Supplementary material

Statistical analyses.

The data — A total of 427 MS patients were included in the current study. Of those, 87.8% completed the trial. Dropouts occurred i) immediately after randomization to treatment vs. control group (14), ii) after completed treatment but before 6-month follow-up (30 patients in the treatment group), and iii) after completion of the baseline questionnaires but before 6-month follow-up (8 patients in the control group). Here, we consider two dropout groups: dropout group 1 missing both baseline and 6-month follow-up questionnaire responses, and dropout group 2 missing 6-month follow-up questionnaire responses, only (see table 1). Besides dropout, missing data occurred in the EQ-VAS score for the first 80 patients included in the study. A single patient (id nr 1036, Control) had missing data for all questionnaire responses at baseline.

At baseline, the data for patient i consisted of the following baseline characteristics: 1) the group variable T indicating whether the patient received treatment ($T = 1$) or was part of the control group ($T = 0$), a set C of six clinical observations, 2) a set D of two demographic variables (sex and age), 3) a variable l for center where patient i was randomized (Ry or Haslev), and 4) the time t for entering the study (in weeks since trial commencement). Furthermore, a set Y_0 of six numerical measurement outcome (index score) variables: the FAMS score, the MSIS-29Phys and MSIS-29Psyc scores, the EQ-VAS score, the EQ-5D-5L score and the 15D score. A corresponding set Y_6 was obtained at 6-month follow-up. Furthermore, for the treatment group, 1 month follow-up (Y_1, immediately after discharge) and 2 month follow-up (Y_2) data were collected for selected scores. Clinical data at baseline consisted of the EDSS score, MS-type (RR, SP, PP), a variable indicating first-time treatment (yes or no), time since diagnosis (in years), reported time since appearance of first symptoms (in years), and a variable indicating whether immuno-treatment was received or not. In all analyses, the numeric variables time, age, EDSS, time since diagnosis and reported time since appearance

of first symptoms were converted to factor variables, as this allows for modelling non-linear relationships between in- and output. The following factor levels were used: Time - (0, 13], (13, 26], …, (117, 130] weeks since trial start; Age - (20, 40], (40, 50], (50, 60], (60, 65] years; EDSS - [0-3.5], [4-5.5], [6-6.5], [7-7.5]; Time since diagnosis - (0, 2], (2, 5], (5, 10], (10, 15], (15, 50] years; Reported time since appearance of first symptoms - (0, 5], (5, 10], (10, 20], (20, 50] years.

Dropout analysis — In order to investigate, whether dropout occurred at different frequencies in the control vs. treatment group, we modelled the i'th patient's membership in either the complete case group ($z = 0$), dropout group 1 ($z = 1$) or dropout group 2 ($z = 2$) as a function of treatment and baseline characteristics using a multinomial logit model (Faraway 2011):

$$\Pr(Z_i = 0, Z_i = 1, Z_i = 2) = \eta^{-1}(\beta_0 + \beta_1 t_i + \beta_2 l_i + \sum_{j=1}^{2} \beta_{Jj} D_{j,i} + \sum_{k=1}^{6} \beta_{Kk} C_{k,i} + \beta_t T_i)$$

where η^{-1} is the inverse logit function. Here, β_t is the coefficient for the treatment effect and β_{Jj} is the coefficient matrix for the j'th demographic variable. We used a step-wise backward model reduction strategy to reach at a minimal adequate model containing only significant model terms at a 5% significance level. Estimated model coefficients were log-transformed in order to obtain estimates of odd ratios for comparisons of group memberships.

Results — The dropout analysis indicated that dropout occurred at significantly different rates in the treatment and control groups (p = 0.0004) with dropouts immediately after randomization (group 1 dropouts) occurring at a higher rate in the control group (treatment/control odds ratios = 0.59, 95% c.i. = [0.19, 1.88]), while dropouts after ended treatment (group 2 dropouts) occurred at a higher rate in the treatment group (treatment/control odds ratios = 4.20, 95% c.i. = [1.83, 9.63]). Furthermore, MS-type (p = 0.037) and age (p = 0.012) were found to have a significant impact on dropout probabilities, with MS-type PP increasing the odds of dropout with a factor of 4.72 (95% c.i. = [1.10, 20.21]) for group 1 dropouts and 2.68 (95% c.i. = [1.02, 7.07]) for group 2 dropouts, when

compared to MS-type RR. Concerning age, dropout frequencies were highest for the youngest age group (20 to 40 years). See Table S1 for more details.

Table S1: Estimated odd ratios and their 95%-confidence intervals.

	Dropout group	
	1	2
Intercept	0.022 (0.003,0.160)	0.005 (0.001,0.048)
Treatment	0.592 (0.186,1.883)	4.198 (1.830,9.628)
Age (20, 40].	3.749 (0.493,28.51)	9.213 (0.965,87.922)
Age (40, 50]	2.59 (0.437,15.346)	7.742 (0.899,66.685)
Age (50, 60]	0.237 (0.019,2.909)	7.284 (0.880,60.26)
MS-type SP	0.816 (0.185,3.597)	0.957 (0.412,2.221)
MS-type PP	4.718 (1.101,20.213)	2.679 (1.016,7.067)

Exploratory Factor Analysis (EFA) — We conducted an EFA using all non-missing, centered and variance-standardized measurement outcomes at baseline (Y_0) in order to extract latent and uncorrelated signals among the outcomes. We used the Varimax rotation. Using the estimated loadings at baseline, factor scores for factor 1 and 2 were obtained for all patients with non-missing measurement outcome data at baseline and 6-month follow-up, and the scores were added to the appropriate sets Y_0 and Y_6.

Results — A priori, we expected the variables in Y to be associated, each being a manifestation of common underlying concepts related to perceived health and disease status. This was confirmed by pairwise comparisons of variables within Y_0 using Pearson's correlations (Table S2). An initial EFA indicated a high uniqueness of the EQ-VAS score (Table S2) and, hence, in combination with EQVAS having a high proportion of missing data at baseline (21.6%), the EQ-VAS score was excluded from the EFA. A mere of two latent factors were identified (Table

S2). This is supported both by sequential chi-square tests using maximum likelihood (2 factors, p = 0.112) and by means of visual inspection using the parallel analysis method described in Reise et al. (2000). The two factors explained a total of 67.5% of the variance in Y_0, with factor 1 alone accounting for 57.4% of the variance. High absolute loadings (>0.75) on factor 1 were found for FAMS and MSIS29Psyc, while factor 2 is defined by moderate absolute loadings (0.6 – 0.7) of EQ-5D-5L and MSIS29Psyc, with 15D loading with comparable strength (0.58, 0.59) on both factors (Table S2).

Table S2: Pairwise Pearson correlations and Exploratory Factor Analysis (EFA) results for the baseline index scores. All Pearson correlations are highly significant ($p < 3.1\times10^{-8}$).

							EFA results			
	Pearson correlation coefficients (95% confidence limits)						Uniqueness[a, I]	Uniqueness[aII]	Loadings[b, II]	
	FAMS	15D	EQ5D5L	MSIS29-Psyc	MSIS29-Phys	EQVAS			Factor 1	Factor 2
FAMS		0.70	0.46	-0.77	-0.56	0.42	0.22	0.22	0.77	0.43
15D	(0.64,0.74)		0.56	-0.62	-0.59	0.40	0.31	0.33	0.58	0.59
EQ5D5L	(0.39,0.54)	(0.49,0.62)		-0.32	-0.55	0.35	0.47	0.46	0.24	0.7
MSIS29-Psyc	(-0.8,-0.72)	(-0.67,-0.55)	(-0.41,-0.24)		0.41	-0.30	0.14	0.15	-0.91	-0.16
MSIS29-Phys	(-0.61,-0.48)	(-0.63,-0.49)	(-0.6,-0.46)	(0.31,0.48)		-0.39	0.43	0.46	-0.33	-0.65
EQVAS	(0.33,0.51)	(0.31,0.49)	(0.25,0.44)	(-0.39,-0.2)	(-0.47,-0.29)		0.74			

[a]: Uniqueness describes the amount of variation not explained by the EFA. [b]: Loading indicates the association of a score with the latent factor.[I]: Initial EFA including EQVAS, n = 335. [II]: EFA excluding EQVAS due to its high uniqueness, n = 412.

Intention-to-treat (ITT) statistical analyses — In order to make valid inference on the population of interest, the effect of missing data and dropouts must be addressed (White et al. 2012). We followed the strategy proposed by White et al. (2012) and performed three analyses: I) (Main analysis) a multiple model-based imputation strategy under the assumption that data are missing at random (MAR) conditionally on modelled baseline characteristics and index score data. II) A sensitivity analysis based on multiple random imputation of the index score data. III) An available case analysis (ACA, i.e. complete-case analysis within each index score). While strategy I seeks to find an honest estimate of the treatment effect under the plausible assumption of MAR, strategy II is in favor of the null hypothesis of no treatment effect and provides a conservative bound on the estimated standard errors by polluting the data with noise proportional to the frequency of missing data. Strategy III is expected to yield

biased estimates and standard errors and is carried out here only for the reason of reference. Under all strategies, the bootstrap (1000 iterations) was used in order to estimate the mean and 95% confidence interval (CI) for all parameters of interest. P-values were calculated by comparing the estimated t-value to the bootstrap t-value distribution under the null hypothsis of no treatment effect resulted from shuffling group membership. For strategy I) we used multiple random forest imputation (Shah et al. 2014). Under strategy II, missing data were imputed by randomly sampling with replacement from the observed data without taking group membership into account.

Under all three strategies, the following models were used at each bootstrap iteration.

i) In order to estimate adjusted population means for the treatment and control group at baseline for each variable h in Y_o, we modelled $Y_{0,h}$ for patient i using the following linear mixed effects model:

$$Y_{0,h,i} = \beta_0 + \sum_{j=1}^{2} \beta_{Jj} D_{j,i} + \sum_{k=1}^{6} \beta_{Kk} C_{k,i} + \beta_t T_i + t_{\tau[i]} + l_{\lambda[i]} + e_i$$

where $t_{\tau[i]}$ is the time class τ that patient i belongs to. Time, locality and residual errors were modelled as uncorrelated random effects, $t_\tau \sim N(0, \sigma_t^2)$, $l_\lambda \sim N(0, \sigma_l^2)$ and $e_i \sim N(0, \sigma_e^2)$. C and D are the sets of clinical and demographic variables.

ii) In order to estimate for the treatment group the change in Y_h from baseline over discharge and 2-months follow-up to 6-month follow-up, we modelled the change in Y_h for patient i relative to the baseline, $\Delta Y_{0\to g,h}$, where *g* indicates the *g*'th follow-up time, as a function of follow-up time *G* and baseline characteristics using the following linear mixed effects model:

$$\Delta Y_{0\to g,h,i} = \beta_0 + \beta_1 G_{g,i} + \beta_2 t_{g,i} + \beta_3 l_{g,i} + \sum_{j=1}^{2} \beta_{Jj} D_{j,i} + \sum_{k=1}^{6} \beta_{Kk} C_{k,i} + i + e_{g,i}$$

where the individual patients and residual errors were modelled as uncorrelated random effects $i \sim N(0, \sigma_i^2)$ and $e_{g,i} \sim N(0, \sigma_e^2)$.

iii) In order to investigate the treatment effect (T) at 6-months follow-up on each variable h in Y we modelled the change in Y_h for patient i from baseline to 6-months follow-up, e.g., $\Delta Y_{0\rightarrow 6,h} = Y_{6,h} - Y_{0,h}$, as a function of $Y_{0,h}$, treatment and baseline characteristics using the following linear mixed effects model:

$$\Delta Y_{0\rightarrow 6,h,i} = \beta_0 + \beta_1 {Y'}_{0,h,i} + \sum_{j=1}^{2} \beta_{jj} D_{j,i} + \sum_{k=1}^{6} \beta_{Kk} C_{k,i} + \beta_t T_i + t_{\tau[i]} + l_{\lambda[i]} + e_i$$

where $t_{\tau[i]}$ is the time class τ that patient i belongs to. Time, locality and residual errors were modelled as uncorrelated random effects, $t_\tau \sim N(0, \sigma_t^2)$, $l_\lambda \sim N(0, \sigma_l^2)$ and $e_i \sim N(0, \sigma_e^2)$. ${Y'}_{0,h}$ stands for the score at baseline, but recoded as a factor variable of 20%-quantile classes. The inclusion of ${Y'}_{0,h}$ seeks to model the ceiling and flooring effect of the bounded scores that by definition show an increased probability for changes toward zero.

In all instances, we used a step-wise backward model reduction strategy to reach at a minimal adequate model containing only significant model terms at a 5% significance level. However, T and ${Y'}_0$ were forced to be retained in the model during this procedure.

Least-square means (i.e. covariate- and imbalance adjusted, model-predicted means, Lenth 2016) were calculated for the treatment groups and follow-up times.

We used standard visualization techniques for checking of model assumptions.

All statistical analyses were carried out using the R environment (R Core Group 2017), with mixed effects models fit using the packages `lme4` (Bates et al. 2015) and `lmerTest` (Kuznetsova et al. 2016), Least-square means calculated using package `lsmeans` (Lenth 2016), and random forest imputations done using package `randomForestSRC` (Liaw & Wiener 2002), with function `impute()` using default settings.

References

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi: 10.18637/jss.v067.i01

Ishwaran H. and Kogalur U.B. (2016). Random Forests for Survival, Regression and Classificati on (RF-SRC), R package version 2.4.1.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. (2016). lmerTest: Tests in Linear Mixed Ef fects Models. R package version 2.0-32. https://CRAN.R-project.org/package=lmerTest

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Reise, S.P., Waller, N.G., Comrey, A.L. (2000). Factor Analysis and Scale Revision. Psychological Assessment Vol. 12, No. 3, 287-297

Lenth, R.V. (2016). Least-Squares Means: The R Package lsmeans. Journal of Statistical Softw are, 69(1), 1-33. doi:10.18637/jss.v069.i01

White, I.R, Carpenter, J., Horton, N.J. (2012). Including all individuals is not enough: lessons f or intention-to-treat analysis. Clin. Trials 9(4): 396-407.