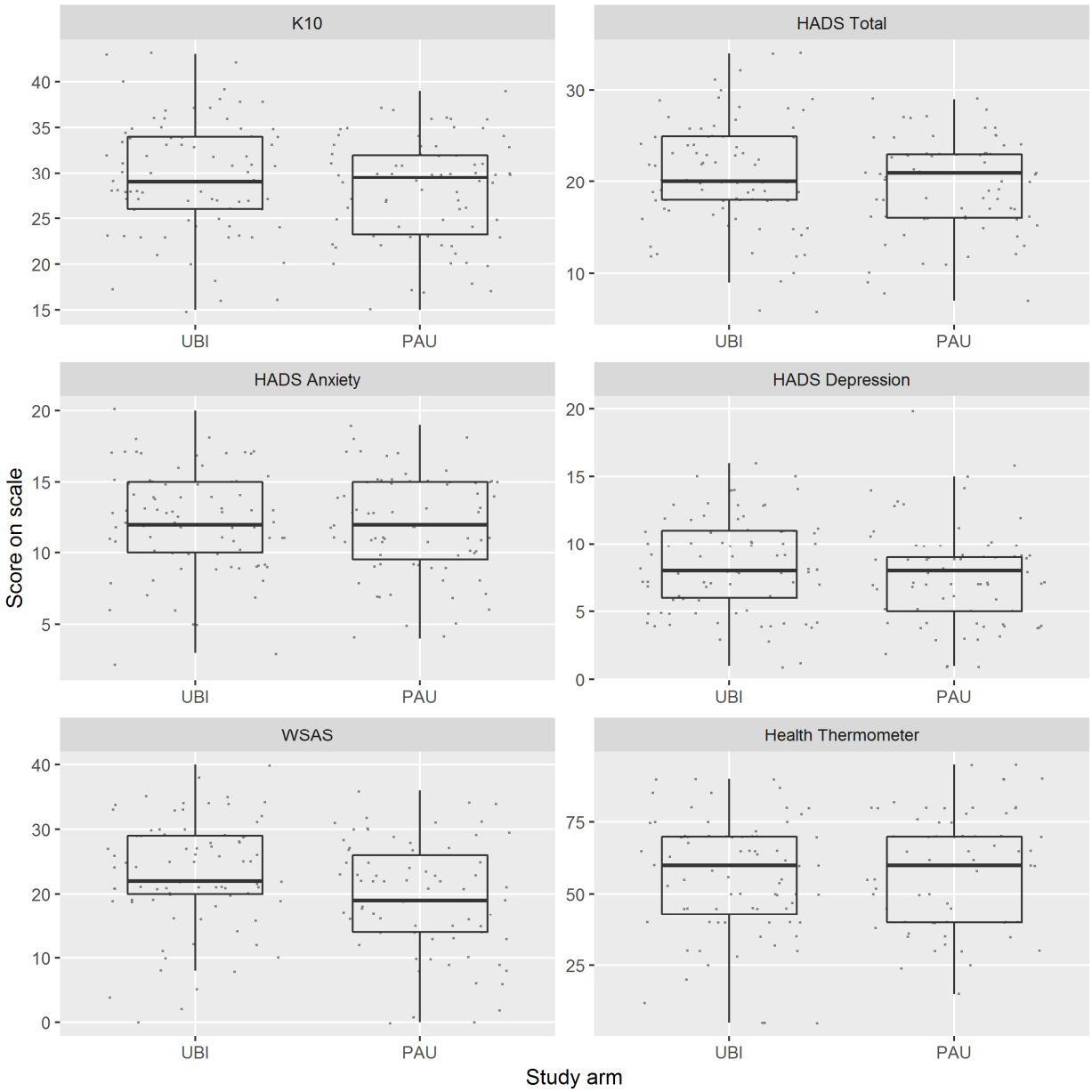
Supplementary Figure R1. Boxplots of baseline scores for each outcome measure (dots show each individual's score on that measure).



Supplementary Table R1. Number of patients recruited into study by GPs in UBI and PAU study arms.

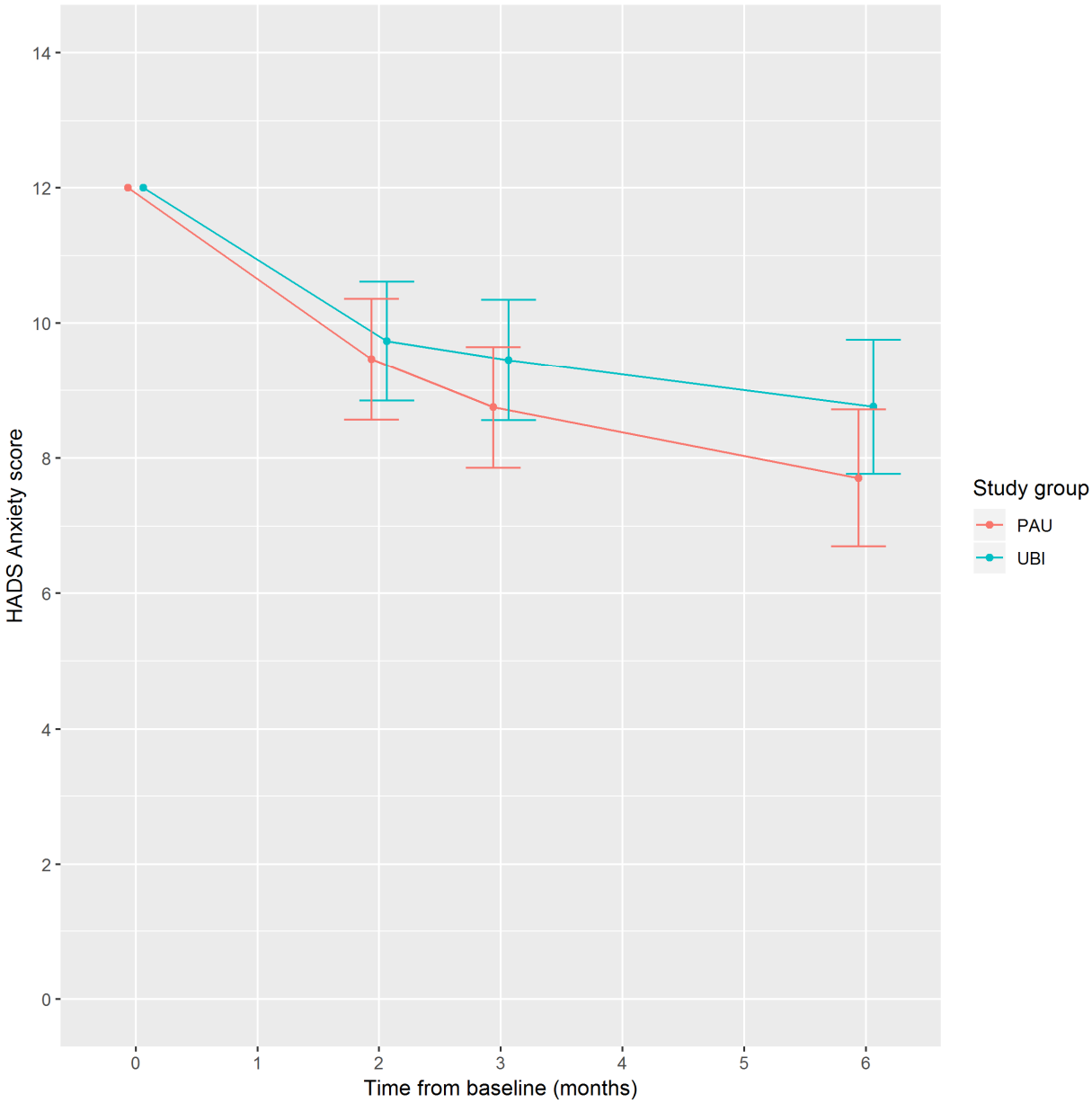
Number of patients recruited by GP	UBI (n GPs*)	PAU (n GPs*)
1	12	8
2	4	2
3	7	5
4	3	0
5	1	2
6	2	0
7	1	0
8	1	1
9	0	2
12	0	1
Total number of GPs	31	21

* Indicates the number of GPs recruiting the stated number of patients (e.g. 12 GPs in the UBI arm recruited one patient each; and five GPs in the PAU arm recruited three patients each).

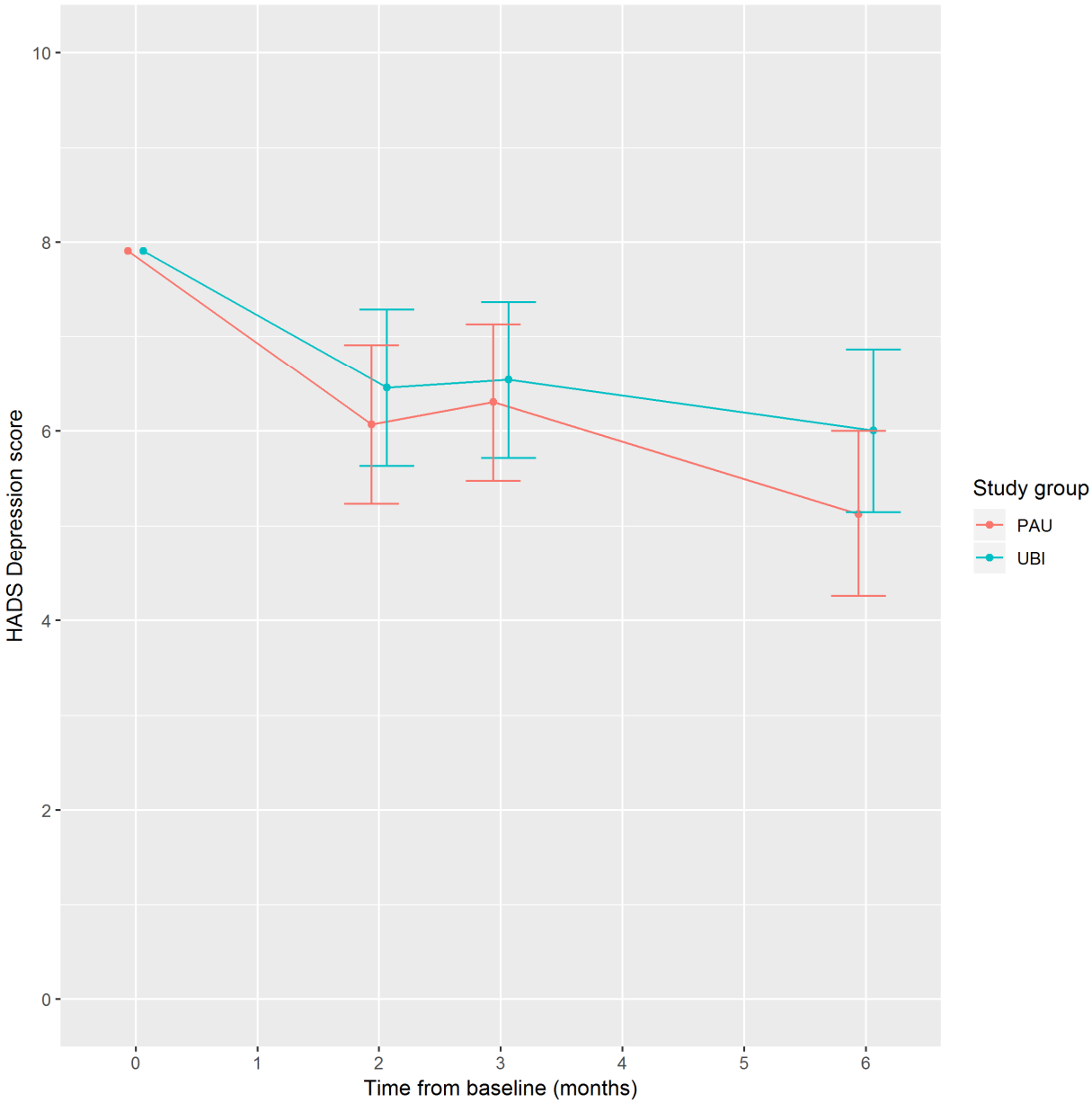
Supplementary Table R2. Mean improvements from baseline to 6 month follow-up for each outcome measure.

Outcome measure	Mean at baseline (both arms)	Mean improvement (95% CI) from baseline to 6 months	
		PAU	UBI
K10	28.8	7.6 (5.5, 9.6)	5.9 (4.0, 7.8)
HADS	19.9	7.0 (5.3, 8.7)	5.2 (3.5, 6.9)
HADS-A	12	4.3 (3.3, 5.3)	3.2 (2.2, 4.2)
HADS-D	7.9	2.8 (1.9, 3.7)	1.9 (1.0, 2.8)
WSAS	21.3	7.7 (5.7, 9.7)	7.2 (5.3, 9.2)
Health Thermometer	57.5	14.0 (9.3, 18.6)	9.0 (4.4, 13.7)

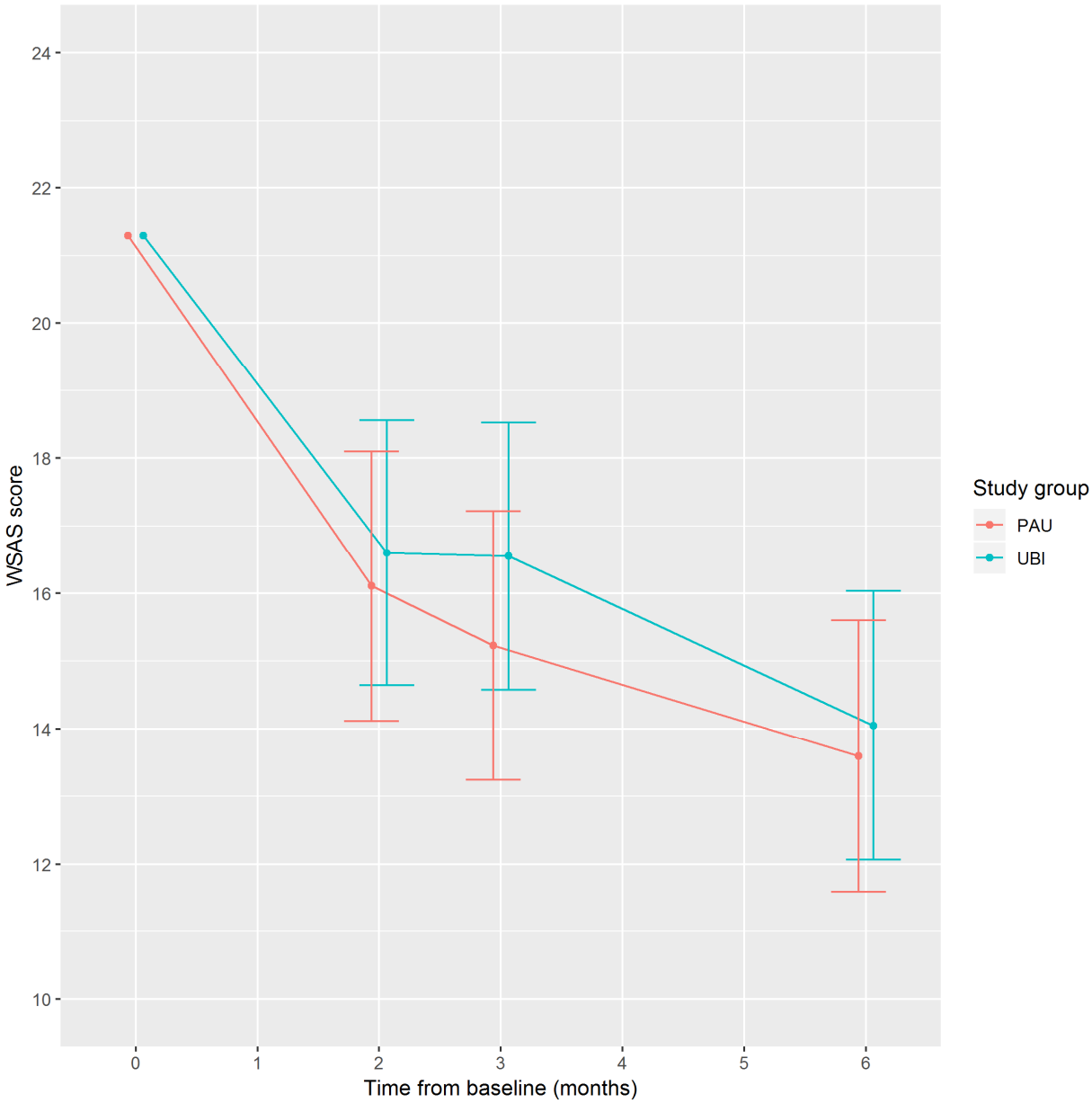
Supplementary Figure R2. Mean HADS Anxiety score (95% CI) at baseline and follow up for UBI and PAU study arms.



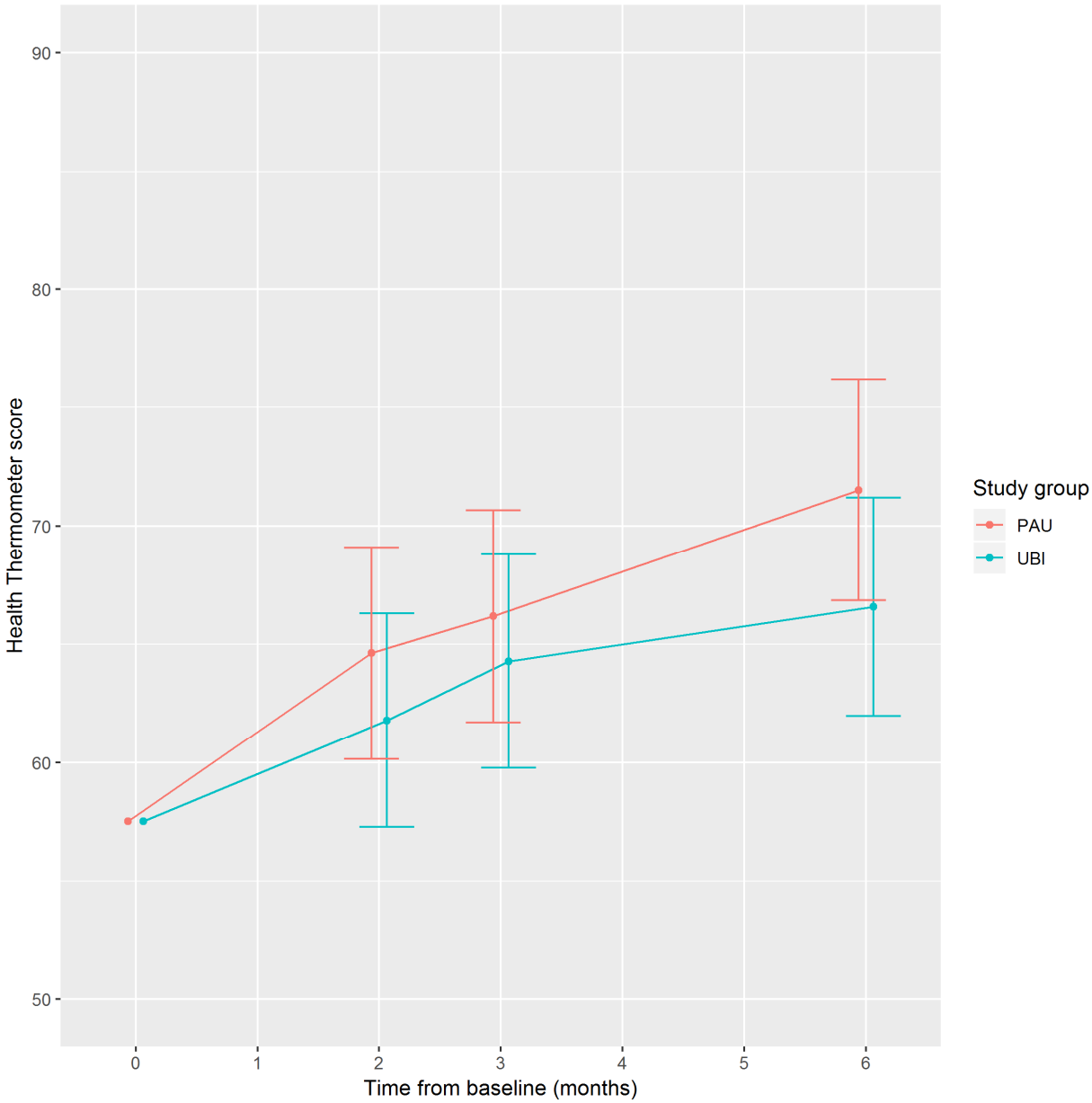
Supplementary Figure R3. Mean HADS Depression score (95% CI) at baseline and follow up for UBI and PAU study arms.



Supplementary Figure R4. Mean WSAS score (95% CI) at baseline and follow up for UBI and PAU study arms.



Supplementary Figure R5. Mean Health Thermometer score (95% CI) at baseline and follow up for UBI and PAU study arms.



Supplementary Table R3. Additional treatment received during UBI trial (from question on 6 month interview)

Type of additional treatment	UBI n (%)	PAU n (%)
Medication status during trial		
no relevant medication	33 (51%)	34 (52%)
on medication prior to entering trial	20 (31%)	16 (24%)
started medication during trial	12 (18%)	16 (24%)
did not complete question*	20	9
Extended GP consultations (n)		
0	68 (100%)	46 (71%)
1-2	0	8 (12%)
3-5	0	9 (14%)
6-10	0	2 (3%)
did not complete question*	17	10
Counselling sessions (n)		
0	44 (75%)	21 (36%)
1-2	4 (7%)	13 (22%)
3-5	2 (3%)	11 (19%)
6-10	7 (12%)	12 (20%)
11+	2 (3%)	2 (3%)
did not complete question*	26	16

* Did not complete 6 month questionnaire and hence no data (UBI n=16; PAU n=9)
Did not answer Meds question at 6 months (UBI: n=4; PAU: n=1)
Did not answer Extended GP question at 6 months (UBI: n=1; PAU: n=1)
Did not answer Counselling question at 6 months (UBI: n=10; PAU: n=7)

Supplementary Methods: Calculation of intra-class correlation coefficients (ICCs) for outcome measures.

ICCs were calculated for each outcome measure in the study to summarise the impact of clustering of outcomes by GPs. These were calculated using simplified mixed linear models with random intercept terms for GPs and no adjustment for covariates. ICCs were calculated in R 3.2.3, using the lme4 package, with their 95% confidence intervals based on 1000 bootstrap resamples calculated using the bootMer() function.

ICCs were also calculated for a scenario where clustering was considered across both the individual GPs (as per the above paragraph) and the practices in which GPs worked. The difference between these two sets of estimates can be considered as the additional impact of clustering of patient responses induced by practices above and beyond clustering induced by GPs. As seen in Supplementary Table R4, there was little impact of this additional clustering on ICCs for the longer health measures (K10 and HADS: minimal difference in ICCs between the two adjustment scenarios) but there appeared to be some additional impact of practice-level clustering for the Work and Social Adjustment Scale (WSAS) and the one-item Health Thermometer.

Supplementary Table R4. Intra-class correlation coefficients (ICCs) for each outcome measure in the study.

	GP clustering only*		GP and Practice clustering**	
Outcome measure	ICC	(95% CI)	ICC	(95% CI)
K10	0.129	(0.045, 0.231)	0.139	(0.006, 0.235)
HADS (total)	0.091	(0.019, 0.189)	0.104	(<0.001, 0.185)
HADS Anxiety	0.098	(0.019, 0.198)	0.106	(<0.001, 0.190)
HADS Depression	0.140	(0.047, 0.250)	0.148	(0.018, 0.233)
WSAS	0.188	(0.081, 0.308)	0.240	(0.076, 0.348)
Health Thermometer	0.088	(0.013, 0.177)	0.135	(0.005, 0.219)

* ICC calculated using only GP-level random effects.

** ICC calculated using random effects for GPs nested within GP practices (joint clustering effect).

Reference for lmer package:

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Supplementary Table R5. Sociodemographic and clinical characteristics at baseline by intervention arm (UBI or Practice as Usual [PAU]) and follow-up status.

Factor	Level	UBI follow-up (FU) status		PAU follow-up (FU) status	
		Lost to FU	some FU	Lost to FU	some FU
Total	All participants	15 (100%)	70 (100%)	6 (100%)	69 (100%)
Ethnicity	NZE Other	10 (67%)	51 (73%)	5 (83%)	49 (71%)
	Māori	5 (33%)	14 (20%)	1 (17%)	13 (19%)
	Pacific	0 (0%)	4 (6%)	0 (0%)	2 (3%)
	Asian	0 (0%)	1 (1%)	0 (0%)	5 (7%)
Age grp	15-24	11 (73%)	44 (63%)	6 (100%)	31 (45%)
	25-34	2 (13%)	14 (20%)	0 (0%)	15 (22%)
	35-44	1 (7%)	2 (3%)	0 (0%)	13 (19%)
	45-54	0 (0%)	5 (7%)	0 (0%)	6 (9%)
	55+	1 (7%)	5 (7%)	0 (0%)	4 (6%)
Gender	Female	7 (47%)	49 (70%)	3 (50%)	54 (78%)
	Male	8 (53%)	21 (30%)	3 (50%)	15 (22%)
NZiDep	0	3 (20%)	15 (21%)	0 (0%)	11 (16%)
	1	2 (13%)	14 (20%)	1 (17%)	16 (23%)
	2	3 (20%)	12 (17%)	2 (33%)	9 (13%)
	3	0 (0%)	10 (14%)	0 (0%)	10 (14%)
	4	2 (13%)	7 (10%)	1 (17%)	11 (16%)
	5	5 (33%)	12 (17%)	2 (33%)	12 (17%)
Education	At least secondary	15 (100%)	63 (90%)	6 (100%)	65 (94%)
	No secondary	0 (0%)	7 (10%)	0 (0%)	4 (6%)
Outcome scores at baseline		mean (sd)	mean (sd)	mean (sd)	mean (sd)
	K10	28.4 (5.9)	29.8 (6.3)	32.2 (5.3)	27.8 (5.6)
	HADS	20.2 (7.5)	20.7 (5.5)	23.0 (3.0)	19.2 (5.1)
	HADS Anxiety	11.9 (4.9)	12.2 (3.2)	13.2 (2.9)	11.7 (3.5)
	HADS Depression	8.3 (3.2)	8.5 (3.5)	9.8 (3.7)	7.5 (3.6)
	WSAS	21.7 (7.8)	23.3 (8.3)	23.8 (5.2)	19.2 (8.7)
	Health Thermometer*	57.4 (16.5)	55.0 (20.6)	50.5 (17.4)	59.5 (18.7)

* Health Thermometer: Lower scores indicate poorer health state.

Supplementary Results Text 1: Mean difference in K10 primary outcome at 6 months, adjusting only for baseline scores.

The protocol for the primary outcome (K10) analysis only specified that linear mixed model would be adjusted for baseline scores. The results from the primary analysis reported in the main paper were also adjusted for baseline sociodemographic variables (repeated in Supplementary Table R6 below from Table 3).

The analysis of K10 scores at 6 months (adjusted solely for baseline K10 scores) returned a slightly smaller mean difference between groups (poorer mean K10 score in UBI compared to PAU: difference = 1.07, 95% CI -1.67, 3.82).

This supplementary analysis draws on all participants with at least one follow-up observation. All other elements of the statistical model (accounting for clustering by GP and repeat observations for the same participant) are handled as per the main analysis (see Methods of main paper).

Supplementary Table R6. Primary outcome (K10) differences between UBI and PAU study arms at 6 months under different covariate adjustment models.

Analysis	Mean difference in K10 at 6 months (95% CI)
Analysis of all participants with some follow-up (n=139)	
Adjusted for baseline covariates *	1.68 (-1.18, 4.55)
Adjusted for baseline K10 score only**	1.07 (-1.67, 3.82)

* Result as reported in Table 3 of main paper.
** Analysis in line with specifications in protocol paper.

Supplementary Methods and Results: Sensitivity analysis to account for participants with no follow-up data.

The following analyses were implemented following initial peer-review, and were not *a priori* components of the analysis plan. Results from analyses are presented in Supplementary Table XX below, following the description of the methods and results. These sensitivity analyses aimed to consider the impact of complete loss-to-follow-up (participants no post-baseline data) on the primary outcome analysis, using two different frameworks assuming data were missing at random (MAR) or missing not at random (MNAR). A discussion of potential impacts of loss-to-follow-up on study results (attrition bias) is available in Bell et al. (2012) and discussion of missing data mechanisms can be read elsewhere (e.g. Bell et al. (2012); Newgard et al. (2015) and Sullivan et al. (2018)).

References for subsequent section:

Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ. 2013;346:e8668.

Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. JAMA. 2015;314:940-1.

Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018;27:2610-26.

van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67.

Imputation of outcomes under the Missing at Random (MAR) assumption.

Imputation was implemented using the mice package in R (van Buuren et al., 2011). All primary and secondary outcomes at all follow-up times were included in the imputation model, along with sociodemographic variables at baseline (gender/sex, age group, ethnicity, education, and NZiDep category: see Table 1 of the main paper for details about the specific sub-groups within each of these variables). Imputation was conducted separately for the intervention (UBI) and control (PAU) groups (Sullivan et al., 2018).

A total of 50 imputation datasets were created; each dataset was analysed for the primary outcome following the methods used for the main analysis in the paper (linear mixed model for K10 score at 6 months, adjusted for baseline K10 score and sociodemographic covariates). The estimates from these 50 models were then combined using Rubin's rules to produce the point estimate and 95% confidence interval (which takes into account variability in the effect estimates across all the imputed datasets.)

The intervention effect at 6 months is presented in Supplementary Table R7: under the assumption that the missing data mechanism was MAR (implemented using multiple imputation) there was a mean difference in K10 scores of 1.78 points (95% CI -0.96, 4.51; positive scores indicate better outcomes in the practice as usual [PAU] arm compared to UBI). This was almost identical to the estimates from the linear mixed model reported in Table 3 (repeated in Supplementary Table R7 for reference) which also assumed an MAR mechanism for missing data (conditional on the adjusted baseline variables in that model), but the analysis in the main results only included participants with at least one post-baseline measurement.

Imputation of outcomes under the Missing Not At Random (MNAR) assumption.

Analysis assuming that outcome values were MNAR was repeated under several conditions to explore the potential impact of different types of missing data mechanisms. These analyses all assumed that participants who did not participate in any follow-up did worse than those who participated in at least one follow-up.

In all scenarios, those who were not lost-to-follow-up (i.e. had at least one follow-up measure) kept either their original K10 scores at 6 months, or their imputed values at 6 months (for those with only partial follow-up: using the same imputed datasets as analysed under the MAR assumption). Imputation under MAR principles was considered reasonable for those with at least one follow-up measurement (but no 6-month measurement), as the follow-up measurements were all timed well after the conclusion of the core interventions delivered as part of the trial.

In MNAR Scenario 1: Individuals with no follow-up data were given a K10 score at six months set to 4 points lower than their imputed score.

In MNAR Scenario 2: Individuals with no follow-up data were given the same K10 score at six months that they had at baseline. This is effectively a "last observation carried forward" analysis for those with no follow-up data.

In MNAR Scenario 3: Individuals with no follow-up data were given a K10 score at six months that was 4 points lower than their baseline score.

The outcome analyses were again repeated on the 50 imputed datasets, and the intervention effect results combined across the resulting estimates.

While the effect sizes were slightly different from the main study result (Supplementary Table R7), these assumptions of data being MNAR had relatively minor impact on effect sizes. The most conservative result was under Scenario 1, assuming outcomes for those with no follow-up data were 4 points worse than imputed, returned a mean difference of 2.03, 95% CI -0.63, 4.70.

Note that the confidence intervals with the MNAR sensitivity analyses are likely to be conservative (i.e. not as wide as they should be) because the differences applied from the imputed or baseline values in each scenario were fixed rather than stochastic quantities (i.e. assumes that the applied difference from the imputed or baseline score was always a fixed quantity for all people).

Supplementary Table R7. Estimates of primary outcome (K10) effect size at 6 months under different assumptions of missing outcome profiles in participants with no follow-up data.

Analysis	Mean difference in K10 at 6 months (95% CI)
Analysis of all participants with some follow-up (n=139)	
Adjusted for baseline covariates (main analysis*)	1.68 (-1.18, 4.55)
Analysis including all randomised participants (n=160)	
Imputed K10 outcome at 6 months (MAR assumption*)	1.78 (-0.96, 4.51)
Imputed K10 outcome at 6 months (MNAR assumptions*)	
1. K10 at 6m set to 4 points worse than imputed	2.03 (-0.63, 4.70)
2. K10 at 6m set to baseline score	1.45 (-0.95, 3.84)
3. K10 at 6m set to 4 point worse than baseline	1.71 (-0.95, 4.37)

* Result as reported in Table 3 of main body of paper.

* MAR (missing at random) and MNAR (missing not at random) assumptions are for the 21 participants lost to follow up (no post-baseline data).