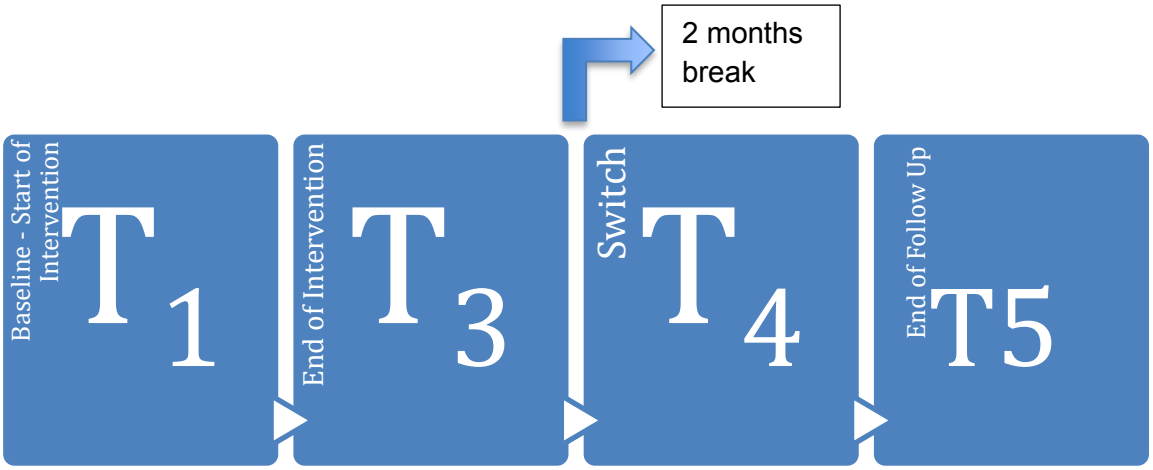
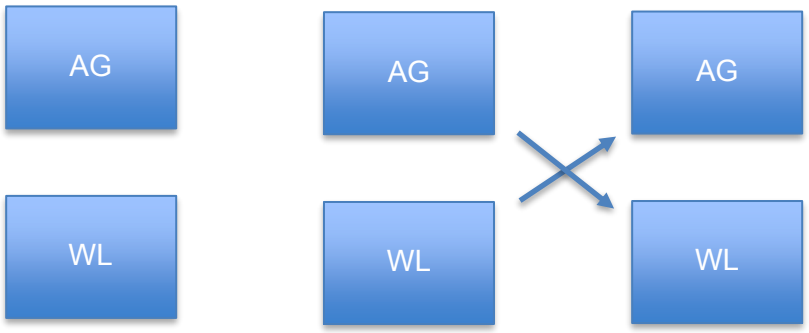
Pre-registration for quantitative part of Art on Prescription

Athens
March, 10, 2024

Design

Stratified randomised controlled trial with two arms: arts intervention (any of the arts interventions, see list), which the investigators call here Active Group (AG) vs waitlist control (WL). Stratifying will happen at each site.





REMEMBER the investigators want to measure AG and WL, then WL becomes AG after 3 months and both the initial AG and the WL → AG will be measured in exactly the same way for another 3 months.

AG for 3 months. Then goes into 3 month follow up.
WL for 3 months. Then this group receives intervention and becomes AG and has measurements for 3 months as the AG did.
There will be a 2-month break between T3 and T4 due to summer vacation

In the beginning, an open invitation to institutions of culture and mental health had been sent. Culture institutions i.e. Opera House, Contemporary Art Museum, National Theater etc. presented their action plans for interventional courses. Investigators did the connection between mental health and culture institutions.

Potential participants declared their preference for the interventional course such as dance, cinema etc. and they listed while an external investigator did the blind randomization into AG or WL group.

Measurement
Participants in both the AG and the WL groups will be measured on the following outcomes.
Primary Outcome: Wellbeing Questionnaire (WEMWBS), PHQ-9, GAD-7
Secondary Outcomes: UCLA-20.

Frequency of Measurement:
→The primary outcome will be measured at time points (for adults):

- T0: (before the randomisation; patient's personal data and demographics)
- T1: (before the start of the intervention for the active group or the entry into the study for the WL group)
- T2: 6 weeks (after the start of the intervention for the active group or the entry into the study for the WL group)
- T3: 12 weeks (after the start of the intervention for the active group or the entry into the study for the WL group)

The secondary outcomes will be measured at time points 0 (T4) and 12 weeks (T5) for both AG and WL.

→The primary outcome will be measured at time points (for children/adolescents):

- T0: (before the randomisation; patient's personal data and demographics)
- T1: (before the start of the intervention for the active group or the entry into the study for the WL group)
- T3: 12 weeks (after the start of the intervention for the active group or the entry into the study for the WL group)

The secondary outcomes will be measured at time points 0 (T4) and 12 weeks (T5) for both AG and WL.

Model
The investigators will follow standard practice for RCTs and apply an ANCOVA model to estimate the effects of treatment. The formalism is:

$$y \sim b0 + b1 \times treatment\_group + b2 \times time + b3 \times treatment\_group \times time + covariates \quad (1)$$

Where, y is each outcome, b1-b3 are the regression coefficients and covariates denote covariates at randomisation (DETERMINE)

The inference of interest for the model concerns b3, the interaction term, as the investigators posit that there will be a difference in slopes between the treatment groups.
The inference is going to happen via an estimation of the t-statistic, defined as follows"

t = b3/SE (2)

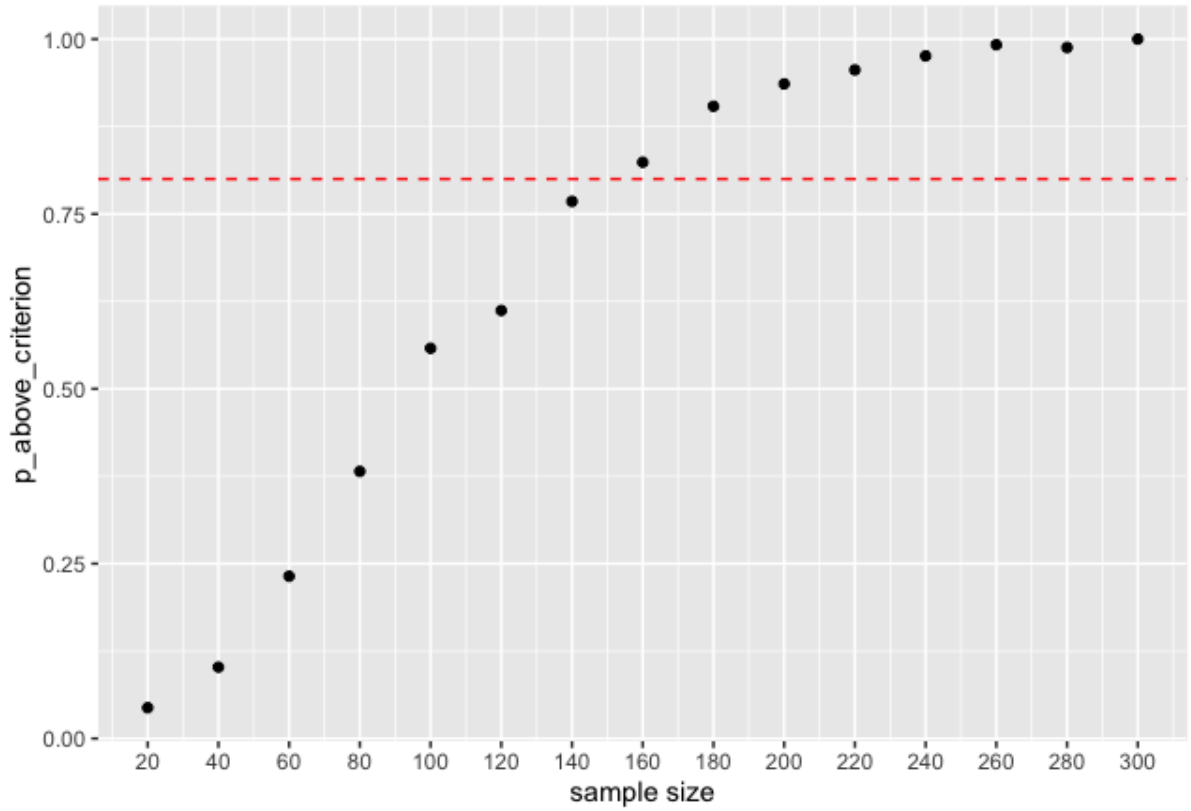
Where SE is the standard error for coefficient b3

t > 1.96 (3)

And 1.96 is the critical value for statistical significance at alpha = 0.05.

Target Sample Size per Arm
Parameters:

- power = 0.8
- alpha = 0.05
- Correlation between time points = 0.6 (a conservative estimate based on https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2222612/)
- Effect size corresponding to d = 0.5




```
library(tidyverse) # for data wrangling
library(faux)      # for simulation
library(broom)     # for tidy analysis results

set.seed(1974)

# I will simulate the following
```

```
# two groups, treatment and WL control.
# they start off with distributions with the same mean and sd (due to randomisation)
# the WL control stays the same, the treatment group changes to arrive at a d = 0.4, which i
# what investigators stipulated in advance
# the shape of the change should not matter that much for this linear model calculation
# what should matter are
# a) sds which I will keep the same as in baseline
# b) number of measurements, stipulated to be n = 4
# c) number of people, sample size, which I will vary for the simulation



# here comes the actual function
ancova_sim_func <- function(df_means, sd_across, n_subjects, number_sims,
correlations_across){

  arts_on_pres_sim <- sim_design(
    n = n_subjects, # this is the per arm n
    between = list(group = c("active", "wl")),
    within = list(time = c("T1", "T2", "T3", "T4")),
    mu= df_means,
    sd = sd_across,
    dv = "score",
    r = correlations_across,
    long = TRUE,
    plot = FALSE,
    rep = number_sims

  )


  temp_results <- lst()
  temp_results_disp <- lst()
  temp_p_values <- 0

  for(i in 1:n_sims){

    temp_results[[i]] <-  aov(score~ time + group + time*group , data =
arts_on_pres_sim$data[[i]])

    temp_results_disp[[i]] <-broom::tidy(temp_results[[i]])

    temp_p_values[i] <-  temp_results_disp[[i]]$p.value[3]


  }
  return(temp_p_values)
}
```

```
# here come the parameters that I will pass to this function
# from https://hqlo.biomedcentral.com/articles/10.1186/1477-7525-10-156
mean_T1 <- 43.5 # the biggest trial in that list, the PEIP
sd_T1 <- 10.4 # from that trial
d <- -0.5 # stipulated effect size of difference
# to derive the mean needed for the final effect size
# cohen_d = mean_T1-mean_T4/sd_T1 # using sd_T_1 as I stipulated this to be the same
throughout
# which converts to
mean_T4 <- mean_T1 + d*sd_T1

# to create equally spaced value, my idealised decline of values
the_treatment_means <- seq(mean_T1, mean_T4, length.out = 4)

# the control means will be the same throughout
the_control_means <- rep(mean_T1, 4)

# investigators assume a correlation between values from each time point r = 0.6
within_corr <- 0.5

# now investigators can simulate using faux.
n_sims <- 500


# the way I have created the function above, I need to create this dataframe as an obejct for
# the faux function
df_means <-data.frame(T1 = c(the_treatment_means[1], the_control_means[1]),
                      T2 = c(the_treatment_means[2], the_control_means[2]),
                      T3 = c(the_treatment_means[3], the_control_means[3]),
                      T4 = c(the_treatment_means[4], the_control_means[4]),
                      row.names = c("active", "wl"))


# Now I will loop this function over a number of different sample sizes

# this is the list of sample sizes
n_per_arm <- seq(20,300, by = 20)

# this is the p-values to extract
p_vals <- lst()
p_above_criterion <- 0
for(i in 1:length(n_per_arm)){

p_vals[[i]] <- ancova_sim_func (df_means = df_means, sd_across = sd_T1,
                n_subjects = n_per_arm[i] , number_sims = n_sims, correlations_across
                = within_corr)
```

```
p_above_criterion[i] <- sum(p_vals[[i]]<0.05)/length(p_vals[[i]])

}

p_above_criterion


# now I can plot the sample sizes
df_to_plot <- data.frame(p_above_criterion = p_above_criterion, n_per_arm = n_per_arm)

df_to_plot %>%
  ggplot(aes(y = p_above_criterion, x = n_per_arm))+
  geom_point()+
  geom_hline(yintercept = 0.8, colour = "red", linetype = "dashed") +
  scale_x_continuous("sample size", breaks = n_per_arm)


# # from https://hqlo.biomedcentral.com/articles/10.1186/1477-7525-10-156
# mean_T1 <- 43.5 # the biggest trial in that list, the PEIP
# sd_T1 <- 10.4 # from that trial
# d <- -0.5 # stipulated effect size of difference
# # to derive the mean needed for the final effect size
# # cohen_d = mean_T1-mean_T4/sd_T1 # using sd_T_1 as I stipulated this to be the same
throughout
# # which converts to
# mean_T4 <- mean_T1 + d*sd_T1
#
# # to create equally spaced value, my idealised decline of values
# the_treatment_means <- seq(mean_T1, mean_T4, length.out = 4)
#
# # the control means will be the same throughout
# the_control_means <- rep(mean_T1, 4)
#
# # investigators assume a correlation between values from each time point r = 0.6
# within_corr <- 0.6
#
# # now we can simulate using faux.
# n_sims <- 50
#
# n_subj_per_arm <- 100
#
#
#####this is an example
# arts_on_pres_sim <- sim_design(
#   n = 10, # this is the per arm n
#   between = list(group = c("active", "wl")),
#   within = list(time = c("T1", "T2", "T3", "T4")),
```

```
#   mu= data.frame(T1 = c(the_treatment_means[1], the_control_means[1]),
#                  T2 = c(the_treatment_means[2], the_control_means[2]),
#                  T3 = c(the_treatment_means[3], the_control_means[3]),
#                  T4 = c(the_treatment_means[4], the_control_means[4]),
#    # sd= sd_T1,
#   row.names = c("active", "wl")),
#   sd = sd_T1,
#   dv = "score",
#   r = within_corr,
#   long = TRUE,
#   rep = 1
#
# )
#
#
# #dim(arts_on_pres_sim$data[[1]])
#
#
# # test <- aov(response~factor(drug)+Error(factor(patient)), data = df)
# #
# #
# # test <-aov(score~ time + group + time*group , data = arts_on_pres_sim$data[[1]])
# # test <-broom::tidy(test)
# # test$p.value[3]
#
# temp_results <- lst()
# temp_results_disp <- lst()
# temp_p_values <- 0
#
# for(i in 1:n_sims){
#
# temp_results[[i]] <-  aov(score~ time + group + time*group , data =
arts_on_pres_sim$data[[i]])
#
# temp_results_disp[[i]] <-broom::tidy(temp_results[[i]])
#
# temp_p_values[i] <-  temp_results_disp[[i]]$p.value[3]
#
#
# }
#
# sum(temp_p_values<0.05)/length(temp_p_values)

# create a function for all the above ------------------------------------

# I will simulate the following
# two groups, treatment and WL control.
# they start off with distributions with the same mean and sd (due to randomisation)
```

the WL control stays the same, the treatment group changes to arrive at a d = 0.4, which i
what investigators stipulated in advance
the shape of the change should not matter that much for this linear model calculation
what should matter are
a) sds which I will keep the same as in baseline
b) number of measurements, stipulated to be n = 4
c) number of people, sample size, which I will vary for the simulation

Brief description of tools:

- Patient's personal data and demographics (i.e. name, age, nationality, religion, educational level, marital status, socio-economic characteristics, medical history, mental health problems, medication, therapies).

- The Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS). This tool was developed in order to measure mental wellbeing in the general population by Tennant et al. (2007). It consists of 14 items responding using a 5-point Likert scale from 1 (never) to 5 (all the time). The overall score is calculated by summing the responses for every item without reversing none. The minimum overall score ranges from 14 to 70. Higher scores indicate increased mental wellbeing. The scale has been widely used nationally and internationally for monitoring, evaluating projects and programmes and investigating the determinants of mental wellbeing. An efficient internal consistency was proven using Cronbach's alpha score which was 0.89 for the student sample and 0.91 for the general population sample. Also, test-retest reliability at one week was high (0.83). The Greek validation showed acceptable internal consistency (Cronbach's alpha score 0.90).
  [Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, Parkinson J, Secker J, Stewart-Brown S. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes. 2007 Nov 27;5:63. doi: 10.1186/1477-7525-5-63. PMID: 18042300; PMCID: PMC2222612.

- The Generalised Anxiety Disorder Assessment (GAD-7) is provided to screen symptom severity for the four most common anxiety disorders (generalized anxiety disorder, panic disorder, social phobia and posttraumatic stress disorder). GAD-7 constructed by Spitzer et al (2006) and validated in Greek by Vogazianos et al (2022). It consists of 7 items. The GAD-7 score is computed by assigning scores of 0, 1, 2, and 3, to the response categories of 'not at all', 'several days', 'more than half the days', and 'nearly every day', respectively, and summing together the scores for the seven questions (scores range from 0 to 21). Higher levels indicate increased anxiety. There are cut-offs for severity of anxiety as: (i) score 0-4: Minimal Anxiety; (ii) score 5-9: Mild Anxiety; (iii) score 10-14: Moderate Anxiety; (iv) score greater than 15: Severe Anxiety. The internal consistency of the GAD-7 was excellent (Cronbach $\alpha = 0.92$) and test-retest reliability was also efficient (intraclass correlation = 0.83) (Spitzer et al., 2006).
  [Spitzer RL, Kroenke K, Williams JB, et al; A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006 May 22;166(10):1092-7].
  [Vogazianos P, Motrico E, Domínguez-Salas S, Christoforou A, Hadjigeorgiou E. Validation of the generalized anxiety disorder screener (GAD-7) in Cypriot pregnant

and postpartum women. BMC Pregnancy Childbirth. 2022 Nov 15;22(1):841. doi: 10.1186/s12884-022-05127-7].

- The Patient Health Questionnaire-9 (PHQ-9) is a questionnaire which measures depression and grade severity of symptoms in general medical and mental health settings using nine DSM-5 criteria for major depression within the last two weeks. PHQ-9 constructed by Kroenke et al. (2001) and validated in Greek by Hyphantis et al. (2011) is using a four-point Likert-type scale. After summing all responses (0=not at all, 3=nearly every day), scores range from 0 to 27, with higher levels indicating increased symptom severity (0–4 no to minimal; 5–9 mild; 10–14 moderate; 15–19 moderately severe; 20–27 severe). The internal reliability of the PHQ-9 was excellent, with a Cronbach's α of 0.89 in the PHQ Primary Care Study (Kroenke et al., 2001) and 0.82 in Greek validation (Hyphantis et al., 2011).
  [Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–613].
  [Hyphantis T, Kotsis K, Voulgari PV, Tsifetaki N, Creed F, Drosos AA. Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders. Arthritis Care Res (Hoboken). 2011 Sep;63(9):1313-21. doi: 10.1002/acr.20505. PMID: 21618450].

- The Revised UCLA Loneliness Scale-20 (UCLA-20) is a questionnaire which measures level of psychological loneliness experienced by a person. UCLA-20 constructed by Russel et al. (1980) and validated in Greek by Anderson & Malikiosi-Loizos (1992). The internal reliability of the UCLA-20 was excellent, with a Cronbach's α of 0.94 (Russel et al., 1980) and 0.89 in Greek validation (Anderson & Malikiosi-Loizos, 1992). This long scale was shortened to 3 questions (UCLA 3-item Loneliness Scale) in order to be used in large surveys (Hughes et al., 2004). It is using a 3-point Likert-type scale (hard ever, some of the time, often). The scores can be added together to give a range of scores from 3 to 9. Higher scores mean most lonely (worse outcome).
  [Russel D, Peplau LA, Cutrona CE. (1980). The Revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence, Journal of Personality and Social Psychology, 39, 472-480.].
  [Anderson LR, Malikiosi-Loizos M. (1992). Reliability data for a Greek translation of the revised UCLA Loneliness Scale: Comparisons with data from the USA. Psychological Reports, 71, 665-666.
  Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. Res Aging. 2004;26(6):655-672. doi: 10.1177/0164027504268574].

- The Strengths and Difficulties Questionnaire (SDQ) is a brief behavioural screening tool for children and adolescents aged 2-17 years. PHQ-9 constructed by Goodman (2001) and validated in Greek by Gomez et al. (2021) is using a 3-point Likert-type scale (“not true,” “somewhat true”, “certainly true”). SDQ consists of 25 positive and negative items. These 25 items are divided between 5 scales: (i) emotional symptoms (5 items), (ii) conduct problems (5 items), (iii) hyperactivity/inattention (5 items), (iv) peer relationship problems (5 items), (v) prosocial behaviour (5 items). “Somewhat

true” is always scored as 1 but the scoring of “not true” and “certainly true” varies and score 0 or 2. The total difficulties score is generated by summing the cores form all the scales except the prosocial scale. The final score can be ranged from 0 to 40. Higher scores mean most difficulties (worse outcome). Cronbach’s alpha of the original study for all scales ranged from 0.35 to 0.64.

[Goodman R (2001) Psychometric properties of the strengths and difficulties questionnaire. J Am Acad Child & Adolesc Psychiatr 40(11):1337–45].

[Gomez R, Motti-Stefanidi F, Jordan S, Stavropoulos V. Greek Validation of the Factor Structure and Longitudinal Measurement Invariance of the Strengths and Difficulties Questionnaire-Self Report (SDQ-SR): Exploratory Structural Equation Modelling. Child Psychiatry Hum Dev. 2021 Oct;52(5):880-890. doi: 10.1007/s10578-020-01065-7].