SUPPLEMENTARY FILES

Supplementary Table 1: MSOAC Workgroups and Members

Supplementary Table 2: Search Terms, Databases, and Time Frame

Supplementary Table 3: Abstract Filtering Process

Supplementary Table 4: Data Extraction Table (separate pdf)

Supplementary Text for the Statistical Analysis Plan

Supplementary Table 1: MSOAC Workgroups and Members

Coordinating Committee

Nicholas LaRocca	**MSOAC Co-Director**; National MS Society
Richard Rudick	**MSOAC Co-Director**; Cleveland Clinic, Biogen
Lynn Hudson	**MSOAC Co-Director**; Critical Path Institute
Elizabeth Merikle	AbbVie
Jane Haley	AbbVie
Steve Hass	AbbVie
Steven Greenberg	AbbVie
Thomas Marshall	AbbVie
Xiaolan Ye	AbbVie
Matthew Sidovar	Acorda
Craig Sherburne	Alberta MS Research Foundation
Bjorn Sperling	Biogen
Claudia Ordonez	Biogen
Gilmore O'Neill	Biogen
Glenn Phillips	Biogen
Jacob Elkins	Biogen
John Richert	Biogen
Sanjay Keswani	Bristol-Myers Squibb
Tanuja Chitnis	Brigham and Women's Hospital
Deborah Miller	Cleveland Clinic
Jeffrey Cohen	Cleveland Clinic
Ann Robbins	Critical Path Institute
Emily Hartley	Critical Path Institute
Enrique Aviles	Critical Path Institute
Gary Lundstrom	Critical Path Institute
Jon Neville	Critical Path Institute
Geoffrey Dunbar	EMD Serono
Tanya Fischer	EMD Serono
Thorsten Eickenhorst	EMD Serono
Maria Isaac	European Medicines Agency
Kathy Smith	Fast Forward (NMSS)
Elektra Papadopoulos	Food and Drug Administration
Indira Hills	Food and Drug Administration
Marc Walton	Food and Drug Administration
Michelle Campbell	Food and Drug Administration
Sarrit Kovacs	Food and Drug Administration
Susan Montenegro	Food and Drug Administration
Wen-Hung Chen	Food and Drug Administration
Irina Antonijevic	Genzyme/Sanofi

Jennifer Panagoulias	Genzyme/Sanofi
Michael Panzara	Genzyme/Sanofi
Maria Davy	GlaxoSmithKline
Paul Thompson	GlaxoSmithKline
Paul Matthews	Imperial College London
Gill Webster	Innate Immunotherapeutics
Simon Wilkinson	Innate Immunotherapeutics
Ellen Mowry	Johns Hopkins University
Peter Calabresi	Johns Hopkins University
Nancy Mayo	McGill
Ed Holloway	MS Society (UK)
Fred Lublin	Mt Sinai School of Medicine
Ursula Utz	National Institute of Neurological Disorders and Stroke
Timothy Coetzee	National MS Society
Weyman Johnson, JD	National MS Society
Laura Balcer	New York University
David Leppert	Novartis
Frank Dahlke	Novartis
Richard Meibach	Novartis
Gordon Francis	Novartis/Consultant
Jeremy Hobart	Plymouth Hospital
Adam Jacobs	Premier Research
Shari Medendorp	Premier Research
Algirdas Kakrieka	Roche/Genentech
Bruno Musch	Roche/Genentech
Donna Masterman	Roche/Genentech
Peter Chin	Roche/Genentech
Giancarlo Comi	Scientific Institute H.S. Raffaele, Italy
Ralph Benedict	State University of NY - Buffalo
Lauren Krupp	Stony Brook Medicine
Joshua Steinerman	Teva
Matthew Davis	Teva
Volker Knappertz	Teva
Elizabeth Morrison	UC Irvine
Raj Kapoor	University College London Institute of Neurology
Gary Cutter	University of Alabama Birmingham
Maria Pia Sormani	University of Genoa
Robert Motl	University of Illinois
Brenda Banwell	University of Pennsylvania
Myla Goldman	University of Virginia
Bernard Uitdehaag	VU University Medical Center

Defining Disability Workgroup (DD)

Jeremy Hobart	**Workgroup Co-Chair**; Plymouth Hospital
Donna Masterman	**Workgroup Co-Chair**; Roche/Genentech
Jane Haley	AbbVie
Matthew Sidovar	Acorda
Glenn Phillips	Biogen
Jacob Elkins	Biogen
Jeffrey Cohen	Cleveland Clinic
Gary Lundstrom	Critical Path Institute
Lynn Hudson	Critical Path Institute
Stephen Coons	Critical Path Institute
David Margolin	Genzyme/Sanofi
Nancy Mayo	McGill
Nicholas LaRocca	National MS Society
Ralph Benedict	State University of NY - Buffalo

COA and Data Analysis Workgroup (CDA)

Jacob Elkins	**Workgroup Co-Chair**; Biogen
Jennifer Panagoulias	**Workgroup Co-Chair**; Genzyme/Sanofi

COA and Data Analysis Workgroup (CDA) – Statistics Team

Jerry Weaver	**Workgroup Co-Chair**; Novartis
Matthew Davis	**Workgroup Co-Chair**; Teva
Adam Jacobs	**Workgroup Co-Chair**; Premier Research
Glenn Phillips	Biogen
Ih Chang	Biogen
Richard Rudick	Biogen
Sophia Lee	Biogen
Stacy Lindborg	Biogen
Jeffrey Cohen	Cleveland Clinic
Bob Stafford	Critical Path Institute
Emily Hartley	Critical Path Institute
Gary Lundstrom	Critical Path Institute
Lynn Hudson	Critical Path Institute
Jeff Palmer	Genzyme/Sanofi
Kathryn Fitzgerald	Johns Hopkins University
Lauren Strober	Kessler Foundation
Nicholas LaRocca	National MS Society
Nikolaos Sfikas	Novartis
Shari Medendorp	Premier Research
Gary Cutter	University of Alabama Birmingham

Maria Pia Sormani	University of Genoa
Bernard Uitdehaag	VU University Medical Center

COA and Data Analysis Workgroup (CDA) – Literature Review Team

Glenn Phillips	**Workgroup Co-Chair**; Biogen
Ralph Benedict	**Workgroup Co-Chair**; State University of NY - Buffalo
Matthew Sidovar	Acorda
Jacob Elkins	Biogen
Richard Rudick	Biogen
William Anthony	Boston University
Jeffrey Cohen	Cleveland Clinic
Gordon Francis	Consultant
Gary Lundstrom	Critical Path Institute
Jon Neville	Critical Path Institute
Lynn Hudson	Critical Path Institute
Jennifer Panagoulias	Genzyme/Sanofi
John DeLuca	Kessler Foundation
Lauren Strober	Kessler Foundation
Wendy Kaye	McKing Consulting
Nicholas LaRocca	National MS Society
Laura Balcer	New York University
Peter Feys	University of Hasslet
Robert Motl	University of Illinois
Robert Naismith	Washington University

COA and Data Analysis Workgroup (CDA) – Voice of the Patient (VOP; formerly MDP) Team

Nicholas LaRocca	**Workgroup Co-Chair**; National MS Society
Deborah Miller	**Workgroup Co-Chair**; Cleveland Clinic
Glenn Phillips	Biogen
Jacob Elkins	Biogen
William Anthony	Boston University
Gordon Francis	Consultant
Ann Robbins	Critical Path Institute
Gary Lundstrom	Critical Path Institute
Lynn Hudson	Critical Path Institute
Sarah Mann	Critical Path Institute
Elektra Papadopoulos	FDA
Michelle Campbell	FDA
Susan Montenegro	FDA
Wen-Hung Chen	FDA
Jennifer Panagoulias	Genzyme/Sanofi
Nancy Chiaravalloti	Kessler Foundation
Nancy Mayo	McGill

Peter Chin	Roche/Genentech
Barbara Vickrey	University of California at Los Angeles
Myla Goldman	University of Virginia
Dagmar Amtmann	University of Washington

Regulatory Advisory Workgroup (RAW)

Ann Robbins	**Workgroup Chair**; Critical Path Institute
John Richert	Biogen
Richard Rudick	Biogen
Jeffrey Cohen	Cleveland Clinic
Alicia West	Critical Path Institute
Gary Lundstrom	Critical Path Institute
Lynn Hudson	Critical Path Institute
Jennifer Panagoulias	Genzyme/Sanofi
Michael Panzara	Genzyme/Sanofi
Paul Matthews	Imperial College London
Nicholas LaRocca	National MS Society
Richard Meibach	Novartis
Peter Chin	Roche/Genentech

Data Standards and Integration Workgroup (DSI)

Jon Neville	**Workgroup Co-Chair**; Critical Path Institute
Ellen Mowry	**Workgroup Co-Chair**; Johns Hopkins University
Christina Casteris	Biogen
Robert Bermel	Cleveland Clinic
Bess LeRoy	Critical Path Institute
Bob Stafford	Critical Path Institute
Emily Hartley	Critical Path Institute
Enrique Aviles	Critical Path Institute
Gary Lundstrom	Critical Path Institute
Lynn Hudson	Critical Path Institute
Jeff Palmer	Genzyme/Sanofi
Paul McGuire	Novartis
Lauren Krupp	Stony Brook Medicine
Jerry Wolinsky	University of Texas
Myla Goldman	University of Virginia
Robert Naismith	Washington University

Placebo Review Board

Lynn Hudson	Review Board Chair; Critical Path Institute
Bjoern Sperling	Biogen
Guido Sabatella	Biogen

Tim Swan	Biogen
Gary Lundstrom	Critical Path Institute
Kathy Smith	Fast Forward
Ellen Mowry	Johns Hopkins
Nicholas LaRocca	National MS Society
Joshua Steinerman	Teva
Mat Davis	Teva

Additional MSOAC Participants

Andrew Blight	Acorda
Diego Cadavid	Biogen
Gilbert Litalien	Bristol-Myers Squibb
Dan Ontaneda	Cleveland Clinic
June Halper	Consortium of Multiple Sclerosis Centers
Phillipe Truffinet	Genzyme/Sanofi
Stephen Lake	Genzyme/Sanofi
Steven Blum	GlaxoSmithKline
Brooke Dennee-Somers	Ironwood
Gwyn Reis	Ironwood
Giampaolo Brichetto	Italian Multiple Sclerosis Society
Paola Zaratin	Italian Multiple Sclerosis Society
Aaron Miller	Mt Sinai School of Medicine
Karen Lee	Multiple Sclerosis Society of Canada
Susan Kohlhaas	Multiple Sclerosis Society UK
Eric Chamot	University of Alabama
John Petkau	University of British Columbia
Andrew Goodman	University of Rochester
Chris Polman	VU University Medical Center

Supplementary Table 2: Search Terms, Databases, and Time Frame

SEARCH #	SEARCH TERMS
S1	multiple PRE/0 sclerosis OR "multiple sclerosis" OR MESH.EXACT("Multiple Sclerosis") OR EMB.EXACT("Multiple Sclerosis") OR tio(MS PRE/0 patient[*1])
S2	MESH.EXACT("Disease Progression") OR MESH.EXACT("Disability Evaluation") or ((disability or impairment or functional) and tio(impact$2 or correlat$3 or predict$4 or association))
S3	(EMB.EXACT.EXPLODE("disability") OR EMB.EXACT("disease course")) and (EMB.EXACT("prediction") or predict$4)
S4	S2 OR S3
S5	S1 AND S4
S6	S5 AND (MESH.EXACT("Patient Outcome Assessment") OR MESH.EXACT("Outcome Assessment (Health Care)") OR MESH.EXACT("Activities of Daily Living") OR MJMESH.EXACT("Quality of Life") OR MESH.EXACT("Treatment Outcome") OR MJMESH.EXACT("Neuropsychological Tests") OR EMB.EXACT("neuropsychological test") OR EMB.EXACT("daily life activity") OR EMB.EXACT("patient assessment") OR EMB.EXACT("clinical assessment tool") EMB.EXACT("neurologic disease assessment") OR EMB.EXACT("motor dysfunction assessment") OR EMB.EXACT("rating scale") OR MESH.EXACT("Psychometrics") OR MESH.EXACT("Reproducibility of Results") OR EMB.EXACT("psychometry") OR EMB.EXACT("test retest reliability") OR EMB.EXACT("predictive value") OR EMB.EXACT("validity") OR tio(validity or reliabil$3))
S7	S6 AND su(human or humans) (SU.EXACT("Multiple Sclerosis") OR multiple pre/0 sclerosis) AND (SU.EXACT("Disabilities") OR SU.EXACT("Disease Course") OR SU.EXACT("Cognitive Assessment") OR SU.EXACT("Cognitive Impairment") OR SU.EXACT("Executive Function") OR SU.EXACT("Cognitive Ability") OR TM(multiple pre/0 sclerosis pre/0 functional pre/0 composite)) AND (MJSUB.EXACT("Quality of Life") OR SU.EXACT("Rating Scales") OR SU.EXACT("Psychometrics") OR SU.EXACT("Test Validity") OR SU.EXACT.EXPLODE("Client Attitudes")) AND POP(human)

S8	S7 OR S8

Consecutive searches (search #S1-S8) were conducted using the Medline, Embase, Embase Alert, PsychInfo, and CINAHL (Cumulative Index of Nursing and Allied Health Literature) databases for the period 1990- 2016.

Research staff at Biogen conducted the search in the first three databases and McKing Consulting conducted the literature search in the CINAHL database. The results from both searches were combined into one database. Every identified article was tracked; none were discarded without explanation.

Supplementary Table 3: Abstract Filtering Criteria and Codes

CODE	CATEGORY
INCLUSION	
1	PerfO
2	Both PerfO and PRO
EXCLUSION	
3	Case study
4	Letters to the Editor/Editorials/Correspondences between authors/etc.
5	Animal (only) Study
6	Pediatric (only) Study
7	Imaging (only)/Evoked Potentials Study – no outcomes
8	Conference Abstract
9	Pharmacokinetics/Pharmacodynamics/Lab/Biomarkers (only)
10	Patient reported outcome only
11	Purely Qualitative no PERFO (e.g., questionnaire design) → err away from this if PERFO or PRO present in substantial amount.
12	Economic Analysis/Risk Benefit Analysis
13	No MS participants
14	Epidemiology
15	News Articles
99	Other
NA	No Abstract – Cannot find abstract
NAX	No Abstract, but not deemed relevant
D	Duplicate – publication already coded and present in database

Each abstract is coded based on the categories listed for inclusion or exclusion criteria. To be included, abstracts were required to reference disabilities, impairments, symptoms and/or impacts of symptoms experienced by people with MS, and also include reference to quantitative performance measures or patient-reported outcome measures (e.g., interviews, focus groups, patient reports, thematic analysis, grounded theory). The searches were limited to publications since 1990 and included multiple languages. Editorials, letters, abstracts, unpublished reports, reviews, and articles published in non-peer-reviewed journals were not used in the analysis.

Supplementary Text for the Statistical Analysis Plan

SAP Part 1: Baseline characteristics: Continuous variables presented with descriptive statistics included the number of observations, mean, standard deviation, median, minimum, and maximum. Categorical variables were presented with the number of observations and % in each category. Each study set included the following characteristics: 1) Age as a continuous variable and in categories of < 35, 35–45, and > 45 years; 2) Sex in categories of male or female; 3) Race in categories of Native American or Alaskan, Asian, black or African American, white, and other; 4) Treatment arm as categories placebo, glatiramer acetate or interferon beta, and other; 4) MS subtype as categories RRMS, SPMS, and PPMS; 5) Duration of disease as a continuous variable and in categories < 10 years and ≥ 10 years; 6) Baseline EDSS score as a continuous variable and in categories 0–3.5 and 4.0–10; 7) Baseline 9HPT score as a continuous variable and in categories below the median and at or above the median; 8) Baseline LCLA score as a continuous variable and in categories below the median and at or above the median; 9) Baseline T25FW score as a continuous variable and in categories below the median and at or above the median; 10) Baseline SDMT score as a continuous variable; 11) Baseline PASAT score as a continuous variable.

SAP Part 2: Descriptive analyses: Descriptive analyses of the SDMT, PASAT, T25FW, 9HPT, and LCLA were carried out separately by timepoint with just 1 record per subject per timepoint (with the exception of the LOWESS smoothed graphs, which use all available data). The following statistics were generated for each timepoint, both for absolute values and change from baseline: 1) Number of observations; 2) Number of missing observations; 3) Mean; 4) Standard deviation; 5) Minimum; 6) 5^{th} centile; 7) 1^{st} quartile; 8) Median; 9) 3^{rd} quartile; 10) 95^{th} centile; 11) Maximum; and 12) Skewness. Analyses were reported for all patients combined and separately by treatment group. Analyses of each variable used the analysis set appropriate to that variable. The distribution of scores for each variable at baseline were presented graphically with a histogram. The absolute values and change from baseline for all variables (SDMT, PASAT, T5FW, 9HPT, and LCLA) over time were depicted in the following graphs: 1) Box and whisker plot of each variable by time, all patients combined; 2) Box and whisker plot of each variable by time, separate boxes for each treatment. The LOWESS graphs and box and whisker plots of SDMT and PASAT by treatment were repeated by subgroups of the baseline score of each variable, with the subgroups defined as below the 40^{th} centile and equal to or above the 40^{th} centile.

SAP Part 3: Analyses of disability worsening: Several definitions of "confirmed disability worsening" using the EDSS have been proposed, but a 1.5-point increase for patients with an EDSS of 0, a 1 point increase for patients with an EDSS of 1.0 to 5.5, or an increase of 0.5 for patients with an EDSS of 6 or greater, sustained for at least 3 or 6 months, are the most common [16,31]. The definition used to define EDSS-based disability worsening for the MSOAC analysis was a 1.5-point increase for patients with an EDSS of 0, a 1 point increase for patients with an EDSS of 1.0 to 5.5, or an increase of 0.5 for patients with an EDSS of 6 or greater, sustained for at least 3 months.

There is no universally established definition of disability worsening by the SDMT, although a number of studies have pointed to a 3- or 4-point change as clinically meaningful [14,15,32–34]. The primary definition used in this analysis was a decrease of 4 points from baseline, sustained for at least 3 months. In addition to the above definition, alternative definitions of SDMT disability worsening using a decrease of 3 points from baseline as well as 10%, 15%, and 20% decreases were used as sensitivity analyses.

The definition of disability worsening for PASAT is a decrease of 20% in the score from baseline. Decreases of 10% and 15% were used as sensitivity analyses. For T25FW, 9HPT, and LCLA, a 20% worsening from baseline (ie an increase for T25FW or 9HPT and a decrease for LCLA) was taken as the definition of disability worsening. Sensitivity analyses were conducted, in which 15% and 10% worsening was used instead of 20% worsening. For EDSS, SDMT, PASAT, T25FW, 9HPT, and LCLA, descriptive statistics (N, mean, SD, median, minimum, and maximum) were presented at each post-baseline timepoint separately for patients who have and who have not met the disability worsening criterion. Kaplan-Meier plots were produced in which the worsening by each outcome measure except EDSS was shown on the same graph as worsening by EDSS. This was done for all patients, and by subgroups according to baseline EDSS score, age, sex, and treatment group. The agreement between disability worsening at endpoint by EDSS and each other measure was assessed by calculating Cohen's kappa coefficient and its 95% confidence interval.

For each outcome measure except EDSS, characteristics of patients (age, sex, baseline EDSS score, and baseline score on the other outcome measure) with different combinations of disability worsening on EDSS and the other outcome measure were tabulated by showing summary statistics for those characteristics according to whether patients have worsened or not by each of the 2 measures (thus for 4 categories in total). Summary statistics are shown overall for patients with concordant and discordant disability worsening at endpoint.

SAP Part 4: Reliability analyses: Test-retest reliability analyses used data from stable patients only over a period not exceeding 6 months from baseline. A patient was defined as stable for as long as the patient's EDSS score did not change. Thus, the reliability analysis was based on all pre-baseline measurements from the first day at which the EDSS score was the same as at baseline and remained the same as at baseline, and continued for all post baseline measurements up to the last day at which the EDSS score remained the same as at baseline or up to 6 months from baseline, whichever was the sooner. The test-retest reliability of the SDMT, PASAT, T25FW, 9HPT, and LCLA was estimated by calculating the intra-class correlation coefficient from a random effects linear regression analysis with a random subject effect and terms to account for practice effects. The fit of the model was examined by regression diagnostic plots of within-subject residuals from the model. The within-subject residual standard deviation of the SDMT, PASAT, T25FW, 9HPT, or LCLA score was also reported. Practice effects were determined by examining the regression coefficients for test number from the above model. Cohen's d was calculated from those terms as a measure of the effect size of practice effects, by dividing the regression coefficients by the pooled standard deviation of all

scores included in the model. This analysis was carried out for all patients combined, and for subgroups of baseline EDSS score, disease duration, and age.

SAP Part 5: Construct validity analyses: To investigate the construct validity of the outcome measures, correlations were investigated between each outcome measure and other outcome measures. For each pair of measures, scatterplots were produced showing the correlation between the measures graphically. Correlation coefficients between each pair of variables were calculated: For baseline, endpoint, and change from baseline at endpoint, both the Pearson and Spearman correlation coefficients were calculated, together with their 95% confidence intervals (using Fisher's transformation). These analyses were done for all patients, and by subgroups of age, baseline EDSS, baseline 9HPT score, baseline LCLA score, and disease duration.

SAP Part 6: Convergent validity analyses: To investigate convergent validity, correlations between the following measures were investigated: 1) SDMT with PASAT; 2) T25FW with EDSS; 3) 9HPT with EDSS; 4) LCLA with high contrast visual acuity, visual functional systems score. The same analyses as described for construct validity were used.

SAP Part 7: Known group analyses: Baseline outcome measure scores were compared between groups of patients with short and long disease duration and between patients with high and low EDSS scores. The baseline score for each outcome measure (SDMT, PASAT, T25FW, 9HPT, and LCLA) was compared between the groups using an ANOVA model adjusting for age in 5-year age bands. The mean difference was presented along with its 95% confidence interval and P value; age effects from this model were also presented.

SAP Part 8: Analysis of sensitivity to change: Sensitivity to change was assessed by comparing scores for SDMT, PASAT, T25FW, 9HPT, and LCLA before and after relapse. Post-relapse scores were included in this analysis if they are measured within 3 months after the start of the relapse. The scores were compared with a paired t-test. Sensitivity to change was also assessed by examining change in groups of patients who either worsen or improve according to their EDSS score. For comparisons based on EDSS score, the baseline score of the relevant measure was compared with the first score on or after the date on which the EDSS score meets the definition of worsening or improvement. The scores from baseline to improvement or worsening were compared with a paired t-test. The above analyses of worsening events (before relapse to during relapse and from baseline to EDSS worsening) were also broken down by treatment groups and by subgroups of the baseline score of the measure being analyzed. In addition to the above analysis, the cumulative distribution of changes was calculated. Thus for each integer change in the score from zero up to a 10-unit change, and thereafter in 5-unit change categories up to the maximum score observed on worsening, (by either relapse or EDSS), the number and percentage of patients (using all patients with scores before and after worsening as the denominator) who have changed by that score or less on worsening were calculated. A similar analysis was done for improvement, calculating the number and percentage of patients who have changed by that score or more on improvement.

SAP Part 9: Determination of a minimum clinically important difference: .this section describes Minimal Important Change (MIC) not MID. The minimum clinically important difference for SDMT, T25FW, 9HPT, and LCLA was determined by reference to the PCS score from the SF-36 (or SF-12). A 5-point change in the PCS score was considered a minimum clinically important difference. To this end, each of those variables was used in a linear regression model in which the change from baseline at endpoint in the PCS score was the dependent variable and the change from baseline at endpoint in the variable being investigated was the independent variable (the variables were taken at the same timepoint, so if only one variable was available at the end of the study, the latest time point where both variables were measured within 7 days of each other was used). That model was used to predict the change from baseline in each variable that corresponds to a change from baseline in PCS of 5 points, and was taken as the minimum clinically important difference. Norman et al (2003)[1] observed that an effect size of 0.5 is generally a good approximation to a clinically meaningful difference. The effect size corresponding to the minimum clinically important change, calculated as described above, was calculated using the standard deviation of baseline scores. An effect size close to 0.5 provides further confidence that the minimum clinically meaningful change has been defined.

[1]Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation. Medical Care. 2003 May;41(5):582–92.