial	Section 2. Evaluates whether Thumos improves resilience and related outcomes in UK medical students in clinical placement years; full rationale in the protocol
Objectives	8	Description of specific objectives or hypotheses	Section 4. Primary: effect on BRS at T3 (4 months). Secondary: confidence, burnout, depression, mental health utility, QALYs and sickness absence at T3 and T4. Hypothesis: adjusted mean difference at T3 corresponding to $d \approx 0.50$
Section 3: study methods			

Trial design	9	Brief description of trial design including type of trial (e.g., parallel group, multi-arm, crossover, factorial) and allocation ratio and may include brief description of interventions	Table 1. Two-arm parallel group superiority RCT, 1:1 individual randomisation to Thumos versus services as usual. T1 baseline, T2 around 6 weeks, T3 4 months (primary), T4 9 months. All questionnaires online via Qualtrics
Randomisation	10	Randomisation details, e.g., whether any minimisation or stratification occurred (including stratifying factors used or the location of that information if it is not held within the SAP)	Section 7. Simple 1:1 via Sealed Envelope, which generates and holds the sequence. No stratification or minimisation. Allocation revealed only after consent and baseline completion
Sample size	11	Full sample size calculation or reference to sample size calculation in protocol (instead of replication in SAP)	Section 6 and Table 5. Target 220 randomised (110 per arm), simulated in powerImm. Protocol v2.3 permits recruiting above target to replace post-randomisation withdrawals, capped at 10% overall
Framework	12	Superiority, equivalence, or noninferiority hypothesis testing framework, including which comparisons will be presented on this basis	Superiority, for all comparisons
Statistical interim analysis and stopping guidance	13a	Information on interim analyses specifying what interim analyses will be carried out and listing of time points	None planned
	13b	Any planned adjustment of the significance level due to interim analysis	N/A
	13c	Details of guidelines for stopping the trial early	Specified in the protocol (sections 10 and 13.3), not the SAP
Timing of final analysis	14	Timing of final analysis, e.g., all outcomes analysed collectively or timing stratified by planned length of follow-up	Section 7. All outcomes analysed collectively after follow-up completion and data lock. Allocation key released only after the dataset is locked and the masked primary results checked
Section 4: statistical principles			
Confidence intervals and p values	15	Level of statistical significance	Section 10. Primary inference two-sided at $\alpha = .05$ for BRS at T3, with emphasis on estimation rather than dichotomous significance
	16	Description and rationale for any adjustment for multiplicity and, if so, detailing how the type 1 error is to be controlled	Section 10. No formal multiplicity adjustment; the primary endpoint is a single test and carries the trial's conclusion. False discovery rate adjusted p-values (Benjamini-Hochberg) reported for secondary endpoints within two families, T3 and T4, for transparency only
	17	Confidence intervals to be reported	95% CIs for all estimates (Section 10)

Adherence and protocol deviations	18a	Definition of adherence to the intervention and how this is assessed including extent of exposure	Section 8. Minimum adherence, per protocol section 9.5, is completion of both group workshops. Operationalised as attendance at both workshops within 6 weeks of randomisation with at least 75% of each session attended. The one-to-one call is not required. Control arm: no receipt of any Thumos component
	18b	Description of how adherence to the intervention will be presented	Section 13. Adherence summarised descriptively by arm. PP analysis excludes non-adherent Thumos participants and major deviations; CACE estimated by two stage least squares with allocation instrumenting take up
	18c	Definition of protocol deviations for the trial	Section 8. Classified minor or major. Major: invalid consent or ineligibility, allocation error or concealment breach, any Thumos receipt in the control arm, no valid T3 BRS within window, data integrity issues
	18d	Description of which protocol deviations will be summarized	Section 8. All deviations logged and summarised by type and arm before the allocation key is released
Section 5: trial population			
Analysis populations	19	Definition of analysis populations, e.g., intention to treat, per protocol, complete case, safety	Section 8. Full analysis set (ITT): all randomised, analysed as allocated. Per protocol: FAS excluding major deviations and non-adherent Thumos participants, requiring a valid T3 BRS in window. CACE: all randomised, with a binary take up indicator instrumented by allocation
Screening data	20	Reporting of screening data (if collected) to describe representativeness of trial sample	Documented in the protocol, not the SAP
Eligibility	21	Summary of eligibility criteria	Table 1. UK medical students in a year containing a clinical placement of one week or longer, recruited nationally (typically Y4 or Y5). Full criteria in protocol sections 5.1 and 5.2
Recruitment	22	Information to be included in the CONSORT flow diagram	Screened where applicable, randomised, received intervention, followed up at T2 to T4, analysed, with reasons for non-participation and loss. Shell in the protocol
Withdrawal or follow-up	23a	Level of withdrawal, e.g., from intervention and/or from follow-up	Withdrawal from intervention and from follow-up summarised separately, by arm and time point
	23b	Timing of withdrawal/lost to follow-up data	Reported against the T2, T3 and T4 windows (Section 8)
	23c	Reasons and details of how withdrawal/lost to follow-up data will be presented	Listed and summarised descriptively in the CONSORT diagram and the missing data summary (Section 11)
Baseline patient characteristics	24a	List of baseline characteristics to be summarized	Age, gender, year of study, medical school, and baseline outcomes (BRS, CCAE, OLBI, PHQ-9, ReQoL-UI)

	24b	Details of how baseline characteristics will be descriptively summarized	Section 10. Continuous: mean and SD, or median and IQR if skewed. Categorical: n and %. Absolute standardised differences between arms reported and flagged above 0.20. Baseline balance is described, not tested
Section 6: analysis			
Outcome definitions	25a	Specification of outcomes and timings. If applicable include the order of importance of primary or key secondary end points (e.g., order in which they will be tested)	Tables 2, 3 and 6. Primary: BRS at T3. Secondary: BRS at T4; CCAE, OLBI (total and subscales), PHQ-9 and ReQoL-UI utility at T3 and T4; QALYs baseline to T4; two sickness absence counts at T3 and T4. No formal testing hierarchy; secondary endpoints are supportive
	25b	Specific measurement and units (e.g., glucose control, hbA1c [mmol/mol or %])	Table 6. BRS mean score (1 to 5); CCAE mean score; OLBI subscale and total means; PHQ-9 total (0 to 27); ReQoL-UI utility index; QALYs from utilities; sickness absence as days lost in the past 4 weeks (two items, analysed separately)
	25c	Any calculation or transformation used to derive the outcome (e.g., change from baseline, QoL score, time to event, logarithm, etc.)	Table 6. Scored per each measure's guidance, with reverse scoring where required. Item level missingness handled by person mean substitution up to the per scale thresholds in Table 6; beyond those the score is set to missing. ReQoL-UI utilities via the published algorithm; QALYs by area under the utility time curve with linear interpolation. No transformations of the analysis scale
Analysis methods	26a	What analysis method will be used and how the treatment effects will be presented	Sections 12 and 13. Primary: linear mixed model for repeated measures (REML) on BRS at T2 to T4 with baseline as covariate, reporting the adjusted Thumos minus SAU difference at T3 with 95% CI, p-value and Cohen's <i>d</i> standardised by the pooled baseline SD. Continuous secondary endpoints mirror this. Counts: Poisson GLMM with negative binomial fallback, reported as adjusted incidence rate ratios with adjusted mean days lost per arm. QALYs: linear regression adjusted for baseline utility and the primary covariate set
	26b	Any adjustment for covariates	Baseline value of the outcome, age, gender and medical school (Section 12)
	26c	Methods used for assumptions to be checked for statistical methods	Sections 10 and 12. Residual, scale-location and influence diagnostics, Q-Q plots, and inspection of estimated covariance parameters and information criteria
	26d	Details of alternative methods to be used if distributional assumptions do not hold, e.g., normality, proportional hazards, etc.	Sections 10, 12 and 13. AR(1) covariance if the unstructured model fails to converge; negative binomial if counts are overdispersed; CR2 cluster robust standard errors (clubSandwich) clustered by participant if diagnostics show clear

			non-normality or heteroskedasticity, with model based and robust SEs reported side by side
	26e	Any planned sensitivity analyses for each outcome where applicable	Sections 11 and 13. Multiple imputation under MAR; pattern mixture delta adjustment (0.10, 0.20, 0.30 BRS points applied to imputed values in the Thumos arm); last post-baseline carried forward; per protocol analysis; CACE
	26f	Any planned subgroup analyses for each outcome including how subgroups are defined	None pre-specified. The CACE (Table 4) targets a principal stratum rather than an observed subgroup
Missing data	27	Reporting and assumptions/statistical methods to handle missing data (e.g., multiple imputation)	Sections 9 and 11. Item level missingness by person mean substitution within thresholds (Table 6). Primary analysis uses all available data and assumes MAR given observed outcomes and model covariates. Missingness patterns and predictors described, including logistic models with missingness as the outcome. Sensitivity analyses per Table 7
Additional analyses	28	Details of any additional statistical analyses required, e.g., complier-average causal effect analysis	Section 13 and Table 8. CACE for BRS at T3 by two stage least squares (single IV fit), with allocation instrumenting a binary take up indicator, reporting the effect among students who would take up Thumos with 95% CI, the take up proportion, and the first stage partial F statistic (below 10 flagged as a weak instrument). Assumptions stated in Table 4
Harms	29	Sufficient detail on summarizing safety data, e.g., information on severity, expectedness, and causality; details of how adverse events are coded or categorized; how adverse event data will be analysed, i.e., grade 3/4 only, incidence case analysis, intervention emergent analysis	Table 8. Number and proportion of participants with at least one AE and at least one SAE by arm, with risk differences and 95% CIs and narrative summaries by category and seriousness. No hypothesis testing. Definitions, coding, reporting routes and stopping guidelines are in protocol sections 10 and 13.3
Statistical software	30	Details of statistical packages to be used to carry out analyses	Table 1 and Section 15. R version 4.5.1 or the latest available at the time of analysis, using lme4, lmerTest, emmeans, mice, AER, clubSandwich and psych. All code version controlled
References	31a	References to be provided for nonstandard statistical methods	Section 18
	31b	Reference to Data Management Plan	N/A, no separate DMP; data handling is specified in protocol sections 14 and 15
	31c	Reference to the Trial Master File and Statistical Master File	Held by the sponsor, University of Manchester

	31d	Reference to other standard operating procedures or documents to be adhered to	Sheffield CTRU SOP ST001, "The Statistical Analysis Plan" (02/07/2025)
--	-----	--	--

Sec 15: Example R code snippets

```
## ACTIVATE SAP Section 15: worked examples on SIMULATED data.
## Verified on R 4.3.3, lme4 1.1-35.1, mice 3.16.0, AER 1.2-12.
## No trial data are used; no result here says anything about Thumos.

pkgs <- c("lme4", "lmerTest", "emmeans", "mice", "broom.mixed", "AER",
          "psych", "GPArotation", "clubSandwich", "tidyr", "dplyr")
new <- setdiff(pkgs, rownames(installed.packages()))
if (length(new)) install.packages(new)
invisible(lapply(pkgs, library, character.only = TRUE))

set.seed(2026)

## ---- 1. Simulate a trial-shaped dataset -----
n <- 220
sim <- tibble(
  pid      = sprintf("P%03d", 1:n),
  treat    = factor(rep(c("SAU", "Thumos"), each = n/2), levels = c("SAU", "Thumos")),
  age      = round(rnorm(n, 23, 3)),
  gender    = factor(sample(c("F", "M", "Other"), n, TRUE, c(.55, .42, .03))),
  school    = factor(sample(paste0("S", 1:12), n, TRUE))
)

u <- rnorm(n, 0, 0.55)                                # participant intercept
brs_items <- sapply(1:6, function(i) pmin(5, pmax(1, round(2.7 + u + rnorm(n, 0,
.8)))))
colnames(brs_items) <- paste0("brs_i", 1:6)
sim$brs_T1 <- rowMeans(brs_items)

eff <- ifelse(sim$treat == "Thumos", 0.35, 0)          # true effect, BRS points
brs <- function(m) pmin(5, pmax(1, 2.75 + u + m*eff + rnorm(n, 0, .45)))
sim$brs_T2 <- brs(0.8); sim$brs_T3 <- brs(1); sim$brs_T4 <- brs(0.8)

frail <- rnorm(n, 0, 0.8)                              # frailty for counts
for (t in c("T1", "T2", "T3", "T4")) {
  sim[[paste0("util_", t)]] <- pmin(1, pmax(-.2,
    .70 + .25*(sim$brs_T1 - 2.7) +
    ifelse(t %in% c("T3", "T4"), .02*(sim$treat == "Thumos"), 0) + rnorm(n, 0,
.12)))
}
```

```

sim[[paste0("days_", t)]] <- rpois(n, exp(-0.2 + frail -
  ifelse(sim$treat == "Thumos" & t %in% c("T3","T4"), 0.35, 0)))
}

sim$took_up <- ifelse(sim$treat == "Thumos", rbinom(n, 1, .70), 0)

d2 <- runif(n) < ifelse(sim$treat == "SAU", .12, .08) # monotone attrition
d3 <- d2 | runif(n) < ifelse(sim$treat == "SAU", .18, .13)
d4 <- d3 | runif(n) < ifelse(sim$treat == "SAU", .15, .12)
for (v in c("brs","util","days")) {
  sim[[paste0(v,"_T2")]][d2] <- NA
  sim[[paste0(v,"_T3")]][d3] <- NA
  sim[[paste0(v,"_T4")]][d4] <- NA
}

wk <- function(mu, lo, hi) pmin(hi, pmax(lo, round(rnorm(n, mu, 3))))
sim$wk_T2 <- wk(6, 4, 10); sim$wk_T3 <- wk(16, 12, 24); sim$wk_T4 <- wk(36, 30,
44)

## ---- 2. Long format and assessment windows -----
long <- sim %>%
  pivot_longer(matches("^(brs|util|days|wk)_T[234]$"),
    names_to = c(".value", "tp"), names_pattern = "(\\w+)_T[234]") %>%
  rename(weeks = wk) %>%
  mutate(time_f = case_when(weeks >= 4 & weeks <= 10 ~ "T2",
    weeks >= 12 & weeks <= 24 ~ "T3",
    weeks >= 30 & weeks <= 44 ~ "T4"),
    time_f = factor(time_f, levels = c("T2","T3","T4"))) %>%
  filter(!is.na(time_f)) %>%
  select(pid, treat, age, gender, school, baseline = brs_T1, util_T1,
    took_up, time_f, weeks, brs, util, days)

## ---- 3. Primary analysis -----
fit <- lmer(brs ~ treat*time_f + baseline + age + gender + school + (1|pid),
  data = long, REML = TRUE)

emm <- emmeans(fit, ~ treat | time_f)
ctr <- confint(pairs(emm, reverse = TRUE))
cat("\n--- adjusted means ---\n"); print(emm)

```

```
cat("\n--- Thumos minus SAU ---\n"); print(ctr)

## ---- 4. Cohen's d, pooled baseline SD -----
sd_bl <- sd(sim$brs_T1, na.rm = TRUE)
d_T3 <- as.data.frame(ctr) %>% filter(time_f == "T3") %>%
  transmute(d = estimate/sd_bl, lcl = lower.CL/sd_bl, ucl = upper.CL/sd_bl)
cat("\n--- Cohen's d at T3 (SD =", round(sd_bl, 3), ") ---\n"); print(d_T3)

## ---- 5. Multiple imputation -----
wide_mi <- sim %>% select(treat, age, gender, school, took_up,
                        brs_T1, brs_T2, brs_T3, brs_T4)
imp <- mice(wide_mi, m = 10, printFlag = FALSE, seed = 2026)

# T3 is the reference level here, so the pooled treatThumos coefficient
# IS the T3 contrast (emmeans does not pool across imputations).
to_long <- function(d) {
  d$pid <- sim$pid
  d %>% pivot_longer(brs_T2:brs_T4, names_to = "time_f",
                    names_prefix = "brs_", values_to = "brs") %>%
    mutate(time_f = factor(time_f, levels = c("T3", "T2", "T4")))
}

mi_fit <- function(data_list)
  pool(as.mira(lapply(data_list, function(d)
    lmer(brs ~ treat*time_f + brs_T1 + age + gender + school + (1|pid),
      data = to_long(d))))))

imp_sets <- lapply(1:imp$m, function(i) complete(imp, i))
cat("\n--- MI under MAR: treatment effect at T3 ---\n")
print(summary(mi_fit(imp_sets), conf.int = TRUE) %>%
  filter(term == "treatThumos") %>%
  select(estimate, std.error, `2.5 %`, `97.5 %`))

## ---- 6. Delta adjustment -----
was_na <- is.na(wide_mi[, c("brs_T3", "brs_T4")])

# arm = "both" penalises every imputed value; arm = "Thumos" penalises the
# intervention arm only. With differential attrition these are NOT
# equivalent: see the two sets of results below.
```

```

delta_sets <- function(delta, arm = c("both","Thumos")) {
  arm <- match.arg(arm)
  cols <- c("brs_T3","brs_T4")
  lapply(imp_sets, function(d) {
    hit <- was_na & (if (arm == "both") TRUE else d$treat == "Thumos")
    d[cols][hit] <- d[cols][hit] - delta
    d
  })
}

cat("\n--- delta adjusted T3 effects ---\n")
for (a in c("both","Thumos")) {
  cat("penalty applied to:", a, "\n")
  for (dl in c(0.10, 0.20, 0.30)) {
    r <- summary(mi_fit(delta_sets(dl, a)), conf.int = TRUE) %>%
      filter(term == "treatThumos")
    cat(sprintf("  delta %.2f:  %.3f (%.3f to %.3f)\n",
                dl, r$estimate, r$`2.5 %`, r$`97.5 %`))
  }
}

## ---- 7. Count outcome -----
fit_p <- glmer(days ~ treat*time_f + age + gender + (1|pid),
               data = long, family = poisson(link = "log"))

disp <- sum(residuals(fit_p, type = "pearson")^2) / df.residual(fit_p)
cat("\n--- dispersion ratio ---\n"); print(round(disp, 2))

fit_cnt <- if (disp > 1.5) {
  # refit only if overdispersed
  glmer.nb(days ~ treat*time_f + age + gender + (1|pid), data = long)
} else fit_p

cat("\n--- rate ratios ---\n")
print(pairs(emmeans(fit_cnt, ~ treat|time_f, type = "response"), reverse = TRUE))
cat("\n--- mean days lost per arm ---\n")
print(emmeans(fit_cnt, ~ treat|time_f, type = "response"))

## ---- 8. QALYs -----
util_long <- sim %>%

```

```

select(pid, util_T1, util_T2, util_T3, util_T4) %>%
  pivot_longer(-pid, names_to = "tp", names_prefix = "util_", values_to = "util")
%>%
  mutate(weeks = c(T1 = 0, T2 = 6, T3 = 16, T4 = 36)[tp]) %>%
  arrange(pid, weeks)

qaly <- util_long %>% group_by(pid) %>%
  mutate(years = (weeks - lag(weeks))/52.18,
         area = ((util + lag(util))/2) * years) %>%
  summarise(qaly = sum(area, na.rm = TRUE), .groups = "drop") %>%
  left_join(select(sim, pid, treat, util_T1, age, gender, school), by = "pid")

cat("\n--- QALY difference ---\n")
print(confint(lm(qaly ~ treat + util_T1 + age + gender + school,
                 data = qaly))["treatThumos", ])

## ---- 9. CACE -----
iv_form <- brs ~ took_up + brs_T1 + age + gender | treat + brs_T1 + age + gender

cace <- suppressWarnings(pool(as.mira(lapply(imp_sets, function(d)
  ivreg(iv_form, data = rename(d, brs = brs_T3))))))
cat("\n--- CACE at T3 (pooled over imputations) ---\n")
print(suppressWarnings(summary(cace, conf.int = TRUE)) %>% filter(term ==
"took_up"))

obs <- filter(long, time_f == "T3")
cat("\n--- weak instrument test, observed cases ---\n")
print(summary(ivreg(brs ~ took_up + baseline + age + gender |
  treat + baseline + age + gender,
  data = obs), diagnostics = TRUE)$diagnostics["Weak
instruments", ])
cat("take up in Thumos arm:", round(mean(obs$took_up[obs$treat == "Thumos"]), 3),
"\n")

## ---- 10. Reliability -----
items <- as.data.frame(brs_items)

rel <- function(d) {
  # alpha and omega for one dataset
  a <- suppressWarnings(suppressMessages(
    psych::alpha(d, warnings = FALSE)$total$raw_alpha))

```

```
invisible(capture.output(o <- suppressWarnings(suppressMessages(
  psych::omega(d, nfactors = 1, plot = FALSE))))
c(alpha = a, omega = o$omega.tot)
}

est <- rel(items)
bs <- replicate(1000, {
  # percentile bootstrap CIs
  d <- items[sample(nrow(items), replace = TRUE), ]
  tryCatch(rel(d), error = function(e) c(alpha = NA, omega = NA))
})

cat("\n--- BRS reliability at T1 ---\n")
print(round(cbind(estimate = est,
  t(apply(bs, 1, quantile, c(.025, .975), na.rm = TRUE))), 3))

## ---- 11. Diagnostics -----
cat("\n--- diagnostics ---\n")
cat("scaled residuals |r| > 3:", sum(abs(resid(fit, scaled = TRUE)) > 3), "\n")

cd <- cooks.distance(influence(fit, group = "pid"))
cat("participants above 4/n Cook's D:", sum(cd > 4/length(cd)), "\n")

## ---- 12. Robust standard errors -----
# Cluster is taken from the model frame, not the original data, because
# lmer drops rows with a missing outcome.
v <- vcovCR(fit, cluster = fit@frame$pid, type = "CR2")
rb <- as.data.frame(coef_test(fit, vcov = v))

cmp <- data.frame(term = rb$Coef,
  model_se = summary(fit)$coefficients[rb$Coef, "Std. Error"],
  robust_se = rb$SE)
cmp$ratio <- cmp$robust_se / cmp$model_se

cat("\n--- model vs robust SE, treatment terms ---\n")
print(cmp[grepl("treat", cmp$term), ], row.names = FALSE)

## ---- 13. FDR adjustment -----
p_T3 <- c(CCAE = .012, OLBI_tot = .048, OLBI_ex = .031, OLBI_dis = .13,
  PHQ9 = .22, ReQoL = .31, days_all = .44, days_work = .61)
```

```
cat("\n--- BH adjusted p-values, T3 family ---\n")  
print(round(p.adjust(p_T3, method = "BH"), 3))
```

Sec 16: Version history

Table 10. SAP version history

Version	Date	Key changes (tracked changes documents kept + available on request)
0.1	27/08/2025	
0.2	28/08/2025	Added completed checklist
0.3	03/09/2025	- Modified in line with Christopher Taylor's comments (draft with comments kept) - Modified to be more in line with previous published examples e.g. SleepWell and iMAPS-2 SAPs - Included more background, specified more detail how to handle missingness in items on measures, clarified things such as missing data approaches, multiplicity adjustment, added code snippets, and added this version history section
0.4	25/09/2025	Modified after further feedback from Chritopher, Judith Johnson, and Lucy Pointon: - The handling of 'intercurrent 'events is narrower, with academic interruptions dropped, and external support is now only explored in sensitivity analyses rather than adjusted directly. - Support access, previously a listed secondary outcome, is no longer specified as such. - The PP/CACE compliance definition is relaxed so that only both workshops ($\geq 75\%$ attendance within 6 weeks) are required, whereas the one-to-one call is optional. - Missing- tem rules are slightly stricter, reverting OLBI back to ≥ 7 of 8 items per subscale (max 1 imputation) rather than the older ≥ 6 with up to 2 imputations. - The primary model covariates are trimmed (dropping "year of study") and clustering sensitivity is simpler (therapist intercept only, without medical school effects). - Section wording is also more extensive, with greater emphasis on SOPs and guidelines.
0.5	15/11/2025	Modified after feedback from our independent statistician. Mainly clarifications, including:

		- Converting standardised effect sizes to real units to further justify sample size analysis - Added much more detail to sample size section - Clarified the use of multilevel models - Labelled tables and provided definitions - Added more clarity and corrections around handling missing item data for outcome measures - Rewritten the missing data section for clarity - Added more detail to analyses sections - Added more citations throughout to justify some of our decisions
0.6	16/12/2025	Addressed the following final comments from Mona, the trial independent statistician, namely: - Added rationale for reliability analyses - Added more clarity in missing data section
1	13/08/2026	- Aligned with protocol v2.3 (15/01/2026). The minimum adherence definition now matches protocol section 9.5 throughout: completion of both group workshops, operationalised as attendance at both within 6 weeks of randomisation with at least 75% of each session attended; the one-to-one call is not required. Earlier wording requiring the call, which had survived in the estimand, additional analyses and checklist sections, has been corrected. - Cohen's d standardiser specified. The standardiser is now stated explicitly as the pooled baseline (T1) SD of the relevant outcome, replacing the previous "model based marginal SD at T3" wording. - Support access is no longer analysed as an efficacy endpoint. It is retained, collected at all four time points, and summarised descriptively by arm and time point as context for the treatment policy strategy. It is not used as a covariate for any other outcome, because it is measured post-randomisation and may itself be affected by allocation.

		- Therapist clustering sensitivity analysis removed. Randomisation is at the individual level and the control arm has no equivalent grouping. The likely magnitude of any clustering is noted in Section 13 instead. - Specified that delta adjustment now applies to the Thumos arm <i>only</i>. Applying the penalty to both arms is not conservative when attrition is differential, since the arm with more missing data is penalised more. - More details on our blinding approach. - Multiplicity wording corrected from Storey q-values to false discovery rate adjusted p-values by the Benjamini-Hochberg procedure (which is more preferable), calculated within the T3 and T4 families separately. - Estimands restructured into three tables (primary, secondary, supplementary) following ICH E9(R1) attributes. - Structural changes: a new trial method summary table replaces the synopsis; all tables numbered and cross-referenced; sections renumbered, so the reporting checklist is now Section 14; the checklist rebuilt against the Gamble et al. (2017) section structure, including the previously missing statistical principles and trial population headings; references split into background references (Section 17) and SAP method references (Section 18), with in text DOIs replaced by author-date citations. - Section 15 rewritten as a single self-contained script that runs end to end on simulated data, so every analysis specified in Sections 10 to 13 has verified syntax before the trial dataset exists. - Minor corrections to typographical errors, the version date in the header table, and cross-references throughout.
--	--	---

Sec 17: Background references

1. Haakenstad, A., et al., *Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019*. The Lancet, 2022. **399**(10341): p. 2129-2154.
2. The Health Foundation, The Kingsfund, and Nuffield Trust. *Closing the gap: Key areas for action on the health and care workforce*. 2019 [cited 2019 20th September]; Available from: https://www.kingsfund.org.uk/sites/default/files/2019-03/closing-the-gap-health-care-workforce-overview_0.pdf
3. Dall'Ora, C., et al., *Nurse staffing levels and patient outcomes: a systematic review of longitudinal studies*. International Journal of Nursing Studies, 2022: p. 104311.
4. Maslach, C. and S.E. Jackson, *The measurement of experienced burnout*. Journal of Occupational Behavior, 1981. **2**(2): p. 99-113.
5. Toppinen-Tanner, S., et al., *Burnout as a predictor of medically certified sick-leave absences and their diagnosed causes*. Behavioral medicine, 2005. **31**(1): p. 18-32.
6. Bae, S.H., et al., *Predictors of actual turnover among nurses working in Korean hospitals: A nationwide longitudinal survey study*. Journal of Nursing Management, 2021. **29**(7): p. 2102-2114.
7. Hodkinson, A., et al., *Associations of physician burnout with career engagement and quality of patient care: systematic review and meta-analysis*. BMJ, 2022. **378**.
8. Vogt, K.S., et al., *Critical care nursing workforce in crisis: A discussion paper examining contributing factors, the impact of the COVID-19 pandemic and potential solutions*. Journal of Clinical Nursing, 2023.
9. Johnson, J., et al., *Reboot coaching programme: a mixed-methods evaluation assessing resilience, confidence, burnout and depression in medical students*. Scottish Medical Journal, 2023: p. 00369330231213981.
10. Rotenstein, L.S., et al., *Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis*. Jama, 2016. **316**(21): p. 2214-2236.
11. Puthran, R., et al., *Prevalence of depression amongst medical students: A meta-analysis*. Medical education, 2016. **50**(4): p. 456-468.
12. Dyrbye, L.N., et al., *Burnout and suicidal ideation among US medical students*. Annals of internal medicine, 2008. **149**(5): p. 334-341.
13. British Medical Association. *Four in ten junior doctors plan to leave the NHS as soon as they can find another job, BMA council chair reveals in New Year's message*. 2022 [cited 2023 23rd June]; Available from: <https://www.bma.org.uk/bma-media-centre/four-in-ten-junior-doctors-plan-to-leave-the-nhs-as-soon-as-they-can-find-another-job-bma-council-chair-reveals-in-new-years-message-to-the-country>.
14. Tolksdorf, K.H., U. Tischler, and K. Heinrichs, *Correlates of turnover intention among nursing staff in the COVID-19 pandemic: a systematic review*. BMC Nursing, 2022. **21**(1): p. 174.
15. Hill, M.R., S. Goicochea, and L.J. Merlo, *In their own words: stressors facing medical students in the millennial generation*. Medical education online, 2018. **23**(1): p. 1530558.
16. Hu, K.S., J.T. Chibnall, and S.J. Slavin, *Maladaptive perfectionism, impostorism, and cognitive distortions: Threats to the mental health of pre-clinical medical students*. Academic Psychiatry, 2019. **43**(4): p. 381-385.
17. Weurlander, M., et al., *How do medical and nursing students experience emotional challenges during clinical placements?* International journal of medical education, 2018. **9**: p. 74.
18. Nteveros, A., et al., *Burnout among medical students in Cyprus: A cross-sectional study*. PLoS One, 2020. **15**(11): p. e0241335.

19. Witt, K., et al., *Effectiveness of universal programmes for the prevention of suicidal ideation, behaviour and mental ill health in medical students: a systematic review and meta-analysis*. *BMJ Ment Health*, 2019. **22**(2): p. 84-90.
20. Gridley, K., *Resilience training is just a band-aid solution for doctor well-being: No*. 2018. **30**(2): p. 261-262.
21. Chew-Graham, C.A., A. Rogers, and N. Yassin, 'I wouldn't want it on my CV or their records': *medical students' experiences of help-seeking for mental health problems*. *Medical education*, 2003. **37**(10): p. 873-880.
22. Dyrbye, L.N., et al., *The impact of stigma and personal experiences on the help-seeking behaviors of medical students with burnout*. *Academic medicine*, 2015. **90**(7): p. 961-969.
23. Johnson, J., Simms-Ellis, R., Janes, G., Mills, T., Budworth, B., Atkinson, L. & Harrison, R., *Can we prepare healthcare professionals for involvement in adverse events? Feasibility study of a resilience training intervention*. *BMC Health Services Research*, 2020. **20**: p. 1094.
24. Janes, G., et al., *Multiple meanings of resilience: Health professionals' experiences of a dual element training intervention designed to help them prepare for coping with error*. *Journal of Evaluation in Clinical Practice*, 2021. **28**: p. 315-323.
25. Vogt, K., et al., *Study protocol for the online adaptation and evaluation of the 'Reboot'(Recovery-boosting) coaching programme, to prepare critical care nurses for, and aid recovery after, stressful clinical events*. *Pilot and Feasibility Studies*, 2022. **8**(1): p. 63.
26. Vogt, K., et al., *Can the Reboot coaching programme support critical care nurses in coping with stressful clinical events? A mixed-methods evaluation assessing resilience, burnout, depression and turnover intentions*. *BMC Health Services Research*, 2024. **In press**

Sec 18: SAP references

1. Kahan BC, Hindley J, Edwards M, Cro S, Morris TP. The estimands framework: a primer on the ICH E9(R1) addendum. *BMJ* 2024;384:e076316.
2. Johnson J, Simms-Ellis R, Grange A, et al. Reboot coaching programme: a mixed-methods evaluation assessing resilience, confidence, burnout and depression in medical students. *Scottish Medical Journal* 2023;68(4):00369330231213981.
3. Cohen J. A power primer. *Psychological Bulletin* 1992;112(1):155-159.
4. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The Brief Resilience Scale: assessing the ability to bounce back. *International Journal of Behavioral Medicine* 2008;15(3):194-200.
5. de Miquel C, Haro JM, van der Feltz-Cornelis CM, Ortiz-Tallo A, Chen T, Sinokki M, Naumanen P, Olaya B, Lima RA. Differential attrition and engagement in randomized controlled trials of occupational mental health interventions in person and online: a systematic review and meta-analysis. *Scandinavian Journal of Work, Environment & Health* 2024;50(8):588-601.
6. Keetharuth AD, Brazier J, Connell J, et al. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *British Journal of Psychiatry* 2018;212(1):42-49.
7. Huisman M. Imputation of missing item responses: some simple techniques. *Quality & Quantity* 2000;34(4):331-351.
8. Parent MC. Handling item-level missing data: simpler is just as good. *The Counseling Psychologist* 2013;41(4):568-600.

9. Jacoby P, Whitehouse A, Leonard H, Saldaris J, Demarest S, Benke T, Downs J. Devising a missing data rule for a quality of life questionnaire: a simulation study. *Journal of Developmental & Behavioral Pediatrics* 2022;43(6):e414-e418.
10. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996;91(434):444-455.
11. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research* 2005;14(4):369-395.
12. Boatman JA, Vock DM, Koopmeiners JS, Donny EC. Estimating causal effects from a randomized clinical trial when noncompliance is measured with error. *Biostatistics* 2018;19(1):103-118.
13. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics* 2005;14(5):487-496.
14. Gamble C, Krishan A, Stocken D, et al. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* 2017;318(23):2337-2343.
15. Sheffield Clinical Trials Research Unit. SOP ST001: The Statistical Analysis Plan. University of Sheffield, 02/07/2025.