Statistical Analysis Plan [SAP]- version 2

Version 1: July, 2019

Version 2: November 30th, 2020

IBBIS

Integrated mental health care and vocational rehabilitation to improve return to work rates for people on sick leave because of common mental disorders (IBBIS)

IBBIS Statistical Analysis Plan [SAP] - version 2.

Authors: Andreas Hoff[1] (corr. author), Jonas Fisker[1], Rie Mandrup Poulsen[1], Carsten Hjorthøj[1,2], and Lene Falgaard Eplov[1]

[1] Copenhagen Research Center for Mental Health [CORE], Mental Health Services Capital Region of Denmark, University of Copenhagen, Gentofte Hospitalsvej 1, Opgang 15-4, DK-2900
[2] University of Copenhagen, Department of Public Health, Section of Epidemiology

Correspondence: andreas.hoff@regionh.dk

Abbreviations:

DREAM: Den Registerbaserede Evaluering Af Marginaliseringsomfanget (Danish)
DPCR: Danish Psychiatric Central Register
IBBIS: Integreret Beskæftigelses- og Behandlingsindsats til Sygedagpengemodtagere (Danish)
LPR: National Patient Register [In Danish: Landspatientregisteret]
RCT: Randomized controlled trial
RTW: Return to work
SDA: Study Design Article

ADMINISTRATIVE INFORMATION

The structure of this SAP is largely aligned with the recommendations by Gamble et. al[1].

1.1 Title and trial registration

This SAP is the detailed statistical analysis plan, expanding the scientific IBBIS protocol of the two IBBIS randomized clinical trials (ClinicalTrials.gov Identifiers: NCT02872051 (RCT1[1]) and NCT02885519 (RCT2[2])):

RCT1: "Integrated Mental Health Care and Vocational Rehabilitation to Individuals on Sick Leave Due to Anxiety and Depression (IBBIS)"

and

RCT2: "Integrated Mental Health Care and Vocational Rehabilitation to Individuals on Sick Leave Due to Stress Disorders (IBBIS)".

Due to extensive methodological similarities between these studies this SAP applies to both, unless differences are mentioned explicitly.

1.2 SAP version

This is the second version of the SAP.

[1] https://clinicaltrials.gov/ct2/show/NCT02872051

[2] https://clinicaltrials.gov/ct2/show/NCT02885519

Differences between version 1 and 2 are explicitly stated through a .docx-version of this newest version, where all changes are tracked, using the *Track changes* function in Microsoft Office Word. This file will be readily mailed through the corresponding author.

In brief, main changes revolves around the 24-month explorative outcomes: after analyses of 6- and 12-month outcomes we realized that results across work outcomes were heterogenous to a higher extent than expected. E.g., while the SAU group tended to have fastest RTW at 6-month follow-up, from explorative proportion over time-curves we realized that they might also tend to have a higher degree of recurrent sick-leave. Despite this, the SAU and INT groups still showed approx. the same number of weeks in work (when stability is disregarded) at 12-month follow-up. Followingly we speculate that the SAU group experiences faster RTW, but more recurrent sick leave. Therefore, we suggest that number of weeks in stable work is a better outcome, since this number is only high if RTW happens early, and if it is stable, and not disrupted by sick leave recurrence. We do though not know what stability threshold we should apply, and to *explore* this, we defined, prior to 24-month analyses, three different outcomes, with three different thresholds, see section 6.4. We plan these as sensitivity analyses.

1.3 Protocol version

Previous to publication of this SAP, plans have been described in both the protocol (published on clinicaltrials.org in the links provided), as well as in two study design articles (SDA), corresponding to the two RCTs[2,3].

2 Introduction: background, rationale and objectives

Described thoroughly in the SDAs[2,3]. Furthermore, the protocol was published[3] on the official webpage of the organization (Mental Health Services, Capital Region of Denmark).

3 Study methods

3.1 Trial design

See SDAs[2,3].

3.2 Randomization

From SDA[2]:

"The allocation ratio between the three arms is 1:1:1. A centralized randomization will take place according to a web-based computer-generated allocation sequence with varying block sizes kept unknown to the assessors. Odense Patient data Explorative Network (OPEN) is responsible for the randomization, administrative personnel in the IBBIS team perform the online randomization and the IBBIS team leader assign the participant to interventions and professionals.

We expect that service delivery can vary from municipality to municipality and the process of gaining a new job from unemployment will take longer time than returning to an existing job. Previous research has shown that diagnosis is a possible predictor of return to work[4]. Thus, the randomization is stratified according to 1) municipality 2) employment status (on sick leave from work vs. on sick leave from unemployment) 3) diagnosis […]"

In RCT 1 diagnosis stratification is depression versus anxiety as primary diagnosis, and in RCT 2 diagnosis stratification is burnout vs. distress vs. adjustment disorder as primary diagnosis.

3.3 Sample size

Replicated from protocol follows:

[3] https://www.psykiatri-regionh.dk/Kvalitet-og-udvikling/udvikling/ibbis/Sider/IBBIS-forskning.aspx

The sample size is based on a sample size calculation, using the 'Power and Sample Size' calculation programme[4].

Type I error (α) risk

In each of the two RCTs we wish to conduct multiple comparisons (between 3 groups), and hence significance level must be as follows, due to Bonferroni correction:

$$\alpha = \frac{0{,}05}{3} = \frac{1}{60} = 0{,}0167$$

Type II error (β) risk

The organizational constellation of the interventions has not yet been trialled, and thus the desired power shall be set to:

$$\beta = 0{,}9$$

If it turns out that we cannot include enough participants, the power could be set to: $\beta = 0{,}8$

Hazard ratio (R)

The mean difference in time for return to work will be calculated as a hazard ratio. We estimate that as sufficient HR is

$$R = 1{,}5$$

since just 50 % faster return to work time in the intervention groups will convey a relevant economic benefit, due to the hence smaller loss of productivity.

Mean time to return to work (M_1)

Number of days from baseline to return to work is conservatively estimated to be 210 days, after an observed range from 104 to 210 days, in the control groups in three Dutch RCTs[5–7], which were comparable to the control groups in the IBBIS RCTs.
Hence,

$$M_1 = 210$$

Inclusion time period (A)

We will include participants through 24 months,

$$A = 730[days]$$

Ratio between groups (m)

Ratio is 1:1:1, and hence m = 1

Follow-up time (F)

We will follow participants up for 365 days, in which they will contribute with risk time in the survival analysis, hence F = 365

Result

In each group, due to the above-mentioned variables, we need

198 *participants per group, and with three groups that yields a need for, with power = 0.9,*

[4] http://ps-power-and-sample-size-calculation.software.informer.com

$$N = 198 \frac{\text{participants}}{\text{group}} \times 3 \frac{\text{groups}}{\text{trial}} = 594 \text{ participants}$$

If, in case of insufficient inclusion possibilities, power could be lowered to 0.8. In such case we would need the following number:

$$N = 153 \frac{\text{participants}}{\text{group}} \times 3 \frac{\text{groups}}{\text{trial}} = 459 \text{ participants}$$

3.4 Statistical interim analyses and stopping guidance

No interim analysis will be performed. We planned no stopping guidance.

3.5 Timing of final analysis

The researchers who will perform the 6- and 12-month outcome analyses (AH and JF) will be blinded from intervention group allocation, until the primary outcome and all 12-month follow-up outcome main analyses are completed. The true randomization group allocation is concealed, with values X, Y and Z reflecting group allocation in the blinded dataset. The randomization allocation variable conversion formula is until unblinding only know and hidden by an administrative co-worker, who will not perform or assist any analysis.

At the time of publication of SAP version 1, baseline distributional analyses, and unadjusted estimated marginal means-analyses of self-reported numerical secondary outcomes at 6-month follow-up (and only these) have been calculated *blinded*, but will not be published, since this was not complying with the SDAs, nor any SAP version.

All 24-month follow-up analyses will be conducted unblinded.

4 Statistical principles

4.1 Confidence intervals and p-values

For all outcomes, the three randomization groups are pairwise compared. Due to these multiple comparisons, we will calculate 98,3% confidence intervals, according to Bonferroni correction of desired α-level of 0,05 in testing of 3 hypotheses:

α-level: $0{,}05 \times \frac{1}{3} \cong 0{,}0167 \quad \Rightarrow$ Confidence Interval: $1 - 0{,}0167 \cong 98{,}33\%$

4.2 Analysis populations

All analyses are performed as *intention-to-treat*, unless otherwise stated.

5 Trial population

5.1 Withdrawal and follow-up

Due to legislative circumstances participants can withdraw consent, and followingly all person sensitive data on these subjects will be deleted, yet participant ID number (not CPR number, but generated for this research project) and randomization result will be stored. In sensitivity analyses these ID numbers will be included, as described in "handling of missing data".

5.2 Baseline patient characteristics

The following will be reported per RCT, per randomization allocation group. For all mean values of numeric variables, standard deviations will be reported.

Total number included in RCT and number in each randomization group
Age (mean, year)
Gender (%)

Bech Depr. Inventory (mean)
Bech Anxiety Inventory (mean)
Work and Social adjustment Scale (mean)
Perceived Stress Scale (mean)
Employment status (%, employed vs. unemployed)
Primary diagnosis (%)
Municipality (%)
Sick leave duration at randomization (mean, days)
Educational level (%, short, moderate, long)

Distributional balances of these covariates (except educational level, since this is only added in SAP v. 2, after primary baseline analyses) will be calculated using one way-ANOVA for numerical data and X^2 for categorical data, and analyses with p≤0,05 will define *imbalanced baseline covariates*.

6 ANALYSIS

The first subsections of this section 6, describes general strategies applying to all analyses unless otherwise specifically stated. Subsection 6.8 contains the separate analysis strategies per outcome in 6.8.x.

6.1 COVARIATE ADJUSTMENT IN GENERAL

Analyses will be adjusted for the three stratification variables, and no other, complying with RCT analysis guidelines from European Medicines Agency[5].

6.2 SENSITIVITY ANALYSES IN GENERAL

As sensitivity analyses, all outcome analyses will be performed adjusted for any unbalanced baseline covariates, as defined in 5.2, *Baseline patient characteristics*.

Results of sensitivity analyses are only interpreted as supplements to the main analysis and will not substitute main results.

6.2.1 SENSITIVITY ANALYSES FOR QUESTIONNAIRE BASED, SELF-REPORTED DATA OUTCOME

As sensitivity analyses, self-reported data outcomes (questionnaire-based) will be calculated with all missing outcome data replaced with a value equalling the mean of the outcome variable ± 2 standard deviations, and participants who withdraw themselves from the study will be included in these analyses with all their data handled as missing.

6.2.2 SENSITIVITY ANALYSES FOR REGISTER DATA BASED OUTCOMES

For register data-based outcomes, sensitivity analyses will be performed including the participants who withdraw themselves from the study, included in these analyses with all their outcomes handled as either the worst possible (never returning to work) vs best possible (returning to work as soon as possible).

Furthermore, all outcomes of number of weeks in stable return to work (outcome number 9, 10, 11 and 12), are sensitivy analyses, exploring the robustness of number of weeks in work (stability disregarded), which is oucome number 13, pre-planned before study commencement.

6.3 SUBGROUP ANALYSES IN GENERAL

All outcomes will be analysed with respect to the following subgroups:

a) per primary diagnosis (in RCT1 anxiety vs. depression; in RCT2 per distress, adjustment disorder, and burnout);
b) per employment status group at baseline (vacant vs. employed);

[5] https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf

c) per IBBIS Team (two teams, Team North and Team Byen)

Furthermore,

d) divided in two groups by relative time of randomization: first and last temporal half of randomized participants.

Finally,

e) we will test for interaction between diagnostic group and treatment allocation group/arm.

No outcomes have other subgroup analyses planned.

6.4 OUTCOME DEFINITIONS

The outcomes are reported as in the study design articles (except for selected outcomes, see alterations to SAP version 1 in appendix). The numbers 1 through 64 denotes the outcome numbers for reference purposes for this SAP section.

PRIMARY AND SECONDARY OUTCOMES and outcome numbering						
Outcome class	**Data source**	**Outcome**		**6-month follow-up**	**12-month follow-up**	**24-month follow-up**
Primary	DREAM database	Time from baseline to RTW			1	
Secondary	DREAM database	Proportion in ordinary work			2	
	DREAM database	Time from baseline to RTW		3		8
	Questionnaire	Depressive symptoms measured by Beck Depression Inventory (BDI)[8]		4		
	Questionnaire	Anxiety symptoms measured by Beck Anxiety Inventory (BAI)[9]		5		
	Questionnaire	Stress symptoms measured by Cohen perceived stress scale (PSS)[10]		6		
	Questionnaire	Social and work related function measured by WSAS[11]		7		

PREDEFINED EXPLORATORY OUTCOMES and outcome numbering					
Outcome class	**Data source**	**Outcome**	**Follow-up**		
			6-month	**12-month**	**24-month**
Pre-defined exploratory outcomes	DREAM database	Weeks in stable work (≥ 4 weeks) from baseline to current follow-up		9	10
		Weeks in stable work (≥8 weeks) from baseline to current follow-up			11
		Weeks in stable work (≥12 weeks) from baseline to current follow-up			12
		Weeks in work from baseline to follow-up			13
		Proportion in ordinary work			14
		Number of recurrent sick leaves			15
	Questionnaires	Symptoms of Distress, anxiety, depression and somatization by Four-Dimensional Symptom Questionnaire (4DSQ)[12]	16	17	18
		Depressive symptoms measured by Beck Depression Inventory (BDI)[9]		19	20
		Anxiety symptoms measured by Beck Anxiety Inventory (BAI)[9]		21	22
		Stress-symptoms measured by Cohen perceived stress scale (PSS)[10]		23	24
		Social and work related function measured by WSAS[11]		25	26
		Burn-out symptoms measured by Karolinska Exhaustion Scale (KES)[13]	27	28	29
		Health-related quality of life measured by EQ-5D-5L[14]	30	31	32
		General Quality of life scale measured by Flanagan's' QOLS[15]	33	34	35
		Self-efficacy concerning symptoms measured by IPQ subscale on personal control[16]	36	37	38
		Return to work self-efficacy measured by RTW-SE[17]	39	40	41

		General self-efficacy measured by General Self-efficacy scale (GSS)[18]	42	43	44
		Client satisfaction with treatment measure measured by CSQ-8[19]	45		
		Presenteeism measured by Stanford Presenteeism Scale (SPS)[20]	46	47	48

HARM MEASURES and outcome numbering					
Outcome class	**Data source**	**Outcome**	**Follow-up**		
			6-month	**12-month**	**24-month**
Harm measures	LPR	Admission to somatic hospital-based (in-patient) health care at least once		49	50
	LPR	Contact with hospital-based out-patient mental health care, at least once		51	52
		Admission to hospital-based in-patient mental health care, at least once		53	54
		Contact with emergency mental health care, at least once		55	56
	LPR	Probable self-harm, at least once		57	58
	LPR	Suicide		61	62
	LPR	Death		63	64

6.5 HYPOTHESES AND NULL-HYPOTHESES

Stated below are the generic versions of all three hypotheses (H_1) and all three null-hypotheses (H_0) that apply to each outcome.

Regarding what is a "better outcome" is listed in section 6.6, defined for each outcome measure, respectively.

6.5.1 HYPOTHESES

This superiority trial hypothesizes that, for all outcomes,

H_{1A} Group 3, "Integrated IBBIS mental health care treatment and vocational rehabilitation"

conveys better outcomes than

Group 2, "IBBIS mental health care (and standard VR)", and

H_{1B} Group 2, "IBBIS mental health care (and standard VR)", and

conveys better outcomes than

Group 1, "Control group, treatment as usual (standard MHC and standard VR)"

and followingly

H_{1C} Group 3, "Integrated IBBIS mental health care treatment and vocational rehabilitation"

conveys better outcomes than

Group 1, "Control group, treatment as usual (standard MHC and standard VR)".

and followingly

Group 3 conveys better outcomes than Group 1,

since if

Group 3 outcome > Group 2 outcome > Group 1 outcome

then

Group 3 outcome > Group 1 outcome.

The groups are thoroughly described in the IBBIS Protocol and the SDAs.

6.5.2 Null-hypotheses

The corresponding null-hypotheses are

H_{oA} Group 3, "Integrated IBBIS mental health care treatment and vocational rehabilitation"

does not convey better outcomes than

Group 2, "IBBIS mental health care (and standard VR)", and

H_{oB} Group 2, "IBBIS mental health care (and standard VR)",

does not convey better outcomes than

Group 1, "Control group, treatment as usual (standard MHC and standard VR)".

and followingly

H_{oC} Group 3 *does not* convey better outcomes than Group 1.

6.6 Outcome benefit direction

Referring to the hypothesis section, this table describes whether a "better outcome" is a higher or lower score on the numeric outcome variables.

Outcome	Is "better outcome" defined by lower or higher numbers?	
Time from baseline to RTW	Lower	
Proportion in stable work		Higher
Time from baseline to RTW	Lower	
Weeks in work (all variations of stability definitions)		Higher
Number of recurrent sick leaves	Lower[6]	
Depressive symptoms measured by Beck Depression Inventory (BDI)[8]	Lower	
Anxiety symptoms measured by Beck Anxiety Inventory (BAI)[9]	Lower	
Stress symptoms measured by Cohen perceived stress scale (PSS)[10]	Lower	
Social and work related function measured by WSAS[11]	Lower	
Symptoms of Distress, anxiety, depression and somatization by Four-Dimensional Symptom Questionnaire (4DSQ)[12]	Lower	
Burn-out symptoms measured by Karolinska Exhaustion Scale (KES)[13]	Lower	
Health-related quality of life measured by EQ-5D-5L[14]		Higher[21]
General Quality of life scale measured by Flanagan's' QOLS[15]		Higher
Self-efficacy concerning symptoms measured by IPQ subscale on personal control[16]		Higher[22]
Return to work self-efficacy measured by RTW-SE[17]		Higher
General self-efficacy measured by General Self-efficacy scale (GSS)[18]		Higher
Client satisfaction with treatment measure measured by CSQ-8[19]		Higher
Presenteeism measured by Stanford Presenteeism Scale (SPS)[20]		Higher[23]

[6] A low number of recurrent sick leave is a positive outcome only if duration of index sick leave is ideally balanced between compared groups.

6.7 Missing data in general

In general, proportion of missing data will be reported per intervention group for all outcomes.

6.7.1 Handling of missing data in registers

For RTW-outcomes (outcomes based on the DREAM register) we expected *no missing data*, due to the nature of the Dream Register, prior to study inception. Missing data should only be in case of a participant moving out of Denmark. We considered these events to be so rare in our data that we would handle such missing data as *missing completely at random*. Thus, no imputation or other correction was considered necessary. We will report proportion of data missing.

We will report number of censored participants per treatment group.

At the of this updated version 2a of the SAP, we have realized that some data were missing due to DREAM database errors, against expectation. We included the cases with missing data in sensitivity analyses to explore the potential impact of the missingness.

6.7.2 Handling of missing data in questionnaire based, self-reported data outcome

For questionnaire-based outcomes, missing data will be handled as *missing at random*. To handle this, 100 multiple imputations will be performed, using following variables: stratification variables: diagnosis, municipality, employment status; age; gender; time to stable RTW; psychometric variables at baseline and all follow-up at outcome time: BDI, BAI, WSAS and PSS.

6.8 Analysis methods per outcome group

This section describes the details of the statistical analyses. Since several outcomes require exact same analysis methods, outcomes are grouped for the following description

6.8.1 Time to return to work-outcomes (outcomes #1, #3 and #8)

This section describes primary outcome *Time from baseline to RTW* at 12-month follow-up (1), and the secondary outcomes *Time from baseline to RTW* at 6- (outcome 3) and 24-month follow-up (outcome 8). The 24-month follow-up outcome will be calculated no earlier than June 2020. The other two, readily after the publication of this SAP, but before unblinding of analysists.

6.8.1.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

Time from baseline to RTW is defined af the number of weeks from randomization date, to stable return to work. Stable return to work is defined as 4 weeks consecutively in work, i.e. with no sick leave benefit those 4 weeks in the Dream register, and a so-called "branch code" in at least some of this 4 week period (benefit codes are week-based, branch codes are month based, and hence a period of 4 weeks may represent only one month, or overlap a two month period; in the latter case, return to work will be attained if at least one of these registrations contains a branch code; a branch code means that the individual received salary from an employer in this period). Time of event is first day of the four weeks.

These events will define censoring: 1) moving out of the country, 2) death, 3) public retirement pension (Da.: "Folkepension"), and 4) voluntary early retirement scheme (Danish: "Efterløn").

At randomization all participants are, according to inclusion criteria, on sick-leave from employment or vacancy. Some participants might be on sick-leave from an employment in a *flexjob*[7], and hence receiving *flexjob benefit* during employment. This benefit is changed to *flexjob sick-leave benefit* similar to regular sick leave benefit for participants not granted flexjob benefit prior to randomization. In these cases (of participants granted flexjob benefit prior to randomization) RTW is defined as either not receiving flexjob sick-leave benefit for four consecutive weeks, along with a registered branch code as above mentioned (or alternatively not receiving flexjob benefit, but an ordinary salary indicated by a branch code during those four weeks).

[7] "Flexjob" is one of the Danish benefit schemes; it is a subsidy granted those with a chronic reduced work capacity

For participants, who at baseline are on sick-leave from vacancy (but not receiving flexjob benefit), RTW can both be defined as above mentioned (four consecutive weeks without sick leave benefits and a branch code during those four weeks) or receiving flexjob benefit for four consecutive weeks and a branch code during those four weeks.

6.8.1.2 Specific analysis method and result presentation

Comparisons of RTW time will be calculated as hazard rate ratios between groups (and corresponding 98,3%CI), using a Cox-regression model.

Kaplan-Meier curves will be presented to illustrate the cumulative incidence of first stable return to work event in each trial-arm.

6.8.1.3 Covariate adjustment

Only for stratification variables, see *6.1 "Covariate adjustment in general"*.

6.8.1.4 Statistical method assumption control

Assumptions for the proportional hazards (~Cox-) regression model are proportional hazards; this will be controlled performing af Schoenfeld (SF) test for residuals and visual inspection.

6.8.1.5 Alternative analysis method in case of assumption fail

If the SF test is positive (p<0,05), the analysis will we performed adjusted for the interaction between time and treatment group allocation. If SF test hereafter is still positive, the analysis will instead be adjusted for the interaction between *quadratic* time (time^2) and treatment group allocation. If SF test hereafter is still positive, the analysis will instead be adjusted for the interaction between *log*(time) and treatment group allocation. If SF test hereafter is still positive, the analysis with the highest p-value will be reported.

6.8.1.6 Sensitivity analyses

See *"6.2.2 Sensitivity analyses for register data based outcomes"*.

6.8.1.7 Reporting and statistical methods to handle missing data

On RWT-outcomes we expect *no missing data*, due to the nature of the Dream Register. Missing data will only be in case of a participant dying or moving out of Denmark. We consider these events to be so rare in our data that we will handle such missing data as *missing completely at random*. Thus, no imputation or other correction is necessary. We will report proportion of data missing.

We will report number of censored participants per treatment group.

6.8.2 Proportion in ordinary work at 12-month follow-up (secondary outcome) and 24-month follow-up (exploratory outcome) (outcome #2 and #14)

6.8.2.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

This outcome is calculated as the share of the treatment allocation group that on the time of follow-up was in stable RTW (≥ 4 weeks). Stable RTW if defined exactly as in the primary outcome, see 6.8.1.1.

6.8.2.2 Specific analysis method and result presentation

Pairwise odds ratios will be calculated using logistic regression.

In addition to the presentation of odds ratios for tests at 12-month follow-up and 24-month follow-up, graphs are presented with the proportions in stable work at each week (week 1-52 for 12-month follow-up and week 1-104 for 24-month follow-up) for each of the three trial-arms. No statistical test will be performed for differences at week 1-51 or week 53-103. These curves are explorative, descriptive analyses.

6.8.2.3 Covariate adjustment

Only for stratification variables, see *6.1 "Covariate adjustment in general"*.

6.8.2.4 Statistical method assumption control

The assumptions of the model are assumed to be acceptable, due to large sample, binary outcome, categorical independent variable.

6.8.2.5 Alternative analysis method in case of assumption fail

No alternative methods are planned, since assumptions are assumed to hold.

6.8.2.6 Sensitivity analyses

See *"6.2.2 Sensitivity analyses for register data based outcomes"*.

6.8.2.7 Reporting and statistical methods to handle missing data

Same as 6.8.1.7.

6.8.3 All self-reported, numerical outcomes, at 6-, 12-, and 24-month follow-up at (secondary outcomes ##4-7 and predefined exploratory outcomes ##16-48)

6.8.3.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

All outcomes are calculated as the sum of scores on the respective measurement scales.

All 6-month follow-up outcome analyses are calculating using baseline and 6-month follow-up observations.

All 12-month follow-up outcome analyses are calculating using baseline and 6- and 12-month follow-up observations.

All 24-month follow-up outcome analyses are calculating using baseline and 6-, 12-, and 24-month follow-up observations.

6.8.3.2 Specific analysis method and result presentation

Linear mixed-effects model with unstructured covariance. Results will be presented in pairwise group differences between outcomes, from the estimated marginal means from the model, and the confidence intervals of these differences.

6.8.3.3 Covariate adjustment

Only for stratification variables, see *6.1 "Covariate adjustment in general"*.

6.8.3.4 Statistical method assumption control

Assumption: normal distribution of scores. Control: Visual inspection by plotting the score residuals.

Assumption: normal distribution of individuals' score differences between baseline and follow-up. Control: Visual inspection by plotting the score difference residuals.

Assumption: Equality and homogeneity of variance. Control: Breusch Pagan test and Bartlett's test are used to identify violations of these assumptions.

6.8.3.5 Alternative analysis method in case of assumption fail

In case of positive tests or visual inspections a robust variance estimator is used to correct standard errors.

6.8.3.6 Sensitivity analyses

See "*6.2.1 Sensitivity analyses for questionnaire based, self-reported data outcome"*.

6.8.3.7 Reporting and statistical methods to handle missing data

Proportion and amount of missing data per outcome variable per follow-up event per treatment group will be reported.

To handle missing data, 100 multiple imputations will be performed, using following variables: stratification variables: diagnosis, municipality, employment status; age; gender; time to stable RTW; psychometric variables at baseline and all follow-up at outcome time: BDI, BAI, WSAS and PSS.

6.8.4 Weeks of work from baseline to 12- and 24-month follow-up (outcomes ##10-14)

6.8.4.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

From baseline to follow-up, the number of weeks in work per participant is calculated. A week is noted as being in work, if no sick leave benefit has been received, *and* if a branch code is registered in the month of that week (branch codes are registered on monthly basis, if an individual has received salary from an ordinary job during that month).

For participants receiving flexible job benefit prior to randomization, and participants on sick leave from vacancy, the same principles apply, as described in 6.8.1.1, in the section "*Time to return to work-outcomes (outcomes #1, #3 and #8)".*

At 24-month follow-up, this analysis is conducted with three variations each applying a different definition of return to work stability as sensitivity analyses. Whereas the first analysis uses the definition of stability from the primary outcome (minimum four weeks see section 6.8.1.1), these sensitivity analyses are conducted with a more conservative approach where stable return to work is defined as minimum 4, 8 and 12 weeks in work respectively.

6.8.4.2 Specific analysis method and result presentation

Severely skewed data is expected for this outcome, why a robust Poisson regression model will be used to test the differences between groups.

6.8.4.3 Covariate adjustment

Only for stratification variables, see *6.1 "Covariate adjustment in general"*.

6.8.4.4 Statistical method assumption control

Assumption: Poisson distribution. Control: X^2 goodness-of-fit test.

6.8.4.5 Alternative analysis method in case of assumption fail

If X^2 goodness-of-fit test is significant, negative binomial regression model will be used instead. If X^2 goodness-of-fit test is significant for this distribution, zero inflated poisson regression will be used.

6.8.4.6 Sensitivity analyses

See *"6.2.2 Sensitivity analyses for register data based outcomes"*.

6.8.4.7 Reporting and statistical methods to handle missing data

See 6.8.1.7

6.8.5 At 24 months: Number of recurrent sick leaves at 24-month follow-up (outcome #15)

6.8.5.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

For each group, the number of persons who have experienced the event 'stable return to work' and followingly experienced the event 'recurring sick leave' is calculated. Recurring sick leave is defined as the first sick leave period starting with the fist week of receiving sickness benefit after a period of stable return to work as defined in paragraph 6.8.1.1.

6.8.5.2 Specific analysis method and result presentation

Only descriptive statistics will be performed for this outcome and no differences between groups will be tested. For each group, the number of persons who have experienced stable return to work and the number of persons who have experienced recurrent sick leave is presented.

6.8.6 Harm measures at 12-, and 24-month follow-up (outcome #49-64)

6.8.6.1 Calculation of the outcome: specific measurement and units (and transformation, where applicable)

For each group, the number of persons who have experienced the harmful event is calculated.

6.8.6.2 Specific analysis method and result presentation

Only descriptive statistics will be performed for this outcome and no differences between groups will be tested.

7 REFERENCES

1. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*. 2017;318(23):2337. doi:10.1001/jama.2017.18556

2. Poulsen R, Fisker J, Hoff A, Hjorthøj C, Eplov LF. Integrated mental health care and vocational rehabilitation to improve return to work rates for people on sick leave because of depression and anxiety (the Danish IBBIS trial): study protocol for a randomized controlled trial. *Trials*. 2017;18:578. doi:10.1186/s13063-017-2272-1

3. Poulsen R, Fisker J, Hoff A, Hjorthøj C, Eplov LF. Integrated mental health care and vocational rehabilitation to improve return to work rates for people on sick leave because of exhaustion disorder, adjustment disorder, and distress (the Danish IBBIS trial): Study protocol for a randomized controlled tri. *Trials*. 2017;18(1). doi:10.1186/s13063-017-2273-0

4. Nielsen MBD, Madsen IEH, Bültmann U, Christensen U, Diderichsen F, Rugulies R. Predictors of return to work in employees sick-listed with mental health problems: Findings from a longitudinal study. *Eur J Public Health*. 2011;21:806-811. doi:10.1093/eurpub/ckq171

5. Oostrom SH Van, Mechelen W Van, Terluin B, Vet HCW De, Knol DL, Anema JR. A workplace intervention for sick-listed employees with distress: results of a randomised controlled trial. 2010. doi:10.1136/oem.2009.050849

6. Vlasveld MC, van der Feltz-Cornelis CM, Ader HJ, et al. Collaborative care for sick-listed workers with major depressive disorder: a randomised controlled trial from the Netherlands Depression Initiative aimed at return to work and depressive symptoms. *Occup Environ Med*. 2013;70(4):223-230. doi:10.1136/oemed-2012-100793

7. Lagerveld SE, Blonk RWB, Brenninkmeijer V, Wijngaards-de Meij L, Schaufeli WB. Work-focused treatment of common mental disorders and return to work: A comparative outcome study. *J Occup Health Psychol*. 2012;17(2):220-234. doi:10.1037/a0027049

8. Beck AT, Brown G, Epstein N, Steer RA. An Inventory for Measuring Clinical Anxiety: Psychometric Properties. *J Consult Clin Psychol*. 1988;56(6):893-897. doi:10.1037/0022-006X.56.6.893

9. Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clin Psychol Rev*. 1988;8(1):77-100. doi:10.1016/0272-7358(88)90050-5

10. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav*. 1983;24(4):385-396. doi:10.2307/2136404

11. Mundt JC, Marks IM, Shear MK, Greist JH. The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *Br J Psychiatry*. 2002;180(MAY):461-464. doi:10.1192/bjp.180.5.461

12. Terluin B, van Marwijk HWJ, Adèr HJ, et al. The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry*. 2006;6:34. doi:10.1186/1471-244X-6-34

13. Saboonchi F, Perski A, Grossi G. Validation of Karolinska Exhaustion Scale: Psychometric properties of a measure of exhaustion syndrome. *Scand J Caring Sci*. 2013;27(4):1010-1017. doi:10.1111/j.1471-6712.2012.01089.x

14. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ*. 2004;13(9):873-884. doi:10.1002/hec.866

15. Burckhardt CS, Anderson KL. The Quality of Life Scale (QOLS): reliability, validity, and utilization. *Health Qual Life Outcomes*. 2003;1(1):60. doi:10.1186/1477-7525-1-60

16. Moss-Morris R, Weinman J, Petrie KJ, Horne R, Cameron LD, Buick D. The revised Illness Perception Questionnaire (IPQ-R). *Psychol Health*. 2002;17(1):1-16. doi:10.1080/08870440290001494

17. Lagerveld SE, Blonk RWB, Brenninkmeijer V, Schaufeli WB. Return to work among employees with mental health problems: Development and validation of a self-efficacy questionnaire. *Work Stress*.

2010;24(4):359-375. doi:10.1080/02678373.2010.532644

18. Schwarzer R, Jerusalem M. Generalized Self-Efficacy scale. In: Wright S, Johnston M, eds. *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*. Windsor, England; 1995:35-37.

19. Attkisson C, Zwick R. The Client Satisfaction Questionnaire: Psychometric properties and correlations with service utilization and psychotherapy outcome. *Eval Program Plann*. 1982;5:233-237.

20. Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: health status and employee productivity. *J Occup Environ Med*. 2002;44(1):14-20.

21. EuroQol Group. EuroQol–a new facility for the measurement of healthrelated quality of life. The EuroQol Group. *Health Policy (New York)*. 1990;16(3):199-208. doi:10.1016/0168-8510(90)90421-9

22. Weinman J, Petrie KJ, Moss-Morris R, Horne R. The illness perception questionnaire: A new method for assessing the cognitive representation of illness. *Psychol Heal*. 1996;11(3):431-445. doi:10.1080/08870449608400270

23. Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: health status and employee productivity. *J Occup Environ Med*. 2002;44(1):14-20. http://www.ncbi.nlm.nih.gov/pubmed/11802460.