Integrated Analysis Plan

Clinical Study Protocol Identification No.	MS200527_0080PPD
Title	A Phase III, Multicenter, Randomized, Parallel Group, Double Blind, Double Dummy, Active Controlled Study of Evobrutinib Compared with Teriflunomide, in Participants with Relapsing Multiple Sclerosis to Evaluate Efficacy and Safety
Study Phase	III
Investigational Medicinal Product(s)	Evobrutinib (M2951)
Clinical Study Protocol Version	Version 6.0 dated 26 (0080 PPD

Integrated Analysis Plan Author

Coordinating Author	
PPD , Merck	PPD
Function	**Author(s) / Data Analyst(s)**
PPD , PPD	PPD
PPD , PPD	PPD

Integrated Analysis Plan Date and Version: 20 November 2023/ Version 8.0

Integrated Analysis Plan Reviewers

Merck/EMD-Serono Reviewers:

Function	Name
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD
PPD	PPD

PPD PPD

PPD Reviewers:

Function	Name
PPD	PPD
PPD	PPD
Medical Responsible	PPD
PPD	PPD

Approval Page

Integrated Analysis Plan: MS200527_0080 PPD

A Phase III, Multicenter, Randomized, Parallel Group, Double Blind, Double Dummy, Active Controlled Study of Evobrutinib Compared with Teriflunomide, in Participants with Relapsing Multiple Sclerosis to Evaluate Efficacy and Safety.

Approval of the IAP by Merck Data Analysis Responsible is documented within the EDMS via eSignature. With the approval within EDMS, the Merck responsible for each of the analysis also takes responsibility that all reviewers' comments are addressed adequately.

By using eSignature, the signature will appear at the end of the document.

1 Table of Contents

Table of In-Text Tables

2 List of Abbreviations and Definition of Terms

9-HPT	9-hole peg test
AdaM	analysis data model
AE	adverse event
AEP	accelerated elimination procedure
AESI	adverse event of special interest
ALT	alanine aminotransferase
ARR	annualized relapse rate
AST	aspartate aminotransferase
ATC	anatomical therapeutic class
BID	twice per day
BLQ	below the limit of quantification
BMI	body mass index
BOA	biostatistics outputs assembly
BSSR	blinded sample size re-estimation
BTK	Bruton's tyrosine kinase
BUN	blood urea nitrogen
BV	brain volume
cCDP	confirmed disability as a composite score
CDI	confirmed disability improvement
CDISC	Clinical Data Interchange Standards Consortium
CDP	confirmed disability progression
CFB	change from Baseline
cIPD	clinically important protocol deviation
CMH	Cochran-Mantel-Haenszel
COVID-19	Coronavirus disease of 2019
CP	conditional power
CR	copy-reference
(e)CRF	(electronic) case report form
CSR	clinical study report
C-SSRS	Columbia-suicide severity rating scale

CUA	combined unique active
CV%	coefficient of variation
DBE	double-blind extension
DBP	diastolic blood pressure
DBTP	double-blind treatment period
DMT	disease modifying therapy
DRM	data review meeting
EAC	endpoint adjudication committee
EAIR	exposure adjusted incidence rate
ECG	electrocardiogram
e-DISH	evaluation of drug-induced serious hepatotoxicity
EDMS	electronic document management system
EDSS	expanded disability status scale
EEA	European Economic Area
EQ-5D-5L	EuroQoL 5 dimension 5 levels
EudraCT	European Union Drug Regulating Authorities Clinical Trials
FAS	full analysis set
Gd	gadolinium
GGT	gamma-glutamyl transferase
GLMM	generalized linear mixed model
HBV	hepatitis B virus
HR	hazard ratio
HRQOL	health related quality of life
HRU	health resource utilization
IA	interim analysis
IAP	integrated analysis plan
ICE	intercurrent event
ICH	International Conference on Harmonization
IDAC	independent data analysis center
IDMC	independent data monitoring committee
Ig	immunoglobulin
IPD	important protocol deviation

IQR	interquartile range
ITT	intent-to-treat
IWRS	interactive web response system
KM	Kaplan-Meier
LFT	liver function tests
LLR	lower limit of reporting
M&S	modelling and simulation
MAR	missing at random
MCID	minimal clinically important difference
MCS	mental component summary
Mean	arithmetic mean
MedDRA	Medical Dictionary for Regulatory Activities
mFAS	modified full analysis set
MI	multiple imputation
MMRM	mixed effect model for repeated measures
MN	Miettinen & Nurminen
MRI	magnetic resonance imaging
MS	multiple sclerosis
n	number of non-missing observations
NB	negative binomial
NCI-CTCAE	National Cancer Institute – Common Terminology Criteria for Adverse Events
NEDA	no evidence of disease activity
NEP	no evidence of progression
NEPAD	no evidence of progression or active disease
NfL	neurofilament light chain
nT1	normalized T1
PA	primary analysis
PBVC	percent brain volume change
PCR	polymerase chain reaction
PCS	physical component summary
PD	pharmacodynamics

PDBTKO	PD Bruton's tyrosine kinase occupancy
PDev	protocol deviation
PDMP	protocol deviations management plan
PF	physical function
PH	proportional-hazards
PIRA	progression independent of relapse activity
PIRMA	progression independent of relapse and brain MRI activities
PK	pharmacokinetics
PKSUB	PK substudy
PRO	patient reported outcome
PROMIS	patient reported outcomes measurement information system
PT	preferred term
Q1	first quartile, 25th percentile
Q3	third quartile, 75th percentile
QoL	quality of life
RMS	relapsing multiple sclerosis
ROW	rest of world
rRR	relapse rate ratio
RRMS	relapsing-remitting multiple sclerosis
SAE	serious adverse event
SAF	safety analysis set
SAP	statistical analysis plan
SAS	statistical analysis system
SBP	systolic blood pressure
SCR	screening analysis set
SD	standard deviation
SDG	standardized drug grouping
SDMT	symbol digit modalities test
SDTM	study data tabulation model
SEL	slowly expanding lesion
SF-36v2	medical outcomes study 36 item short form health survey
SI	System International

SMC	safety monitoring committee
SMQ	standardized MedDRA query
SOC	system organ class
SPMS	secondary progressive multiple sclerosis
STERM	study termination
T1 Gd+	T1 weighted gadolinium-enhancing
T25-FW	timed 25-foot walk
TEAE	treatment-emergent adverse event
TLF	tables, listings, and figures
TTERM	treatment termination
VAS	visual analog scale
WHO-DD	World Health Organization Drug Dictionary

3 Modification History

Unique Identifier for Version	Date of IAP Version	Author	Changes from the Previous Version
1.0	20 January 2021	PPD	Initial Version
2.0	04 August 2022	PPD	**Updates regarding Protocol Versions 2 and 3:** New exploratory endpoints (PRL and CBF [Sections 5 and 14.3]) Removed the IA for unblinded sample size re-estimation and include optional IA for BSSR with details regarding methodology [Sections 6, 9.13 and 14] **Updates regarding Protocol Version 4:** Revision of the treatment duration from fixed 96 weeks to variable duration up to 156 weeks, and removal of OLE period (throughout the document). Revised hierarchy of secondary endpoints in the primary analysis (time to first occurrence of CDI and Week 12 NfL concentration new secondary endpoints [Section 5 and Section 14.2], and new strategy to control the Type I error at study and pooling level [Section 9.13]). Modification of secondary PROMIS endpoints from CFB "at 96 weeks to CFB "over 96 weeks" and modification of population level-summary to difference of average least-squares means, with average taken over a set of timepoints ending with Week 96 [Section 5 and Section 14.2.3]. Modification of secondary MRI endpoints to account for all available MRI scans [Section 5 and Section 14.2.4]. Inclusion of intercurrent event "Ukraine crisis" and strategy for handling in all primary and secondary endpoints. Summarize the outcome of the optional IA for BSSR and information for ARR provided to adequately power the endpoint with corresponding trigger for PA [Section 6]. Inclusion of modified analysis sets (mFAS and mSAF) [Section 8]. Population PK/PD modeling section revised to specify that all analyses will be described in a pharmacometric modeling analysis plan and reported separately [Section 16.3]. **Updates regarding draft Protocol Version 5:** Addition of exploratory endpoints time to PIRA, time to PIRMA, NEP at Week 48 or 96 and level of anti-SARS-CoV-2 antibodies [Sections 14.3 and 15.13]. **Added details to sensitivity and supplementary analyses for primary and secondary efficacy endpoints as well as safety endpoints** [Sections 14.1.2, 14.2, 15.1.1.4, Appendix 18.9]. **Addition of exploratory endpoints** time to confirmed PROMIS PF improvement [Section 16.4.5] and CFB in PROMIS Fatigue score independent of relapse activity [Section 16.4.6]. **Addition of** LTF analyses [Section 15.3]. **Throughout the document:** Updates for consistency and clarification.

Unique Identifier for Version	Date of IAP Version	Author	Changes from the Previous Version
3.0	16 February 2023	PPD	**Updates regarding Protocol Version 5:** OLE treatment period added back (removed following Protocol Version #4) [Sections 5 and 6]. **Updates regarding FDA feedback on IAP Version 2.0:** Change in definition of participants at risk for time to event analyses, all participants from mFAS are now included [Sections 14.2.1, 14.2.2, 14.3.4, 14.3.5, 14.3.14]. Removal of the AE sensitivity analysis evaluating the potential impact of COVID-19 related ICEs [Section 15.1.1.4]. Change in the derivation of COVID-19 vaccine related AEs [Section 15.2.4]. **Other updates:** Added Section 9.11 to define the analysis study day for Week 48 endpoints. Updated DMD to be DMT [Section 11.3.3 and throughout document]. Aligned S1/S2 IgG antibodies to SARS-CoV-2 analysis with the 0086 study analysis and added additional summaries [Section 15.13]. Updated the description for censoring due to the ICE of 'Ukraine crisis' to differentiate between censoring for mFAS and exclusion from mFAS [Sections 8.1 and 14]. Updated sensitivity analysis 1 to apply a range of delta values [Sections 14.1.2, 14.2.1.3, 14.2.2, 14.2.3.3 and 14.2.4.3]. Changed EAIR to be presented in 100 participant years, from 1000 participant years [Section 15.1.1.3]. Added median line plots for immunoglobulin parameters [Section 15.8]. Added analyses up to Week 120 for NfL concentration [Section 16.2.2]. Updated abbreviations etc to follow the latest Global Style Guide [throughout the document].
4.0	13 July 2023	PPD	**Updates regarding Protocol Version 6:** Updated PA trigger description [Sections 6.2 and 7] Added DBE analysis [Sections 6.0 and 6.3] Updated the definition of on-treatment period to consider only the DBTP [Section 9.10] The overview table summarizing the results of the multi-stage testing procedure was updated to include estimates [Section 9.16] Removed reference to long-term follow-up study and added DBE [Section 10.1] Updated definition of a treatment completer for the PA [Section 10.1] Update to Tier 1 AEs definition, to reflect the recent update to the risk profile of evobrutinib (i.e. important identified risk of drug-induced liver injury) [Section 15.1.1.1].

Unique Identifier for Version	Date of IAP Version	Author	Changes from the Previous Version
			Criteria for temporary treatment discontinuation and rechallenge was updated [Section 15.3] Replaced "PA trigger" with "PA cutoff" [Throughout] **Updates due to FDA feedback:** Added additional sensitivity analysis to censor data at the originally projected date of the PA trigger, for ARR and 12-week CDP endpoints [Sections 14.1.2 and 14.2.1.3] **Other updates:** Corrected the PDBTKO analysis set to include only participants who received at least 1 dose of evobrutinib [Section 8.1] Data handling after cutoff was updated to follow the current SDTM process [Section 9.3] Study day definition was clarified, derived from start of study intervention [Section 9.5] Updated exposure and compliance calculations to use dose (mg) instead of number of tablets. [Section 13] Updated descriptive statistics for MRI lesion measures to be by timepoint only [Sections 14.2.4.1 and 14.2.4.2] Added repeats of composite endpoints on all relapses [Sections 14.3.15, 14.3.18, 14.3.20, 14.3.21 and 16.4.8] Added details for ICE handling in exploratory endpoint sections [Sections 14.3 and 16.4] Removed 12-week CDI status at Week 96 [Sections 14.3.5, 5 and 7] Added analysis to evaluate clinically meaningful improvement in SDMT [Section 14.3.19] PIRA up to Week 156 was updated to focus on qualified relapses, instead of investigator reported relapses [Section 14.3.30] Repeat lab summary by visit on only central lab data and updated paragraph on boxplots for laboratory parameters [Section 15.3] Description of presentation of summary statistics of PK parameter data added [Section 9.2] A flag will be added for concomitant administration of proton pump inhibitors, H2 blockers, and antacids [Section 16.1.1] Additional details regarding handling of duplicate PK samples included [Section 16.1.1] Visit window information added for PK collections [Section 16.1.1] Estimation of PK parameters for evobrutinib and MSC2729909A from Day 1 and Week 4 of PK substudy added [Section 16.1.2] Description of the presentation of PK parameters in outputs added [Section 16.1.3] **Throughout the document:** Updates for consistency and clarification.
5.0	28 September 2023	PPD	**8.1** - Update to definition of fully operational status for sites in Ukraine

Unique Identifier for Version	Date of IAP Version	Author	Changes from the Previous Version
6.0	31 October 2023	PPD	Updated the source of baseline EDSS to be the vendor data, not the eCRF [Section 8.2 and throughout] Added baseline tables on subset of subjects with baseline EDSS ≥2 [Section 11.1 and 11.3.1] Added Section 15.14 on Focused Genetic Testing Added steps in case the primary model fails to converge [Section 14.1.1] Total Carbon Dioxide worst on-treatment direction was corrected [Section 18.6] **Throughout the document:** Updates for consistency and clarification.
7.0	8 November 2023	PPD	Actualized date of PA trigger 02 October 2023 (Monday) from 01 October 2023 (Sunday) [Section 7] Added race and ethnicity in subgroups [Section 8.2] Clarified steps in case the primary model fails to converge [Section 14.1.1] Corrected inconsistency [Section 14.2.5]
8.0	20 November 2023	PPD	Updated the definition of FAS and mFAS [Section 8.1] Updated FAS as primary analysis population throughout the efficacy section [Section 14] Changed ITT sensitivity analysis to mFAS sensitivity analysis throughout the efficacy section [Section 14] The updates listed above were made in accordance with the FDA request from 13/15 November 2023 to conduct the primary analysis in all randomized participants (i.e. the intention to treat population), and that pre-specified sensitivity analyses be conducted in other defined analysis populations to evaluate the impact of the stated concerns (e.g., Ukraine Conflict, protocol violations, etc.).

4 Purpose of the Integrated Analysis Plan

The purpose of this IAP is to document technical and detailed specifications for the primary analysis of data collected for protocols MS200527_0080 and PPD. Results of the analyses described in this IAP will be included either in the CSR or separate reports. Additionally, the planned analyses identified in this IAP may be included in regulatory submissions or future manuscripts. Any post-hoc, or unplanned analyses performed to provide results for inclusion in the CSR but not identified in this prospective IAP will be clearly identified in the CSR.

This IAP is based upon Section 9 (Statistical Considerations) of the study protocols and protocol amendments and is prepared in compliance with ICH E9. It describes analyses planned in the protocols and protocols amendments.

5 Objectives and Estimands

Primary and secondary objectives and estimands for the DBTP are summarized in Table 1. Tertiary/exploratory objectives and endpoints for the DBTP are summarized in Table 2.

Objectives and endpoints for the DBE and the OLE periods will be specified in separate IAPs. The DBE IAP will be finalized before the primary analysis database lock.

Table 1 Objectives and Estimands for the DBTP:

Objectives	Estimand Attributes	IAP section
Primary		
To demonstrate superior efficacy with evobrutinib compared to teriflunomide in terms of ARR	**Endpoint:** ARR based on qualified relapses up to 156 weeks in participants with RMS **Population:** Patients with RMS as defined by inclusion/exclusion criteria **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death attributable to MS or treatment: composite variable • Death unattributable to MS or treatment: while alive • Ukraine crisis: hypothetical **Population-Level Summary:** rRR and CI ([NB model), with test of treatment effect based on test of rRR	14.1.1
Secondary		
To demonstrate the efficacy of evobrutinib relative to that of teriflunomide on disability progression	**Endpoints:** • Time to first occurrence of 12-week CDP as measured by the EDSS up to 156 weeks • Time to first occurrence of 24-week CDP as measured by the EDSS up to 156 weeks **Population:** see Primary **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death attributable to MS or treatment: composite variable • Death unattributable to MS or treatment: while alive • Ukraine crisis: hypothetical **Population-Level Summary:** hazard ratio and CI based on a Cox model for progression hazard, with test of treatment effect based on logrank test	14.2.1
To demonstrate the efficacy of evobrutinib relative to that of teriflunomide on disability improvement	**Endpoints:** • Time to first occurrence of 24-week CDI as measured by the EDSS up to 156 weeks **Population:** Patients with RMS as defined by inclusion/exclusion criteria and who have baseline EDSS ≥ 2.0 **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death (any cause): while alive • Ukraine crisis: hypothetical **Population-Level Summary:** hazard ratio and CI based on a Cox model, with test of treatment effect based on logrank test	14.2.2

Objectives	Estimand Attributes	IAP section
To demonstrate the efficacy of evobrutinib relative to that of teriflunomide on patient reported symptoms and functional status	**Endpoints:** • CFB in PROMIS PF score over 96 weeks • CFB in PROMIS Fatigue score over 96 weeks **Population:** see Primary **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death (any cause): while alive • Ukraine crisis: hypothetical **Population-Level Summary:** difference of average least-squares means of score CFB (average over weeks 72, 84 and 96 for PROMIS PF and average over weeks 48, 60, 72, 84 and 96 for PROMIS Fatigue) and CI based on a MMRM, with test of treatment effect based on difference of average least-squares means	14.2.3
To demonstrate the efficacy of evobrutinib relative to that of teriflunomide on MRI lesion parameters	**Endpoints:** • Total number of T1 Gd+ lesions based on all available MRI scans • Number of new or enlarging T2 lesions on the last available MRI scan relative to the baseline MRI scan **Population:** see Primary **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death (any cause): while alive • Ukraine crisis: hypothetical **Population-Level Summary:** lesion rate ratio and CI based on a NB model, with test of treatment effect based on test of lesion rate ratio	14.2.4
To demonstrate the efficacy of evobrutinib relative to that of teriflunomide by evaluating response on NfL concentration in serum	**Endpoints:** • NfL concentration at 12 weeks **Population:** see Primary **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Treatment discontinuation: treatment policy • Death (any cause): while alive • Ukraine crisis: hypothetical **Population-Level Summary:** ratio of geometric means at 12 weeks and CI based on a MMRM, with test of treatment effect based on difference of least-squares means	14.2.5
To characterize the safety and tolerability of evobrutinib.	Safety as assessed by the nature, severity, and occurrence of AEs and AESIs; vital signs; ECGs; absolute concentrations and CFB in Ig levels; and clinical laboratory safety parameters up to the end of the Safety Follow-up period **Population:** see Primary **Treatment:** Evobrutinib vs. teriflunomide **Intercurrent Event Strategy:** • Ukraine crisis: hypothetical **Population-Level Summary:** not applicable	15

Table 2 Objectives and Endpoints for the DBTP:

Objectives	Endpoints (Outcome Measures)	IAP section
Tertiary/Exploratory		
To evaluate the effect of evobrutinib compared to teriflunomide on clinical parameters	• ARR based on qualified relapses at Weeks 48 and 96 • Time to first qualified relapse over 156 weeks • Qualifying relapse-free status at Week 96 • 12-week confirmed EDSS progression free status at Week 96 • 24-week confirmed EDSS progression free status at Week 96 • 24-week confirmed disability improvement status at Week 96	14.3
To evaluate the efficacy of evobrutinib relative to that of teriflunomide on MRI parameters	• T1 Gd+ lesion free status at Week 96 based on assessments up to Week 96 • Change in volume of T1 Gd+ lesions from Baseline to Week 96 based on assessments up to Week 96 • New or enlarging T2 lesion free status at Week 96 based on assessments up to Week 96 • Change in volume of T2 lesions from Baseline to Week 96 based on assessments up to Week 96 • CUA lesion free status at Week 96 based on assessments up to Week 96 • Total number of CUA lesions based on all available MRI scans • Total number of new T1 hypo-intense lesions based on all available MRI scans • Percentage change BV from Week 24 to Week 96 based on assessments up to Week 96 • Percentage change in thalamic volume from Week 24 to Week 96 based on assessments up to Week 96 • Percentage change in cortical grey matter volume from Week 24 to Week 96 based on assessments up to Week 96 • Change in normalized T1 intensity within pre-existing nonenhancing T2 weighted lesion volume from Baseline to Week 96 based on assessments up to Week 96 • Volume of SELs based on scans at Weeks 24, 48 and 96 • Change in number of PRL at Weeks 24, 48 and 96 • Change in Cerebral Blood Flow in normal-appearing white matter at Weeks 24, 48 and 96	14.3
To evaluate the effect of evobrutinib compared to teriflunomide on functional parameters	• Time to ≥ 20% increase (confirmed at 12 weeks) in T25-FW up to 156 weeks • Time to ≥ 20% increase (confirmed at 12 weeks) in 9-HPT up to 156 weeks	14.3.14

Objectives	Endpoints (Outcome Measures)	IAP section
To evaluate the effect of evobrutinib compared to teriflunomide on composite parameters	• NEDA-3 at Week 48, Week 96 defined by: o Qualifying relapse-free status o T1 Gd+ lesions free status and new or enlarging T2 lesion free status o 12-week confirmed EDSS progression free status • Time to first occurrence of 12-week CDP up to 156 weeks based on a composite score defined by: o 12-week confirmed EDSS progression (at least 0.5- or 1.0- point change, depending on the baseline EDSS) or; o 12-week confirmed worsening (≥ 20%) in T25-FW versus Baseline or; o 12-week confirmed worsening (≥ 20%) in 9-HPT versus Baseline. • NEP at Week 48, Week 96 as defined by: o No 12-week CDP on EDSS o No 12-week confirmed worsening (≥ 20%) in 9-HPT score o No 12-week confirmed worsening (≥ 20%) in T25-FW time • NEPAD at Week 48, Week 96 as defined by: o No protocol-defined relapses on treatment o No 12-week CDP on EDSS o No 12-week confirmed worsening (≥ 20%) in 9-HPT score o No 12-week confirmed worsening (≥ 20%) in T25-FW time o No new or enlarging T2 lesions and no T1 Gd+ lesions on MRI	14.3
To evaluate the efficacy of evobrutinib relative to that of teriflunomide on progression independent of relapse activity	• Time to PIRA, where disability progression is defined for each of 2 endpoints as follows: o 12-week CDP on EDSS o 12-week confirmed worsening as a composite of 3 endpoints (EDSS, 9-HPT, T25-FW)	14.3.20
To evaluate the efficacy of evobrutinib relative to that of teriflunomide on progression independent of relapse and brain MRI activity	• Time to PIRMA, where disability progression is defined for each of 2 endpoints as follows: o 12-week CDP on EDSS o 12-week confirmed worsening as a composite of the 3 endpoints (EDSS, 9-HPT, T25-FW)	14.3.21
To evaluate the efficacy of evobrutinib relative to that of teriflunomide on cognitive function	• Change from baseline SDMT score at Week 48, Week 96	14.3.19

Objectives	Endpoints (Outcome Measures)	IAP section
To evaluate the efficacy of evobrutinib relative to that of teriflunomide on patient reported symptoms and functional status	• Change from baseline in PROMIS PF score at Week 48, Week 96, and Week 144 • Change from baseline in PROMIS fatigue score at Week 48, Week 96, and Week 144 • Time to first occurrence of 12-week confirmed PF deterioration compared to baseline (decrease of at least 2.7 points on PROMIS PF score) up to 156 weeks • Change from baseline in Patient Reported Outcome scores at Week 48, Week 96 and Week 144: o SF-36v2 o EQ-5D-5L	16.4
To evaluate the effect of evobrutinib on HRU relative to that of teriflunomide over 96 weeks	Absolute values of HRU, including but not limited to doctor/home/emergency visits, hospitalizations, paid assistance, and missed work	16.5
To characterize the PK profile of evobrutinib in participants with MS and to describe the exposure-response relationship between evobrutinib and efficacy endpoints and safety endpoints	PK parameters: CL/F, Vz/f, Cmax, and AUC after a single dose (Day 1 data), and at steady state (Weeks 2, 4, 8, 12, 24, 48, 72, and 96 data)	16.1, 16.3
To describe the exposure-response relationship between evobrutinib and PD biomarkers	PD biomarker endpoints at Baseline/Screening, Weeks 2, 4, 8, 12, and 24	16.2, 16.3
To assess relationship between candidate disease biomarker and disease activity or treatment response	• Level of biomarkers of disease with disease activity/treatment response at Baseline, Weeks 12, 24, 48, 72, 96; and upon relapse/disease progression (unscheduled) • NfL concentration at Weeks 24, 48, 72, and 96	16.2
To measure the antibody levels to SARS-CoV-2 vaccine administered in participants	Level of anti-SARS-CoV-2 antibodies to vaccines pre- and post-vaccination in a subset of participants who have received anti-COVID vaccinations during the study	15.13
To evaluate the relationship of the novel biomarkers of liver function e.g. protein, miRNA, etc. compared to standard clinical chemistry endpoints (ALT)	Levels of novel biomarkers of hepatic function compared to ALT at Baseline, and on treatment	NA – analyses will be conducted post-PA and be described in a separate IAP
To assess the effect of evobrutinib on gene expression in whole blood	Gene expression at Baseline, Weeks 12, 24, 48, and 96	16.2.3

6 Overview of Planned Analyses

4 analyses performed by the Sponsor are planned for the studies: (1) an optional IA for BSSR based on the secondary endpoint 12-week CDP, with data pooled across studies 0080 and PPD, triggered when enrollment is nearing completion, (2) a PA with timing and endpoint evaluation as

described in Section 6.2, (3) an analysis of all data from blinded periods at the end of the DBE Period, performed after the corresponding DBE database lock, and (4) a final analysis, triggered when 100% of participants enrolled in the OLE Period complete, or discontinue prematurely from the OLE Period.

Details of the IDMC analyses will be described in the IDMC SAP (See Appendix 18.10) and the IDMC Charter and related documentation.

6.1 Optional Interim Analysis for Blinded Sample Size Re-estimation

Per Protocol Version 3, when enrollment was nearing completion, an optional IA for BSSR based on the secondary endpoint 12-week CDP was performed by the Sponsor, using blinded data pooled across studies. Only data relevant to 12-week CDP were cleaned for the IA.

Parametric distributions for time to 12-week CDP and time to censoring were fitted using blinded data pooled across studies. The resulting parameter estimates were used to estimate expected number of events given the originally planned sample size. If the expected number of events was less than the number required to power the 12-week CDP endpoint at 80%, consideration was to be given to increasing the sample size up to 25% so that the required number of events was achieved (Friede 2019).

Performed in August 2021, the outcome of this IA for BSSR was a decision to not increase sample size. In the setting of superiority trials, an IA for BSSR has a minimal or non-existent effect on type-1 error inflation (Friede 2019). The analysis datasets required for the interim analysis of 12-week CDP were specified in the same way as the analysis datasets required for the primary analysis of 12 week CDP (Section 14.2.1.1).

Prior to the IA, a detailed BSSR SAP was provided.

6.2 Primary Analysis

Per protocol, the Primary Analysis will be triggered for both studies simultaneously when all of the following conditions are met, based on an analysis of accumulating blinded data:

- The 2 Phase 3 studies have individually collected ≥ 32.3 units of information for the primary ARR endpoint, where "Information" is defined in Protocol Section 9.2.1, Equation 9.2.1. At this timepoint both studies will provide ≥ 90% power for the detection of a 43.5% relative treatment effect on ARR in the primary statistical test at a 1-sided alpha-level of 0.025.
- In data pooled from the 2 studies, ≥ 149 12-week CDP events have been observed. At this timepoint, the combined studies have collected sufficient information to provide ≥ 71.5% power for the detection of 34% (hazard ratio = 0.66) relative reduction in risk of 12-week CDP in a log-rank test at a 1-sided alpha-level of $0.025\text{-}0.025^2 = 0.024375$.
- All participants have been treated for 24 weeks or discontinued prematurely.

After protocol deviations are determined, and the database is locked for the PA, the drug codes will be broken and made available for the primary data analysis. All endpoints based on data from the DBTP (see Section 9.10) will be evaluated.

6.3 DBE Analysis

The DBE analysis will occur at the end of the DBE period, when the protocol deviations are determined, and the database is locked for the DBE period.

6.4 Final Analysis

The final analysis will occur only when 100% of participants enrolled in the OLE complete the OLE, or discontinue prematurely from the OLE, the protocol deviations are determined, and the database is locked for the final analysis.

7 Changes to the Planned Analyses in the Clinical Study Protocol

- Primary Analysis:

In light of a newly identified drug induced liver injury, it appears essential to balance this with the need to have the statistical power to better elucidate the benefit-risk assessment. The current review of blinded event rates for 12-week CDP showed that 149 events (the minimum number required by the protocol) across the two studies were reached on 01 June 2023. To increase the power to 75-80% for the key disability endpoint, 12-week CDP, the PA trigger was shifted to 02 October 2023. This decision was made without unblinding any data of the two Phase 3 studies and was communicated to the FDA in May 2023.

- Analysis sets and subgroups

In the protocol, it was defined that the mFAS would be used for the PA of efficacy. In accordance with the FDA request from 13/15 November 2023 to conduct the primary analysis in all randomized participants, it was decided that FAS defined as all randomized participants will be used for the PA of efficacy and mFAS will be used for sensitivity analyses of primary and secondary efficacy endpoints. In that context, the definition of mFAS was revised to exclude participants with certain cIPDs related to eligibility violations, randomization in error or failure to receive study intervention, per ICH E9. In addition, the mFAS excludes participants from sites lacking data robustness throughout the study, regardless of their country. This was a change from the definition in the Protocol, which stated that only participants from Ukraine, Russian Federation, and Belarus lacking data robustness were to be excluded. The mSAF analysis set will not be defined as no safety data robustness issues have been identified that would warrant exclusion of a SAF participant from the primary analysis of safety.

The PDBTKO analysis set has been defined for all analyses related to BTKO data.

Ethnic origin is listed as a subgroup in the Protocol but will not be included as a subgroup in the IAP.

- CDP derivation

CDP will be derived based on EDSS scores only. Assessment whether the increase in EDSS score is not attributable to another etiology will be disregarded in the CDP derivation because this information is not collected in the eCRF.

- CDI analysis

As Time to 12-week CDI is not a relevant endpoint and not included in the protocol, the exploratory endpoint "12-week confirmed disability improvement status at Week 96" will not be analyzed as part of the PA.

- MRI distribution analysis

The supplementary analysis model for MRI lesion measures is described as Poisson-distributed in the protocol. Due to the large sample size, negative binomial modelling is more appropriate, and the modelling details are described in Section 14.2.4.3.

- AEs severity

The severity and grading criteria were removed for the AEs of special interest: opportunistic infections, by the Safety Team. This is reflected in an updated file, refer to Appendix 18.5.

8 Analysis Sets and Subgroups

8.1 Definition of Analysis Sets

The analysis sets are specified in Table 3 below. The final decision to exclude participants from any analysis set will be made during a blinded data review meeting prior to database lock and unblinding. Exclusion from PK set will be performed after unblinding post-database lock as only participants who received at least one dose of evobrutinib are considered.

Depending on the endpoints, the analysis set will be defined at study level (i.e. for each 0080 and PPD separately) and/or at the pooled level.

Table 3 Analysis Sets

Analysis Set	Description
Screening (SCR)	All participants who provided informed consent, regardless of the participant's randomization and study intervention status in the study.
Full Analysis Set (FAS)	All participants who were randomized to study intervention. Participants will be analyzed per the intervention group to which they were randomized (i.e. intention-to-treat principle). This analysis set will be used for the primary analysis of efficacy.
Modified FAS (mFAS)	Per ICH E9, all participants in FAS except participants • with eligibility violations • who are screen failures randomized in error • who did not receive study intervention • from sites lacking data robustness (i.e. inability to verify data source) throughout the course of the study This analysis set will be used for sensitivity analyses of primary and secondary efficacy endpoints.
Safety (SAF)	All participants who received at least one dose of study intervention. Participants will be analyzed per the actual study intervention they received. This analysis set will be used for all analyses of safety.
Modified Safety (mSAF)	Not defined. While the protocol allows for the definition of mSAF, no safety data robustness issues have been identified that would warrant exclusion of a SAF participant from the primary analysis of safety.
PK	All participants who have received at least 1 dose of evobrutinib and have at least 1 quantifiable evobrutinib and/or metabolite plasma concentration at a scheduled PK time point postdose without any important deviations or events that may impact the quality of the data or alter the evaluation of PK. Participants who receive active control will not be included. Participants will be analyzed per the actual study intervention they received.
PD	All participants who have received at least 1 dose of evobrutinib or active control, have a predose assessment, and at least one postdose assessment, without deviations or important events affecting PD, and who provide evaluable PD data. Participants will be analyzed per the actual study intervention they received. This analysis set is only applicable for the MS200527_0080 study.
PDBTKO	All participants who have received at least 1 dose of evobrutinib, have a predose assessment, and have at least 1 measured BTK occupancy value at a scheduled PD time point post baseline without deviation or important events affecting PD, and who provide evaluable PD data. This analysis set is only applicable for the MS200527_0080 study.
PKSUB	A subset of participants from the PK analysis set who are enrolled at sites selected for the PK substudy. This analysis set is only applicable for the MS200527_0080 study.

It is anticipated that a subset of sites in Ukraine, Russian Federation and Belarus will not remain fully operational post start of Ukraine crisis. For participants from these sites (considering site at time of randomization and new site in case of transfer)), the ICE "Ukraine crisis" will be handled via the Hypothetical strategy, wherein data post start of crisis will be censored in the analysis of efficacy (see Section 14).

"Fully operational" is defined with the following criteria:1. Sites open for participant onsite visits, 2. Sites with no issues with study intervention dispensation/resupply, 3. Sites with no issues with collecting/processing biological samples either centrally or locally. If there is a single month after

the start of the crisis during which a site's status is not fully operational, then data from participants at that site will be assumed to lack robustness starting with the date at which the site lost fully operational status.

A site is considered to lack data robustness throughout the study if source data verification cannot be performed. Participants from such sites are excluded from the mFAS.

Sites with lack of data robustness throughout the study (i.e. excluded from mFAS) and not continuously operational post start of the crisis (i.e. censored in primary efficacy analyses) will be identified before DBL and unblinding by conducting a blinded longitudinal analysis of site operational status. The list of the corresponding sites will be included in an IAP Appendix prior to DBL.

PDs leading to exclusion from the mFAS analysis set will be encoded as cIPDs in the database. See Section 10.2 for details.

For the PDBTKO analysis set, assay deviation for BTKO data will be identified in the raw data received from the vendors.

Actual Treatment Assignment

A participant who received 2 different types of study intervention regimen over the course of study intervention should be tabulated according to the study intervention regimen received most frequently (based on number of kits received). If there is a "tie", then evobrutinib will be chosen.

The following Table 4 summarizes the use of the analysis sets in the different analyses.

Table 4 Use of Analysis Sets

	Analysis Set						
Analyses	**Full Analysis Set**	**modified FAS (mFAS)**	**Safety**	**PK**	**PD**	**PDBTKO**	**PKSUB**
Baseline Characteristics*	✓						
Previous and Concomitant Therapies			✓				
Compliance and Exposure			✓				
Efficacy: Primary	✓	✓					
Efficacy: Secondary	✓	✓					
Efficacy: Tertiary/Exploratory	✓						
Safety and Tolerability			✓				
Pharmacokinetics				✓			✓
Pharmacodynamics					✓	✓	
Biomarkers					✓		
PRO	✓						

*Presented on both FAS/mFAS as well as SAF if there is a difference in the analysis sets greater or equal to 10%.

8.2 Subgroup Definition and Parameterization

Subgroup analyses will be performed on the primary estimand ARR and on the secondary estimands time to 12-week CDP, time to 24-week CDP and time to 24-week CDI as defined below. All subgroup analyses will be exploratory, no adjustment for multiplicity will be performed.

For the definition of subgroup level, data as documented in the eCRF or vendor data will be taken. The category "missing" will not be included in any subgroup analysis.

The following subgroups will be defined and re-grouping determined at time of analysis in the event of too few participants in some levels:

- Baseline EDSS (vendor data):
  - EDSS < 4.0 (reference level)
  - EDSS ≥ 4.0
- Region (eCRF):
  - North America
  - Western Europe (reference level)
  - Eastern Europe
  - ROW
- Sex:
  - Male (reference level)
  - Female
- Age:
  - Age < 40 years (reference level)
  - Age ≥ 40 years
- Race:
  - White (reference level),
  - Black or African American,
  - Asian,
  - American Indian or Alaska Native,
  - Native Hawaiian or Other Pacific Islander

Participants with multiple races, race not collected at site, or 'Other' will not be included in the subgroup analysis. Categories with 5 participants or less in one or both treatment arms will not be included in the subgroup analysis. For this subgroup, the stratum Region will not be included in any modeling.

- Ethnicity:
  - Hispanic or Latino,
  - Not Hispanic or Latino (reference level)

For this subgroup, the stratum Region will not be included in any modeling.

- Number of relapses in previous 2 years before Screening:
  - ≤ 2 (reference level)
  - > 2
- Presence of T1 Gd+ lesions within 6 months prior to randomization:
  - 0 (reference level)
  - ≥ 1
- Any DMT before study entry:
  - Yes (reference level)
  - No
- T2 lesion number at Baseline:
  - ≥ 9 (reference level)
  - < 9
- T2 lesion volume at Baseline:
  - ≤ median (reference level)
  - > median
- Time (years) since initial MS diagnosis date (derived based on informed consent date):
  - ≤ 3 (reference level)
  - > 3 - ≤ 10
  - > 10
- NfL concentration at Baseline:
  - ≤ median (reference level)
  - > median
- Membership in initial or second recruitment cohort:
  - Initial cohort (reference level)
  - Second cohort

There were 2 recruitment cohorts during study conduct. The first cohort was recruited July 2020 – October 2021. The second cohort was recruited starting in July 2022. This subgroup analysis will

be defined if the number of participants recruited in the second recruitment cohort represents at least 10% of the FAS participants.

9 General Specifications for Data Analyses

This section describes any general specifications not included in subsequent sections.

9.1 Study intervention

Study interventions in the double blind, double dummy period are defined as indicated in Table 5:

Table 5 Study Interventions and Regimens

	Regimen[a]	Study Intervention Presentation
1	Evobrutinib 45 mg BID	Evobrutinib
2	Teriflunomide 14 mg QD	Teriflunomide

[a] The study will have a double dummy design.

9.2 Presentation of continuous and qualitative variables

Continuous variables will be summarized using the following descriptive statistics:

- number and percentage of participants with non-missing values
- number and percentage of participants with missing values
- mean, SD
- median, 25th Percentile - 75th Percentile (Q1-Q3)
- minimum, and maximum

The number of digits for non-derived and derived data, presented in outputs or available in ADaM datasets, is specified in the BOA document. For efficacy endpoints from Section 14, median and SD will be presented with 1 more digit compared to the original data, whereas mean, min, max, Q1-Q3 will be presented with the same number of digits as the original data.

For both continuous and qualitative variables, percentages such as 0% or 100% should be reported with the same format used for the column, together with the count of observations. For example, if the count of observations is zero, then display '0 (0.0)'; if the count of observations is 100% then display 'xx (100.0)'.

Qualitative variables will be summarized by counts and percentages. The "Missing" category should always be displayed at Baseline – even when there are no missing data at Baseline. At timepoints other than Baseline, the "Missing" category should only be displayed when there are missing data.

Unless otherwise stated, the calculation of proportions will be based on the number of participants in the analysis set of interest. Therefore, counts of missing observations will be included in the denominator and presented as a separate category.

The total of missing and non-missing observations at each timepoint will reflect the population still in the study at that time. For example, if a participant is still in the study at the timepoint but with missing data, that participant should be counted in the number of missing observations.

Visits will be presented in by-visit descriptive summaries if there are at least 10 participants still in the study at that timepoint.

Presentation of PK Concentration Data

Pharmacokinetic concentration data will be descriptively summarized using: number of n, mean, SD, CV%, minimum, median, and maximum.

Descriptive statistics will only be calculated for $n > 2$ in which a measurement of BLQ represents a valid measurement and will be taken as zero for summary statistics. In $n \leq 2$, only n, minimum, and maximum will be presented in summary tables.

Descriptive statistics of PK concentration data will be calculated using values with the same precision as the source data and rounded for reporting purposes only. PK concentrations will be carried over with full precision as provided in the source data without any rounding applied to CDISC SDTM PC and ADaM PC domains.

The following conventions will be applied when reporting descriptive statistics of PK concentration data:

n:	0 decimal place
mean, minimum, median, maximum:	3 significant digits
SD:	4 significant digits
CV%:	1 decimal place

Presentation of PK Parameter Data

PK parameter data will be descriptively summarized using: n, mean, SD, CV%, minimum, median, maximum, geometric mean, the geometric coefficient of variation and the 95% confidence interval for the geometric mean. For PK parameters related to time (e.g. t_{max}), only n, minimum, median, and maximum may be reported.

Descriptive statistics will only be calculated for a PK parameter when $n > 2$. In case $n \leq 2$, individual data will be presented (minimum, maximum) in summary tables.

PK parameters read directly from the measurements (i.e. C_{max}) will be reported with the same precision as the source data. All other PK parameters will be reported to 3 significant figures. Descriptive statistics of PK parameter data will be calculated using full precision and rounded for reporting purposes only.

PK parameters will be provided with full precision, without any rounding applied to CDISC SDTM PP and ADaM PP domains.

The following conventions will be applied when reporting descriptive statistics of PK parameter data:

n	0 decimal place
mean, minimum, median, maximum, geometric mean, 95% CI:	3 significant digits
SD:	4 significant digits
CV%, geometric CV%:	1 decimal place

9.3 Data Handling After Cutoff Date

For the PA, the trigger date will be used as the cutoff date.

Data obtained after the cutoff date will not be displayed in any listings or used for summary statistics, i.e. laboratory values of samples taken after data cutoff, AEs with onset date after data cutoff, etc. will not be included in any analysis or listing.

The cutoff date will be applied at the SDTM level (refer to *SDTM Cut-off date implementation rules_BTKi(M2951)-Compound_v3.0.docx*).

9.4 Definition of Baseline and Change from Baseline

Definition of Baseline

Unless otherwise specified, baseline is defined as the last non-missing value on the day of or prior to first administration of study intervention in the DBTP. Of note, time of first dosing administration and assessment of baseline (both done on the same Baseline day) will not be used in the determination of baseline value, except for ECGs.

If both central and local labs are collected, the baseline will be derived based only on the central lab collected data. Only central read ECGs data will be used to derive baseline as well as subsequent visits.

Definition of CFB

CFB and percent CFB at a given post Baseline Visit will be computed as follows:

- CFB = visit value – baseline value
- Percent CFB = 100 * (visit value – baseline value) / baseline value

At the Baseline Visit, the CFB will be equal to zero and the percent CFB will be missing.

9.5 Study Day

Day 1 is the day of start of study intervention, the day before is Day -1 (no Day 0 is defined). Study day is defined relative to Day 1.

9.6 Definition of Duration and 'time since' Variables

Duration in days will be calculated as the difference between start and stop dates plus 1 (e.g. AE duration (days) = AE end date - AE start date + 1).

The time since an event (e.g. time since first diagnosis) will be calculated as reference date minus date of event.

9.7 Conversion Factors

The following conversion factors will be used to convert days into months or years:

1 week = 7 days, 1 month = 30.4375 days, 1 year = 365.25 days.

The following conversion factors will be used to convert substance units:

1 oz = 30 mL; 20 cigarettes = 1 pack

9.8 Time Window

Assessments may be made at times other than the nominal times of planned visits, due to participant scheduling issues, unscheduled visits for neurological worsening or relapse assessment, or early treatment termination visits.

For by-visit analyses of efficacy and safety (except coagulation), each measurement will be assigned an analysis visit number according to the prespecified time window, up to Week 156 Visit. All visit windowing will be done based upon study day, and any assessments completed on the same day as study intervention start will be considered for baseline, except in the case of Day 1 ECGs, which require a determination of predose and post-dose. The analysis visit will then be used for missing data imputations, analysis variable derivations, statistical calculations and presentations.

All Safety Follow-up Visit data will be presented under the Safety Follow-up Visit, regardless of whether the participant is an early treatment discontinuer or a treatment completer.

For by-visit summary of the coagulation safety endpoints, visit windowing will not be performed as this data only being collected at Screening and End of DBTP Visits. Assessments will be presented by nominal visit (as collected in the database).

For participant data listings by time point, the nominal visit as well as the analysis visit will be displayed.

For the calculation of analysis visit using the time windows, 1 month is expressed as 30 days.

As the schedule of assessments is different for each visit, specific time windows must be used for each endpoint. They are defined in Appendix 18.4.

If there are multiple assessments within a same visit window, then the one closest to the target day will be used. If 2 assessments have the same difference with target day, the earlier one will be used.

9.9 Repeated and Unscheduled Measurements

Repeated and unscheduled measurements are included in the listings. Data collected at both unscheduled and scheduled visits will be used for shift tables. For summary statistics, figures, or inferential analysis, data are re-allocated to an analysis visit using time window calculations (Section 9.8).

9.10 Definition of On-treatment Period

For the PA, on-treatment values are results of assessments done from the first study intervention administration on Day 1 until end of the DBTP (completion, early termination including Safety Follow-up, or the cutoff date, whichever is earliest). Values reported in the DBE or OLE periods are not defined as on-treatment for the PA.

9.11 Definition of 48 weeks Post Randomization Period

The 48 weeks post randomization period used in efficacy analysis is defined from the day of randomization until Day 344, which corresponds to the study day of Week 48 Visit + 7 days visit window defined in the protocol.

9.12 Definition of 96 weeks Post Randomization Period

The 96 weeks post randomization period used in efficacy analysis is defined from the day of randomization until Day 680, which corresponds to the study day of Week 96 Visit + 7 days visit window defined in the protocol.

9.13 Definition of 156 weeks Post Randomization Period

The 156 weeks post randomization period used in efficacy analysis for early treatment discontinuers is defined from the day of randomization until Day 1100, which corresponds to the study day of Week 156 Visit + 7 days visit window defined in the protocol.

9.14 Imputation of Missing Data

Presentation of missing data

For efficacy analysis, methods of missing data handling are detailed in Section 14.

In all participant data listings, partial dates, which are not to be imputed according to this IAP, will be presented using the format "_____YYYY". When presented, imputed dates will be flagged (i.e. D for day, M for month).

Missing statistics, e.g. when they cannot be calculated, should be presented as 'nd', with 'nd' standing for 'not done'. For example, if n = 1, the measure of variability (SD) cannot be computed and should be presented as 'nd'.

In case of zero records available for presentation in a given TLF, an empty output with 0 occurrence or a sentence stating that there are no data will be provided. For tables of AEs and Deaths (outputs required for EudraCT and/or clinicaltrial.gov), if there is no observation, the output must contain the first line 'Participant with…' or 'Participant who died' displayed with 0 occurrence.

If a SOC or ATC term is missing/not coded yet, then 'Uncoded SOC' (or 'Uncoded ATC') will be indicated at the output level. When a PT is missing, it will be set to 'Uncoded PT:' TEAE verbatim text.

Handling of missing or partial AEs dates

For defining the TEAE flag, missing or partial AE dates will be imputed as follows:

- In case the onset date is missing completely or missing partially but the onset month and year, or the onset year are equal to the start of study intervention, then the onset date will be replaced by the minimum of start of study intervention and AE resolution date.
- In all other cases the missing onset day or missing onset month will be replaced by 1.
- Incomplete stop dates will be replaced by the last day of the month (if day is missing only), if not resulting in a date later than the date of participant's death or the cutoff date. In the latter case the date of death, or cutoff date, will be used to impute the incomplete stop date.
- In all other cases the incomplete stop date will not be imputed.
- Special cases: AEs may be split into multiple records if the toxicity Grade and/or the seriousness changes (see Section 15.1 for more details). The end date of the previous record and the start date of the new record are expected to be the same. In the event that the month and year are indeed the same, but the day is missing, it will be imputed by 15. In case the imputed start date of the new record is after the end date of this same record, no imputation will be performed.

Imputed dates will only be used for defining the TEAE flag for the on-treatment period.

Handling of partially missing MS dates

For time since MS first attack, MS diagnosis or date of conversion from RRMS, a missing onset day/month will be replaced by 1 for the duration derivation.

9.15 Scoring of HRQOL Data

Unless otherwise specified, HRQOL questionnaires will be scored using their published administration and scoring manual. For items with missing responses, the response will be managed as per the scoring manual. See Appendices 18.1, 18.2 and 18.3 for details.

9.16 Control of Multiplicity

As detailed in Section 9.4.4.3 of the clinical study protocol.

To control trial-wise and family-wise type I error at the 1-sided 0.025 level in the presence of multiple endpoint testing, a graphical approach to sequentially rejective multiple testing will be employed (Bretz 2009, Hung 2013, Bretz 2019).

The 1-sided null hypotheses in the graph are as follows, where "Study 1" denotes study 0080, and "Study 2" study PPD :

Primary Endpoint Null Hypotheses:

H_{011}: $rRR_1 \geq 1$, where rRR_1 denotes qualified relapse rate ratio comparing evobrutinib to teriflunomide in Study 1

H_{012}: $rRR_2 \geq 1$, where rRR_2 denotes qualified relapse rate ratio comparing evobrutinib to teriflunomide in Study 2

Secondary Endpoint Null Hypotheses:

H_{02}: $S_{12e}(t) \leq S_{12c}(t)$, where $S_{12e}(t)$ denotes the survival function for time to 12-week CDP in the experimental (evobrutinib) group based on pooled data, $S_{12c}(t)$ denotes the survival function for time to 12-week CDP in the comparator (teriflunomide) group based on pooled data, and the variable t denotes time since randomization.

H_{03}: $S_{24e}(t) \leq S_{24c}(t)$, where $S_{24e}(t)$ and $S_{24c}(t)$ denotes the survival functions for time to 24-week CDP based on pooled data

H_{04}: $S_{24e}(t) \geq S_{24c}(t)$, where $S_{24e}(t)$ denotes the survival function for time to 24-week CDI in the experimental (evobrutinib) group based on pooled data from participants with baseline EDSS $\geq$ 2.0, $S_{24c}(t)$ denotes the survival function for time to 24-week CDI in the comparator (teriflunomide) group based on pooled data from participants with baseline EDSS $\geq$ 2.0, and the variable t denotes time since randomization.

H_{05}: $\Delta_{PF} \leq 0$, where Δ_{PF} denotes difference in PROMIS Physical Function score CFB over 96 weeks least-squares mean, comparing evobrutinib to teriflunomide, based on pooled data (higher score corresponds to improved physical function).

H_{06}: $\Delta_{Fatigue} \geq 0$, where $\Delta_{Fatigue}$ denotes difference in PROMIS Fatigue score CFB over 96 weeks least-squares mean, comparing evobrutinib to teriflunomide, based on pooled data (higher score corresponds to more fatigue).

H_{021}: $lRR_{21} \geq 1$, where lRR_{21} denotes T1 Gd+ lesion rate ratio comparing evobrutinib to teriflunomide in Study 1

H_{022}: $lRR_{22} \geq 1$, where lRR_{22} denotes T1 Gd+ lesion rate ratio comparing evobrutinib to teriflunomide in Study 2

H_{031}: $lRR_{31} \geq 1$, where lRR_{31} denotes new or enlarging T2 lesion rate ratio comparing evobrutinib to teriflunomide in Study 1

H_{032}: $lRR_{32} \geq 1$, where lRR_{32} denotes new or enlarging T2 lesion rate ratio comparing evobrutinib to teriflunomide in Study 2

H_{041}: $\Delta_{NfL1} \geq 0$, where Δ_{NfL1} denotes a difference in NfL Concentrations at 12 weeks comparing evobrutinib to teriflunomide in Study 1

H_{042}: $\Delta_{NfL2} \geq 0$, where Δ_{NfL2} denotes a difference in NfL Concentrations at 12 weeks comparing evobrutinib to teriflunomide in Study 2

At the PA, the primary efficacy endpoint, ARR, will be tested at the 0.025 (1-sided) level in each study. Appendix 18.8 shows the multiple testing procedure involving study-specific endpoints and pooled endpoints.

At the PA, the 12-week CDP pooled endpoint will be tested at the $0.025\text{-}0.025^2 = 0.024375$ level, 1-sided, only if ARR is significant in both studies at the 0.025 level, 1-sided.

If the 12-week CDP pooled endpoint is significant at the 0.024375 level, 1-sided, the subsequent pooled endpoints (24-week CDP, 24-week CDI, PROMIS PF CFB over 96 weeks, PROMIS Fatigue CFB over 96 weeks) will be tested in a hierarchical order at 0.024375 level, 1-sided.

If the primary efficacy endpoint, ARR up to 156 weeks, is significant within a study at the 0.025 level, 1-sided, the subsequent single-study endpoints (total number of T1 Gd+ lesions based on all available MRI scans, number of new/enlarging T2 lesions on the last available MRI scan relative to the baseline scan, Week 12 NfL concentration) will be tested in a hierarchical order at the 0.025 level, 1-sided.

An overview table summarizing the results of the multi-stage testing procedure will be provided. This table will include the estimates and p-values for each comparison of the evobrutinib intervention group to the teriflunomide intervention group, for the hypothesis family corresponding to the primary endpoint, and for the hypothesis families corresponding to each of the secondary endpoints. The critical value used to assess each p-value for significance will be reported in footnotes, with an annotation of the p-value indicating significance, illustrating the point in the procedure at which the hypothesis testing halted. This overview table will be provided in addition to tables that summarize the analysis of each endpoint separately.

9.17 Software

All statistical analyses will be performed using SAS® (Statistical Analysis System, SAS-Institute, Cary, North Carolina Windows Version 9.4 or higher). Graphics will be prepared with SAS Version 9.4, or higher.

The computer program Phoenix® WinNonlin® Version 8.3, or higher (Certara, L.P., Princeton, New Jersey, USA) will be used to review PK data for possible exclusion and for parameter estimation in the PK substudy.

9.18 Unblinding

Details regarding the unblinding process are available in the latest version of unblinding plans for each study.

9.19 Terminology Conventions

Generally, the term 'participant' will be used instead of 'subject' or 'patient'. However, in tables and listings the term 'subject' will be used to match CDISC requirements, except for in-text tables where 'participant' will be used to match the CSR and protocol templates.

Similarly, the term 'study intervention' will be used in this document instead of 'treatment' to match protocol and CSR templates, however, tables and listings will use 'treatment' for brevity reasons. Exceptions from this rule are commonly used terms like "on-treatment", "treatment-emergent", "treatment policy", "subject-years", "by-subject", or names of eCRF pages like "Treatment Termination" page.

10 Study Participants

The subsections in this section include specifications for reporting participant disposition and study treatment/study discontinuations. Additionally, procedures for reporting protocol deviations are provided.

10.1 Disposition of Participants and Discontinuations

A table on screened participants describing the number and percent of participants in each of the following disposition categories will be produced by study intervention group as indicated in Table 6, for each study and on pooled data:

Table 6 Disposition

• Total number of screened participants, i.e. participants that gave informed consent (overall summary only)
• Number of participants who discontinued prior to randomization and reason (overall summary only)
• Number of randomized participants
• Number of randomized participants who did not start treatment and reasons

• Number of DBTP treatment ongoing participants, i.e. participants who neither permanently discontinued treatment during DBTP nor completed DBTP before the data cutoff*
• Number of randomized participants who completed DBTP**
○ Number of randomized participants who completed DBTP and entered Double Blind Extension
○ Number of randomized participants who completed DBTP and entered Open Label Extension
○ Number of randomized participants who completed DBTP plus safety follow-up period according to protocol
○ Number of randomized participants who discontinued from study after completion of DBTP during safety follow-up and reason
○ Number of randomized participants who discontinued from study after completion of DBTP during safety follow-up due to COVID-19
○ Number of randomized participants who discontinued from study after completion of DBTP during safety follow-up due to the Ukraine crisis
• Number of randomized participants who permanently discontinued treatment during DBTP and reason
○ Number of randomized participants who permanently discontinued treatment due to COVID-19
○ Number of randomized participants who permanently discontinued treatment due to the Ukraine crisis.
○ Number of randomized participants who completed safety follow-up period
○ Number of randomized participants who discontinued from study during safety follow-up and reason
○ Number of randomized participants who discontinued from study during safety follow-up due to COVID-19
○ Number of randomized participants who discontinued from study during safety follow-up due to the Ukraine crisis.

* No End of Treatment eCRF page completed

** A participant is considered to have completed the DBTP if he/she completed the End of Treatment Visit and was assigned the status 'Completed'.

Reason for receiving no study intervention for randomized participants is retrieved from the STERM eCRF page. The reason 'did not meet eligibility criteria' will apply to all participants for whom reason in STERM is entered as 'Other' with specification beginning with 'randomized in error'.

Participants' information on informed consent, Screening and randomization will be listed. Participants discontinued from study intervention or study will be listed with their reason for withdrawal (from study intervention or study).

Participants with study participation impacted by COVID-19 will be listed with a description of the different events and impacts. This could include AEs (related to COVID-19 or to COVID-19 vaccines), protocol deviations, study intervention discontinuation or study discontinuation, death, medical history, concomitant medications or procedures (including COVID-19 vaccinations). Participants impacted by COVID-19 do not correspond necessarily to the participants experiencing the ICEs of 'COVID-19 infection' and 'COVID-19 vaccination' defined in efficacy analyses (see Section 14). Depending on the MedDRA version available at the time of analysis, applicable advice from ICH M1 Points to Consider Working Group and MedDRA Maintenance and Support Services Organization Communication on Coronavirus or more recent official guidance will be applied to coding and reporting Coronavirus concepts.

Participants impacted by the Ukraine crisis will also be listed with a description of the different events and impacts that the participant experienced. This could include study intervention discontinuation, study discontinuation, site transfer, protocol deviations, missed doses and missed visits. In addition, a corresponding summary statistics table will be produced presenting the number and percentage of participants with study intervention interruption, site transfer due to the Ukraine crisis as well as number of missed doses, number of participants with permanent treatment discontinuation and median time to treatment discontinuation by country. Only events with cause attributed to the Ukraine crisis will be included. These are identified by the text "BCP22" added as a prefix to explanations in eCRF free-text fields or the protocol deviation description, explained in the study specific eCRF completion guidelines and the MS200527-0080 PPD _PDMP.pdf document.

Participants impacted by the Ukraine crisis do not correspond necessarily to the participants experiencing the ICE of 'Ukraine crisis' defined in efficacy analyses (see Section 14).

A listing of sites not continuously operational due to the Ukraine crisis will be produced, detailing the site ID, country, number of randomized participants at the site, date the site was deemed not operational and reason.

Listings for home visits and telephone contacts will be produced.

A table based on screened participants, describing the number and percent of participants in each analysis set by intervention group, will be produced.

The following summaries will be provided:

- Number of participants in each analysis set by region, country and site
- Number of randomized participants (from IWRS) by region and country
- Number of randomized participants (IWRS) by randomization strata
- Cross tabulation: stratum by IWRS versus stratum by eCRF/vendor data
- Cross tabulation: participants randomized (Teriflunomide/evobrutinib) vs. treated (Teriflunomide/evobrutinib/no)
- By-visit number and percentage of participants with on-site versus at home versus telephone visits/contact
- Number of randomized participants by age category.

Re-screened participants will be presented in a listing.

Disposition of PROMIS Questionnaires

The disposition of the PROMIS questionnaires (PROMIS PF and Fatigue) will be described separately at each scheduled visit in terms of completion and compliance rates as described below:

- Number and percentage of randomized participants who returned the questionnaire at the visit as expected (compliance rate using all expected questionnaires in the denominator)

- Number and percentage of randomized participants with at least one evaluable questionnaire at the visit (completion rate using all expected questionnaires in the denominator)

- Population-level summary (per questionnaire):

  - $\%\ \text{Compliance} = 100 \times \frac{\text{number of participants who returned the PRO questionnaire}}{\text{number of participants for whom a PRO questionnaire is expected}}$
  - $\%\ \text{Completion} = 100 \times \frac{\text{number of participants with at least one evaluable PRO}}{\text{number of participants for whom a PRO questionnaire is expected}}$

A questionnaire is considered returned if at least one item from the questionnaire is answered. A questionnaire is considered evaluable if PROMIS T-scores can be derived.

Compliance and completion rates for each PROMIS questionnaire will be provided at Baseline and subsequent scheduled visits by study intervention group.

The reasons for non-compliance and non-completion will be described at Baseline and subsequent scheduled visits by study intervention group as follows:

- Number and percentage of randomized participants with missing questionnaire at an expected visit by reason for non-compliance:
  - Death
  - Lost to follow-up
  - Withdrawal by subject
  - Other reason for study termination
- Number and percentage of randomized participants with missing questionnaire at the visit by reason while expected – reasons for non-completion:
  - Subject felt too ill
  - Clinician or nurse felt the subject was too ill
  - Subject unable to come to site
  - Site was closed
  - Investigator decision
  - Subject did not come for unknown reasons
  - Other
  - COVID-19 related (yes/no/unknown)
  - Ukraine crisis related (yes/no/unknown)

10.2 Protocol Deviations / Exclusion from Analysis Populations

10.2.1 Important Protocol Deviations

IPDs are PDevs that might significantly affect the completeness, accuracy, and/or reliability of the study data or that might significantly affect a participant's rights, safety, or well-being. Important protocol deviations are defined in a separate document (latest version of MS200527-0080 PPD _ PDMP.pdf).

All deviations will be identified by either site monitoring, medical review processes or programming and confirmed prior to or at the DRM, which will occur before the database lock. cIPDs are a subset of important protocol deviations that could impact the key objectives of the study. A small number of participants may be excluded from the mFAS on the basis of cIPDs related to eligibility, randomization error or failure to receive study intervention, per ICH E9 (i.e. cIPDs with PD codes equal to PDEV04 till PDEV50, or PDEV71 or PDEV73 or PDEV74 as defined in PDMP). Participants from sites in Ukraine, Russian Federation and Belarus lacking data robustness throughout the study may be excluded from mFAS on the basis of assessment of site operationality. When evaluating mFAS exclusion prior to DBL and unblinding, sites from all countries will be reviewed.

As described in Section 8.1, the protocol allows the flexibility to exclude SAF participants from mSAF, if sites with lack of safety data robustness throughout the study are identified. At the time of finalization of this IAP, no such site has been identified.

The outcome of the DRM will document the IPDs (including cIPDs) as well as the finalization of the analysis populations in a memo. IPDs will be documented in CDISC SDTM whether identified through sites monitoring, medical review and/or programming.

COVID-19 related PDevs will include all PDevs including the mention "COVID-19" at the beginning of their description.

Ukraine crisis related PDevs will include all PDevs including the mention "BCP22" at the beginning of their description.

The following summary tables and listings of IPDs/PDevs will be provided on all randomized participants:

- Table providing frequency for each type of IPD/cIPD:
    - COVID-19 related IPDs will be summarized as a separate category within the table.
    - Ukraine crisis related IPDs will also be summarized as a separate category within the table.
- Table providing frequency for each type of minor COVID-19 related PDev
- Table providing frequency for each type of minor Ukraine crisis related PDev

- Listing of IPDs
- A specific listing for COVID-19 related IPDs/PDevs
- A specific listing for Ukraine crisis related IPDs/PDevs

10.2.2 Reasons Leading to the Exclusion from an Analysis Population

A frequency table per reason of exclusion from the FAS, mFAS, SAF, PK, PKSUB, PD, and PDBTKO populations as well as a listing will be provided.

11 Demographics and Other Baseline Characteristics

Demographics and baseline characteristics will be summarized on FAS. They will be presented by intervention group and overall, for each study and on pooled data. All listed data will be on FAS with an additional column or flag for participants in mFAS.

11.1 Demographics

- Demographic characteristics:
  - Sex: male, female
  - Race: White, Black or African American, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Multiple (Combinations included as sub-summarization), Not collected at this site, Other
  - Ethnicity: Hispanic or Latino, Not Hispanic or Latino
  - Age (years) at informed consent: summary statistics
  - Age (years) at informed consent categories: < 40, ≥ 40
- Geographic Region: North America, Western Europe, Eastern Europe, ROW
- EEA

Specifications for computation:

- Age (years):
  - (date of given informed consent - date of birth + 1) / 365.25
- In case of missing day for date of birth, but month and year available:
  - For the derivation of age, the day of birth will be set to 1 and the formula above will be used
- In case of missing month for at date of birth, but year available:

- For the derivation of age, the day and the month of birth will be set to 1 and the formula above will be used

Country will be utilized to identify the geographic region and EEA membership.

Geographic region and EEA membership for participants from Ukraine who moved to another country after the start of the Ukraine crisis are derived based on the original country the participants were randomized to (i.e. Ukraine).

Demographic summary will also be presented by the subgroup 'membership in initial or second recruitment cohorts' (i.e. initial cohort vs. second cohort).

Demographic summary will also be presented by the subset 'EDSS $\geq$ 2 at baseline'.

11.2 Medical History

The medical history will be summarized from the "Medical History Details" eCRF page, using the most recent MedDRA version at time of database lock, PT as event category and SOC body term as Body System category. The MedDRA version used will be indicated in footnote. Each participant will be counted only once within each PT or SOC.

Medical history will be displayed in terms of frequency tables based upon FAS: ordered by primary SOC and PT in alphabetical order. All medical history data will be listed.

11.3 Other Baseline Characteristics

11.3.1 Disease History

Information on MS baseline disease characteristics, based on data collected on Day 1 predose and during Screening, will be summarized by study intervention group. Descriptive statistics will be presented for:

- Type of MS, either RRMS or SPMS with relapses
    - Time (years) since conversion from RRMS to SPMS (for participants with SPMS)
- Time (years) since onset of symptoms
- Time (years) since initial MS diagnosis date
- System(s) affected by first attack
- Number of relapse(s) in the last 2 years before Screening
- Number of relapse(s) from last year prior to randomization
- Fulfills McDonald criteria 2017 at Screening

- Presence of at least 1 T1 Gd+ lesion within 6 months prior to randomization

Time (years) since event derivation is described in Section 9.6.

Disease history summary will also be presented by the subset 'EDSS $\geq$ 2 at baseline'.

11.3.2 Baseline Lesion Measures, Neurological Assessments, PROs and NfL Concentration

Baseline neurological measures based on MRI and PRO assessments will be summarized by study intervention group. Descriptive statistics will be presented for:

- Number of T1 Gd+ lesions
- Volume of T1 Gd+ lesions (cc)
- Number of T2 lesions
- Volume of T2 lesions (cc)
- Normalized Brain Volume (cm^3)
- Normalized Cortical Grey Matter Volume (cm^3)
- Thalamic Volume (cm^3)
- EDSS score (from vendor data)
- Scores for each of the 7 Functional Systems (Visual, Brainstem, Pyramidal, Cerebellar, Sensory, Bowel/Bladder, Cerebral) and score for Ambulation used to derive EDSS score.
- PROMIS Physical Functioning score
- PROMIS Fatigue score
- EQ-5D-5L Index and VAS
- SF-36v2 normalized score for each of the 8 health domain scales
- SF-36v2 PCS score and MCS score
- T25-FW score
- 9-HPT score
- SDMT score

- NfL concentration

11.3.3 History of DMT

Information on DMTs, as collected on the "HISTORY OF DISEASE MODIFYING DRUGS DETAILS" eCRF page, will be summarized by study intervention. Descriptive statistics will be presented for:

- Any DMT before study entry: yes/no
- Prior DMT type:
  - Aubagio (teriflunomide)
  - Rebif / Avonex / Plegridy (IFN beta-1a)
  - Betaseron / Betaferon / Extavia (IFN beta-1b)
  - Copaxone / Glatopa (glatiramer acetate)
  - Gilenya (fingolimod) / Mayzent (siponimod) / Zeposia (ozanimod) / Ponvory (ponesimod)
  - Lemtrada (alemtuzumab)
  - Mavenclad (cladribine)
  - Novantrone (mitoxantrone)
  - Ocrevus (ocrelizumab) / Kesimpta (ofatumumab) / Rituxan (rituximab) / Ublituximab (BRIUMVI)
  - Tysabri (natalizumab)
  - Tecfidera (Dimethyl fumarate) / Vumerity (diroximel fumarate)
  - Investigational drug
  - Other
- Reason for discontinuation from last prior DMT:
  - Lack of Efficacy
  - AE
  - Other
- Number of prior treatments (1, 2, $\geq$ 3)
- Prior treatment duration – combined (sum of all prior DMT); as a continuous variable and categorically:
  - $\leq$ 1 month
  - > 1 - 6 months
  - > 6 - 12 months
  - > 12 - 24 months

- > 24 - 36 months
- > 36 - 48 months
- > 48 – 60 months
- > 60 months

Although only month and year for DMT start and end dates are collected, duration will be calculated as described in Section 9.6 (definition of duration) and converted into months, as described in Section 9.7 (conversion factors), after the following imputations are performed:

Start date:

- Day will be imputed as 1
- In case the month is missing, the month will be imputed as January

End date:

- If the month and year are prior to the month and year of first study intervention, then day will be imputed as the last day of the month
- If the month and year are equal to the month and year of first study intervention, then day will be imputed as the first day of the month
- In case the month is missing:
  - If the year is prior to the year of the first study intervention, the month will be imputed as December
  - If the year is the same as the year of the first study intervention, the month will be imputed as the month prior to the month of first study intervention. If the month of the first study intervention is January, the date imputed will be 01 January

If the start or end date is completely missing, no imputation will be performed and the duration will be missing.

11.3.4 Other Screening Characteristics

Other characteristics like viral serology, QuantiFERON®-TB test and Ferritin and transferrin saturation will be listed only.

11.3.5 Baseline Substance Use Characteristics

Conversion factors for exposure will be as per specifications in Section 9.7.

Nicotine usage will be summarized by study intervention group as follows:

- Nicotine use status for each exposure type (never / current / former user)
- Exposure type (cigarettes / chewing tobacco / nicotine gum / e-cigarettes and vapor)
- Exposure summary statistics per week for each exposure type except cigarettes

Cigarettes will be summarized by packs per day.

Alcohol usage will be summarized by study intervention group as follows:

- Alcohol use status for each exposure type (yes / no)
- Exposure type (beer / wine / spirits)
- Exposure summary statistics per week for each exposure type

Caffeine usage will be summarized by study intervention group as follows:

- Caffeine use status (no use / occasional use (not daily use) / regular use (1-2 cup(s)/day) / regular use (> 3 cups/day))

All nicotine, alcohol, and caffeine usage data will be listed as collected on the relevant eCRF pages.

12 Previous or Concomitant Medications/Procedures

12.1 Previous or Concomitant Medications

Based on "RELEVANT PREVIOUS MEDICATIONS Details" & "CONCOMITANT MEDICATIONS Details" eCRF pages.

For the PA, previous and concomitant medications will be summarized separately by intervention group on SAF for each study. Data from the on-treatment period including Safety Follow-up data will be included.

Statistical analysis

The ATC-2nd level and PT will be tabulated as given from the WHO-DD current version. In case multiple ATCs are assigned to a drug, only ATC-2nd level related to the indication will be used for reporting.

The number and proportion of participants with previous or concomitant medications will be separately summarized by study intervention group and will be presented by descending frequency of ATC 2nd level term and then by descending frequency of PT in the evobrutinib column. If multiple ATCs/PTs have the same frequency, they will be sorted alphabetically. The WHO-DD version used will be indicated in footnote.

Previous or concomitant medications will be also listed.

Previous medications are medications, other than study interventions, which either:

- started and stopped before first administration of any study intervention (teriflunomide or evobrutinib).
- started prior to the first administration of study intervention (teriflunomide or evobrutinib) and are taken by participants on or after the first administration of study intervention during the treatment period (including Safety Follow-up for early discontinued participants).

Concomitant medications are medications, other than study interventions, which either:

- started on or after the first administration of any study intervention in the DBTP.
- started prior to the first administration of study intervention in the DBTP and are taken by participants on or after the first administration of study intervention (including Safety Follow-up for early discontinued participants).

Partial dates will be handled as follows:

- For previous medications, in case the date values will not allow a medication to be unequivocally allocated to previous medication, the medication will be considered as previous medication.
- For concomitant medications, in case the date values will not allow a medication to be unequivocally allocated to concomitant medication, the medication will be considered as concomitant medication.

12.2 Prior or Concurrent Procedures

Based on the "RELEVANT PREVIOUS PROCEDURES Details" & "CONCOMITANT PROCEDURES Details" eCRF page. Prior and concurrent procedures will be summarized the same way as medications (see Section 12.1).

Number of participants with prior/concurrent procedures overall and by SOC and PT will be summarized by intervention group, using current version of MedDRA dictionary. SOC terms will be sorted alphabetically. PTs within each SOC will be sorted by descending frequency of the evobrutinib intervention group, and then alphabetically if multiple PTs have the same frequency.

Prior and concurrent procedures will be also listed.

12.3 COVID-19 Vaccinations

COVID-19 vaccinations will be identified according to the SDGs subgroup "Vaccines for COVID-19" and corresponding SDG subcategories of the latest version of the WHO-DD. The summary table will include counts for vaccines that were given prior to first administration of any study intervention as well as concomitant.

Participants experiencing the ICE "COVID-19 vaccination" will be identified as those receiving a COVID-19 vaccination concomitant with study intervention.

13 Study Treatment: Compliance and Exposure

Exposure, cumulative/calculated total dose and compliance will be summarized by intervention group on SAF. Data from the on-treatment period will be analyzed.

13.1 Exposure Calculation

Dose interruptions/changes will not be considered in exposure calculation. Dose interruptions/changes with the associated reason will only be listed.

Details describing study therapy dosing and administration are provided in Table 7.

Table 7 Dosage and administration

Treatment Group	Number of Tablets per day 45 mg Evobrutinib	Number of Tablets per day 45 mg Placebo	Number of Tablets per day 14 mg Teriflunomide	Number of Tablets per day 14 mg Placebo
Evobrutinib 45 mg twice daily	2	0	0	1
Teriflunomide 14 mg once daily	0	2	1	0

Treatment duration in weeks will be calculated according to the following formula:

$$Duration\ of\ Evobrutinib\ or\ Teriflunomide\ (weeks) = \frac{(date\ of\ last\ dose\ -\ date\ of\ first\ dose\ +\ 1)}{7}$$

First dose refers to the first administration of active study intervention in the on-treatment period. Last dose refers to the last administration of active study intervention in the on-treatment period. Both dates of first and last dose will be retrieved from SDTM EX domain ("Evobrutinib / Placebo Administration Details" and "Teriflunomide / Placebo Administration Details" eCRF pages). If the end date of the last dose in EX is missing for participants who discontinued treatment prematurely, the date of last dose will be replaced by the end of treatment date recorded in TTERM eCRF page or End of Study date recorded in STERM eCRF page.

Treatment duration will be presented by summary statistics and according to the following categories:

- ≤ 1 week
- > 1 to 12 weeks
- > 12 to 24 weeks
- > 24 to 36 weeks
- > 36 to 48 weeks
- > 48 to 60 weeks
- > 60 to 72 weeks
- > 72 to 84 weeks
- > 84 to 96 weeks
- > 96 to 108 weeks

- > 108 to 120 weeks
- > 120 to 132 weeks
- > 132 to 144 weeks
- > 144 to 156 weeks
- > 156 weeks

The expected total dose (mg) per participant for the on-treatment period will also be summarized for the intervention groups, based on the actual study intervention the participant receives. The expected total dose is defined according to the following formula:

$$Expected\ total\ dose\ for\ Evobrutinib\ or\ Teriflunomide\ (mg) = Treatment\ duration\ (days) \times Daily\ dose$$

where the daily dose for each intervention is provided in Table 7.

The cumulative actual dose (mg) per participant for the on-treatment period will also be summarized for the active intervention groups. The cumulative actual dose is defined according to the following formula:

$$Cumulative\ actual\ dose\ for\ Evobrutinib\ or\ Teriflunomide\ (mg) = Sum\ of\ dose\ per\ day\ for\ Evobrutinib\ or\ Teriflunomide\ records$$

where the dose per day is derived in the SDTM EX domain, from the number of tablets ingested per day, recorded on the "Evobrutinib / Placebo Administration Details" and "Teriflunomide / Placebo Administration Details" eCRF pages, and the dosage information in Table 7.

Study drug administrations will be listed by intervention group, and participant, with start/end dates of administration, and reason for dose change or no dose (if applicable).

13.2 Compliance Calculation

For the on-treatment period, compliance with study intervention is defined as the cumulative actual dose during a period divided by the expected total dose for that period, multiplied by 100 to yield a percentage, i.e.:

$$Compliance = 100 \times \left(\frac{Cumulative\ actual\ dose\ (mg)}{Expected\ total\ dose\ (mg)}\right)$$

Compliance with study intervention will be tabulated by intervention group from first intake to last intake during the on-treatment period.

Compliance with study intervention will also be presented into categories as follows:

- < 80% (Insufficient compliance)
- ≥ 80% to ≤ 100%
- > 100% to ≤ 110%
- > 110%

For each analysis, the following listings will be provided:

- listing of kit numbers with date of dispense, date of return, amount dispensed and amount returned. This listing will be repeated based on participants who received active treatment or placebo that they were not randomized to, or received both active treatments.
- listing of start/end dates with dose (mg) per day, reasons for missed/modified dose
- listing with exposure time, cumulative dose (total and actual), and compliance

14 Efficacy Analyses

This section includes specifications for analyzing efficacy endpoints.

Data handling for efficacy endpoints is described in Table 8 for the IA for BSSR and the PA.

Table 8 Data Handling for Efficacy Analysis

Analysis	Analysis sets	Period covered	Treatment groups
IA for BSSR	FAS	12-week CDP data from the treatment period (Safety Follow-up included) through data cutoff for BSSR	Not applicable (Blinded)
PA	FAS/mFAS	All data from the DBTP (Safety Follow-up included) until the PA cutoff date	Evobrutinib 45 mg twice daily Teriflunomide 14 mg once daily

The secondary objectives based on disability progression and disability improvement, and patient reported symptoms and functional status will be evaluated based on pooled data from study MS200527_0080 and study PPD. Analysis based on pooled data will also be presented for each study.

The randomization strata used in efficacy analyses are those based on vendor data for EDSS and eCRF for region.

Further details regarding imputations and statistical methods will be included in the Statistical Methodology document (Appendix 18.9).

14.1 Primary Estimand: Annualized Relapse Rate up to 156 Weeks

This section provides detailed information related to the analysis of the primary efficacy estimand, including primary, sensitivity, and subgroup analyses.

14.1.1 Primary Objective: Derivation and Analysis of ARR up to 156 Weeks

This section details the 5 attributes of the primary efficacy estimand.

The 5 estimand attributes according to ICH E9(R1) are:

- **Endpoint**: ARR based on qualified relapses up to 156 weeks.

  The patient-level data required to address the clinical question are: (1) number of qualified relapse events, as adjudicated by the EAC, observed up to Week 156 and (2) the time (in years) over which those events were observed up to Week 156.

- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.

- **Population**: Participants with RMS as defined by inclusion/exclusion criteria.

- **Strategies used to address ICEs**:

  - Treatment discontinuation: Treatment Policy strategy.
  - Death attributable to MS or treatment: Composite variable strategy
  - Death unattributable to MS or treatment: While alive strategy
  - Ukraine crisis: Hypothetical strategy
  - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.

- **Population-level summary** for the endpoint, providing a basis for assessing study intervention effect: Relapse rate ratio, based on a NB model for qualified relapse count, with terms for intervention group and randomization strata, with offset equal to log follow-up time (in years) over which the qualified relapses experienced by a participant are observed.

  The NB regression will be computed with the SAS® GENMOD procedure, using the dist=NB option in the MODEL statement. If the model fails to converge, the following steps will be applied sequentially until the convergence is reached: use of a more efficient algorithm in the GLIMMIX procedure, removal of the term for the region randomization stratum, removal of the term for the baseline EDSS randomization stratum, use of the Poisson model with the robust variance estimation, see Table 9 for details.

Table 9 Details of Steps to be Applied When the Primary Model Fails to Converge

(0)*	PROC GENMOD	Model includes terms for treatment and for both region and baseline EDSS randomization strata / dist = NB
(1)	PROC GLIMMIX	Model includes terms for treatment and for both region and baseline EDSS randomization strata / dist = NB
(2)	PROC GLIMMIX	Model includes terms for treatment and the baseline EDSS randomization stratum / dist = NB
(3)	PROC GLIMMIX	Model includes only the term for treatment / dist = NB
(4)	PROC GLIMMIX	Model includes only the term for treatment / dist = POISSON / option 'empirical' in 'PROC GLIMMIX' statement

* Initial model

Missing Data Handling

For participants who have not been treated and followed until the planned end of treatment (i.e. PA cutoff or 156 weeks post randomization (Section 9.13), whichever occurs first), the missing data will be assumed to be MAR; treatment discontinuers will be assumed to experience relapse through the planned end of treatment at the same rate as participants with DBTP treatment ongoing/completed within the same intervention group and stratum.

Handling of ICEs

In accordance with Treatment Policy strategy, for participants experiencing the ICE premature treatment discontinuation, any qualified relapses occurring through Safety Follow-up will be included in the analysis up to the planned end of treatment (i.e. PA cutoff date or 156 weeks post randomization, whichever occurs first). If a participant is a treatment completer, i.e. does not experience the ICE of premature treatment discontinuation, then the Treatment Policy strategy for handling this ICE does not apply, and only data collected on-treatment (i.e. no data from Safety Follow-up Visit) will be included in the analysis (see Section 9.10).

In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as a qualified relapse. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, the death will not be counted as a qualified relapse.

Deaths with primary reason entered as 'progressive disease and/or disease related condition' in the eCRF Death Form will be considered as attributable to MS. Deaths with primary reason set to 'event related to study treatment' will be considered as attributable to study intervention. Other deaths (i.e. primary reason equal to 'event unrelated to study treatment' and 'unknown') will be considered as unattributable to MS or treatment.

Participants at sites not fully operational due to the Ukraine crisis, post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data from such participants will be censored at the time at which the site lost

fully operational status. Censoring at the end of the pre-crisis period is non-informative for the outcome of interest (relapse).

Sites not fully operational post start of the crisis along with corresponding censoring date will be identified before DBL and listed in an IAP Appendix where the longitudinal analysis of site status is documented.

In accordance with Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination', or 'emergency unblinding', the occurrence of these ICEs is considered irrelevant in the primary analysis of qualified relapse.

Inferential analysis

The population-level summary comparing the intervention groups is qualified relapse rate ratio, based on a NB model for qualified relapse count over 156 weeks, with terms for intervention group and covariates defined by randomization strata, with offset equal to log follow-up time for a participant during the 156 weeks post randomization. The adjusted qualified relapse rate ratio comparing evobrutinib to teriflunomide estimated from the NB model, 95% 2-sided CI, and 1-sided p-value will be reported, together with adjusted ARR for each intervention group and associated 95% 2-sided CI. In addition, the ARR ratio will be expressed as percentage reduction relative to teriflunomide. The analysis of ARR will be based on FAS data from a single study (0080 or PPD).

The follow-up time, in years, is defined as follows:

- For participants with treatment ongoing at time of PA cutoff date:

$$\frac{(date\ of\ last\ relapse\ assessment\ before\ the\ PA\ cutoff\ date\ -randomization\ date)+1}{365.25}$$

- For participants who completed 156 weeks of treatment:

$$\frac{(End\ of\ Treatment\ date\ -randomization\ date)+1}{365.25}$$

- For early treatment discontinuers:

$$\frac{(date\ of\ last\ relapse\ assessment*\ -randomization\ date)+1}{365.25}$$

- For participants with no post randomization relapse assessment before the PA cutoff date: 0.1/365.25

* within 156 weeks post randomization or before the PA cutoff date, whichever occurs first.

The MAR assumption underpins the rate ratio estimator implemented by the NB model that uses observed qualified relapse events and observed follow-up time for each participant, regardless of treatment completed/discontinued/ongoing status.

The covariates defined by randomization strata are region (4 levels: North America, Western Europe, Eastern Europe, ROW) and baseline (Day 1) EDSS (2 levels: < 4.0, ≥ 4.0). The NB model assumes a common dispersion parameter for all participants, independent of intervention group or baseline covariates. The adjusted qualified relapse rate ratio estimate is given by exponentiation of the estimate for the treatment coefficient from the NB model. In the PA, only observed events and observation time will be included in the analysis; there will be no imputation of events for participants discontinuing study early.

Descriptive analysis

Descriptive statistics of the relapses will be provided by study intervention group on FAS, including the number and percentage of relapses requiring hospitalization, number and percentage of relapses treated with corticosteroids (as reported on the RELAPSEDT eCRF page) out of the total number of relapses, and this will be repeated out of the total number of qualified relapses.

Descriptive statistics for ARR up to 156 weeks will be provided by study intervention group on FAS. No censoring will be applied for descriptive statistics. At the treatment-group-level, total number of qualified relapses, total number of relapses, follow-up (in participant-years), and unadjusted ARR (as number of all relapses divided by person-time (in years)), with 95% 2-sided CI (calculated using the Garwood method), will be reported. In addition, descriptive statistics of number of participant-level qualified relapses and participant-level ARR (number of qualified relapses experienced by a participant up to 156 weeks, divided by the follow-up experienced by a participant up to 156 weeks) will be summarized for each study intervention group, in terms of mean, SD, median, Q1, Q3, min, and max.

Number and percentages of participants experiencing intercurrent events will also be presented.

ARR up to 156 weeks will be presented as a by-treatment group bar chart, with a vertical line segment for each bar representing 95% 2-sided CI, and with the p-value for the comparison between evobrutinib and teriflunomide (based on NB model) displayed above the bar pair.

14.1.2 Sensitivity Analyses of ARR

The following sensitivity analyses for ARR up to 156 weeks are planned, with the same handling of ICEs as in the primary analysis, unless otherwise specified:

1. Analysis to evaluate the potential influence of informative treatment discontinuation according to discontinuation reason.

   This sensitivity analysis will use the same NB model used in the primary analysis and report the same inferential results. In addition, a forest plot will present the qualified relapse rate ratio, 95% CI and p-value for each delta.

The following applies to participants in the evobrutinib intervention group. If a participant's discontinuation is unrelated to treatment (i.e. unrelated to efficacy or safety issues), the participant's relapse count will be multiply imputed using ARR in the comparator group (adjusted for the participant's stratum) based on CR MI method, for the interval between discontinuation and planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first). If a participant's discontinuation is related to treatment, relapse count will be multiply imputed at a rate higher than ARR in the comparator group (adjusted for the participant's stratum), for the interval between discontinuation and planned end of treatment. The comparator rate after withdrawal will be multiplied by delta, ranging from 1 to 3.1 (~tripling of relapse rate) in steps of 0.3. For participants with discontinuation of treatment (all reasons) in the teriflunomide group, the relapse count will be multiply imputed under the MAR assumption.

MI methods for count data will be applied following the approach from Keene 2014, see Appendix 18.9 for details.

Treatment termination with primary reason entered as 'adverse event', 'lack of efficacy' and 'death' in the eCRF TTERM form will be considered as related to treatment. Other treatment terminations (i.e. primary reason equal to 'lost to follow-up', 'protocol non-compliance', 'withdrawal by subject' and 'other') will be considered as unrelated to treatment.

2. Analysis in which early treatment discontinuation is imputed as qualified relapse.

   This sensitivity analysis will use the same NB model used in the primary analysis of ARR up to 156 weeks and report the same inferential results.

   If a participant discontinued treatment early during the on-treatment period without qualified relapse in the 30 days prior to discontinuation, the participant is assumed to have a qualified relapse event at the date of discontinuation.

3. Analysis to evaluate the potential impact of model misspecification.

   This sensitivity analysis will add the following covariates to the NB model: number of relapses in previous 2 years before Screening, presence/absence of T1 Gd+ lesions within 6 months prior to randomization, prior DMT (y/n), and age, as defined in Section 8.2. The same inferential results will be reported as for the primary analysis.

4. Analysis to evaluate the potential impact of COVID-19 infection.

   This sensitivity analysis will handle the ICE of COVID-19 infection with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the infection will be censored. This is in contrast to the primary analysis in which any COVID-19-related ICEs are ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of infection will be assumed to be non-informative for the event of interest (relapse).

   The ICE 'COVID-19 vaccination' will be handled via Treatment Policy strategy as in primary analysis. Participants experiencing the ICE 'COVID-19 infection' will be identified as those with AEs related to COVID-19 as defined in Section 15.2.4. The hypothetical scenario envisaged is that COVID-19 infection does not occur but COVID-19 vaccination does.

5. Analysis to evaluate the potential impact of emergency unblinding of a participant prior to study discontinuation of the participant.

   This sensitivity analysis will handle the ICE of emergency unblinding with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the emergency unblinding will be censored. This is in contrast to the primary analysis in which the ICE of emergency unblinding is ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of emergency unblinding will be assumed to be noninformative for the event of interest (relapse).

   Participants experiencing the ICE emergency unblinding will be identified as those having a Yes in the question 'Was randomization code broken by the site?' in the STERM eCRF page.

6. Analyses to evaluate the potential impact of the war in Ukraine.

   In the primary analysis of ARR, only participants at sites not fully operational post start of the Ukraine crisis, will be considered to have experienced the ICE 'Ukraine crisis'. The data from such participants will be censored at the date at which the site lost fully operational status.

   In the first sensitivity analysis of ARR related to the impact of the war in Ukraine, all participants from sites in Ukraine will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting only Ukraine participants) had not occurred. In a second sensitivity analysis related to the impact of war in Ukraine, all participants from sites in Ukraine, Russian Federation, and Belarus will have their data censored at the start of the crisis. The hypothetical scenario envisaged is one in which the crisis (affecting participants from Ukraine, Russian Federation, and Belarus) had not occurred.

7. mFAS analysis

   In this sensitivity analysis, the primary analysis of ARR will be repeated on the mFAS.

8. Analysis on all relapses.

   This sensitivity analysis will use the same NB model used in the primary analysis of ARR up to 156 weeks and report the same inferential results.

   All relapses will be counted, regardless of qualification status.

9. Analysis to evaluate the potential impact of the second recruitment cohort.

   This sensitivity analysis will add the following covariate to the NB model: membership in second recruitment cohort (yes/no), as defined in Section 8.2. The same inferential results will be reported as for the primary analysis with the p-value of the additional covariate reported.

10. Analysis with a censoring date of 01 June 2023 for all participants still ongoing at that date.

    To assess the impact of the PA trigger delay, participants will be censored on the originally projected date of the PA trigger: 01 June 2023.

14.1.3 Subgroup Analyses of ARR

The analysis of the primary endpoint assessment, as described in Section 14.1.1, will be repeated for all subgroup levels defined in Section 8.2 "Subgroup Definition and Parametrization". Additionally, a larger model that adjusts for subgroup and intervention-by-subgroup interaction will be fitted, and the p-value of the interaction reported.

14.2 Secondary Estimands

This section provides detailed information related to the analysis of the secondary efficacy estimands, including primary, sensitivity, supplementary, and subgroup analyses.

14.2.1 Time to Confirmed Disability Progression

The 5 estimand attributes of secondary efficacy estimands time to first occurrence of 12-week and 24-week CDP are:

- **Endpoint**: Time to first occurrence of 12-week/24-week CDP as measured by the EDSS up to 156 weeks
- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.
- **Population**: Participants with RMS as defined by inclusion/exclusion criteria.
- **Strategies used to address ICEs**:
    - Treatment discontinuation: Treatment policy strategy.
    - Death attributable to MS or treatment: Composite variable strategy
    - Death unattributable to MS or treatment: While alive strategy
    - Ukraine crisis: Hypothetical strategy
    - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.
- **Population-level summary** comparing the intervention groups is the hazard ratio and CI based on a stratified Cox model for hazard rate, with term for intervention group, and strata defined by randomization strata and study ID. Test is based on stratified logrank test with strata defined by randomization strata and study ID. Data are pooled from the FAS of the studies 0080 and PPD .

EDSS is assessed in all participants by the Examining Investigator at Screening and Baseline Visit, and every 12 weeks up to and including the Safety Follow-up Visit. Unscheduled EDSS

assessments for individual participants may be performed between scheduled assessments (i.e. during a MS relapse).

Disability progression is defined as an increase of ≥ 1.0 point from the baseline EDSS score when the baseline score is 5.0 or less and an increase of ≥ 0.5 when the baseline score is 5.5. Initial progression can happen at any visit during the DBTP until the PA cutoff date.

14.2.1.1 Time to 12-week CDP

Disability progression is considered sustained for 12 weeks when the initial increase in the EDSS is confirmed at a regularly scheduled visit at least 12 weeks after the initial documentation of neurological worsening.

Confirmation of disability progression must occur at the regularly scheduled visit that is at least 12 weeks (84 days) after initial progression. If a participant has a missing EDSS at the scheduled visit occurring at least 84 days after an initial progression or the scheduled visit occurs several days before the 84-day window after an initial progression, confirmation of the disability progression must be on the basis of the assessment at the next scheduled visit.

EDSS assessments at unscheduled or scheduled visits that are less than 84 days after the initial progression are considered non-confirmatory EDSS assessments. For disability progression to be confirmed at a regularly scheduled visit at least 12 weeks after initial progression, any non-confirmatory EDSS assessments should be at least as high as the minimum change required for progression. This means that after a scheduled or unscheduled visit at which the participant fulfills the initial disability progression (IDP), all subsequent EDSS assessments (scheduled and unscheduled) need to also fulfill the worsening criteria until the worsening can be confirmed at the first scheduled visit that occurs at least 84 days after onset of the worsening.

Time to 12-week CDP is defined as the time from randomization date to the IDP date when an event is present, and from randomization date to censoring date when an event is not present. Censoring occurs in those participants who did not experience a 12-week confirmed disability progression up to PA cutoff or 156 weeks post randomization, whichever occurs first; this includes participants who had a "tentative" disability progression that could not be confirmed due to an early discontinuation or any other reason. The censoring date is the date of the last EDSS assessment. Participants with no post randomization EDSS assessment will be censored on the day of randomization.

Missing Data Handling

Participants who did not experience 12-week CDP by 156 weeks post randomization, by time of PA cutoff, by time of early study discontinuation, or before being lost to follow up, will be censored at the date of the last EDSS assessment during the 156 weeks post randomization. In the primary analysis of 12-week CDP, censoring will be assumed to be noninformative, conditional on intervention group, randomization stratum, and study ID.

Handling of ICEs

In accordance with Treatment Policy strategy, for participants experiencing the ICE premature treatment discontinuation, all EDSS data through Safety Follow-up will be included in the analysis up to the planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first).

In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as 12-week CDP. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, participants will be censored at time of death. (i.e. death will not be counted as 12-week CDP).

Deaths with primary reason entered as 'progressive disease and/or disease related condition' in the eCRF Death Form will be considered as attributable to MS. Deaths with primary reason set to 'event related to study treatment' will be considered as attributable to study intervention. Other deaths (i.e. primary reason equal to 'event unrelated to study treatment' and 'unknown') will be considered as unattributable to MS or treatment.

Participants at sites not fully operational post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data from such participants will be censored at the time at which the site lost fully operational status. Censoring at the end of the pre-crisis period is non-informative for the outcome of interest (CDP).

In accordance with the Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination' or 'emergency unblinding' the occurrence of these ICEs is considered irrelevant in the primary analysis of CDP.

Assessment of pooling validity

Prior to pooling data from studies 0080 and PPD , the validity of pooling data will be assessed by reviewing consistency of demographics, baseline characteristics, ARR and 12-week CDP results. For consistency, the treatment effect on ARR and 12-week CDP must be in the same direction in both studies.

Inferential analysis

The PA of time to 12-week CDP will be based on data pooled from the FAS of the 2 Phase 3 studies using a stratified log rank test with intervention group as main effect and stratification variables used in randomization and study ID (i.e. Study MS200527_0080 or PPD) as covariates. The 1-sided p-value (α=0.024375) from the stratified log rank test for comparison between evobrutinib vs. teriflunomide will be reported. In addition, the hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata and study ID (i.e. Study MS200527_0080 or PPD), along with 95% 2-sided CI, and (alpha-compatible) 95.125% 2-sided CI. An expanded model with a treatment-by-study interaction will be fit to the data and between-study heterogeneity will be

tested as the type-3 test of the treatment-by-study interaction; the corresponding p-value will be provided.

The assumption of non-informative censoring underpins the hazard ratio estimator implemented by the stratified Cox model that uses observed follow-up time for each participant, regardless of treatment completed/discontinued/ongoing status, where participants are censored for 12-week CDP at the EDSS assessment.

The trial-wise and family-wise-type-1 error will be preserved at the 0.025 1-sided level in the presence of multiple testing due to multiple endpoints (primary and secondary), as described in Section 9.16. In the setting of superiority trials, an IA for BSSR has a minimal or non-existent effect on type-1 error rate inflation (Friede 2019), so there is no need to adjust for the IA for BSSR conducted in August 2021.

Descriptive analysis

KM estimates (product-limit estimates) of the cumulative probability of experiencing 12-week CDP over time will be presented by intervention group, together with a summary of percentiles (5^{th}, 10^{th}, 15^{th}) and corresponding 2-sided 95% CIs. The KM curves will also be provided for each intervention group.

Number and percentages of participants experiencing intercurrent events will also be presented.

By-visit descriptive statistics for EDSS absolute and CFB values treated as continuous variables (i.e. mean, median, and range) will be reported. Descriptive statistics for categorical EDSS CFB, i.e. number and proportion improving, stable, or worsening, will also be reported for each time point. Here improvement is defined as a decrease of 1.0 point or more, stable condition as a change of no more than half a point in either direction, and worsening as an increase of 1.0 point or more.

14.2.1.2 Time to 24-week CDP

Disability progression is considered sustained for 24 weeks when the initial increase in the EDSS is confirmed at a regularly scheduled visit at least 24 weeks after the initial documentation of neurological worsening.

The rules for confirming disability progression for 24-week CDP are the same as those for 12-week CDP, with "168 days" replacing "84 days".

In all other respects, the analysis of the 24-week CDP endpoint will be the same as that of the 12-week CDP endpoint.

14.2.1.3 Sensitivity Analyses: Time to Confirmed Disability Progression

The following sensitivity analyses for time to 12-week CDP are planned, with the same handling of ICEs as in the primary analysis unless otherwise specified, based on pooled data from the 2 Phase 3 studies:

1. Analysis to evaluate the potential influence of informative treatment discontinuation according to discontinuation reason or occurrence of initial progression event at last assessment.

   This sensitivity analysis will use the same Cox model used in the primary analysis of 12-week CDP and report the resulting HR, 95% CI and p-value. In addition, a forest plot will present the hazard ratio, 95% CI and p-value for each delta.

   The following applies to participants in the evobrutinib intervention group. If a participant's discontinuation is unrelated to treatment (i.e. unrelated to efficacy or safety issues) and no initial progression event occurred, the participant's time to CDP will be multiply imputed using hazard rate in the comparator group (adjusted for participant's stratum), for the interval between discontinuation and planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first). If a participant's discontinuation is related to treatment or an initial unconfirmed progression event occurred, the participant's time to CDP will be multiply imputed using hazard rate higher than that in the comparator group (adjusted for participant's stratum), for the interval between discontinuation and planned end of treatment. The comparator hazard rate after withdrawal will be multiplied by delta, ranging from 1 to 3.1 (~tripling of hazard rate) in steps of 0.3. For participants with discontinuation of treatment (all reasons) in the teriflunomide group, the time to CDP will be multiply imputed under the MAR assumption.

   MI methods for time-to-event data will be applied following the approach from Lipkovich 2016 (delta adjustment), see Appendix 18.9 for details.

   Treatment termination with primary reason entered as 'adverse event', 'lack of efficacy' and 'death' in the eCRF TTERM form will be considered as related to treatment. Other treatment terminations (i.e. primary reason equal to 'lost to follow-up', 'protocol non-compliance', 'withdrawal by subject' and 'other') will be considered as unrelated to treatment.

2. Analysis in which initial disability progression not confirmed due to early treatment discontinuation is imputed as confirmed disability progression.

   This sensitivity analysis will use the same Cox model used in the primary analysis of 12-week CDP and report the resulting HR and 95% CI.

   If a participant discontinued treatment early during the on-treatment period after having an initial progression event, but prior to 12-week confirmation, the participant is assumed to have a 12-week CDP event at the date of initial progression.

3. Analysis to evaluate the potential impact of model misspecification.

   This sensitivity analysis will add the following covariates to the Cox model: number of relapses in previous 2 years before Screening, presence/absence of T1 Gd+ lesions within 6 months prior to randomization, prior DMT (y/n), and age. The resulting HR for treatment effect and 95% CI will be reported.

4. Analysis to evaluate the potential impact of COVID-19 infection.

   This sensitivity analysis will handle the ICE of COVID-19 infection with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the infection will be censored. This is in contrast to the primary analysis in which any COVID-19-related ICEs are ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of the infection will be assumed to be noninformative for the outcome of interest (CDP).

   The ICE 'COVID-19 vaccination' will be handled via Treatment Policy strategy as in primary analysis. Participants experiencing the ICE 'COVID-19 infection' will be identified as those with AEs related to COVID-19 as defined in Section 15.2.4. The hypothetical scenario envisaged is that COVID-19 infection does not occur, but COVID-19 vaccination does.

5. Analyses to evaluate the potential impact of the war in Ukraine.

   In the first sensitivity analysis of CDP related to the impact of the war in Ukraine, all participants from sites in Ukraine will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting only Ukraine participants) had not occurred. In a second sensitivity analysis related to the impact of war in Ukraine, all participants from sites in Ukraine, Russian Federation, and Belarus will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting participants from Ukraine, Russian Federation, and Belarus) had not occurred.

6. mFAS analysis.

   In this sensitivity analysis, the primary analysis of CDP will be repeated on the mFAS.

7. Analysis to evaluate the potential impact of the second recruitment cohort.

   This sensitivity analysis will add the following covariate to the Cox model: membership in second recruitment cohort (yes/no), as defined in Section 8.2. The resulting HR for treatment effect and 95% CI will be reported, along with the p-value for the additional covariate.

The sensitivity analyses above will be repeated for time to 24-week CDP.

8. Analysis with a censoring date of 01 June 2023 for all participants still ongoing at that date.

   To assess the impact of the PA trigger delay, participants will be censored on the originally projected date of the PA trigger: 01 June 2023. This sensitivity analysis will be performed for time to 12-week CDP only.

14.2.2 Time to 24-Week Confirmed Disability Improvement

The 5 estimand attributes of secondary efficacy estimands time to first occurrence of 24-week CDI are:

- **Endpoint**: Time to first occurrence of 24-week CDI as measured by the EDSS up to 156 weeks
- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.
- **Population**: Participants with RMS as defined by inclusion/exclusion criteria who have baseline EDSS ≥ 2.
- **Strategies used to address ICEs**:
  - Treatment discontinuation: Treatment policy strategy.
  - Death (any cause): While alive strategy
  - Ukraine crisis: Hypothetical strategy
  - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.
- **Population-level summary** comparing the intervention groups is the hazard ratio and CI based on a stratified Cox model for hazard rate, with term for intervention group, and strata defined by randomization strata and study ID. Test is based on stratified logrank test with strata defined by randomization strata and study ID. Data are pooled from the FAS of the studies 0080 and PPD.

This endpoint will be analyzed for the subgroup of participants from the FAS pooled studies with baseline EDSS score ≥ 2.0.

For participants with a baseline EDSS score ≥ 2 and ≤ 6, disability improvement is defined as a reduction in EDSS score ≥ 1.0 compared to baseline EDSS score. For participants with a baseline EDSS score ≥ 6.5 and ≤ 9.5, disability improvement is defined as a reduction in EDSS score of ≥ 0.5.

The same approach to the timing of 24-week confirmation of disability improvement is applied as for 24-week disability progression. Confirmation of disability improvement must occur at the regularly scheduled visit that is at least 24 weeks (168 days) after initial improvement. If a participant has a missing EDSS at the scheduled visit occurring at least 168 days after an initial improvement or the scheduled visit occurs several days before the 168-day window after an initial improvement, confirmation of the disability improvement must be based on the assessment at the next scheduled visit.

EDSS assessments at unscheduled or scheduled visits that are less than 168 days after the initial improvement are considered non-confirmatory EDSS assessments. For disability improvement to be confirmed at a regularly scheduled visit at least 24 weeks after initial improvement, any non-confirmatory EDSS assessments should be at least as low as the minimum change required for improvement. This means that after a scheduled or unscheduled visit at which the participant fulfills the initial disability improvement, all subsequent EDSS assessments (scheduled and unscheduled) need to also fulfill the improvement criteria until the improvement can be confirmed at the first scheduled visit that occurs at least 168 days after onset of the improvement.

Time to 24-week CDI is defined as the time from randomization date to the initial disability improvement date when an event is present, and from randomization date to censoring date when an event is not present. Censoring occurs in those participants who did not experience a 24-week confirmed disability improvement during 156 weeks post randomization; this includes participants who had a "tentative" disability improvement that could not be confirmed due to an early discontinuation or any other reason. The censoring date is the date of the last EDSS assessment. Participants with no post randomization EDSS assessment will be censored on the day of randomization.

Missing Data Handling

Participants who did not experience 24-week CDI by 156 weeks post randomization, by time of PA cutoff, by time of early study discontinuation, or before being lost to follow up, will be censored at the date of the last EDSS assessment during the 156 weeks post randomization. In the primary analysis of 24-week CDI, censoring will be assumed to be noninformative, conditional on intervention group, randomization stratum, and study ID.

Handling of ICEs

In accordance with Treatment Policy strategy, for participants experiencing the ICE premature treatment discontinuation, all EDSS data through Safety Follow-up will be included in the analysis up to the planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first).

In accordance with the While Alive strategy, for participants experiencing the ICE death (any cause), participants will be censored at time of death.

Participants at sites not fully operational post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data from such participants will be censored at the time at which the site lost fully operational status.

In accordance with the Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination' or 'emergency unblinding', the occurrence of these ICEs is considered irrelevant in the primary analysis of CDI.

Inferential analysis

The PA of time to 24-week CDI will be based on data pooled from the FAS of the 2 Phase 3 studies using a stratified log rank test with intervention group as main effect and stratification variables used in randomization and study ID (i.e. Study MS200527_0080 or _PPD) as covariates. The 1-sided p-value (α=0.024375) from the stratified log rank test for comparison between evobrutinib vs. teriflunomide will be reported. In addition, the hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata and study ID (i.e. Study MS200527_0080 or _PPD), along with 95% 2-sided CI, and (alpha-compatible) 95.125% 2-sided CI.

The assumption of non-informative censoring underpins the hazard ratio estimator implemented by the stratified Cox model that uses observed follow-up time for each participant, regardless of treatment completed/discontinued/ongoing status, where participants are censored for 24-week CDI at the third to last EDSS assessment.

Descriptive analysis

KM estimates (product-limit estimates) of the cumulative probability of experiencing 24-week CDI over time will be presented by intervention group, together with a summary of percentiles (5^{th}, 10^{th}, 15^{th}) and corresponding 2-sided 95% CIs based. The KM curves will also be provided for each intervention group.

Number and percentages of participants experiencing ICEs will also be presented.

Sensitivity analysis

Sensitivity analyses described for time to 12-week and 24-week CDP in Section 14.2.1.3 will be applied to time to 24-week CDI except for the sensitivity analyses 2 and 8.

For sensitivity analysis #1, the hazard rate after withdrawal will be divided by delta, ranging from 1 to 3.1 (reduction of hazard rate) in steps of 0.3.

14.2.3 PROMIS Scores

PF is assessed with the PROMISnq Short Form v2.0 – Physical Function – Multiple Sclerosis 15a (PROMISnq PF(MS) 15a). Fatigue is assessed with the PROMIS Short Form v1.0 – Fatigue – Multiple Sclerosis 8a (PROMIS Fatigue (MS) 8a). PROMIS PF and Fatigue scores are assessed in all participants at Baseline Visit, and every 12 weeks until the end of the study period up to 156 weeks.

14.2.3.1 PROMIS PF Score CFB over 96 Weeks

The 5 estimand attributes of secondary efficacy estimands CFB in PROMIS PF score are:

- **Endpoint**: CFB in PROMIS PF score over 96 weeks

- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.
- **Population**: Participants with RMS as defined by inclusion/exclusion criteria.
- **Strategies used to address ICEs**:
  - Treatment discontinuation: Treatment Policy strategy.
  - Death (any cause): While alive strategy
  - Ukraine crisis: Hypothetical strategy
  - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.
- **Population-level summary** comparing the intervention groups is the difference of average least-squares means of score CFB (average over Weeks 72, 84 and 96) and CI based on a MMRM, with test of treatment effect based on difference of average least-squares means. Data are pooled from the FAS of the studies 0080 and PPD .

Measures from the PROMIS PF item bank are scored on a T-score metric (higher scores = higher PF).

Missing Data Handling

For participants with missing PROMIS PF data, the missing data will be assumed to be MAR; such participants will be assumed to have the same mean PROMIS PF score CFB trajectory through 96 weeks as participants with available data within the same intervention group and stratum, and having the same baseline score and study ID.

Handling of ICEs

In this study, the ICE of treatment discontinuation is handled by the Treatment Policy strategy, in which data post premature treatment discontinuation, i.e. at Safety Follow-up Visit, are used in the analysis. However, PROMIS scores are not assessed at the Safety Follow-up Visit. Thus, the data to include in the analysis of PROMIS scores are the same, regardless of whether the participant is a treatment discontinuer or has treatment ongoing/completed: all available assessments while on treatment over 96 weeks post randomization are included in the analysis.

In accordance with the While Alive strategy, for participants experiencing the ICE death (any cause), all PROMIS PF data up to the time of death will be used in the analysis.

Participants at sites not fully operational post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data

from such participants will be censored at the time at which the site lost fully operational status. Censoring at the end of the pre-crisis period is noninformative for the outcome of interest (PROMIS PF Score CFB).

In accordance with the Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination' or 'emergency unblinding', the occurrence of these ICEs is considered irrelevant in the primary analysis of PROMIS PF Score CFB.

Inferential analysis

The population-level summary comparing the intervention groups will be the difference in average least-squares means, based on a MMRM for CFB, with average taken over Weeks 72, 84 and 96, where the model includes terms for intervention group, visit, intervention group by visit interaction, baseline score, baseline score by visit interaction, randomization strata, and study ID, and where the data are pooled from the FAS of studies 0080 and PPD . Data up to and including analysis visit Week 96 will be included in the model. The unstructured covariance matrix will be considered. Denominator degrees of freedom will be computed using Kenward and Roger's method (Kenward 1997). If the model fails to converge, the baseline score by visit interaction will be omitted.

The difference (comparing evobrutinib and teriflunomide) in average least-squares mean CFB, 95% 2-sided CI, (alpha-compatible) 95.125% CI and 1-sided p-value will be reported. For each intervention group, the adjusted average least-squares mean score CFB and associated 95% 2-sided CI will be reported. In addition, the results of the MMRM will be presented graphically by intervention group with points for the estimates per visit and error bars for the 95% CIs.

The MAR assumption underpins the estimator of difference of average LS means implemented by the MMRM that uses observed score CFB data from each participant, regardless of treatment completed/discontinued/ongoing status.

Descriptive analysis

A figure will be provided describing the distribution of average CFB taken over Weeks 72, 84 and 96, with one curve for each intervention group. The estimated empirical cumulative distribution function curve for average CFB will display proportion of participants having a value up to average CFB.

Descriptive statistics for PROMIS scores and CFB at each visit (including visits post Week 96) will be provided by intervention group.

Number and percentages of participants experiencing ICEs will also be presented.

- In addition, the raw mean score over time will be presented as a by-visit boxplot for each study intervention group in a single figure.

14.2.3.2 PROMIS Fatigue Score CFB over 96 Weeks

Measures from the fatigue item bank are scored on a T-score metric (higher scores = higher fatigue).

- The secondary estimand, CFB in PROMIS Fatigue score over 96 weeks, will be analyzed in the same manner as CFB in PROMIS PF score over 96 weeks, except that the average will be taken over Weeks 48, 60, 72, 84 and 96.

14.2.3.3 Sensitivity Analyses: PROMIS Score CFB

The following sensitivity analyses for PROMIS PF score CFB over 96 weeks are planned, with the same handling of ICEs as in the primary analysis, unless otherwise specified, based on pooled data from the 2 Phase 3 studies:

1. Analysis to evaluate the potential influence of informative treatment discontinuation according to discontinuation reason.

   This sensitivity analysis will use the same MMRM model used in the primary analysis of PROMIS PF Score CFB averaged over Weeks 72, 84 and 96 and report the resulting difference in average LS means due to treatment, 95% CI, and 1-sided p-value for the significance of the treatment effect. In addition, a forest plot will present the difference in average LS means, 95% CI and p-value for each delta.

   The following applies to participants in the evobrutinib intervention group. If a participant's discontinuation is unrelated to treatment (i.e. unrelated to efficacy or safety issues), the participant's score CFB at each timepoint will be multiply imputed using CFB distribution at that timepoint in the comparator group (adjusted for the participant's stratum and baseline score), for the interval between discontinuation and 96 weeks post randomization or PA cutoff, whichever occurs first. If a participant's discontinuation is related to treatment, the participant's score CFB at each timepoint will be multiply imputed using a worse CFB distribution at that timepoint than that in the comparator group (adjusted for the participant's stratum and baseline score), for the interval between discontinuation and 96 weeks post randomization or PA cutoff, whichever occurs first. The comparator CFB after withdrawal will be shifted by delta ranging from 1 to a maximum of 10 in steps of 1 (stopped at the point at which the difference comparing evobrutinib and teriflunomide is not statistically significant). The shift indicating worse CFB distribution is given as decreased score for PROMIS PF and as increased score for PROMIS Fatigue. For participants with discontinuation of treatment (all reasons) in teriflunomide group, the CFB will be multiply imputed under the MAR assumption.

   Control-based pattern imputation and delta-adjusted pattern imputation will be applied (Little 1996, O'Kelly 2014, Ratitch 2013) see Appendix 18.9 for details.

   Treatment termination with primary reason entered as 'adverse event', 'lack of efficacy' and 'death' in the eCRF TTERM form will be considered as related to treatment. Other treatment terminations (i.e. primary reason equal to 'lost to follow-up', 'protocol non-compliance', 'withdrawal by subject' and 'other') will be considered as unrelated to treatment.

2. Analysis to evaluate the potential impact of model misspecification.

   This sensitivity analysis will add the following covariates to the mixed model: number of relapses in the 2 years before Screening, presence/absence of T1 Gd+ lesions within 6 months prior to randomization, prior DMT (y/n), and age, as defined in Section 8.2. The same inferential results will be reported as for the primary analysis.

3. Analysis to evaluate the potential impact of COVID-19 infection.

   This sensitivity analysis will handle the ICE of COVID-19 infection with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the infection will be censored. This is in contrast to the primary analysis in which any COVID-19-related ICEs are ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of infection will be assumed to be non-informative for the outcome of interest (PROMIS score CFB).

   The ICE 'COVID-19 vaccination' will be handled via Treatment Policy strategy as in primary analysis. Participants experiencing the ICE 'COVID-19 infection' will be identified as those with AEs related to COVID-19 as defined in Section 15.2.4. The hypothetical scenario envisaged is that COVID-19 infection does not occur, but COVID-19 vaccination does.

4. Analysis to evaluate the potential impact of the war in Ukraine.

   In the first sensitivity analysis of PROMIS score CFB related to the impact of the war in Ukraine, all participants from sites in Ukraine will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting only Ukraine participants) had not occurred. In a second sensitivity analysis related to the impact of war in Ukraine, all participants from sites in Ukraine, Russian Federation, and Belarus will have their data censored at the start of the crisis. The hypothetical scenario envisaged is one in which the crisis (affecting participants from Ukraine, Russian Federation, and Belarus) had not occurred. The censored data will not be imputed; censoring at the time of the start of the crisis will be assumed to be noninformative for the outcome of interest (PROMIS Score CFB).

5. mFAS analysis.

   In this sensitivity analysis, the primary analysis of PROMIS score CFB will be repeated on the mFAS.

6. Analysis to evaluate the potential impact of the second recruitment cohort.

   This sensitivity analysis will add the following covariate to the mixed model: membership in second recruitment cohort (yes/no), as defined in Section 8.2. The same inferential results will be reported as for the primary analysis with the p-value of the additional covariate reported.

The sensitivity analyses for PROMIS Fatigue score CFB averaged over Weeks 48, 60, 72, 84 and 96 will be similar.

14.2.4 MRI Lesion Measures

The 5 estimand attributes of secondary MRI lesion measure efficacy estimands are:

- **Endpoint**:
  - Total number of T1 Gd+ lesions based on all available MRI scan.
  - Number of new or enlarging T2 (active T2) lesions based on the last available MRI scan relative to the baseline MRI scan.
- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.
- **Population**: Participants with RMS as defined by inclusion/exclusion criteria.
- **Strategies used to address ICEs**:
  - Treatment discontinuation: Treatment policy strategy.
  - Death (any cause): While alive strategy
  - Ukraine crisis: Hypothetical strategy
  - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.
- **Population-level summary** comparing the intervention groups is lesion rate ratio and CI based on NB model, with test of treatment effect based on test of lesion rate ratio, with offset defined in Sections 14.2.4.1 and 14.2.4.2.

Brain MRIs are scheduled at Screening and Weeks 24, 48, 96, 156 and early treatment discontinuation. Various MRI derived parameters have been related to disease activity, including T1 Gd+ lesions and active T2 lesions.

14.2.4.1 Total Number of T1 Gd+ lesions Based on All Available MRI Scans

The total number of T1 Gd+ lesions for a participant will be calculated as the sum of new T1 Gd+ lesions from non-missing post baseline scans. Log number of non-missing scans contributing to the total is used as the offset in NB modeling. Analyses are presented for each study separately based on FAS.

Missing Data Handling

For participants missing one or more scans, the missing data will be assumed to be MAR; such a participant will be assumed to experience lesions through planned end of treatment (i.e. PA cutoff or 156 weeks (Section 9.13), whichever occurs first) at the same rate as participants with available data within the same intervention group and stratum and having the same baseline lesion activity (presence/absence).

Handling of ICEs

In this study, the ICE of treatment discontinuation is handled by the Treatment Policy strategy, in which data post premature treatment discontinuation, i.e. at Safety Follow-up Visit, are used in the analysis. However, MRI scans are not scheduled at the Safety Follow-up Visit. Thus, the data to include in the analysis of T1 Gd+ lesions are the same, regardless of whether the participant is a treatment discontinuer or has treatment ongoing/completed: all available scans while on treatment are included in the analysis.

In accordance with the While Alive strategy, for participants experiencing the ICE death (any cause), all T1 Gd+ lesion data up to the time of death will be used in the analysis.

Participants at sites not fully operational, post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data from such participants will be censored at the time at which the site lost fully operational status. Censoring at the end of the pre-crisis period is non-informative for the outcome of interest (T1 Gd+ lesions).

In accordance with the Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination' or 'emergency unblinding', the occurrence of these ICEs is considered irrelevant in the primary analysis of T1 Gd+ lesions.

Inferential analysis

The population-level summary comparing the intervention groups is lesion rate ratio, based on a NB model for lesion count, with terms for intervention group, randomization strata, and baseline lesion activity (i.e. presence/absence of T1 Gd+ lesions at baseline), with offset equal to log number of total available scans over which the lesions experienced by a participant are observed. The adjusted lesion RR comparing evobrutinib to teriflunomide estimated from the NB model, 95% 2-sided CI, and 1-sided p-value will be reported, together with adjusted lesion rate for each intervention group and associated 95% 2-sided CI. In addition, the lesion RR will be expressed as percentage reduction relative to teriflunomide. The analysis of total number of T1 Gd+ lesions will be based on data from a single study (0080 or PPD).

The MAR assumption underpins the rate ratio estimator implemented by the NB model that uses observed lesion count and number of scans giving rise to that count for each participant, regardless of treatment completed/discontinued/ongoing status.

Descriptive analysis

Descriptive statistics for the number of new T1 Gd+ lesions by timepoint up to 156 weeks will be provided by intervention group.

Number and percentages of participants experiencing ICEs will also be presented.

Mean number of T1 Gd+ lesions at a given visit will be presented as a by-visit line plot for each intervention group, with a vertical line segment at each visit representing ± SE (jittered if needed for legibility), and with both intervention groups included in a single figure. Proportion of participants who are T1 Gd+ lesion-free will be presented as a by-visit line plot for each intervention group (vertical line segment at each visit representing 95% CI may be omitted for legibility), and with both intervention groups included in a single figure. In these figures, the spacing of tick marks on the horizontal axis should reflect the time interval between visits (i.e. visits that are nonuniformly spaced in time should be nonuniformly spaced on the horizontal axis).

Mean number of T1 Gd+ lesions per scan estimated based on the NB model, will be presented as a grouped bar chart by intervention group, with a vertical line segment for each bar representing 95% 2-sided CI, and with the p-value for the comparison with teriflunomide (based on NB model) displayed above the grouped bar chart pair.

14.2.4.2 Number of new or enlarging T2 lesions on the Last Available MRI Scan Relative to the Baseline MRI Scan

The number of new/enlarging T2 lesions on the last available scan relative to the baseline scan will be calculated as the sum of new/enlarging T2 lesions from non-missing scans. For each scan collected, NeuroRx reports the number of new/enlarging T2 lesions relative to the previous scan. Thus, for example, summing the new/enlarging T2 lesions from the Week 24 and Week 48 scans provides the sum of new/enlarging T2 lesions at Week 24 relative to baseline (i.e. Screening) and new/enlarging T2 lesions at Week 48 relative to Week 24.

If a participant has scans only at Week 24 and Week 96, the NeuroRx MRI reader will identify new/enlarging T2 lesions at Week 96 relative to the previously available scan at Week 24. Thus, summing the new/enlarging T2 lesions from the Week 24 and Week 96 scans provides the sum of new/enlarging T2 lesions at Week 24 relative to baseline (i.e. Screening) and new/enlarging T2 lesions at Week 96 relative to Week 24.

The time between the last available scan and the baseline scan (in years) is used as the offset in the NB modeling. The time (in years) from Screening scan will be calculated as (date of last scheduled MRI scan with a non-missing value for the number of new or enlarging T2 lesions – date of Screening scan +1)/365.25.

The same missing data handling and ICE strategy used in the analysis of Total number of T1 Gd+ lesions will be used in the analysis of Number of new/enlarging T2 lesions.

Inferential analysis

The population-level summary comparing the intervention groups is lesion rate ratio, based on a NB regression model for lesion count, with terms for intervention group, randomization strata, and baseline volume of T2 lesion (as continuous covariate), with offset equal to log time between the last available scan and baseline scan (in years). The adjusted lesion RR comparing evobrutinib to teriflunomide estimated from the NB model, 95% 2-sided CI, and 1-sided p-value will be reported, together with adjusted lesion rate for each intervention group and associated 95% 2-sided CI. In addition, the lesion RR will be expressed as percentage reduction relative to teriflunomide. The analysis of total number of new or enlarging T2 lesions will be based on data from a single study (0080 or PPD).

Descriptive analysis

Descriptive statistics for the number of new/enlarging T2 lesions by timepoint up to 156 weeks will be provided by intervention group.

Number and percentages of participants experiencing ICEs will also be presented.

Mean number of new/enlarging lesions at a given visit will be presented as a by-visit line plot for each intervention group, with a vertical line segment at each visit representing ± SE (jittered if needed for legibility), and with both intervention groups included in a single figure. Proportion of participants who are new/enlarging T2 lesion-free will be presented as a by-visit line plot for each intervention group (vertical line segment at each visit representing 95% CI may be omitted for legibility), and with both intervention groups included in a single figure. The spacing of tick marks on the horizontal axis should reflect the time interval between visits.

Mean number of new/enlarging T2 lesions per year estimated based on the NB model will be presented as a grouped bar chart by intervention group, with a vertical line segment for each bar representing 95% 2-sided CI, and with the p-value for the comparison with teriflunomide (based on NB model) displayed above the grouped bar chart pair.

14.2.4.3 Sensitivity / Supplementary Analyses: MRI Lesion Measures

The following sensitivity analyses for total number of T1 Gd+ lesions are planned, with the same handling of ICEs as in the primary analysis, unless otherwise specified:

1. Analysis to evaluate the potential influence of informative treatment discontinuation according to discontinuation reason.

   This sensitivity analysis will use the same NB model used in the primary analysis of total T1 Gd+ lesions based on all available MRI scans. The lesion rate ratio for treatment effect will be reported, together with 95% 2-sided CI, and 1-sided p-value for the significance of the treatment effect. In addition, a forest plot will present the lesion rate ratio, 95% CI and p-value for each delta.

The following applies to participants in the evobrutinib intervention group. If a participant's discontinuation is unrelated to treatment (i.e. unrelated to efficacy or safety issues), the participant's lesion count will be multiply imputed using the rate in the comparator group (adjusted for the participant's stratum and baseline lesion activity) based on CR MI method, for the interval between discontinuation and planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first). If a participant's discontinuation is related to treatment, lesion count will be multiply imputed at a rate higher than that in comparator group (adjusted for the participant's stratum and baseline lesion activity), for the interval between discontinuation and planned end of treatment. The comparator rate after withdrawal will be multiplied by delta, ranging from 1 to 3.1 (~tripling of relapse rate) in steps of 0.3. For participants with discontinuation of treatment (all reasons) in the teriflunomide group, the relapse count will be multiply imputed under the MAR assumption.

MI methods for count data are applied following the approach from Keene 2014, see Appendix 18.9 for details.

Treatment termination with primary reason entered as 'adverse event', 'lack of efficacy' and 'death' in the eCRF TTERM form will be considered as related to treatment. Other treatment terminations (i.e. primary reason equal to 'lost to follow-up', 'protocol non-compliance', 'withdrawal by subject' and 'other') will be considered as unrelated to treatment.

2. Analysis to evaluate the potential impact of COVID-19 infection.

   This sensitivity analysis will handle the ICE of COVID-19 infection with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the infection will be censored. This is in contrast to the primary analysis in which any COVID-19-related ICEs are ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of infection will be assumed to be non-informative for the outcome of interest (T1 Gd+ lesions).

   The ICE 'COVID-19 vaccination' will be handled via Treatment Policy strategy as in primary analysis. Participants experiencing the ICE 'COVID-19 infection' will be identified as those with AEs related to COVID-19 as defined in Section 15.2.4. The hypothetical scenario envisaged is that COVID-19 infection does not occur but COVID-19 vaccination does.

3. Analysis to evaluate the potential impact of the war in Ukraine.

   In the first sensitivity analysis of T1 Gd+ lesions related to the impact of the war in Ukraine, all participants from sites in Ukraine will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting only Ukraine participants) had not occurred. In a second sensitivity analysis related to the impact of war in Ukraine, all participants from sites in Ukraine, Russian Federation, and Belarus will have their data censored at the start of the crisis. The hypothetical scenario envisaged is one in which the crisis (affecting participants from Ukraine, Russian Federation, and Belarus) had not occurred. The censored data will not be imputed; censoring at the time of the start of the crisis will be assumed to be noninformative for the outcome of interest (T1 Gd+ lesions).

4. mFAS analysis.

   In this sensitivity analysis, the primary analysis of T1 Gd+ lesions will be repeated on the mFAS.

5. Analysis to evaluate the potential impact of the second recruitment cohort.

   This sensitivity analysis will add the following covariate to the NB model: membership in second recruitment cohort (yes/no), as defined in Section 8.2. The same inferential results will be reported as for the primary analysis with the p-value of the additional covariate reported.

The following supplementary analysis is planned to evaluate the potential impact of model misspecification:

Analysis wherein lesion count at a single scan are considered as NB-distributed longitudinal data and analyzed using a mixed effect model, with a random effect for participant, and terms for intervention group, randomization strata, and baseline lesion activity. This model will compare treatment on the basis of conditional adjusted lesion rate ratio, where conditioning is on the random effect. See Appendix 18.9 for details.

Sensitivity analyses for number of new or enlarging T2 lesions on the last available MRI scan relative to the baseline MRI scan will be similar, except that the offset in the model will be log time from baseline scan to last available scan (in years), in place of log numbers of scans. In the supplementary analysis in which new/enlarging T2 lesion count at a single scan is modeled as NB-distributed, the offset parameter will be equal to the time (in years) between the scan in question and the previous available scan, representing the time interval over which new/enlarging T2 lesions accrued.

14.2.5 Neurofilament Light Chain Concentration at Week 12

The 5 estimand attributes of this secondary efficacy estimand are:

- **Endpoint**: NfL concentration at 12 weeks
- **Treatment condition of interest**: Evobrutinib 45 mg twice daily for up to 156 weeks; alternative treatment condition to which comparison will be made: teriflunomide 14 mg once daily for up to 156 weeks.
- **Population**: Participants with RMS as defined by inclusion/exclusion criteria.

- **Strategies used to address ICEs**:
    - Treatment discontinuation: Treatment policy strategy.
    - Death (any cause): While alive strategy
    - Ukraine crisis: Hypothetical strategy
    - COVID-19 infection, COVID-19 vaccination and emergency unblinding: Treatment Policy strategy.
- **Population-level summary** comparing the intervention groups is the ratio of geometric means at 12 weeks and CI based on a MMRM model for log(NfL concentration), with test of treatment effect based on difference of least-squares means.

Serum NfL concentrations are assessed in all participants at Baseline Visit and Weeks 12, 24, 48, 72, 96, 120, 144, 156, end of DBTP and early treatment discontinuation.

Missing Data Handling

For participants with missing NfL concentration data, the missing data will be assumed to be MAR; such participants will be assumed to have the same mean NfL concentration trajectory through planned end of treatment (i.e. PA cutoff or 156 weeks, whichever occurs first) as participants with available data within the same intervention group and stratum, and having the same baseline concentration.

Handling of ICEs

In this study, the ICE of treatment discontinuation is handled by the Treatment Policy strategy, in which data post premature treatment discontinuation, i.e. at Safety Follow-up Visit, are used in the analysis. However, NfL assessments are not scheduled at the Safety Follow-up Visit. Thus, the data to include in the analysis of NfL concentration are the same, regardless of whether the participant is a treatment discontinuer or has treatment ongoing/completed: all available assessments while on treatment are included in the analysis.

In accordance with the While Alive strategy, for participants experiencing the ICE death (any cause), all NfL concentration data up to the time of death will be used in the analysis.

Participants at sites not fully operational post start of Ukraine crisis will be considered to have experienced the ICE 'Ukraine crisis'. This ICE will be handled using the Hypothetical strategy, where the hypothetical scenario envisaged is one in which the crisis had never happened. The data from such participants will be censored at the time at which the site lost fully operational status. Censoring at the end of the pre-crisis period is non-informative for the outcome of interest (NfL concentration).

In accordance with the Treatment Policy strategy, for participants experiencing the ICEs 'COVID-19 infection', 'COVID-19 vaccination' or 'emergency unblinding', the occurrence of these ICEs is considered irrelevant in the primary analysis of NfL concentration.

Inferential analysis

The population-level summary comparing the intervention groups will be the ratio of geometric means, based on a MMRM for log(NfL concentration), where the model includes terms for intervention group, visit, intervention group by visit interaction, log(baseline concentration) and randomization strata. The unstructured covariance matrix will be considered. Denominator degrees of freedom will be computed using Kenward and Roger's method (Kenward 1997). The analysis of NfL concentration will be based on FAS data from a single study (0080 or PPD).

The ratio (comparing evobrutinib and teriflunomide) of geometric means, 95% 2-sided CI, and 1-sided p-value will be reported. For each intervention group, the adjusted geometric mean at Week 12 and associated 95% 2-sided CI will be reported. In addition, the results of the MMRM will be presented graphically by intervention group with points for the estimates per visit and error bars for the 95% CIs.

The MAR assumption underpins the estimator of ratio of geometric means through Week 12 implemented by the MMRM that uses observed NfL concentration data from each participant, regardless of treatment completed/discontinued/ongoing status.

Descriptive analysis

The mean and median (min, max) observed values for each intervention group will be plotted versus visit till the planned end of treatment (i.e. PA cutoff or 156 Weeks post randomization, whichever occurs first). Median percent change from baseline line plot will also be provided.

Descriptive statistics for NfL concentration data will be provided by intervention group.

Number and percentages of participants experiencing ICEs will also be presented.

14.2.5.1 Sensitivity Analyses: Neurofilament Light Chain Concentration

The following sensitivity analyses for NfL concentration are planned, with the same handling of ICEs as in the primary analysis, unless otherwise specified:

1. Analysis to evaluate the potential impact of COVID-19 infection.

   This sensitivity analysis will handle the ICE of COVID-19 infection with the Hypothetical strategy, i.e. all of a participant's data on the day of and after the infection will be censored. This is in contrast to the primary analysis in which any COVID-19-related ICEs are ignored in the analysis (i.e. Treatment Policy strategy). Censoring at the time of infection is assumed to be noninformative for the outcome of interest (NfL concentration).

The ICE 'COVID-19 vaccination' will be handled via Treatment Policy strategy as in primary analysis. Participants experiencing the ICE 'COVID-19 infection' will be identified as those with AEs related to COVID-19 as defined in Section 15.2.4. The hypothetical scenario envisaged is that COVID-19 infection does not occur but COVID-19 vaccination does.

2. Analysis to evaluate the potential impact of the war in Ukraine.

   In the first sensitivity analysis of NfL concentration related to the impact of the war in Ukraine, all participants from sites in Ukraine will have their data censored at the start of the crisis (i.e. 24 February 2022). The hypothetical scenario envisaged is one in which the crisis (affecting only Ukraine participants) had not occurred. In a second sensitivity analysis related to the impact of war in Ukraine, all participants from sites in Ukraine, Russian Federation, and Belarus will have their data censored at the start of the crisis. The hypothetical scenario envisaged is one in which the crisis (affecting participants from Ukraine, Russian Federation, and Belarus) had not occurred. The censored data will not be imputed; censoring at the time of the start of the crisis will be assumed to be noninformative for the outcome of interest (NfL concentration).

3. mFAS analysis.

   In this sensitivity analysis, the primary analysis of NfL concentration will be repeated on the mFAS.

4. Analysis to evaluate the potential impact of the second recruitment cohort.

   This sensitivity analysis will add the following covariate to the MMRM model: membership in second recruitment cohort (yes/no), as defined in Section 8.2. The same inferential results will be reported as for the primary analysis with the p-value of the additional covariate reported.

14.2.6 Subgroup Analyses for Secondary Estimands

The analysis of the secondary estimands, 12-week CDP, 24-week CDP, and 24-week CDI, as described in Sections 14.2.1.1, 14.2.1.2 and 14.2.2, will be repeated for all subgroup levels defined in Section 8.2 "Subgroup Definition and Parametrization". Additionally, a larger model that adjusts for subgroup and intervention-by-subgroup interaction will be fitted, and the p-value of the interaction reported.

14.3 Tertiary/Exploratory Endpoints

Analyses of tertiary/exploratory endpoints will be based on the FAS. The handling of ICEs will be analogous to the primary and secondary endpoints in Sections 14.1.1 and 14.2. Details are provided in each subsection below.

14.3.1 ARR at Weeks 48 and 96

The analysis of ARR up to 48 weeks (96 weeks) will follow that of ARR up to 156 weeks described in Section 14.1.1. For participants discontinuing treatment, any qualified relapses occurring through Safety Follow-up will be included in the analysis up to 48 (96 respectively) weeks post randomization (see Sections 9.11 and 9.12 respectively).

14.3.2 Time to First Qualified Relapse over 156 Weeks

Time to first qualified relapse is defined as the time from randomization date to the first qualified relapse date. The handling of ICEs will be analogous to the primary endpoint, described in Section 14.1.1. Participants who did not experience qualified relapse by 156 weeks post randomization, by time of PA cutoff, by time of early study discontinuation, or before being lost to follow up, will be censored at the date of the last relapse assessment during the 156 weeks post randomization. Censoring will be assumed to be noninformative for qualified relapse, conditional on intervention group, and randomization stratum.

The hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata, along with 95% 2-sided CI. A test of treatment effect will be based on a stratified log rank test with strata defined by randomization strata.

The cumulative distribution function will be estimated by KM method by intervention group.

14.3.3 Qualifying Relapse-free Status at Week 96

The proportion of participants qualifying relapse-free at Week 96 (measured from randomization, see Section 9.12) will be compared between intervention groups using Cochran-Mantel-Haenszel (CMH) χ^2 test stratified by randomization strata.

Participants who discontinue study prior to Week 96, without having a qualified relapse will be counted as not being qualified relapse-free. The handling of ICEs will be analogous to the primary endpoint, described in Section 14.1.1.

For each intervention group, the proportion will be reported, with 95% 2-sided CI.

In addition, the estimated probability of being qualified relapse-free at Week 96, based on the KM method (corresponding to the analysis in Section 14.3.2) will be reported, with 95% 2-sided CI.

14.3.4 12-week (24-week) Confirmed EDSS Progression-Free Status at Week 96

The proportion of participants that are 12-week confirmed EDSS progression-free (i.e. 12-week CDP-free) at Week 96 (measured from randomization, see Section 9.12) will be compared between intervention groups using CMH χ^2 test, stratified by randomization strata and study ID, based on data pooled from both Phase 3 studies.

The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.1.1. Participants who discontinue study prior to Week 96 or having not reached Week 96 at time of planned end of treatment, without having 12-week confirmed EDSS progression will be counted as not being 12-week confirmed EDSS progression-free.

For each intervention group, the proportion will be reported, with 95% 2-sided CI.

In addition, the estimated probability of being 12-week confirmed EDSS progression-free at Week 96, based on the KM method (corresponding to the analysis in Section 14.2.1.1) will be reported, with 95% 2-sided CI.

The same approach will be used for the analysis of 24-week confirmed EDSS progression-free status.

14.3.5 24-week Confirmed Disability Improvement Status at Week 96

This endpoint will be analyzed for the subgroup of participants from the pooled studies with baseline EDSS score ≥ 2.0. The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.2.

For participants with a baseline EDSS score ≥ 2 and ≤ 6, disability improvement is defined as a reduction in EDSS score ≥ 1.0 compared to baseline EDSS score. For participants with a baseline EDSS score ≥ 6.5 and ≤ 9.5, disability improvement is defined as a reduction in EDSS score of ≥ 0.5.

The same approach to the timing of 24-week confirmation of disability improvement is applied as for 24-week disability progression. Confirmation of disability improvement must occur at the regularly scheduled visit that is at least 24 weeks (168 days) after initial improvement. If a participant has a missing EDSS at the scheduled visit occurring at least 168 days after an initial improvement or the scheduled visit occurs several days before the 168-day window after an initial improvement, confirmation of the disability improvement must be on the basis of the assessment at the next scheduled visit.

EDSS assessments at unscheduled or scheduled visits that are < 168 days after the initial improvement are considered non-confirmatory EDSS assessments. For disability improvement to be confirmed at a regularly scheduled visit at least 24 weeks after initial improvement, any non-confirmatory EDSS assessments should be at least as low as the minimum change required for improvement.

A participant without confirmed disability improvement will be counted as not improved, independent of follow-up time. The intervention group comparison will be based on a CMH χ^2 test stratified according to randomization strata and study ID.

14.3.6 Lesion-Free Status at Week 96

The proportion of participants having lesion-free status at Week 96 will be compared between intervention groups using CMH χ^2 test stratified by randomization strata. The status at Week 96 will be based on measurements performed from randomization up to day 694 (which corresponds to the study day of Week 96 Visit [673] + 21 days). If the Week 96 scan was performed prior to day 666 (day 673 – 7 days) and was not repeated prior to day 694, it will be considered as missing for this analysis.

The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.4.1.

Participants who discontinued study early and/or are missing one or more of the 3 scans planned through Week 96 (i.e. at Weeks 24, 48, 96), will be counted as not lesion-free, independent of follow-up time.

For each intervention group, the proportion will be reported, with 95% 2-sided CI.

This analysis will be done for the following endpoints:

- New T1 Gd+ lesion-free status
- New or enlarging T2 lesion-free status
- CUA lesion-free status

14.3.7 Total Number of Lesions Based on All Available MRI Scans

The analysis of total number of CUA lesions based on all available scans, will follow that of the number of new/enlarging T2 lesions described in Section 14.2.4.2 including the handling of ICEs. Baseline lesion activity will be adjusted for using the following covariates: baseline T1 Gd+ lesions (absence/presence) and baseline T2 lesion volume (continuous). For participants discontinuing treatment, all available lesion data will be included in the analysis up to the planned end of treatment (i.e. PA cutoff or 156 weeks post randomization, whichever occurs first).

A similar analysis will be applied to total number of new T1 hypo-intense lesions based on all available MRI scans, with baseline lesion activity adjusted for using baseline T1 Gd+ lesions (absence/presence).

14.3.8 Change in Volume of Lesions from Baseline to Week 96

Volume of lesions is assessed in all participants at Screening, Weeks 24, 48, 96, 156 and early discontinuation of treatment. The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.4.1.

Cube-root transformed lesion volume CFB will be modeled via MMRM, with intervention group, visit, intervention group by visit interaction, baseline value of cube-rooted volume, interaction between baseline value and visit, and covariates defined by randomization strata as fixed effects. Estimation of treatment effect (and 95% 2-sided CI) will be based on adjusted difference in

least-squares means from the model. All available volume of lesions data (irrespective of discontinuation) will be included in the analysis up to and including the Week 96 analysis visit.

This analysis will be done for the following endpoints:

- change in volume of T1 Gd+ lesions from Baseline to Week 96
- change in volume of T2 lesions from Baseline to Week 96

Descriptive statistics by visit and intervention group will be provided.

14.3.9 Percentage Change in Brain Volume from Week 24 to Week 96

Percent change in volume (brain volume, thalamic volume, or cortical grey matter volume), relative to Week 24, is assessed in all participants at Weeks 48, 96, 156, and early discontinuation of treatment, during the double-blind period.

Brain volume is recorded as an absolute "normalized" value at the Baseline Visit then recorded at subsequent visits as a percentage change relative to the absolute value at the Baseline Visit. Therefore, brain volume at Week 24 will be calculated as the brain volume at the Baseline Visit multiplied by 1 + ([percentage change in brain volume from Baseline visit to Week 24]/100). This value will be used to determine the percentage change in brain volume at Weeks 48 and 96 relative to Week 24.

The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.4.1.

PBVC from Week 24 will be modeled via MMRM, with intervention group, visit, intervention group by visit interaction, Week 24 value of volume, interaction between Week 24 value and visit, and covariates defined by randomization strata as fixed effects. An unstructured covariance matrix will be assumed. Should the model fail to converge, consideration will be given to using a covariance structure with fewer parameters (i.e. compound symmetry), or omitting the baseline value-visit interaction from the model. Denominator degrees of freedom will be computed using Kenward and Roger's method (Kenward 1997). Missing scan assessments will be handled via the MAR assumption. Continuous covariates will be centered by the mean, where mean is based on all participants in FAS.

Estimation of treatment effect (and 95% 2-sided CI) and p-value comparing intervention groups, based on adjusted difference in PBVC least-squares means will be reported for Week 48 and Week 96 timepoints. In addition, the adjusted PBVC least-squares means estimate (and 95% 2-sided CI) for each intervention group will be reported at Week 48 and Week 96. All available data (irrespective of discontinuation) will be included in the analysis up to and including the Week 96 analysis visit.

This analysis will be done for the following endpoints:

- percentage change in BV from Week 24 to Weeks 48 and 96
- percentage change in thalamic volume from Week 24 to Weeks 48 and 96
- percentage change in cortical grey matter from Week 24 to Weeks 48 and 96

Descriptive statistics by visit and intervention group will be provided.

14.3.10 Change in Normalized T1 Intensity within Pre-Existing Non-Enhancing T2 Weighted Lesion Volume from Baseline to Week 96

Change in nT1 intensity within pre-existing non-enhancing T2 weighted lesion volume, from baseline to Week 96, is a "time course" variable reported in normalized T1 units. These units represent a scale of relative intensity where the median intensity of normal appearing grey matter from a given participant is mapped to zero and the median intensity of normal appearing white matter is mapped to 1. A value below zero represents a normalized intensity that is lower than the participant's normal appearing grey matter and likewise a value above 1 is a normalized intensity above that of the participant's normal appearing white matter. The intensity at each assessment visit is defined once at the end of the period. Participants require a minimum of 2 scans to have an assessment of nT1 intensity. Missing values will not be imputed.

The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.4.1.

Change from baseline to post baseline visit in nT1 intensity for post baseline at Week 24, Week 48 and Week 96 Visits will be derived. The Hodges-Lehman estimate (and 95% 2-sided CI) of shift in distribution of Week 96 nT1 intensity CFB between intervention groups, stratified according to randomization strata and baseline T2 lesion volume category (based on tertiles) will be reported. The Wilcoxon rank sum test, stratified according to randomization strata and baseline T2 lesion volume category (based on tertiles), will be used to compare the intervention groups. All available data (irrespective of discontinuation) will be included in the analysis up to and including the Week 96 analysis visit.

In addition, intervention group differences across the distribution of the change from baseline in nT1 intensity will be presented based on median regression run on the 50^{th} quantile.

Descriptive statistics will be provided by study intervention group. In addition, boxplots of the change from baseline by intervention group and timepoint will be provided.

14.3.11 Volume of SELs Based on Scans at Weeks 24, 48, and 96

Volume of SELs based on scans through Week 96 is a "time course" variable reported in mm^3 units: a single value is available per participant for the DBTP, either at Week 96 if participant completed that visit or at Early Discontinuation Visit if participant discontinued treatment early. Participants require a minimum of 3 scans to have a SEL assessment. Missing values will not be imputed.

SELs are identified as areas of pre-existing T2 lesions of at least 10 contiguous voxels (voxel size is ~ 3 mm^3) showing gradual and constant concentric expansion (Elliott 2019). SELs may qualify as a potential read-out for progressive accumulation of irreversible neural tissue damage and especially axonal loss (van Walderveen 1998; Filippi 2012).

The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.4.1.

The Hodges-Lehman estimate (and 95% 2-sided CI) of shift in distribution of SEL volume between intervention groups, stratified according to randomization strata and baseline T2 lesion volume category (based on tertiles) will be reported for Week 96 (Early Discontinuation Visit assessment would be included only if it falls in the Week 96 analysis visit window). The Wilcoxon rank sum test, stratified according to randomization strata and baseline T2 lesion volume category will be used to compare the intervention groups at Week 96. All available data (irrespective of discontinuation) will be included in the analysis up to and including the Week 96 analysis visit.

In addition, intervention group differences across the distribution of SEL volume will be presented based on median regression run on the 50th quantiles.

SELs volume as a percentage of baseline T2 lesion volume will be derived and analyzed in the same way than volume of SELs. Percent SEL volume is based on the ratio of SEL volume to baseline T2 lesion volume.

Descriptive statistics will be provided by study intervention group.

14.3.12 Change in Number of Phase Rim Lesions at Weeks 24, 48 and 96

The change in number of Phase Rim Lesions endpoint will not be analyzed as part of the PA but will be analyzed later. The corresponding analysis will therefore be described in a separate IAP.

14.3.13 Change in Cerebral Blood Flow in Normal-Appearing White Matter at Weeks 24, 48 and 96

The Cerebral Blood Flow endpoint will not be analyzed as part of the PA but will be analyzed later. The corresponding analysis will therefore be described in a separate IAP.

14.3.14 Time to ≥ 20% Increase (Confirmed at 12 Weeks) in T25-FW (9-HPT) up to 156 Weeks

Time to ≥ 20% increase, confirmed at 12 weeks, in T25-FW is defined as the time from randomization date to the date of the initial worsening of ≥ 20%. The worsening is defined based on the baseline value. In general, the value of T25-FW is taken to be the average scores of 2 T25-FW attempts. If the test result is missing due to "physical limitation", the maximum possible value on the scale is to be imputed (i.e. 180 seconds). If the test result of one attempt is missing, but not due to a "physical limitation," the result from the other attempt will be used to impute the missing value.

The handling of ICEs will be analogous to the secondary endpoint, time to 12-week CDP, described in Section 14.2.1.1. In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as 12-week ≥ 20% increase. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, participants will be censored at time of death. (i.e. death will not be counted as 12-week ≥ 20% increase).

Participants who did not experience ≥ 20% increase, confirmed at 12 weeks, by 156 weeks post randomization (Section 9.13), by time of PA cutoff, by time of early study discontinuation, or before being lost to follow up, will be censored at the date of the last T25-FW assessment during the 156 weeks post randomization. Censoring will be assumed to be noninformative for 12-week confirmed worsening in T25-FW, conditional on intervention group, randomization stratum, and study ID. Confirmation of ≥ 20% increase must occur at the regularly scheduled visit that is at least 12 weeks (84 days) after initial worsening. All subsequent assessments (scheduled and unscheduled) from initial worsening need to also fulfill the worsening criteria until the worsening can be confirmed at the first scheduled visit that occurs at least 84 days after onset of the worsening.

Based on data pooled from both Phase 3 studies, the hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group, strata defined by randomization strata, study ID, and baseline T25-FW score as continuous covariate, along with 95% 2-sided CI. A test of treatment effect will be based on a stratified log rank test.

The cumulative distribution function will be estimated by KM method by intervention group.

A similar analysis will be applied to Time to ≥ 20% increase (confirmed at 12 weeks) in 9-HPT. In general, the value of 9-HPT is taken to be the average scores of the 4 9-HPT attempts (2 attempts on the right hand and 2 attempts on the left-hand). If the test result is missing due to "physical limitation", the maximum possible value on the scale is to be imputed (i.e. 300 seconds). If the test result of one attempt is missing, but not due to a "physical limitation," the result from the other attempt on the same hand will be used to impute the missing value. If results from both attempts of the same hand are missing, then the average score from the other hand will be used to impute the missing value.

14.3.15 NEDA-3 at Week 48, Week 96

NEDA at a timepoint is defined as meeting the following 3 criteria at that timepoint:

- Qualifying relapse-free status
- New T1 Gd+ lesions free status and new or enlarging T2 lesion free status
- 12-week confirmed EDSS progression free status

The handling of ICEs will be analogous to the primary (for relapses) and secondary (for MRI lesions and 12-week CDP) endpoints, described in Sections 14.1.1 and 14.2.1.1, respectively.

Participants who reach the timepoint of interest will be considered as having evidence of disease activity if at least one qualified relapse, at least one MRI scan showing disease activity (defined as Gd+ T1 lesions, or new or enlarging T2 lesions), or at least one 12-week CDP event, was reported up to the timepoint of interest. Otherwise the participant will be considered to have NEDA.

Participants who discontinue study early with at least one event before early discontinuation will be considered as having evidence of disease activity.

Even if an event was not reported before early study discontinuation, the participant will be considered to show evidence of disease activity if the reason for early discontinuation is lack of efficacy or death; otherwise, the participant's NEDA status will be considered a missing observation.

The proportion of participants with NEDA will be compared between intervention groups using the CMH χ^2 test stratified by randomization strata.

Descriptive statistics will be provided by intervention group.

The analysis will be repeated considering all relapses, not only qualified relapses.

14.3.16 Time to 12-week Confirmed Disability Based on Composite Score

Time to first occurrence of 12-week CDP up to 156 weeks based on a composite score is defined by 3 components:

- 12-week CDP on EDSS or;
- 12-week confirmed progression (≥ 20%) in T25-FW versus baseline or;
- 12-week confirmed progression (≥ 20%) in 9-HPT versus baseline.

The handling of ICEs will be analogous to the secondary (for 12-week CDP) and tertiary (for 12-week confirmed progression in T25-FW and 9-HPT) endpoints, described in Sections 14.2.1.1 and 14.3.14, respectively.

This endpoint is analyzed for participants pooled from both Phase 3 studies. The hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata and study ID (i.e. Study MS200527_0080 or _PPD), along with 95% 2-sided CI. The test of treatment effect will be based on a stratified logrank test of distribution of time to event with strata defined by randomization strata and study ID. Cumulative distribution function will be estimated via KM method by intervention group.

The assumption of non-informative censoring underpins the Cox model and KM estimator.

Descriptive statistics will indicate proportion of composite events due to each of the 3 component events, by study intervention group.

14.3.17 NEP at Week 48, Week 96

NEP at a timepoint is defined as meeting the following 3 criteria at that timepoint:

- No 12-week CDP on EDSS
- No 12-week confirmed progression (≥ 20%) in 9-HPT score
- No 12-week confirmed progression (≥ 20%) in T25-FW time

The handling of ICEs will be analogous to the secondary (for 12-week CDP) and tertiary (for 12-week confirmed progression in T25-FW and 9-HPT) endpoints, described in Sections 14.2.1.1 and 14.3.14, respectively.

Participants who reach the timepoint of interest will be considered as having evidence of progression if at least one 12-week CDP event, at least one 12-week confirmed 9-HPT score worsening event or at least one 12-week confirmed T25-FW score worsening event was reported up to the timepoint of interest. Otherwise the participant will be considered to have NEP.

Participants who discontinue study early with at least one event before early discontinuation and the timepoint of interest will be considered as having evidence of progression.

Even if an event was not reported before early study discontinuation and the timepoint of interest, the participant will be considered to show evidence of progression if the reason for early discontinuation is lack of efficacy or death; otherwise, the participant's NEP status will be considered a missing observation.

The proportion of participants with NEP will be compared between intervention groups using the CMH χ^2 test stratified by randomization strata.

Descriptive statistics will be provided by intervention group.

14.3.18 NEPAD at Week 48, Week 96

NEPAD at a timepoint is defined as meeting the following 5 criteria at that timepoint:

- No qualified relapses on treatment
- No 12-week CDP on EDSS
- No 12-week confirmed progression (≥ 20%) in 9-HPT score
- No 12-week confirmed progression (≥ 20%) in T25-FW time
- No new or enlarging T2 lesions and no new T1 Gd+ lesions on MRI

The handling of ICEs will be analogous to the primary (for relapses), secondary (for 12-week CDP and MRI lesions) and tertiary (for 12-week confirmed progression in T25-FW and 9-HPT) endpoints, described in Sections 14.1.1, 14.2.1.1, 14.2.4.1 and 14.3.14, respectively.

Participants who reach the timepoint of interest will be considered as having evidence of progression or active activity if at least one qualified relapse, at least one 12-week CDP event, at least one 12-week confirmed 9-HPT score worsening event, at least one 12-week confirmed T25-FW score worsening event, or at least one MRI scan showing disease activity (defined as Gd+ T1 lesions or new or enlarging T2 lesions) was reported up to the timepoint of interest. Otherwise the participant will be considered to have NEPAD.

Participants who discontinue study early with at least one event before early discontinuation and timepoint of interest will be considered as having evidence of progression or active disease.

Even if an event was not reported before early study discontinuation and timepoint of interest, the participant will be considered to show evidence of progression or active disease if the reason for early discontinuation is lack of efficacy or death; otherwise, the participant's NEPAD status will be considered a missing observation.

The proportion of participants with NEPAD will be compared between intervention groups using the CMH χ^2 test stratified by randomization strata.

Descriptive statistics will be provided by intervention group.

The analysis will be repeated considering all relapses, not only qualified relapses.

14.3.19 CFB in SDMT score at Week 48 and Week 96

SDMT scores are assessed in all participants at Screening, Baseline Visit, and every 12 weeks until the end of the DBTP at Week 156.

Based on data pooled from both Phase 3 studies, the analysis of SDMT score CFB at Week 48 and Week 96 analysis visits, will follow that of PROMIS score CFB at a given timepoint (Section 16.4.1), including the handling of ICEs.

In addition, analyses on improvement or worsening status will be performed (Benedict 2021). Participants will be classified into the following categories, at Week 48 and Week 96 based on the initial date of improvement or worsening:

- Sustained improvement: defined as a $\geq$ 4-point increase from baseline sustained for 6 months until end of DBTP, regardless of worsening at earlier visits. Hence all results must meet the criteria for 6 months and onwards, until the End of Treatment assessment or PA cutoff, whichever occurs first.

- Sustained worsening: defined as a $\geq$ 4-point decrease from baseline sustained for 6 months until End of Treatment or PA cutoff, whichever occurs first.

- Stable: defined as not meeting sustained improvement or worsening.

If a participant is missing the baseline result or post-baseline results at least 6 months after the baseline measurement, their status will be considered missing.

The handling of ICEs will be analogous to the secondary endpoint, time to 12-week CDP, described in Section 14.2.1.1. In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as sustained worsening. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, participants will be censored at time of death (i.e. death will not be counted as sustained worsening).

Descriptive statistics will be provided for all categories by intervention group.

The proportion of participants with sustained improvement/worsening will be compared between intervention groups using the CMH $\chi 2$ test stratified by randomization strata.

14.3.20 PIRA up to Week 156

PIRA up to Week 156 is defined during the on-treatment period for these 2 progression criteria separately:

- 12-week CDP on EDSS
- 12-week cCDP defined by 3 components (12-week CDP on EDSS, 12-week confirmed worsening in T25-FW and 9-HPT versus baseline, see Section 14.3.16)

PIRA (Lublin 2022) is defined as a 12-week CDP event (respectively 12-week cCDP) with either no prior qualified relapse, or an onset more than 30 days after the start date of the last qualified relapse (irrespective of the EDSS confirmation). In addition, to qualify as PIRA, no relapse must occur within 30 days before or after the EDSS confirmation. If a relapse with incomplete recovery occurred (meaning that the EDSS is not the same as before the relapse), the baseline (i.e. the reference EDSS value) is reset > 30 days after the relapse onset to identify the next PIRA. In an individual participant, the baseline could be reset multiple times (i.e. after each relapse) until either a PIRA event is discovered, or until the individual EDSS profile ended. In case of missing EDSS before or after the relapse, the relapse will be assumed to be recovered and the baseline will not be reset.

The handling of ICEs will be analogous to the primary (for relapses), secondary (for 12-week CDP) and tertiary (for 12-week confirmed progression in T25-FW and 9-HPT) endpoints, described in Sections 14.1.1, 14.2.1.1 and 14.3.14, respectively.

This endpoint is analyzed for participants pooled from both Phase 3 studies. Time to PIRA is defined as the time from randomization date to the first 12-weeks CDP (cCDP respectively) event date when an event is present, and from randomization to censoring date when an event is not present. Censoring occurs in those participants who did not experience any of the 12-week confirmed disability progression events up to the PA cutoff or 156 weeks post randomization, whichever occurs first; this includes participants who had a "tentative" disability progression that could not be confirmed due to an early discontinuation or any other reason. The censoring date is the date of the last assessment among EDSS. For cCDP, the censoring date is the minimum of the last available EDSS assessment date, the last available 9-HPT assessment date and the last available T25-FW assessment date.

The hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata and study ID (i.e. Study MS200527_0080 or _PPD), along with 95% 2-sided CI. The test of treatment effect will be based on a stratified logrank test of distribution of time to event with strata defined by randomization strata and study ID. Cumulative distribution function will be estimated via KM method by intervention group.

The assumption of non-informative censoring underpins the Cox model and KM estimator.

Descriptive statistics will indicate proportion of composite events due to each of the 3 component events, by intervention group.

The analysis considering the time to CDP will be repeated considering all relapses, not only qualified relapses.

In addition, a sensitivity analysis will be considering time to PIRA as the time to CDP (cCDP) in the hypothetical scenario for the ICE 'relapse'. Considering that the qualified relapse would not happen, the analyses described above will be repeated censoring participants with a qualified relapse at the time of the first qualified relapse.

14.3.21 PIRMA up to Week 156

PIRMA up to Week 156 is defined as meeting the following criteria during the on-treatment period:

- 12-week CDP on EDSS
- 12-week cCDP defined by 3 components (12-week CDP on EDSS, 12-week confirmed worsening in T25-FW and 9-HPT versus baseline, see Section 14.3.16)

PIRMA is defined as a 12-week CDP event (respectively 12-week cCDP) with either no prior qualified relapse and no new or enlarging T2 lesions and no new T1 Gd+ lesions, or an onset (i.e. IDP) more than 30 days after the start date of the last qualified relapse or the date new or enlarging T2 lesions or new T1 Gd+ lesions are identified (irrespective of the EDSS confirmation). In addition, to qualify as PIRMA, no relapse and no new or enlarging T2 lesions and no new T1 Gd+ lesions on MRI must occur within 30 days before or after the 12-week confirmation. If a relapse with incomplete recovery occurred (meaning that the EDSS is not the same as before the relapse), the baseline (i.e. the reference EDSS value) is reset > 30 days after the relapse onset to identify the next PIRMA. In an individual participant, the baseline could be reset multiple times (i.e. after each relapse) until either a PIRMA event is discovered, or until the individual EDSS profile ended. In case of missing EDSS before or after the relapse, the relapse will be assumed to be recovered and the baseline will not be reset.

The handling of ICEs will be analogous to the primary (for relapses), secondary (for 12-week CDP and MRI lesions) and tertiary (for 12-week confirmed progression in T25-FW and 9-HPT) endpoints, described in Sections 14.1.1, 14.2.1.1, 14.2.4.1 and 14.3.14, respectively.

This endpoint is analyzed for participants pooled from both Phase 3 studies. Time to PIRMA is defined as the time from randomization date to the first 12-weeks CDP (cCDP respectively) event date when an event is present, and from randomization to censoring date when an event is not present. Censoring occurs in those participants who did not experience any of the 12-week confirmed disability progression events up to the PA cutoff or 156 weeks post randomization, whichever occurs first; this includes participants who had a "tentative" disability progression that could not be confirmed due to an early discontinuation or any other reason. The censoring date is the date of the last assessment among EDSS. For cCDP, the censoring date is the minimum of the last available EDSS assessment date, the last available 9-HPT assessment date and the last available T25-FW assessment date.

The hazard ratio comparing evobrutinib to teriflunomide will be estimated via a stratified Cox model with a term for intervention group and strata defined by randomization strata and study ID (i.e. Study MS200527_0080 or _PPD), along with 95% 2-sided CI. The test of treatment effect will be based on a stratified logrank test of distribution of time to event with strata defined by randomization strata and study ID. Cumulative distribution function will be estimated via KM method by intervention group.

The assumption of non-informative censoring underpins the Cox model and KM estimator.

Descriptive statistics will indicate proportion of composite events due to each of the 3 component events, by intervention group.

The analysis considering the time to CDP will be repeated considering all relapses, not only qualified relapses.

In addition, a sensitivity analysis will be considering time to PIRMA as the time to CDP (cCDP) in the hypothetical scenario for the ICEs 'relapse', 'new or enlarging T2 lesions' and 'T1 Gd+ lesions'. Considering that the ICEs would not happen, the analyses described above will be repeated censoring participants with a relapse at the time of the first relapse or with new or enlarging T2 lesions at the time of the lesion or with T1 Gd+ lesions at the time of the lesion, whichever occur first.

15 Safety Analyses

This section includes specifications for summarizing safety endpoints that are common across clinical studies such as AEs, laboratory tests and vital signs. If no participants are excluded from the SAF due to lack of safety data robustness, then there will be no distinct mSAF analysis set, and all safety analyses will be presented for each study separately on the SAF. If an mSAF is defined via exclusion of SAF participants due to lack of safety data robustness, safety analyses will be presented for each study separately on the mSAF and SAF.

Data handling for safety endpoints is described in Table 10 for the PA.

Table 10 Data Handling for Safety Analysis

Data summarized	Analysis	Analysis sets	Period covered	Treatment groups
AEs Laboratory data Vital signs ECGs Other safety evaluations	PA	SAF	All data from the on-treatment period until the PA cutoff date (Safety Follow-up included)	Evobrutinib 45 mg twice daily Teriflunomide 14 mg once daily

15.1 AEs

All analyses described in this section will be based on TEAEs if not otherwise specified.

TEAEs for the on-treatment period will be defined as:

- AEs starting on or after first study intervention administration of any study intervention until the PA cutoff date
- or if it was present prior to any study intervention administration but exacerbating after.
- Any AE which started before first administration of any study intervention but improved during the on-treatment period will not be counted as TEAE.

AEs with changes in toxicity grade/severity, seriousness or outcome of AEs are recorded as separate entries in the eCRF with associated end and start dates (start date equals end date of previous entry, supported in eCRF by 'AENEWID' in SUPPAE). Records of the same AE will be considered as one event in the analysis. If the severity of the reported event worsens after start of study intervention, the TEAE flag will be re-evaluated for the worse and the subsequent records as per the TEAE definition. If the worse record starts outside of the on-treatment period, it will not appear on the summaries/listings of TEAEs, unless otherwise specified. These events will be kept as separate records in the database in order to maintain the full detailed history of the events. When such AEs are listed, start and end date should be provided together with change date, toxicity grade/severity, outcome and seriousness per episode.

15.1.1 All AEs

AEs will be coded according to the latest MedDRA version available at the time of analysis. The severity of AEs will be graded using NCI-CTCAE version 5.0 toxicity grades. AEs with missing classification concerning study intervention relationship will be considered related to the study intervention.

15.1.1.1 3 Tier Approach

The 3-tier approach is a systematic way to summarize and analyze AEs in clinical studies (Crowe 2009). AEs in different tiers are analyzed using different levels of statistical analyses.

The AEs defined by the Benefit Risk Action Team for Tier 1 reporting in this study are in Table 11.

Table 11 Tier 1 AEs

Tier 1 AE (PT)	Within group summary	Between group comparison (Evobrutinib - Teriflunomide) 95% CI
Drug Induced Liver Injury	n (%)	Δ [xx%, xx%]

Tier 1 AE: DILI is an umbrella term and will be identified using sponsor-defined list of search term as defined in Appendix 18.5 and all corresponding PTs will be included in the Tier 1 summary.

This table will be repeated for AEs starting within 6 months (180 days) of treatment start.

All TEAEs (including those in Tier 1) will be classified into Tier 2 or Tier 3 based on the Rule-of-4. If there are 4 or more participants with the reported term in any study intervention group, that term will be included in Tier 2. Otherwise, it will be included in Tier 3.

The Tier 1 and Tier 2 AEs will be assessed with a 95% CI for between-group comparisons. For the difference in crude rates, expressed in percentages, the CIs will be based on MN method (Miettinen & Nurminen 1985). For Tier 1 events which were observed for less than 4 participants the CI will not be displayed in the Tier 1 table. The analyses will be done for all Tier 2 TEAEs.

EAIR of TEAEs will also be separately presented for each Tier, by PT (see Section 15.1 for multiple events handling). Difference in EAIR will be summarized. For Tier 1 and Tier 2 AEs, the CIs will be calculated using the Poisson based method. EAIR are calculated as number of participants with AE divided by the sum of the individual times of all participants in the safety population from start of treatment to first onset of AE. If a participant has multiple events, the exposure period of the first event after first administration of study intervention is used. For a participant with no event, the exposure period is censored at the last follow-up time for the AE summarization period. The incidence rate multiplied with 100 would give the number of AEs expected in 100 participants within 1-time unit (for example 1 year). An example of SAS code is available in the Statistical Methodology document noted in Appendix 18.9.

No multiplicity adjustment will be applied for Tier 1 and 2 AEs. The Tier 3 AEs will be assessed via summary statistics and risk differences without any CIs.

For the comparison of evobrutinib with teriflunomide, forest trees for Tier 1 and Tier 2 AEs, and EAIR AEs will be provided as well, displaying the incidence rate and associated 95% CI of difference.

15.1.1.2 Overview of TEAEs

2 types of summary table of TEAEs will be provided as described in Table 12.

Table 12 Summary Tables of TEAEs

Summary	Modalities
Overview of TEAEs	• Any TEAE • Related TEAE • Serious TEAE • Related serious TEAE • TEAE with NCI-CTCAE Grade ≥1 • Related TEAE with NCI-CTCAE Grade ≥1 • TEAE with NCI-CTCAE Grade ≥2 • Related TEAE with NCI-CTCAE Grade ≥2 • TEAE with NCI-CTCAE Grade ≥3 • Related TEAE with NCI-CTCAE Grade ≥3 • TEAE with NCI-CTCAE Grade ≥4 • Related TEAE with NCI-CTCAE Grade ≥4 • TEAE leading to death • Related TEAE leading to death
Overview of TEAEs leading to Actions: • Change in dose • Study intervention interruption or withdrawal • Administration of medication • Procedure • Study termination	• TEAE with no change of dose • Related TEAE with no change of dose • TEAE leading to interruption of study intervention • Related TEAE leading to interruption of study intervention • TEAE leading to withdrawal of study intervention • Related TEAE leading to withdrawal of study intervention • TEAE leading to administration of concomitant medication • Related TEAE leading to administration of concomitant medication • TEAE leading to concomitant procedure • Related TEAE leading to concomitant procedure • TEAE leading to study termination • Related TEAE leading to study termination

15.1.1.3 Tabulation of AEs by SOC and PT

The TEAE tables to be prepared are listed below:

Table 13 TEAE Tables

	Overall frequency	By primary SOC and PT	By PT only	By primary SOC, PT and worst grade
TEAE overview summary	✓	*NA*	*NA*	*NA*
TEAE leading to discontinuation of study intervention /study overview summary	✓	*NA*	*NA*	*NA*
All TEAEs	✓	✓	✓	✓
Serious TEAEs	✓	✓		
Non-serious TEAEs*	✓	✓		
TEAEs leading to study intervention withdrawal	✓	✓		
TEAEs leading to study termination	✓	✓		
TEAEs leading to death	✓	✓		
Treatment-related TEAEs	✓	✓	✓	✓
Treatment-related serious TEAEs	✓	✓		
Treatment-related TEAEs leading to treatment withdrawal	✓	✓		
Treatment-related TEAE leading to study termination	✓	✓		
Treatment-related TEAEs leading to death	✓	✓		

(*): 2 tables will be provided: 1) a table with all non-serious TEAEs by SOC and PT, and 2) a table with only TEAEs exceeding a frequency of 5% on the PT level in at least one of the intervention groups (> 5%), by SOC and PT.

EAIR will also be summarized as described in Section 15.1.1.3. The EAIR of TEAEs to be prepared are listed below:

Table 14 EAIR of TEAE Tables

	Overall frequency	By primary SOC and PT	By primary SOC, PT and worst grade
All TEAEs	✓	✓	✓
Serious TEAEs	✓	✓	*NA*
Study intervention-related TEAEs	✓	✓	*NA*
Study intervention-related Serious TEAEs	✓	✓	*NA*

Specific rules for SOC/PT tabulation

All AEs recorded during the study (i.e. assessed from signature of informed consent until the end of the Follow-up/End of Study Visit) will be coded according to the MedDRA and assigned to a SOC and PT.

SOC terms will be sorted alphabetically. For the PA, each intervention group will be displayed in the table. PTs within each SOC will be sorted by descending frequency of the evobrutinib intervention group, and then alphabetically if multiple PTs have the same frequency.

If a participant experiences more than one occurrence of the same TEAE (same SOC and same PT) during the study, the participant will be counted only once for that study intervention (the worst severity and the worst relationship to study intervention will be tabulated).

In case a participant had events with missing and non-missing grades, the maximum of the non-missing grades will be displayed.

Incomplete AE-related dates will be handled as specified in Section 9.14.

EAIR of TEAEs will be presented by SOC and PT.

Participant data listings

A listing of TEAEs leading to withdrawal of study intervention and leading to study termination, if any, will be provided as well.

15.1.1.4 Sensitivity Analyses of AEs

There will be 2 sensitivity analyses to evaluate the potential impact of the war in Ukraine on TEAE/serious TEAE EAIR. The first sensitivity analysis will consider that all participants in Ukraine experience the ICE 'Ukraine crisis' on 24 February 2022. The ICE will be handled by the

Hypothetical strategy (i.e. all Ukraine participants will have their data censored post start of crisis). The second sensitivity analysis will consider that all participants in Ukraine, Russian Federation, and Belarus experience the ICE 'Ukraine crisis' on 24 February 2022. Again, the ICE will be handled by the Hypothetical strategy (i.e. all participants from Ukraine, Russian Federation, and Belarus will have their data censored post start of crisis).

15.2 Deaths, Other Serious AEs, and Other Significant AEs

15.2.1 Deaths

A summary of deaths will be provided including (clinicaltrials.gov requirement):

- Number and percentage of (all) deaths
- Number and percentage of the primary cause of death (categories: progressive disease and/or disease related condition, event unrelated to study intervention, event related to study intervention, unknown)

The tabulation of TEAEs leading to death is described in Section 15.1.1.3. A listing of deaths, if any, will be provided.

In case there is no death in the study, only the summary of death required by clinicaltrials.gov will be performed, neither tabulation of TEAE leading to death will be produced, nor the listing of death.

15.2.2 Serious AEs

The tabulation of serious TEAEs is described in Section 15.1.1.3. A participant listing of serious TEAEs will be provided.

15.2.3 TEAEs of Special Interest

The following events are defined as AESI:

- Liver related AEs
- Infections:
  - All Infections reported as Serious AEs or Grade ≥ 3
  - All opportunistic infections
- Amylase and lipase elevations including acute pancreatitis
- Seizures

AESI will be identified using SMQ if available or sponsor-defined list of search term (refer to Appendix 18.5).

An overview of AESI will be presented by intervention group showing number and percentage of participants experiencing AESI. Study intervention related AESI will also be presented. AESIs will be flagged in the AE listings.

15.2.4 Other Significant TEAEs

Tables and a listing of participants with COVID-19 related TEAEs will be produced. TEAEs related to COVID-19 will be retrieved from COVID-19 SMQ (narrow scope). TEAEs related to COVID-19 vaccinations (see definition below) will not be considered as related to COVID-19.

Additionally, tables of participants with TEAEs associated with COVID-19 vaccinations will be produced by intervention group. TEAEs associated with COVID-19 vaccinations will be defined as TEAEs meeting all of the following criteria:

1. High level term equal to "Vaccination related complications" OR "Vaccination site reactions" OR lowest level term equal to "Vaccination adverse reaction"

2. Flagged in SDTM AE: "Causality factors other than study treatment" free text field containing specific terms related to COVID-19 vaccination

The TEAE tables to be prepared are listed below:

Table 15 TEAE Tables to be produced

	By primary SOC and PT	By primary SOC, PT and worst grade
TEAEs related to COVID-19	✓	✓
TEAEs related to COVID-19 leading to withdrawal of study intervention	✓	*NA*
Serious TEAEs related to COVID-19	✓	*NA*
Study intervention-related TEAEs related to COVID-19	✓	✓
TEAEs associated with vaccinations for COVID-19	✓	✓
Serious TEAEs associated with vaccinations for COVID-19	✓	*NA*
Study intervention-related TEAEs associated with vaccinations for COVID-19	✓	*NA*

15.3 Clinical Laboratory Evaluation

The following laboratory parameters will be measured during the study as part of the safety evaluation:

- Hematology
- Biochemistry, including the following LFT:
  - Aspartate aminotransferase
  - Alanine aminotransferase
  - Alkaline phosphatase
  - γ-Glutamyl-transferase
  - Total Bilirubin
- Urinalysis
- Coagulation

The clinical laboratory safety tests to be measured in this study are provided in the protocol (refer to Appendix 5 of the CTP). Parameters from Appendix 5 of the CTP to be summarized and listed in the TLFs are provided in Appendix 18.6.

Listings and summary statistics at each assessment time will be presented using the SI units. Normal ranges will be provided by the central laboratory, and out of range flags will be calculated based on the normal ranges. Laboratory data not transferred from the central laboratory in SI units will be converted to SI units before processing. Both original units and SI units will be provided in the SDTM domain.

For hematology, biochemistry and coagulation descriptive statistics by visit, data from central labs will be utilized for all tables and figures. If the central lab is missing for a visit but a local lab is present, the local lab values will be used. The local lab results will be assigned to analysis visits as described in Section 9.8. Normal ranges provided by the local lab will be used for further derivations. Both central and local lab data will be listed.

For urinalysis, local lab data will be utilized for TLFs. A specific listing for urine microscopy will be provided (central lab data).

Continuous clinical laboratory data (hematology, biochemistry, urinalysis, coagulation) will be summarized by study intervention using descriptive statistics at baseline and each post baseline visit for absolute value and changes from baseline (see Section 9.2 and Appendix 18.6). Hematology, biochemistry and urinalysis data will be presented by analysis visit, while coagulation data will be presented by nominal visit (see Section 9.8 and Appendix 18.4).

As a sensitivity analysis, descriptive statistics by visit will be repeated using only central lab data for the following parameters: Neutrophil count, Amylase, Lipase, ALT, AST.

Some laboratory results will be classified according to NCI-CTCAE Version 5.0. Some of the toxicity gradings are based on laboratory measurements in conjunction with clinical findings. Non-numerical qualifiers (except for fasting flags) will not be taken into consideration in the derivation of CTCAE criteria (e.g. hypokalemia Grade 1 and Grade 2 are only distinguished by a non-numerical qualifier and therefore Grade 2 will not be derived). In case a laboratory parameter has bi-directional toxicities (e.g. Potassium) both directions will be presented for the given parameter (i.e. Potassium Low and Potassium High).

Additional laboratory results that are not part of NCI-CTCAE will be presented according to the categories: below normal limit, within normal limits and above normal limit (according to the laboratory normal ranges). Laboratory results containing a modifier such as "<" or ">=" will be handled as per Appendix 18.7 for summary statistics and both results as collected in the database and as imputed will be reported in participant data listings.

Shift tables of baseline versus post baseline will be presented by intervention group for hematology and biochemistry.

- Laboratory gradable parameters part of NCI-CTCAE will be used to summarize changes from baseline Grade to worst on-treatment grade.
    - In case of gradings involving baseline measurements for the identification of grades during the on-treatment period, the shift table will present baseline normal and abnormal. Normal will include measurements below and within normal range (direction increase), or measurements within and above normal range (direction decrease).
- Laboratory parameters that are not part of NCI-CTCAE will be used to summarize changes from baseline finding to worst on-treatment finding.

Participant data listings will be provided, with a flag for abnormal values, along with corresponding normal ranges:

- Laboratory gradable parameters part of NCI-CTCAE will be presented according to the categories based on normal ranges along with the grade. Abnormal values will be flagged according to the direction of toxicity as detailed in Appendix 18.6 (e.g. for a parameter such as Potassium Low, only values below the LLN will be flagged).
- Laboratory parameters that are not part of NCI-CTCAE will be presented according to the categories based on normal ranges: below normal limits (Low), within normal limits (Normal), and above normal limits (High). Values that are either above ULN or below LLN will be flagged.

Boxplots of the laboratory values by intervention group and time point will be provided for the following parameters included in Table 16:

Table 16 Laboratory Parameters

Category	Laboratory parameter	Conventional Unit	SI Units
Hematology	Hemoglobin	g/dL	g/L
	Red blood cell count	10^6/µL	10^{12}/L
	Reticulocyte count	10^3/µL	10^9/L
	White blood cell count	10^3/µL	10^9/L
	Neutrophil count	10^3/µL	10^9/L
	Lymphocyte count	10^3/µL	10^9/L
	Platelet count	10^3/µL	10^9/L
Biochemistry	Alanine aminotransferase (ALT)	U/L	U/L
	Albumin	g/dL	g/L
	Aspartate aminotransferase (AST)	U/L	U/L
	Gamma-glutamyl transferase (GGT)	U/L	U/L
	Alkaline phosphatase	U/L	U/L
	Total bilirubin	mg/dL	µmol/L
	Amylase	U/L	U/L
	Lipase	U/L	U/L
	eGFR	mL/min/1.73m^2	mL/min/1.73m^2
	Blood urea nitrogen (BUN)	mg/dL	mmol/L

Boxplots for laboratory parameters where toxicity grades are defined based on the ratio of the parameter values and the ULN will not be displayed using the unit of measurement but instead using the ratio of the measured value over ULN. This comprises ALP, ALT, AST, bilirubin, GGT, Lipase and Amylase. Where feasible, a reference line for CTCAE grade ≥ 3 should be added to graphical displays.

Liver function tests: ALT, AST, and total bilirubin are used to assess possible drug induced liver toxicity. The ratios of test result over ULN will be calculated and classified for these 3 parameters during the on-treatment period.

Summary of liver function tests will include the following categories. The number and percentage of participants with each of the following categories during the on-treatment period will be summarized by study intervention group:

- ALT $\geq 3\times$ULN, ALT $\geq 5\times$ULN, ALT $\geq 8\times$ULN, ALT $\geq 10\times$ULN, ALT $\geq 20\times$ULN
- AST $\geq 3\times$ULN, AST $\geq 5\times$ULN, AST $\geq 8\times$ULN, AST $\geq 10\times$ULN, AST $\geq 20\times$ULN
- (ALT or AST) $\geq 3\times$ULN, (ALT or AST) $\geq 5\times$ULN, (ALT or AST) $\geq 8\times$ULN, (ALT or AST) $\geq 10\times$ULN, (ALT or AST) $\geq 20\times$ULN
- TBILI $\geq 1.5\times$ULN, TBILI $\geq 2\times$ULN

- Concurrent ALT ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent ALT ≥ 3×ULN and TBILI ≥ 1.5×ULN
- Concurrent AST ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent AST ≥ 3×ULN and TBILI ≥ 1.5×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 1.5×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN and ALP > 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 1.5×ULN and ALP > 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN and ALP ≤ 2×ULN or missing
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 1.5×ULN and ALP ≤ 2×ULN or missing

In addition, the following categories will be summarized in a separate table, limited to the first 6 months (180 days) after treatment start date:

- ALT ≥ 5×ULN
- ALP ≥ 2×ULN
- Concurrent ALT ≥ 5×ULN and TBILI ≥ 2×ULN
- Concurrent ALP ≥ 2×ULN and TBILI ≥ 2×ULN

Concurrent measurements are those occurring on the same date.

Categories will be cumulative, i.e. a participant with an elevation of AST ≥ 10×ULN will also appear in the categories ≥ 8×ULN , ≥ 5×ULN and ≥ 3×ULN.

A listing will be produced, displaying all ALT, AST, TBILI and ALP results for participants with any LFT elevation (AST ≥ 3×ULN, ALT ≥ 3×ULN or TBILI ≥ 1.5×ULN).

In addition, the following graphical displays will be provided for the on-treatment period:

- ALT values over time will be presented in a spaghetti plot for participants with ALT Grade 2 or higher elevation (i.e. Grade ≥ 2) by intervention group
- Time in weeks from first dose to first ALT occurrence of Grade 2 or higher elevation (i.e. Grade ≥ 2) for all participants by intervention group (KM plot)
- Time in weeks from first dose to first eGFR occurrence equal to less than 60 mL/min/1.73m^2 for participants with baseline eGFR ≥ 60 mL/min/1.73m^2 by intervention group (KM plot)
- e-DISH figure, as described in Merz et al (2014) will be presented including 2 panels (for PA): one for each intervention group.

For the 2 time-to-event figures, at the time of PA, a participant ongoing and without experiencing the event will have the event time right-censored at the time of the last ALT/eGFR assessment

before cutoff. To be included in a time-to-event figure, a group must have at least one event. Below the horizontal axis of the time-to-event figure, number of participants at risk will be displayed. The vertical axis may be restricted to 0.50 – 1.0, if none of the curves reach 50%.

In these studies, clinically significant lab abnormalities were recorded as AEs. In lieu of a listing of clinically significant lab abnormalities for each domain, the following by-participant lab value listings will be provided:

- Listing of post baseline Grade ≥ 3 hematology values
- Listing of post baseline Grade ≥ 3 biochemistry values
- Listing of post baseline urinalysis values with an increase of "++" for non-gradable parameters when applicable.
- Listing of post baseline Grade ≥ 2 AST, ALT or Bilirubin. For each parameter, when a participant has an increase to Grade 2, all posterior values will be included in the listing.

Per Section 7.1 of the MS200527-0080/PPD Protocol Version 6, study intervention is to be temporarily discontinued in the following scenarios:

- If a participant had an increase in ALT or AST to > 3 x to < 5 ULN. A recheck of the value within 72 hours is required, to confirm the reported elevation.
- If AST and/or ALT > 3 × ULN to < 5 × ULN are confirmed within 72 hours and have increased by more than 50% (compared to latest confirmed value), the interruption of the study intervention continues, with an additional recheck 72 hours later.
    - If the rechecked value in the second recheck has decreased to < ULN or < participant's baseline, rechallenge can be considered in the absence of hyperbilirubinemia.
    - In all other cases, study intervention is to be permanently discontinued.
- If AST and/or ALT > 3 × ULN to < 5 × ULN are confirmed within 72 hours and have increased by less than 50% (compared to latest confirmed value), the interruption of the study intervention continues with an additional recheck in 1 week.
    - If rechecked value in the second recheck has decreased to < ULN or < participant's baseline, rechallenge can be considered in the absence of hyperbilirubinemia.
    - In all other cases, study intervention is to be permanently discontinued.

To support assessment of the fraction of temporarily discontinued participants who could reinitiate treatment if allowed 1 week, 2 weeks, 3 weeks, 4 weeks to recover, the time to first ALT and AST meeting rechallenging criterion is to be characterized among participants who discontinued treatment (temporarily or permanently) for ALT or AST > 3 x ULN based on SAF.

(1) For a participant who has at least one date where ALT or AST > 3 x ULN (local lab ALT/AST assessments taken into consideration), x_1 is defined as the earliest elevation date.

(2) The treatment discontinuation date x_2 is defined as the earliest date among (i) "No dose" date ≥ x_1 from 'Drug Administration' CRF, (ii) "Start" date ≥ x_1 for AE record indicating "Drug Interrupted" for "Action taken with study treatment" from 'AE' CRF, and (iii) "last

administration date" $\geq x_1$ from 'Treatment Termination' CRF. Participants for whom both x_1 and x_2 can be defined comprise the subgroup of interest for time to event analysis.

A participant is considered to have "ALT or AST > 3 x ULN" if (a) ALT > 3 x ULN, (b) AST > 3 x ULN, or (c) both ALT > 3 x ULN and AST > 3 x ULN. For the participant to be considered recovered, the participant must have both ALT ≤ 3 x ULN and AST ≤ 3 x ULN.

Participants in the subgroup of interest who have ALT and AST recovery by the planned end of treatment, will have the time to first ALT and AST recovery calculated as "(x_1 earliest date at which both ALT ≤ 3 x ULN and AST ≤ 3 x ULN) – (x_2 date of treatment discontinuation due to ALT or AST elevation) + 1". Note that the "earliest" date of recovery is relative to the date of treatment discontinuation due to ALT or AST elevation.

Participants in the subgroup of interest who do not have ALT and AST recovery by the PA cutoff date or prior to study termination, or are lost to follow up, will be censored at the date of the last available ALT or AST assessment. The censoring time will be calculated as "(date of censoring) – (date of treatment discontinuation due to ALT or AST elevation) + 1".

KM estimates (product-limit estimates) of the cumulative probability of first meeting rechallenge criterion over time will be presented by intervention group, together with a summary of percentiles (25th, 50th, 75th) and corresponding 2-sided 95% CIs. Summary statistics (mean, SD, median, minimum and maximum) will be presented for follow-up time. The KM curves will also be provided for each intervention group. The KM figure and supportive KM tables will be repeated based on the following 2 subgroups of participants:

- Participants who had ALT/AST elevation > 3 - < 5 x ULN at treatment discontinuation
- Participants who had ALT/AST elevation ≥ 5 x ULN at treatment discontinuation

Participants with ALT/AST values ≤ 3 x ULN at treatment discontinuation are included in the subgroup ALT/AST > 3 - < 5 x ULN and reported in footnotes.

Treatment gap analysis will explore whether longer treatment pause in study intervention affects efficacy. This analysis will also be produced on pooled data from both studies. Treatment gaps of interest will be categorized as following:

(1) Gap between End of Treatment and End of Study Visits (gap 1):

   a. Early treatment discontinuation followed by Safety Follow-up (gap 1a)

   b. Planned End of Treatment followed by Safety Follow-up (gap 1b).

   In case End of Treatment reason is "Death" or "Lost to follow-up", it will not be considered as a gap.

(2) Treatment gap after first study intervention administration due to any reason, obtained from CRF as following (gap 2):

   a. 'Drug Administration' records where "No dose", with reason "Adverse event" or "Other", is selected in response to the question "Is there a change in dose?" (gap 2a).

   b. 'AE' records where "Drug Interrupted" or "Drug Withdrawn" is selected for "Action taken with study treatment" (gap 2b).

Gap start for category 1 (i.e. gap 1) is End of Treatment date and gap end date will be imputed as Safety Follow-up Visit end date even though drug is not administered during this visit. To determine non-overlapping gaps, gap start for category 2 (i.e. gap 2) is defined as the earliest of "No dose" date (i.e. gap 2a) and AE start date (i.e. gap 2b) and ends with first study intervention administration record after the gap start date. In case gap 1 and gap 2 occur on the same day, information for gap 1 will be presented. Multiple non-overlapping gaps can occur for category gap 2, all gaps from category 2a and all gaps from category 2b will be presented.

Participants in the subgroup of interest will have the gap duration in days calculated as the difference between start and stop dates plus 1 (gap duration [days] = gap end date - gap start date + 1). Only gaps with duration longer than 7 days and duration shorter or equal to 30 days will be presented.

Reason for gap is premature End of Treatment reason for gap 1a, treatment completion for gap 1b, no dose administration reason for gap 2a and AE for gap 2b. In case gap 2a and gap 2b start on the same day, AE is presented as the reason for gap.

Efficacy assessments (e.g. relapse, EDSS and MRI) at start of gap and at end of gap will be presented for the participants in the subgroup of interest. For relapse/EDSS assessment/MRI result at start of gap, the last non-missing relapse/EDSS assessment/MRI result on or before the start of gap is selected. For relapse/EDSS assessment/MRI result at end of gap, the first non-missing relapse/EDSS assessment/MRI result on or after the end of gap is selected.

Details of treatment gaps and corresponding efficacy information before/after treatment gap will be presented in a listing based on the SAF. Summary tables will also be produced to describe number of gaps and reason, as well as treatment gap impact on efficacy assessment.

Efficacy Evaluation of Participants who Restarted Study Intervention after Interruption due to ALT/AST Elevation

These analyses are meant to investigate the effect of interruption of treatment on ARR and EDSS progression, by presenting 3 subgroups side by side:

(1) Participants who interrupted IMP due to ALT/AST elevations and restarted treatment after the rechallenging criterion were met;

(2) The complementary FAS subgroup (participants in FAS that are not in (1));

(2a) A subset of (2) who did not experience LFT elevations.

The analyses will be performed using the FAS. ARR (descriptive statistics and results of negative binomial model) and EDSS (descriptive statistics and time to 12-week confirmed progression) will be displayed by study intervention group. Further details regarding the modelling will be included in the Statistical Methodology document (Appendix 18.9).

During the DBTP, teriflunomide level determination might be completed for safety reasons only, these teriflunomide levels will be presented in a listing.

15.4 Vital Signs

Vital signs (height (cm), weight (kg), BMI (kg/m²), body temperature (°C), SBP (mmHg), DBP (mmHg), respiratory rate (breaths/min) and pulse rate (beats/min)) will be summarized by intervention group using descriptive statistics (see Section 9).

The descriptive statistics will be presented as follows:

- The baseline will be presented first
- Then each scheduled time point will be presented on participants who have reached this time point: absolute values and CFB will be displayed (except for height and BMI as they are only present at Screening visit).

$$\text{BMI (kg/m}^2) = \frac{weight\ [kg]}{height[cm]^2} \times 10000$$

Body temperature, SBP, DBP, respiratory rate and pulse rate will be analyzed with shift tables of maximum CFB using the categories defined in Table 17:

Table 17 Vital Signs Categories

Parameter	Unit	Shift	Baseline categories	Post baseline categories (absolute change)
Temperature	°C	Increase	<37 / ≥37 - <38 / ≥38 - <39 /≥39 - <40 / ≥ 40	≤0* / >0 - <1 / ≥1 - <2 /≥2 - <3 / ≥3
Pulse rate	bpm	Increase and decrease	<100 / ≥100	≤0* / >0 - ≤20 / >20 - ≤40 / >40
SBP	mmHg	Increase and decrease	<140 / ≥140	≤0* / >0 - ≤20 / >20 - ≤40 / >40
DBP	mmHg	Increase and decrease	<90 / ≥90	≤0* / >0 - ≤ 20 / >20 - ≤40 / >40
Respiratory rate	breaths/min	Increase and decrease	<20 / ≥20	≤0* / >0 - ≤5 / >5 - ≤10 / >10

* This category will include the participants with no changes or decrease/increase in the increase/decrease part of the table respectively.

A listing of maximum CFB and a listing of all vital signs data will be provided.

15.5 12-Lead ECG

Only local ECG reads were collected at the study start, however all local reads will now be sent to be read centrally. Hence for analyses purposes (including change from baseline computation), central read data will be used as source data.

Descriptive statistics (absolute and change from baseline) over time for 12-lead ECG parameters will be provided. Parameters to be included are: QTcF, HR, PR and QRS. Baseline assessments as defined in Section 9.4 will be considered to be pre-dose.

Additionally, a summary table of abnormal values (outliers) for 12-lead ECG parameters will be presented:

- HR:
  - Absolute < 50 bpm, < 40 bpm, < 30 bpm
  - Change from baseline > 20 bpm, > 30 bpm, > 40 bpm
  - Absolute < 50 bpm and decrease from baseline > 25%
  - Absolute > 100 bpm and increase from baseline > 25%
- PR:
  - Absolute > 200 msec and > 220 msec
  - Increase from baseline > 30 msec
  - Absolute > 200 msec and increase from baseline > 25%
- QRS:
  - Absolute > 110 msec
  - Absolute > 120 msec and increase from baseline > 25%
- QTcF:
  - Absolute ≤ 450 msec, > 450 msec and ≤ 480 msec, > 480 msec and ≤ 500 msec, and > 500 msec
  - Increase from baseline ≤ 30 msec, > 30 msec and ≤ 60 msec, and > 60 msec

By-visit frequency tables by diagnosis term will be provided, displaying all terms, only the first term and only the second term separately. The percentage of diagnosis terms will be based on the total number of participants with results at the visit.

A shift table of morphological assessments, from baseline to the worst observation of the on-treatment period, of the number and percentage of participants for each interpretation category (Normal, Abnormal-Insignificant, Abnormal-Significant, Missing, and Total) will also be provided. Results recorded as "incomplete analysis" or "uninterpretable" will be counted under the "Missing" category.

Listings of ECG quantitative values, morphological and rhythm results will be produced including results from both central and local reads.

15.6 Physical Examination

No summary table will be provided since physical examination findings during Screening will be recorded as medical history events and findings during the study as AEs.

15.7 Pregnancy Test

Results of pregnancy test (urine or serum) will be listed.

15.8 Safety/Pharmacodynamic Endpoints: Immunoglobulin levels

Boxplots of Serum IgG, IgA and IgM levels (g/L) by intervention group and time point will be provided, as well as line plots of median value and median percent change from baseline by time point and intervention group.

Descriptive statistics (absolute value, change from baseline and percent change from baseline) by intervention group and time point will be presented as well.

Line plots of median value and median percent change from baseline by time point and intervention group will also be provided for serum IgE and IgG subclass levels (where applicable).

Serum IgG, IgA, IgM, IgE and IgG subclass levels (where applicable) will be listed by intervention group, participant and time point (where applicable). IgG values < 3 g/L (severe hypogammaglobulinemia) and IgG values < 6 g/L (hypogammaglobulinemia) will be flagged in the listing.

15.9 Urinalysis Microscopic Evaluation

Urinalysis Microscopic Evaluation data will be listed by intervention group and time point (where applicable).

15.10 C-SSRS

The C-SSRS is a numerical score derived from 10 categories. The C-SSRS assesses the suicidal behavior and suicidal ideation in participants.

The C-SSRS outcome categories are provided below. Each category has a binary response (yes/no) and are numbered, providing a score. A score of 0 is assigned if there is no suicidal ideation or suicidal behavior present.

Ideation:

1. Wish to be Dead
2. Non-specific Active Suicidal Thoughts
3. Active Suicidal Ideation with Any Methods (Not Plan) without Intent to Act
4. Active Suicidal Ideation with Some Intent to Act, without Specific Plan
5. Active Suicidal Ideation with Specific Plan and Intent

Behavior:

6. Preparatory Acts or Behavior
7. Aborted Attempt
8. Interrupted Attempt
9. Actual Attempt (non-fatal)
10. Completed Suicide

Occurrence of suicidal behavior is defined as having answered "yes" to a least 1 of the 4 suicidal behavior subcategories for Baseline/Screening assessments and to at least 1 of the 5 suicidal behavior subcategories for post baseline "since last visit" assessments. Note that non-suicidal self-injurious behavior is not considered in the derivation of suicidal behavior and occurrence will be reported separately.

Occurrence of suicidal ideation is defined as having answered "yes" to at least one of the suicidal ideation sub-categories.

Occurrence of suicidality is defined as having at least one occurrence of suicidal ideation or at least one occurrence of suicidal behavior.

The number and percentage of participants with occurrence of suicidal behavior at any time during treatment, occurrence of suicidal ideation at any time during treatment, occurrence of suicidality at any time during treatment and occurrence of self-injurious behavior without suicidal intent at any time during treatment will be summarized.

Shifts from baseline (maximum score from "Baseline/Screening" assessment) to the worst on treatment outcome (maximum score from "since last visit" assessments) will be presented. The following categories will be summarized:

- No ideation (score 0)
- Non-serious suicidal ideation (score 1-3)
- Serious suicidal ideation (score 4-5)
- Suicidal behavior (score 6-10).

Participant data listings will be also provided.

15.11 HBV DNA

HBV DNA PCR testing results will be listed by intervention group and time point (where applicable). A shift table from baseline to highest post baseline values including the categories < 20 IU/mL versus ≥ 20 IU/mL will be provided. This summary will be based on participants with negative Hepatitis B surface antigen at Screening but positive Hepatitis B surface antibody or Hepatitis B core antibody.

15.12 Local Laboratory Results

Local laboratory results that are not described in Section 15.3 will be listed by intervention group and time point (where applicable).

15.13 Anti-SARS-CoV-2 Antibodies Levels

All participants in the FAS analysis set with a full COVID-19 vaccination and signed informed consent for Protocol Version 5 will be included in the analysis regardless of the vaccine type. Comirnaty, Spikevax, Valneva COVID-19 Vaccine or VLA2001, Nuvaxovid or Covovax, Vaxzevria or Covishield require 2 doses to be considered fully vaccinated while Jcovden requires only 1.

In addition, participants need to have exposure to evobrutinib or teriflunomide ≥ 2 weeks at the time of first COVID-19 vaccination and to remain on treatment at the time of the second COVID-19 vaccination (if 2 doses are required) and at the time a booster dose is received (as applicable).

Pre-vaccination timepoint corresponds to the last assessment before the date of the first vaccine dose and can occur prior to first administration of active study intervention. Post-vaccination timepoint corresponds to the first assessment after the last dose of full vaccination + at least 28 days for all COVID-19 vaccines. For Jcovden, this is the date of the first vaccine dose + at least 28 days.

Post-booster timepoint corresponds to the first assessment after a booster vaccination + at least 7 days.

SARS-CoV-2 vaccination dates are recorded on the Concomitant Medication eCRF page. In case of multiple dates, the first one will be used to derive the first dose of SARS-CoV-2 vaccination and the second vaccine date will be considered for all analyses post vaccination. Additional vaccine dates after the full vaccination will be considered as booster doses.

Values of IgG anti-S1/S2 levels for SARS CoV-2 above and below the LLR of 3.8 AU/mL will be handled as follows in the analyses:

- Values > LLR will be used as observed
- All non-missing values ≤ LLR will be set as LLR value (i.e. they will be set as 3.8 AU/mL)
- Values of IgG anti-S1/S2 levels reported as ">2,900 AU/mL" (upper limit of reporting) will be set to a numeric value of "2,900+1".

The following subgroups will be used for analysis, where specified:

- Serostatus at pre-vaccination:
    - Seronegative (pre-vaccination anti-IgG antibodies to SARS-CoV-2 < 15.0 AU/mL)
    - Seropositive (pre-vaccination anti-IgG antibodies to SARS-CoV-2 ≥ 15.0 AU/mL)

- Duration of time between the full COVID-19 vaccination cycle and the post-vaccination sample antibody assessment:
    - >4 to ≤ 8 weeks
    - >8 to ≤ 12 weeks
    - >12 weeks

Where duration will be calculated as the difference between the date of the full COVID-19 vaccination and the date of the post-vaccination sample plus 1, divided by 7.

- Duration of time between the booster vaccination and the post-booster sample antibody assessment:
    - ≥ 1 to <4 weeks
    - ≥ 4 to ≤ 8 weeks
    - >8 to ≤ 12 weeks
    - >12 weeks

Where duration will be calculated as the difference between the date of the booster COVID-19 vaccination and the date of the post-vaccination sample plus 1, divided by 7.

- Type of vaccine:
    - mRNA: including Comirnaty or Spikevax vaccines
    - Non-mRNA: including Valneva COVID-19 Vaccine or VLA2001, Nuvaxovid or Covovax, Vaxzevria or Covishield and Jcovden vaccines
    - If any participants receive a mix of mRNA and non-mRNA vaccines for the full vaccination cycle, a third group will be created: Both mRNA and non-mRNA.

The disposition of participants will be summarized by descriptive statistics: Number of participants who received the complete vaccination cycle, overall, by country and by type of vaccine, number of participants who received the booster dose, overall and by type of vaccine.

To assess the demographics and baseline characteristics of participants included in this analysis, the following tables will be repeated on all participants in the FAS analysis set with a full COVID-19 vaccination and signed informed consent for Protocol Version 5:

- Demographic characteristics
- RMS baseline disease characteristics
- History of DMT

In order to assess the duration of treatment exposure at the time of COVID vaccination, the duration of treatment at the time of COVID-19 vaccination (weeks) is derived as follows:

$$\textit{Duration of Evobrutinib or Teriflunomide (weeks)} = \frac{(\textit{date of first COVID-19 vaccination} - \textit{date of first dose} + 1)}{7}$$

The duration of exposure to evobrutinib or teriflunomide at the time of first COVID-19 vaccination will be analyzed by summary statistics and by frequencies according to the following categories:

- ≥ 2 to 12 weeks
- > 12 to 24 weeks
- > 24 to 36 weeks
- > 36 to 48 weeks
- > 48 to 60 weeks
- > 60 to 72 weeks
- > 72 to 84 weeks
- > 84 to 96 weeks
- > 96 to 108 weeks
- > 108 to 120 weeks
- > 120 to 132 weeks
- > 132 to 144 weeks
- > 144 to 156 weeks
- > 156 weeks

With respect to the time between the full COVID-19 vaccination cycle and the post-vaccination sample antibody assessment (in weeks) the following derivation rule applies:

$$\textit{Time between full vaccination cycle and antibody assessment (weeks)} = \frac{(\textit{date of post-vaccination sample} - \textit{date of full COVID-19 vaccination cycle} + 1)}{7}$$

The duration of the time between the full COVID-19 vaccination cycle and the post-vaccination sample antibody assessment will be analyzed by summary statistics and by frequencies according to the following categories:

- > 4 to ≤ 8 weeks
- > 8 to ≤ 12 weeks
- > 12 weeks

With respect to the time between the booster COVID-19 vaccination and post-booster vaccination sample antibody assessment (in weeks) the following derivation rule applies:

$$\textit{Time between booster vaccination and antibody assessment (weeks)} = \frac{(\textit{date of post-booster vaccination sample} - \textit{date of booster vaccination} + 1)}{7}$$

The duration of the time between the booster vaccination and the post-booster sample antibody assessment will be analyzed by summary statistics and by frequencies according to the following categories:

- $\geq$ 1 to < 4 weeks
- $\geq$ 4 weeks to $\leq$ 8 weeks
- > 8 to $\leq$ 12 weeks
- > 12 weeks

A listing will include participant identifier, age, sex, race, study intervention start date, pre-vaccination antibody assessment date, COVID-19 vaccination dates, post-vaccination antibody assessment date, post-booster sample collection date, duration of study intervention at the time of first COVID-19 vaccination, time between the full COVID-19 vaccination cycle and the post-vaccination sample antibody assessment, as well as time between the booster vaccination and the post-booster sample antibody assessment.

To characterize SARS-CoV-2 vaccine response, i.e. level of anti-S1/S2 IgG antibodies, in participants with RMS, summary statistics (including geometric mean and corresponding 95% CIs) of anti-S1/S2 IgG antibody levels at pre-vaccination, post-vaccination and post-booster timepoints will be provided by study intervention arm. The pre-vaccination and post-vaccination timepoints will be summarized separately for all participants who received a booster vaccination. 2-sided 95% CIs for geometric mean will be calculated as follows: the log-transformed levels of anti-S1/S2 IgG antibodies will be used and the arithmetic mean and upper and lower limits of the 2-sided 95% CIs for the arithmetic mean will be computed by use of percentiles of t-distribution; the exponentiate of these limits and the arithmetic mean itself are the geometric mean and limits of the CI for the geometric mean, respectively.

For a visual representation of anti-S1/S2 IgG antibodies pre-vaccination, post-vaccination and post-booster boxplots will be provided in addition. The figure will be repeated considering only participants with COVID-19 vaccination booster.

An exploratory analysis will be conducted to investigate whether anti-S1/S2 IgG antibody levels after a full vaccination cycle are impacted by covariates. A linear regression model with the following independent variables will be estimated: intervention group, age, gender, anti-S1/S2 IgG antibody levels before vaccination (i.e. sero-positive or sero-negative), duration of exposure to teriflunomide / evobrutinib at the time of the first COVID-19 vaccination, time between the full vaccination cycle and antibody assessment. In case diagnostic methods (e.g. QQ-plots) show that antibody levels are not normally distributed, log-transformed values of antibody levels will be used for the analysis.

The fold change is the ratio between the post-vaccination IgG anti-S1/S2 antibody levels and the pre-vaccination anti-S1/S2 IgG antibody levels.

Summary statistics of the fold changes will be provided.

Additionally, the number and percentage of participants in each of the following fold change categories will be presented:

- ≤ 1x - participants having a fold increase ≤ 1
- > 1x - ≤ 10x; participants having a fold increase greater than 1 and less or equal than 10
- > 10x - ≤ 30x; participants having a fold increase greater than 10 and less or equal than 30
- > 30x - ≤ 60x; participants having a fold increase greater than 30 and less or equal than 60
- > 60x - ≤ 100x; participants having a fold increase greater than 60 and less or equal than 100
- > 100x; participants having a fold increase > 100

A listing of anti-S1/S2 IgG antibody levels with pre- and post-vaccination, and if available post-booster samples will be presented per participant. The listing will contain participant number, parameter name, date of sample taken, visit, serostatus at pre-vaccination, result, fold change to post-vaccination, fold change to post-booster vaccination, fold change from post-vaccination to post-booster vaccination, vaccination dates of the participant and whether the sample is a pre- or post-vaccination sample, type and name of vaccine.

All described analyses will be repeated on the subgroups "Serostatus at pre-vaccination", "time between the full COVID-19 vaccination cycle and the post-vaccination sample antibody assessment", "time between the booster vaccination and the post-booster vaccination sample antibody assessment" as well as for "type of vaccine" as defined above.

Percentages for the subgroup analyses will be based on the number of participants within the subgroup of interest.

For subgroup analyses by "time between the full COVID-19 vaccination and the post-vaccination sample", only pre- and post-vaccination timepoints will be considered. For subgroup analyses by "time between booster COVID-19 vaccine and post-booster sample", only pre-booster and post-booster timepoints will be considered. For the subgroup "Serostatus at pre-vaccination" a scatterplot of anti-S1/S2 IgG antibody levels at pre-vaccination, post-vaccination and post-booster timepoints will be also provided.

Immunoglobulin levels (serum IgG, IgA and IgM levels [g/L]) will be summarized at the time of the first COVID-19 vaccination dose, by study intervention group. The summaries will be split by the timing of the assessment before the first COVID-19 vaccination: up to and including 3 months before and between 3 and 6 months before. The timing will be derived as described in Section 9.6. If the assessment occurs on the same date as the first COVID-19 vaccination, the assessment will be included in the "up to and including 3 months before" group.

15.14 Focused Genetic Testing

For participants with results of focused genetic testing performed as per the Protocol Section 7.1 for variants in the High Iron Fe (human hemochromatosis protein; HFE) gene (C282Y, H63D), the summary of number of participants with detected variants will be produced, based on the SAF analysis set. The results will be also listed.

16 Analyses of Other Endpoints

16.1 PK

16.1.1 General Specifications for PK Concentration and PK Parameter Data

PK samples will be collected according to the Schedule of Activities (Section 1.3.1 of the protocol) from all participants in the main studies and in a PK substudy (MS200527_0080 only).

The following checks for evobrutinib and metabolite MSC2729909A concentrations will be performed. This is not a comprehensive list and additional checks may be required.

- If an episode of vomiting occurs on the day of PK collection, the concentration results for that day will be flagged and identified in a listing. If it is known that the time of vomiting is after the time of evobrutinib dosing, and prior to collection of a PK sample on the day of the PK collection, the concentration results following the time of vomiting will be excluded from the calculation of summary statistics and estimation of PK parameters.
- Meal information in relation to evobrutinib dosing on the day of PK collection: If evobrutinib is not taken with a meal on the day of the PK collection, concentration results will be flagged and identified in a listing, but no data will be excluded from the calculation of summary statistics or the estimation of PK parameters. Concentration results will also be flagged when the dose is taken >1 hour prior to a meal or the dose is taken >2 hours after a meal.
- Meal information in relation to evobrutinib dosing on the previous day of PK collection: If evobrutinib is not taken with a meal on the previous day of the PK collection, pre-dose concentration results will be flagged and identified in a listing, but no data will be excluded from the calculation of summary statistics. Pre-dose concentration results will also be flagged when the previous dose is taken >1 hour prior to a meal or the previous dose is taken >2 hours after a meal.
- Dosing: If the most recent evobrutinib dose was taken >18 hours or <6 hours prior predose sampling time, the predose PK sample will be flagged and identified in a listing, but no data will be excluded from the calculation of summary statistics. There will be no flag for postdose concentrations if the dose on the day of PK sample is taken according to protocol.
- Concomitant medication: If PK concentration data is collected while a moderate or strong CYP3A inducer or inhibitor is taken, or shortly after discontinuation of the moderate or strong

CYP3A inducer or inhibitor, affected concentration results and any estimated parameters will be flagged and excluded from the calculation of summary statistics. If PK concentration data is collected when a proton pump inhibitor, H2 blocker, or antacid is taken, affected concentration results and any estimated parameters will be flagged and identified in a listing but not excluded from summary statistics.

- PK blood collection: Designated predose PK samples collected just after dosing of evobrutinib and samples collected out of order (as specified in the protocol) will be excluded from the calculation of summary statistics. The 2nd duplicate PK sample in a sequence will be excluded from the calculation of descriptive statistics and estimation of PK parameters. If a PK sample is collected outside of the collection windows defined below, see Table 18 and Table 19, the affected concentration results and any estimated parameters will be flagged and identified in a listing but will not be excluded from calculation of summary statistics.

Table 18 PK Sampling Time Windows – Main Study

	Day 1	Week 4, 24, 96[a]	Week 12, 48, 72[a]	Week 156[a]
Scheduled Time	**Window**			
Predose	Must be collected prior to dosing	Must be collected prior to dosing	Must be collected prior to dosing	Must be collected prior to dosing
0.5 h	0 – 1 h	N/A	N/A	N/A
0.75 – 1.5 h[b]	No window	No window	N/A	N/A
1 h	0.5 h – 2 h	N/A	N/A	N/A
1 h – 6 h	N/A	N/A	No window	N/A
2 h	1 h – 4 h	N/A	N/A	N/A
4 h	2 h – 12 h	N/A	N/A	N/A

a As specified in the protocol, visit window is ± 3 days for all visits from Week 2 to Week 44, inclusive. Visit window is ± 7 days for all visits from Week 48 to Week 156 inclusive. Exception: Week 4 has a visit window of ± 2 days.
b To be collected on Day 1, Week 4, Week 24, and Week 96. Sample to be collected within 30 minutes after ECG assessment which should be recorded approximately 45 minutes to 60 minutes after dosing.
N/A = not applicable

Table 19 PK Sampling Time Windows – Substudy

	Day 1, Week 2, 4, 24[a, b]	Week 8, 12, 48, 72[b]	Week 96[b]
Scheduled Time	**Window**		
Predose	Must be collected prior to dosing	Must be collected prior to dosing	Must be collected prior to dosing
0.25 h	0 – 0.5 h	N/A	N/A
0.5 h	0.25 h – 1 h	N/A	N/A
0.75 – 1.5 h[c]	No window	N/A	No window
1 h	0.5 h – 2 h	N/A	N/A
1 h – 6 h	N/A	No window	N/A
2 h	1 h – 4 h	N/A	N/A
4 h	2 h – 6 h	N/A	N/A
6 h	4 h – 12 h	N/A	N/A

a PK samples will only be collected up to 4 h postdose for Week 24.
b As specified in the protocol, visit window is ± 3 days for all visits from Week 2 to Week 44, inclusive. Visit window is ± 7 days for all visits from Week 48 to Week 156 inclusive. Exception: Week 4 has a visit window of ± 2 days.
c To be collected after ECG assessment on Day 1, Week 4, Week 24, and Week 96. Sample to be collected within 30 minutes after ECG assessment which should be recorded approximately 45 minutes to 60 minutes after dosing.
N/A = not applicable

Data listings, summaries of evobrutinib and MSC2729909A concentrations, and associated figures will be the responsibility of IQVIA, Overland Park, Kansas. For the descriptive statistical presentation of evobrutinib and MSC2729909A concentrations, all concentrations that are BLQ will be assigned a numerical value of zero prior to calculation of statistics. For days when a PK sample is scheduled to be collected between 1 and 6 hours postdose, the statistical presentation of concentrations will use the following time intervals for summarization: predose, ≥1 to 2 hours postdose, >2 to 4 hours postdose, and >4 to 6 hours postdose. Graphical displays of individual and mean (± SD for linear profiles only) plasma concentration-time profiles for Day 1 (main studies) and Day 1, Week 2, Week 4, and Week 24 (PK substudy) will be presented on linear and semi-logarithmic scales. Individual and mean (linear only) trough (predose) concentration-day profiles will be presented.

16.1.2 Estimation of PK Parameters

Evobrutinib and MSC2729909A PK parameters will be calculated for Day 1 and Week 4 of participants in the PK substudy using the actual elapsed time since dosing. In cases where the actual sampling time is missing, calculations may be performed using the nominal time. Details (e.g. number of samples, participants affected) will be described in the CSR. In case actual dosing time is missing, scheduled time might be used for NCA after performance of adequate plausibility checks and agreement with the sponsor. Decision and rationale should be included in the CSR. Otherwise, there will be no further imputation of missing data.

The following plasma PK parameters will be calculated for evobrutinib and MSC2729909A for Day 1 and Week 4 where appropriate:

Symbol	Definition	Calculation
C_{max}	Maximum observed concentration	
t_{max}	The time to reach the Cmax in a dosing interval	In case of multiple/identical Cmax values, the first occurrence will be used.
AUC_{0-6}	The partial AUC from time zero (= dosing time) to 6 hours after dosing	Calculated using the mixed log-linear trapezoidal rule (linear up, log down). In cases where the actual observation time is not equal to the scheduled observation time AUC_{0-6} will be calculated based on the estimated concentration at 6 hours and not the concentration at the actual observation time.
AUC_{0-12}	The partial AUC from time zero (= dosing time) to 12 hours after dosing	Calculated using the mixed log-linear trapezoidal rule (linear up, log down). AUC_{0-12} will be calculated based on the estimated concentration at 12 hours.
$AUC_{0-tlast}$	The AUC from time zero (= dosing time) to the time of the last quantifiable concentration (t_{last}).	Calculated using the mixed log-linear trapezoidal rule (linear up, log down).
$AUC_{0-\infty}$	The AUC from time zero (= dosing time) up to infinity with extrapolation of the terminal phase.	$AUC_{0-\infty} = AUC_{0-tlast} + C_{last\ pred} / \lambda_z$, where $C_{last\ pred}$ is the predicted concentration at the t_{last}, calculated from the log-linear regression line for λ_z determination.
$t_{½}$	The terminal half-life.	$t_{½} = \ln(2) / \lambda_z$, where λ_z is the terminal first order (elimination) rate constant determined from the terminal slope of the log-transformed concentration curve using linear regression on terminal data points of the curve.
CL/F	The apparent total body clearance following extravascular administration.	CL/F = Dose / $AUC_{0-\infty}$ For Day 1
CL_{ss}/F	The apparent body clearance at steady state following extravascular administration.	CL_{ss}/F = Dose / AUC_{0-12} For Week 4.

Symbol	Definition	Calculation
V_z/F	The apparent volume of distribution during the terminal phase following extravascular administration.	V_z/F = Dose / ($AUC_{0-\infty} \times \lambda_z$) For Day 1 V_z/F = Dose / ($AUC_{0-12} \times \lambda_z$) For Week 4
M/P(C_{max})	Molecular weight-corrected ratio of metabolite (MSC2729909A) C_{max} to parent (evobrutinib) C_{max}.	M/P(C_{max}) = (C_{max}metabolite × MWparent) / (C_{max}parent × MWmetabolite)
M/P(AUC_{0-6})	Molecular weight-corrected ratio of metabolite (MSC2729909A) AUC_{0-6} to parent (evobrutinib) AUC_{0-6}	M/P(AUC_{0-6}) = (AUC_{0-6}metabolite × MWparent) / (AUC_{0-6}parent × MWmetabolite)
M/P(AUC_{0-12})	Molecular weight-corrected ratio of metabolite (MSC2729909A) AUC_{0-12} to parent (evobrutinib) AUC_{0-12}	M/P(AUC_{0-12}) = (AUC_{0-6}metabolite × MWparent) / (AUC_{0-12}parent × MWmetabolite)
M/P($AUC_{0-\infty}$)	Molecular weight-corrected ratio of metabolite (MSC2729909A) $AUC_{0-\infty}$ to parent (evobrutinib) $AUC_{0-\infty}$	M/P($AUC_{0-\infty}$) = ($AUC_{0-\infty}$metabolite × MWparent) / ($AUC_{0-\infty}$parent × MWmetabolite)

Additional PK parameters may be calculated where appropriate.

Units for PK parameter output will be based on concentration and dose units used in the study, unless otherwise specified. In case concentration data units change within the study, PK parameters will be reported using consistent units throughout study outputs.

The parameters C_{max} and t_{max} will be obtained directly from the concentration-time profiles.

In cases where the actual observation time is not equal to the scheduled observation time, AUC_{0-6} will be calculated by estimated concentration at 6 h, provided λ_z is estimable. In case suitable regression cannot be performed, partial areas may be calculated using the concentration at actual sampling time if actual sampling time is within 10% of the nominal sampling time.

At steady state, if AUC_{0-12} is not estimable via extrapolation, the pre-dose concentration may be duplicated and used as the trough concentration for calculation of AUC_{0-12}. In such cases, $AUC_{0-tlast}$ shall be estimated based on observed C_{last}.

- In cases BLQ concentrations occur at the end of the collection interval, these concentrations might be set to missing for calculations of partial AUCs after agreement with the Sponsor. This is an exemption to the general BLQ handling rule and should be applied in case its application would result in estimation of implausible partial AUC values. Implausibility is considered in cases partial AUCs are greater than $AUC_{0-\infty}$.

The following PK parameters will be calculated for diagnostic purposes and listed, but will not be summarized:

- First (λ_z low) and last (λ_z up) time point of the time interval of the log-linear regression to determine λ_z.
- Number of data points ($N\lambda$) included in the log-linear regression analysis to determine λ_z.
- Goodness of fit statistic (adjusted Rsq) for calculation of λ_z.
- AUC from time t_{last} extrapolated to infinity given as percentage of $AUC_{0-\infty}$. ($AUC_{extra\%}$)
- Span ratio of interval over which $t_{½}$ was estimated / $t_{½}$

The regression analysis should contain data from at least 3 different time points in the terminal phase consistent with the assessment of a straight line on the log-transformed scale. Phoenix WinNonlin "best fit" methodology will be used as standard. If warranted, further adjustment may be made by the pharmacokineticist, after agreement with the Sponsor. The last quantifiable concentration >LLOQ should always be included in the regression analysis, while the concentration at t_{max} and any concentrations BLQ which occur after the last quantifiable data point >LLOQ should not be used.

If $AUC_{extra\%}$>20% and/or the coefficient of correlation (Rsq_{adj}) of λ_z is <0.8 and/or the observation period over which the regression line is estimated (span ratio) is less than 2-fold the resulting $t_{½}$, the rate constants and all derived parameters (e.g. $t_{½}$, $AUC_{0-\infty}$, CL etc.) will be listed, flagged and included in the parameter outputs. Should more than 10% of participants be flagged for $AUC_{extra\%}$ and/or Rsq_{adj} (for a particular analyte), a sensitivity analysis excluding flagged parameters may be performed after discussion with the Sponsor.

The molecular weights of evobrutinib and MSC2729909A to be used in the calculation of M/P ratios are as follows:

- evobrutinib = 429.5213
- MSC2729909A = 463.5289

The PK of evobrutinib will also be characterized using population modeling. Please see Section 16.3 for details.

16.1.3 Presentation of PK Parameter Data

Data listings and summaries of evobrutinib and MSC2729909A PK parameters will be the responsibility of IQVIA, Overland Park, Kansas. A table of descriptive statistics of PK parameters will be produced and displayed by analyte and day using PKSUB analysis set. In addition, listings of individual PK parameters and diagnostics will be produced and displayed by analyte and day/week using SAF analysis set. The Phoenix WinNonlin NCA Core Output will be included. No figures will be produced for PK parameters.

16.2 PD

16.2.1 BTK Occupancy at Baseline and Weeks 2, 4, 8, 12 and 24

Measurement of BTK (free and total) will be assessed as part of the PD substudy as indicated in the protocol for MS200527_0080, Schedule of Assessments. Data on these variables will be summarized descriptively in tabular and/or graphic format and listed as appropriate for the data based on the PDBTKO analysis set.

BTK endpoints (free BTK, total BTK, and % BTK occupancy as calculated with the formula below) will be summarized using the PDBTKO analysis set. Descriptive statistics for BTK endpoints will be presented as described in Section 9.2.

Percent BTKoccupancy is calculated as follows and will be retrieved from external data:

$$\%\ Occupancy_{BTK} = \left(1 - \frac{\left(\frac{BTKF}{BTKT}\right)}{\frac{BTKF_{baseline}}{BTKT_{baseline}}}\right) * 100$$

with BTKF denoting free BTK, BTKT denoting total BTK, and baseline defined as last non-missing measurement prior to the first dose of study intervention.

Given that BTK occupancy is a value lying in the unit interval (0, 1), it is preferable to model the data as beta-distributed, rather than normal-distributed. The mean and 95%, 90%, and 80% CIs for the mean will be estimated using a beta model for BTK occupancy at each time point. As a supportive analysis, the conventionally estimated mean and CIs, in which BTK occupancy values are treated as normally distributed, will also be provided.

The BTK occupancy median (5th and 95th percentiles) values will be plotted versus time point.

The BTK occupancy data may also be included in exposure-response analyses (refer to Section 18.3).

16.2.2 NfL Concentration at Baseline and Weeks 24, 48, 72, 96, and 120

The analysis of NfL concentrations at Weeks 24, 48, 72, 96 and 120 will follow that of NfL at Week 12 described in Section 14.2.5 and will be presented in the same output. For participants discontinuing treatment, all available NfL concentration data will be included in the analysis up to 156 weeks post randomization.

16.2.3 Other PD Endpoint Analyses at Baseline and Weeks 2, 4, 8, 12 and 24

PD endpoints including cellular function (i.e. B-cell activation and FcγRIII stimulation), immunophenotyping (i.e., TBNK, and B-cell & plasma cell subset panels) and serum cytokines (N=16)) will be assessed in a subset of participants from selected sites in the PD substudy as indicated in the protocol Schedule of Assessments.

The Absolute values, and Absolute and Percentage Change from baseline of each biomarker parameter will be summarized descriptively for PD analysis set by visit. In addition, the number of participants with a value below LLOQ will be presented by visit together with the LLOQ limit, if available. No imputation will be performed for values below LLOQ. Descriptive statistics for these variables will be presented as described in Section 9.2.

The median (5^{th} and 95^{th} percentiles) observed values for each intervention group will be plotted versus time point in a line plot format for the PD endpoints. Median percent change from baseline line plots will also be provided.

Gene expression will not be analyzed as part of the PA but may be analyzed later. The corresponding analysis will therefore be described in a separate IAP.

16.3 Population PK Modeling and Exposure-Response Analyses

The population PK modeling and exposure-response analyses will be the responsibility of the Sponsor (may be conducted by a different external vendor). The primary objectives of the population PK modeling and exposure-response analyses are: 1) to describe the PK concentration time profile following the administration of evobrutinib in patients with relapsing MS, and 2) to explore the relationship between evobrutinib exposure and selected efficacy/safety endpoints. Secondary objectives include:

- To estimate typical population parameters and magnitude of the inter- and intraindividual variability for evobrutinib PK parameters in patients with relapsing multiple sclerosis
- To identify intrinsic and extrinsic covariates (e.g. demographic factors, meal etc.) that are significant predictors of variability in PK.

The analysis will include: a) population PK modeling and b) exposure-response analyses conducted on the primary efficacy endpoint ARR and other select clinically relevant efficacy and safety endpoints.

The PK data of the current study will be combined with the corresponding data of other studies (i.e. the already completed studies MS200527_0019, MS200527_0017, MS200527_0086, and others), during the model development process (e.g. to help estimate the parameters of an adequate structural model).

Relevant evobrutinib PK parameters and exposures (e.g. steady state AUC) will be predicted for all patients with evaluable PK using the population PK modeling approach. Not all PK parameters (e.g. V_z/F) may be computed for every patient.

There will be an integrated exposure-response analysis of the combined data of the 2 Phase 3 studies for all endpoints analyzed. The results will be summarized separately in dedicated pharmacometric report/s. Details of the planned M&S work for studies MS200527_0080 and PPD will be developed in Pharmacometric Modeling Analysis Plans by the external vendor that will perform these M&S tasks under Merck Clinical Pharmacometry guidance and supervision. The Pharmacometric Modeling Analysis Plans will be finalized and signed-off before the PA DB lock of the 2 Phase 3 studies.

Early access of the PK data using masked patient IDs from the Study MS200527_0080 and PPD by an external vendor is planned prior to the database lock to facilitate population PK model development. The Sponsor will remain blinded to the data during this period. Prior to enabling early data access, this process will be documented in an unblinding memo.

16.4 PRO – Exploratory Endpoints

This section discusses only those PRO endpoints that are exploratory. PRO endpoints that are secondary are discussed in Section 14.2. For the analysis in the following subsections, data selection for PA will be handled as indicated in Table 8. Analyses presented at pooled level will also be produced for each study.

16.4.1 Change from Baseline in PROMIS Score at Week 48, Week 96 and Week 144

The analysis of change from baseline in PROMIS score (PF or fatigue) at Week 48 (Week 96/Week 144) will be implemented via the same model (based on pooled data) used for the analysis of change from baseline in PROMIS score (PF or fatigue) over 96 weeks (Section 14.2.3), by selecting an appropriate contrast from the model specific to Week 48 (Week 96/Week 144), and reporting the treatment effect measure, 95% 2-sided CI, and p-value. The handling of ICEs will be analogous to the secondary endpoint, described in Section 14.2.3.1.

For each PROMIS score CFB at Week 48 (Week 96/Week 144), a figure will be provided describing the distribution of CFB at Week 48 (Week 96/Week 144), one curve per intervention group, as described in Section 14.2.3.

16.4.2 Change from Baseline in SF-36v2 score at Week 48, Week 96 and Week 144

The analysis of change from baseline in SF-36v2 score at Weeks 48, 96 and 144 will be based on pooled data, and follow the same approach used for the analysis of change from baseline in PROMIS score at Week 96 (Section 14.2.3). This applies to the individual domain scores, as well as the MCS and PCS. The handling of ICEs will also be analogous to the PROMIS secondary endpoint, described in Section 14.2.3.1.

The analysis of change from baseline in SF-36v2 score at Weeks 48, 96 and 144 will be implemented via the same model described above, by selecting an appropriate contrast from the model specific to Weeks 48, 96 and 144, and reporting the treatment effect measure, 95% 2-sided CI, and p-value.

For each intervention group, descriptive statistics will be presented for absolute (raw) value, CFB, and percent CFB value for each visit, from baseline to Week 156 (Safety Follow-up Visits included).

For the SF-36 MCS and PCS endpoints, a figure will be provided describing the distribution of CFB at Week 48, one curve per intervention group. This will be repeated for Week 96 and Week 144. This will follow the same approach described in Section 14.2.3.

For each SF-36 endpoint (domain scores, MCS, PCS), the number (proportion) of participants not worsening, and number (proportion) of participants worsening (decrease $\geq$ 5), between baseline and Week 48, baseline and Week 96 and baseline and Week 144 will be summarized by intervention group.

Mean score CFB will be presented as a by-visit line plot for each intervention group, with both intervention groups included in a single figure, and horizontal axis extending to Week 144.

A bar chart figure of Mean score CFB at Week 48 presenting each SF-36 sub-domain score (PF, RP, BP, GH, VT, SF, RE, MH) for each intervention will be provided. This will be repeated for Week 96 and Week 144.

16.4.3 Change from Baseline in EQ 5D-5L VAS and Index at Week 48, Week 96 and Week 144

The analysis of change from baseline in EQ 5D-5L VAS at Week 48, Week 96 and Week 144 will be based on pooled data and follow the same approach used for analysis of CFB in SF-36v2 at Weeks 48, 96 and 144 (Section 16.4.2), including the handling of ICEs. The analysis of EQ 5D-5L Index at Week 48, Week 96 and Week 144 will also follow this approach.

For each intervention group, descriptive statistics will be presented for absolute (raw) value, CFB, and percent CFB value for each visit, from baseline to Week 156 (Safety Follow-up Visits included).

16.4.4 Time to First Occurrence of 12-week Confirmed PF Deterioration Compared to Baseline up to 156 Weeks

The MCID for the PROMISnq PF(MS) 15a is reported to be 2.3 – 2.7 (Kamudoni 2022). A MCID cutoff of 2.7 will be used. PF deterioration is defined as a reduction in PROMIS PF T-score $\geq$ 2.7 compared to baseline PROMIS PF T-score. Confirmed PF deterioration is a deterioration sustained for at least 12 weeks.

Confirmation of deterioration must occur at the regularly scheduled visit that is at least 12 weeks (84 days) after initial deterioration. If a participant has a missing PF score at the scheduled visit occurring at least 84 days after an initial deterioration or the scheduled visit occurs several days before the 84-day window after an initial deterioration, confirmation of the deterioration must be on the basis of the assessment at the next scheduled visit.

PF score assessments at unscheduled or scheduled visits that are less than 84 days after the initial deterioration are considered non-confirmatory PF score assessments. For deterioration to be confirmed at a regularly scheduled visit at least 12 weeks after initial deterioration, any non-confirmatory PF score assessments should be at least as much worsening as that required for deterioration.

The analysis of time to first occurrence of 12-week confirmed PF deterioration compared to baseline will be based on pooled data and follow the same approach used for the analysis of time to 12-week CDP (Section 14.2.1.1).

The handling of ICEs will also be analogous to the secondary endpoint, time to 12-week CDP, described in Section 14.2.1.1. In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as confirmation of deterioration. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, participants will be censored at time of death. (i.e. death will not be counted as confirmation of deterioration).

Descriptive statistics will follow the same approach used for the 12-week CDP.

16.4.5 Time to First Occurrence of 24-week Confirmed PF Improvement Compared to Baseline up to 156 Weeks

PROMIS PF improvement is defined as an increase of at least 2.7 points on PROMIS PF score relative to baseline. Confirmed PF score improvement events are defined as PROMIS PF improvement sustained for at least 24 weeks.

Confirmation of improvement must occur at the regularly scheduled visit that is at least 24 weeks (168 days) after initial improvement. If a participant has a missing PF score at the scheduled visit

occurring at least 168 days after an initial improvement or the scheduled visit occurs several days before the 168-day window after an initial improvement, confirmation of the improvement must be on the basis of the assessment at the next scheduled visit.

PF score assessments at unscheduled or scheduled visits that are less than 168 days after the initial improvement are considered non-confirmatory PF score assessments. For improvement to be confirmed at a regularly scheduled visit at least 24 weeks after initial improvement, any non-confirmatory PF score assessments should be at least as low as the minimal change required for improvement.

The analysis of time to first occurrence of 24-week confirmed PF improvement compared to baseline will be based on pooled data and follow the same approach used for the analysis of time to 24-week CDI (Section 14.2.2). The handling of ICEs will also be analogous to the secondary endpoint.

Descriptive statistics will follow the same approach used for the 24-week CDI.

16.4.6 Time to First Occurrence of 12-week Confirmed Fatigue Deterioration Compared to Baseline up to 156 Weeks

The MCID for the PROMIS Fatigue (MS) 8a is reported to be 3.4 - 4 (Kamudoni 2021). A MCID cutoff of 4 points will be used. PROMIS fatigue deterioration is defined as an increase of at least 4 points on PROMIS fatigue score relative to baseline. Confirmed fatigue score deterioration events are defined as deteriorations in PROMIS fatigue sustained for at least 12 weeks.

The analysis of time to first occurrence of 12-week confirmed fatigue deterioration compared to baseline will be based on pooled data and follow the same approach used for the analysis of time to 12-week confirmed PF deterioration (Section 16.4.4).

The handling of ICEs will also be analogous to the time to 12-week confirmed PF deterioration (Section 16.4.4). In accordance with the Composite Variable strategy, for participants experiencing the ICE death attributable to MS or treatment, the death will be counted as confirmation of deterioration. In accordance with the While Alive strategy, for participants experiencing the ICE death unattributable to MS or treatment, participants will be censored at time of death. (i.e. death will not be counted as confirmation of deterioration).

Descriptive statistics will follow the same approach used for the 12-week confirmed PF deterioration.

16.4.7 Time to First Occurrence of 24-week Confirmed Fatigue Improvement Compared to Baseline up to 156 Weeks

PROMIS fatigue improvement is defined as a decrease of at least 4 points on PROMIS fatigue score relative to baseline. Confirmed fatigue score improvement events are defined as improvements in PROMIS fatigue sustained for at least 24 weeks.

The analysis of time to first occurrence of 24-week confirmed fatigue improvement compared to baseline will be based on pooled data and follow the same approach used for the analysis of time to 24-week confirmed PF improvement (Section 16.4.5). The handling of ICEs will also be analogous.

Descriptive statistics will follow the same approach used for the 24-week confirmed PF improvement.

16.4.8 PROMIS Fatigue Score CFB over 96 Weeks Independent of Relapse Activity

The primary analysis of CFB in PROMIS Fatigue score over 96 weeks defined in Section 14.2.3.2 will be repeated considering the hypothetical scenario for the ICE 'relapse'.

The handling of ICEs will be analogous to the primary (for relapses) and secondary (for PROMIS fatigue) endpoints, described in Sections 14.1.1 and 14.2.3.1, respectively.

Participants with qualified relapse will be censored at the time of the onset of the first qualified relapse. Descriptive statistics and inferential analysis (MMRM) will be presented.

The analysis will be repeated considering all relapses, not only qualified.

16.4.9 PROMIS Response up to Week 156

Response in PROMIS (PF or fatigue) will be analyzed based on differences in proportions of responders between intervention groups.

Responder definitions will be based on MCID criteria for improvement or deterioration from baseline as described in Sections 16.4.4, 16.4.5, 16.4.6 and 16.4.7. The handling of ICEs will also be analogous.

The following analysis will be repeated for the 4 definitions of response (PF improvement, PF deterioration, fatigue improvement and fatigue deterioration):

A logistic regression model with repeated measures will be applied, with terms for intervention group, visit, intervention group by visit interaction, baseline score, baseline score by visit interaction, randomization strata and study ID, and where the data are pooled from the FAS of studies 0080 and PPD . The unstructured covariance matrix will be considered. The estimated odds ratio at Week 48, Week 96, Week 144 and Week 156, together with an associated 2-sided 95% CI of the estimated odds ratio will be presented.

16.5 HRU

HRU parameter data will be pooled from both Phase 3 studies and summarized by visit via descriptive statistics (mean, SD, median, minimum/maximum, 25th and 75th percentile; 95% CI) for the on-treatment period. The following parameters will be summarized:

- Number of doctor/clinic visits
- Number of home visits
- Number of emergency room visits
- Number of days in hospital
- Number of participants who paid someone to assist them, n (%):
    - Average number of days per week of paid assistance (restricted on participants who paid someone to assist them)
    - Average number of hours per day of paid assistance (restricted on participants who paid someone to assist them)
- Number of participants with a relative or friend who missed work because of MS, n (%):
    - Number of days of missed work (restricted on participants with a relative or friend who missed work because of MS)
- Number of participants who missed at least one full day of work because of MS, n (%):
    - Number of days of missed work (restricted on participants who missed at least one full day of work because of MS)
- Number of participants who missed at least one partial day of work because of MS, n (%):
    - Average number of hours per day missed from work (restricted on participants who missed at least one partial day of work because of MS)
- Number of participants who accomplished less at work due to MS, n (%)
- Percent that best indicates the amount of work the participant was able to accomplish despite MS

HRU data will also be listed.

17 References

Benedict RHB, Tomic D, Cree BA, et al. Siponimod and Cognition in Secondary Progressive Multiple Sclerosis: EXPAND Secondary Analyses. Neurology. 2021;96(3):e376-e386. doi:10.1212/WNL.0000000000011275

Bretz F, Maurer W, Brannath W, et al. A graphical approach to sequentially rejective multiple test procedures. Stat Med. 2009;28:586-604.

Bretz F, Maurer W, Xi D. Replicability, reproducibility, and multiplicity in drug development. CHANCE. 2019;32:4,4-11. Clinical Study Measures: https://www.nationalmssociety.org/For-Professionals/Researchers/Resources-for-Researchers/Clinical-Study-Measures

Crowe BJ, Xia HA, Berlin JA, et al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. Clin Trials. 2009;6(5):430-40.

Elliott C, Belachew S, Wolinsky JS, Hauser SL, Kappos L, Barkhof F, Bernasconi C, Fecker J, Model F, Wei Wei, Arnold DL. Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis. Brain 2019: 142; 2787-2799

Filippi M, Rocca MA, Barkhof F, Bruck W, Chen JT, Comi G, et al. Association between pathological and MRI findings in multiple sclerosis. Lancet Neurol 2012; 11: 349-60.

Friede T, Pohlmann H, Schmidli H. Blinded sample size restimation in event-driven clinical trials: Methods and an application in multiple sclerosis. Pharm Stat. 2019;18:351-65.

Hung HMJ and Wang S. Multiple comparisons in complex clinical trial designs. Biometrical Journal. 2013;55(3):420-9.

Kamudoni P, Amtmann D, Johns J, et al. The validity, responsiveness, and score interpretation of the PROMISnq Physical Function – Multiple Sclerosis 15 a short form in multiple sclerosis. Multiple Sclerosis and Related Disorders . 2022;Vol. 62:103753 [https://doi.org/10.1016/j.msard.2022.103753]

Kamudoni P, Henke C, Barrett A, et al. A new PROMIS physical function short form for use in relapse and progressive multiple sclerosis types. Qual Life Res. 2018;27(1):S1–S190.

Kamudoni P, Johns J, Cook KF et al. Standardizing fatigue measurement in multiple sclerosis: the validity, responsiveness and score interpretation of the PROMIS SF v1.0 – Fatigue (MS) 8a. Multiple Sclerosis and Related Disorders . 2021;Vol. 54:103117 [https://doi.org/10.1016/j.msard.2021.103117]

Keene O.N., Roger J.H., Hartley B.F., Kenward M.G. Missing data sensitivity analysis for recurrent event data using controlled imputation. Pharm. Stat. 2014;13:258-264 [DOI : 10.1002/pst.1624]

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics, 53(3), 983–997. https://doi.org/10.2307/2533558

Lipkovich I., Ratitch B., O'Kelly M. Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. Pharm. Stat. 2016;15(3):216-29 [DOI : 10.1002/pst.1738].

Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics 1996;2(4):1324-33.

Lublin F, Häring D, Ganjgahi H et al. How patients with multiple sclerosis acquire disability. Brain, 2022; 145(9), 3147–3161 [https://doi.org/10.1093/brain/awac016].

O'Kelly M, Ratitch B. Clinical trials with missing data: a guide for practitioners. 10.1002/9781118762516.

Merz, M., Lee, K.R., Kullak-Ublick, G.A. et al. Methodology to Assess Clinical Liver Safety Data. Drug Saf 37 (Suppl 1), 33–45 (2014).

Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med 1985;4(2):213-26 Ratitch B, O'Kelly M, Tosiello B. Missing data in clinical trials: from clnical assumptions to statistical analysis using pattern mixture models. Pharm. Stat. 2013;12(6):337-347 [DOI : 10.1002/pst.1549].

Van Walderveen MA, Kamhorst W, Scheltens P, van Waesberghe JH, Ravid R, Valk J, et al. Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis. Neurology 1998; 50: 1282-8.

Ware J, Kosinski M, Bjorner J., Turner-Bowker D, Gandek B, Meruish M. In User's manual for the SF-36v2 health survey, Lincoln, RI, Quality Metric Incorporated, 2007.

Ware J, Sherbourne C. The MOS 36-ltem Short-Form Health Survey (SF-36) I. Conceptual Framework and Item Selection. Med Care. 1992;30(6):473-83.

18 Appendices

18.1 PROMIS Scores

PROMIS PF(MS)15a

The PROMISnq Short Form v2.0 - Physical Function - Multiple Sclerosis 15a [PROMISnq PF(MS) 15a] is scored on a T score metric, with a mean of 50 and a SD of 10. Higher scores indicate higher physical function. The T score metric allows for easy interpretation of scores by facilitating comparisons to the "average" person, as referenced to the US general population. For example, a person with a T score of 60 is one SD above average physical function. This short form has an MCID of 2.3 to 2.7.

PROMIS Fatigue (MS)8a

The PROMIS Short Form v1.0 - Fatigue - Multiple Sclerosis 8a [PROMIS Fatigue (MS) 8a] is scored on a T score metric, with a mean of 50 and a standard deviation (SD) of 10. Higher scores indicate higher fatigue. The T score metric allows for easy interpretation of scores by facilitating comparisons to the "average" person, as referenced to the US general population. For example, a person with a T score of 60 is one SD above average fatigue. This short form has an MCID of 3.4 to 4 (Kamudoni 2021)

T-scores for the PROMIS PF (MS) and the PROMIS Fatigue (MS) are calculated using/applying an item-response theory (i.e. graded response model) approach, based on "response pattern scoring", which utilizes individuals' item responses and the PROMIS item bank graded response model parameters for each item. These include a discrimination (slope) parameter and an item-response difficulty parameter (level of underlying trait value at which a response option has a 50% chance of being selected), for each item, respectively.

The scoring will be performed using the Healthmeasures Assessment Centre Application Programming Interface (accessible at: http://www.healthmeasures.net), which evaluates the pattern of individuals' responses to the administered items and generates an estimated score based on this pattern (thus the name "response pattern scoring"). Response pattern scoring is especially useful when there are missing responses (i.e. skipped items); a person's score (i.e. PF/fatigue level) can be estimated using available response. Only questionnaires with at least 8 items (for PF) or 4 items (fatigue) representing 50% of responses will be evaluable.

18.2 36-Item Short Form Survey (SF-36) Scoring Instructions

The SF-36 is a validated 36-item, participant-reported indication of overall health status not specific to any age, disease, or intervention group (Ware 1992).

The SF-36 includes multi-item scales measuring each of the following 8 health concepts: (1) physical functioning; (2) role limitations because of physical health problems; (3) bodily pain; (4) social functioning; (5) general mental health (psychological distress and psychological wellbeing); (6) role limitations because of emotional problems; (7) vitality (energy/fatigue); and (8) general health perceptions (see Table 20). These are summarized in 2 summary measures of physical and mental health: the PCS and MCS.

Questions in the standard version of the SF-36 refer to a 4-week time period. Scales are scored according to the Likert method. Lower scores equate to higher disability and higher scores equate to lower disability.

The SF-36v2 multi-item scales yield a health profile (8 scores) or can be aggregated into 2 summary scores, the PCS score and MCS score obtained through a linear combination of weighted transformed scores from the 8 subscales. PCS and MCS are standardized, with an average of 50 and a standard deviation of 10 in the general American population. PCS and MCS are computed only if all of the 8 scale scores are available. This Appendix details how these 2 scores should be calculated, as described in Ware (2007).

Table 20 SF-36 – Abbreviated Item Content for the SF-36v2 Health Domain Scales

Scale	Original item#	Item# in the MS200527-0080 study	Abbreviated Item Content
Physical Functioning (PF)	3a	3	Vigorous activities, such as running, lifting heavy objects, or participating in strenuous sports
	3b	4	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf
	3c	5	Lifting or carrying groceries
	3d	6	Climbing several flights of stairs
	3e	7	Climbing one flight of stairs
	3f	8	Bending, kneeling, or stooping
	3g	9	Walking more than a mile
	3h	10	Walking several hundred yards
	3i	11	Walking one hundred yards
	3j	12	Bathing or dressing oneself
Role-Physical (RP)	4a	13	Cut down the amount of time spent on work or other activities
	4b	14	Accomplished less than you would like
	4c	15	Limited in kind of work or other activities
	4d	16	Had difficulty performing work or other activities (e.g. it took extra effort)

Scale	Original item#	Item# in the MS200527-0080 study	Abbreviated Item Content
Bodily Pain (BP)	7	21	Intensity of bodily pain
	8	22	Extent pain interfered with normal work
General Health (GH)	1	1	Is your health: excellent, very good, good, fair, poor
	11a	33	Seem to get sick a little easier than other people
	11b	34	As healthy as anybody I know
	11c	35	Expect my health to get worse
	11d	36	Health is excellent
Vitality (VT)	9a	23	Feel full of life
	9e	27	Have a lot of energy
	9g	29	Feel worn out
	9i	31	Feel tired
Social Functioning (SF)	6	20	Extent health problems interfered with normal social activities
	10	32	Frequency health problems interfered with social activities
Role-Emotional (RE)	5a	17	Cut down the amount of time spent on work or other activities
	5b	18	Accomplished less than you would like
	5c	19	Did work or other activities less carefully than usual
Mental Health (MH)	9b	24	Been very nervous
	9c	25	Felt so down in the dumps that nothing could cheer you up
	9d	26	Felt calm and peaceful
	9f	28	Felt downhearted and depressed
	9h	30	Been happy
Self-Evaluated Transition (SET)	2	2	How health is now compared to 1 year ago

Step 1: Recoding Item Response Values

Some of the SF-36v2 items will be re-coded so that across all questions, a higher score will indicate a better health state. Questions 2, 3a-3j, 4a-4d, 5a-5c, 9b, 9c, 9f, 9g, 9i, 10, 11a, 11c will be scored as recorded; the other questions will have the scores transformed as shown in Table 21. If multiple answers are given to the same item, then the item score will be left as missing.

Table 21 SF-36 – Recoding

Question	Original code and re-code response				
Question number: 1					
Original response	1	2	3	4	5
Re-coded response	5	4.4	3.4	2	1
Questions numbers: 6, 11b, 11d					
Original response	1	2	3	4	5
Re-coded response	5	4	3	2	1

Question	Original code and re-code response					
Question number: 7						
Original response	1	2	3	4	5	6
Re-coded response	6	5.4	4.2	3.1	2.2	1
Question number: 8 (if question number 7 is answered)						
Original response to #8	1	1	2	3	4	5
Original response to #7	1	2-6	1-6	1-6	1-6	1-6
Re-coded response	6	5	4	3	2	1
Question number: 8 (if question number 7 is NOT answered)						
Original response	1	2	3	4	5	
Re-coded response	6	4.75	3.5	2.25	1	
Questions number: 9a, 9d, 9e, 9h						
Original response	1	2	3	4	5	
Re-coded response	5	4	3	2	1	

Step 2: Determining Health Domain Scale Scores (0-100 Scores)

After item recoding, a total raw score is computed for each health domain scale. The total raw score is the simple algebraic sum of the final response values for all items in a given scale, as shown in Table 22. The total raw score for each scale is transformed to a 0-100 scale score using the following formula:

$$Total\ raw\ score = 100 \times \frac{(\text{Raw score} - \text{Lowest possible raw score})}{\text{Possible raw score range}}$$

Table 22 SF-36 – Values used in Transforming SF-36v2 Health Survey Health Domain Scale Total Raw Scores on the 0-100 Scale

Scale	Sum of Final Response Values	Lowest and highest possible total raw scores	Possible total raw score range
PF	3a+3b+3c+3d+3e+3f+3g+3h+3i+3j	10, 30	20
RP	4a+4b+4c+4d	4, 20	16
BP	7+8	2, 12	10
GH	1+11a+11b+11c+11d	5, 25	20
VT	9a+9e+9g+9i	4, 20	16
SF	6+10	2, 10	8
RE	5a+5b+5c	3, 15	12
MH	9b+9c+9d+9f+9h	5, 25	20

Raw and transformed scale scores are not calculated for the Reported Health Transition (HT) item.

As recommended by the developers of the questionnaire, missing item responses will be treated using the "Half-scale rule", which states that a score can be calculated if the respondent answers at least 50% of the items in a multi-item scale. In such cases, the missing item data will be replaced by the mean of the answered items of its scale. If more than 50% of the items are missing within a scale, the scale score will be missing.

Step 3: Calculating Normalized Health Domain Scores

The normalized scale scores will then be calculated using the following formulas:

$$Health\ Domain\ Z_{\text{score}} = {(Health\ Domain_{0-100\ score} - a)}/{b}$$

$$Normalized\ Health\ Domain\ Score = 50 + (Health\ Domain\ Z_{\text{score}} \times 10)$$

where a and b are the Mean and Standard Deviation of the Health Domain scale in the 1998 U.S. general population as shown in Table 23.

Table 23 SF-36 – 1998 General US Population Means and Standard Deviations used to Calculate Normalized Health Domain Scores

Health Domain Scales	Mean	Standard Deviation
PF	83.29094	23.75883
RP	82.50964	25.52028
BP	71.32527	23.66224
GH	70.84570	20.97821
VT	58.31411	20.01923
SF	84.30250	22.91921
RE	87.39733	21.43778
MH	74.98685	17.75604

The advantages of the normalization of the 8 health domain scales are that results for one health domain scale can be meaningfully compared with those from the other scales and that domain scores have a direct interpretation in relation to the distribution of scores in the 1998 U.S. general population.

Step 4: Scoring the Physical and Mental Component Summary Measures

The PCS and MCS measures are scored using a 3-step procedure:

1. First, the 8 health domain scale scores are standardized using means and standards deviations from the 1998 U.S. general population (see Table 23).

2. Second, these Z-scores are aggregated using weights (factor score coefficient) from the 1990 U.S general population.

3. Third, aggregate PCS and MCS scores are standardized by multiplying the standardized scale by 10 and adding 50.

U.S. general population statistics used in the standardization and in the aggregation of SF-36v2 Health Survey health domain scale scores are presented in Table 24.

Table 24 SF-36 – Factor Score Coefficients used to Calculate PCS and MCS Scores for the SF-36v2

	Summary component measure factor score coefficients	
Scales	**PCS**	**MCS**
PF	0.42402	-0.22999
RP	0.35119	-0.12329
BP	0.31754	-0.09731
GH	0.24954	-0.01571
VT	0.028877	0.23534
SF	-0.00753	0.26876
RE	-0.19206	0.43407
MH	-0.22069	0.48581

Example:

Let's consider the following answers from a participant:

Table 25 SF-36 – Numerical Example – Raw Data

#	Item description	Answer
1	In general, would you say your health is:	1 – Excellent
2	Compared to one year ago, how would you rate your health in general now?	2 – Somewhat better now than one year ago
3	Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	2 – Yes, limited a little

#	Item description	Answer
4	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	2 – Yes, limited a little
5	Lifting or carrying groceries	2 – Yes, limited a little
6	Climbing several flights of stairs	2 – Yes, limited a little
7	Climbing one flight of stairs	2 – Yes, limited a little
8	Bending, kneeling or stooping	1 – Yes, limited a lot
9	Walking more than a mile	2 – Yes, limited a little
10	Walking several hundred yards	2 – Yes, limited a little
11	Walking one hundred yards	2 – Yes, limited a little
12	Bathing or dressing yourself	2 – Yes, limited a little
13	Cut down on the amount of time you spent on work or other activities	2 – Most of the time
14	Accomplished less than you would like	3 – Some of the time
15	Were limited in the kind of work or other activities	2 – Most of the time
16	Had difficulty performing the work or other activities (for example, it took extra effort)	3 – Some of the time
17	Cut down on the amount of time you spent on work or other activities	2 – Most of the time
18	Accomplished less than you would like	3 – Some of the time
19	Don't do work or other activities as carefully as usual	2 – Most of the time
20	During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups?	2 – Slightly
21	How much bodily pain have you had during the past 4 weeks?	3 – Mild
22	During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?	2 – A little bit
23	Did you feel full of life?	4 – A little of the time
24	Have you been a very nervous person?	3 – Some of the time
25	Have you felt so down in the dumps that nothing could cheer you up?	2 – Most of the time
26	Have you felt calm and peaceful?	4 – A little of the time
27	Did you have a lot of energy?	3 – Some of the time
28	Have you felt downhearted and low?	2 – Most of the time
29	Did you feel worn out?	4 – A little of the time
30	Have you been a happy person?	3 – Some of the time
31	Did you feel tired?	2 – Most of the time
32	During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?	4 – A little of the time
33	I seem to get ill more easily than other people	3 – Don't know
34	I am as healthy as anybody I know	2 – Mostly true
35	I expect my health to get worse	3 – Don't know
36	My health is excellent	2 – Mostly true

After having recoded the answers, the 8 domain scores are equal to:

Table 26 SF-36 – Numerical Example – Domain Scores

	Raw Score	Raw Score$_{0-100}$	Health Domain Z-Score	Normalized Health Domain Z-Score
PF	19	100 x (19-10) /20 = 45	45 - 83.29094 / 23.75883 = -1.61	-1.61 x 10 + 50 = 33.9
RP	10	100 x (10-4) /16 = 37.5	37.5 - 82.50964 / 25.52028 = -1.76	-1.76 x 10 + 50 = 32.4
BP	8.2	100 x (8.2-2) /10 = 62	62 - 71.32527 / 23.66224 = -0.39	-0.39 x 10 + 50 = 46.1
GH	19	100 x (19-5) /20 = 70	70 - 70.84570 / 20.97821 = -0.04	-0.04 x 10 + 50 = 49.6
VT	11	100 x (11-4) /16 = 43.75	43.75 - 58.31411 / 20.01923 = -0.73	-0.73 x 10 + 50 = 42.7
SF	8	100 x (8-2) / 8 = 75	75 - 84.30250 / 22.91921 = -0.41	-0.41 x 10 + 50 = 45.9
RE	7	100 x (7-3) / 15 = 27	27 - 87.39733 / 21.43778 = -2.82	-2.82 x 10 + 50 = 21.8
MH	12	100 x (12-5) / 20 = 35	35 - 74.98685 / 17.75604 = -2.25	-2.25 x 10 + 50 = 27.5

Finally, PCS and MCS are provided below using Table 24:

- PCS = -1.61 x 0.42402 + … + -2.25 x -0.22069 = -0.41
- Normalized PCS = 10 x PCS + 50 = 45.9
- MCS = -1.61 x -0.22999 + … + -2.25 x 0.48581 = -2.62
- Normalized MCS = 10 x MCS + 50 = 23.8

18.3 EQ-5D-5L

The EQ-5D questionnaire comprises 5 questions (items) relating to current problems in the dimensions mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Responses in each dimension are divided into 5 ordinal levels coded: (1) no problems, (2) slight problems, (3) moderate problems, (4) severe problems, (5) extreme problems. This part, called the EQ-5D descriptive system, provides a 5-dimensional description of health status.

A unique health state is defined by combining one level from each of the 5 dimensions. A total of 3125 possible health states is defined in this way. Each state is referred to in terms of a 5-digits code. For example, state 11111 indicates no problems on any of the 5 dimensions, while state 12345 indicates no problems with mobility, slight problems with self-care, moderate problems with doing usual activities, severe pain or discomfort, and extreme anxiety or depression.

The EQ-5D-5L descriptive system is followed by a VAS (EQ VAS), similar to a thermometer, ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). The EQ VAS records the respondent's self-rated valuation of health state, i.e. a value which is based on the respondent's preferences (EQ VAS score).

Based on the EQ-5D-5L, the numeric EQ-5D-5L index can be derived, using validated translations provided by EuroQoL presented in the following XLS file. The EQ-5D-5L index is frequently used in economic evaluations: it represents societal preference values for the full set of EQ-5D-5L health states with the best state (perfect health) and "death" being assigned values of 1 and 0, respectively.

The United Kingdom value set of this questionnaire will be used for all participants regardless of their country.

EQ-5D-5L_Crosswalk_
Index_Value_Calculato

Example:

Let's consider a French participant who answered the questionnaire according to Table 27.

Table 27 EQ-5D-5L – Example

Item #	Description	Answer
1	MOBILITY	1 – I have no problems in walking about
2	SELF-CARE	3 – I have moderate problems washing or dressing myself
3	USUAL ACTIVITIES	4 – I have severe problems doing my usual activities
4	PAIN / DISCOMFORT	2 – I have slight pain or discomfort
5	ANXIETY / DEPRESSION	2 – I am slightly anxious or depressed

The health state is then defined as 13422, and the Index value as 0.403.

18.4 Analysis Visit Windows

See Section 9.5 for the definition of study day 1, which is used to determine the analysis visit day in the tables below.

Table 28 Analysis visit windows for Neurological Evaluation EDSS, T25-FW, 9-HPT, SDMT, C-SSRS, PROMIS, EQ-5D-5L and HRU through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 12	85	[2, 127)
Week 24	169	[127, 211)
Week 36	253	[211, 295)
Week 48	337	[295, 379)
Week 60	421	[379, 463)
Week 72	505	[463, 547)
Week 84	589	[547, 631)
Week 96	673	[631, 715)
Week 108	757	[715, 799)
Week 120	841	[799, 883)
Week 132	925	[883, 967)
Week 144	1009	[967, 1051)
Week 156	1093	[1051, 1100)

Table 29 Analysis visit windows for Vital Signs and Weight through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 12	85	[2, 127)
Week 24	169	[127, 211)
Week 36	253	[211, 295)
Week 48	337	[295, 379)
Week 60	421	[379, 463)
Week 72	505	[463, 547)
Week 84	589	[547, 631)
Week 96	673	[631, 883)
Week 156	1093	[883, 1100)

Table 30 Analysis visit windows for MRI through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Screening/Baseline	-28 to 1	≤1
Week 24	169	[2, 253)
Week 48	337	[253, 505)
Week 96	673	[505, 883)
Week 156	1093	[883, 1100)

Table 31 Analysis visit windows for SF-36v2 through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 24	169	[2, 253)
Week 48	337	[253, 421)
Week 72	505	[421, 589)
Week 96	673	[589,757)
Week 120	841	[757,925)
Week 144	1009	[925,1051)
Week 156	1093	[1051,1100)

Table 32 Analysis visit windows for 12-lead ECG through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	[1, dose time)
Day 1 Postdose	1	[dose time, 1]
Week 4	29	[2, 149)
Week 24	269	[149, 471)
Week 96	673	[471, 883)
Week 156	1093	[883, 1100)

Table 33 Analysis visit windows for Biochemistry (except tests noted as Supplement LFTs) and Hematology through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 12	85	[2, 127)
Week 24	169	[127, 253)
Week 48	337	[253, 421)
Week 72	505	[421, 589)
Week 96	673	[589, 715)

Week 108	757	[715, 841)
Week 132	925	[841, 1009)
Week 156	1093	[1009, 1100)

Table 34 Analysis visit windows for Biochemistry tests noted as Supplement LFTs through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 2	15	[2, 22)
Week 4	29	[22, 36)
Week 6	43	[36, 50)
Week 8	57	[50, 64)
Week 10	71	[64, 78)
Week 12	85	[78, 92)
Week 14	99	[92, 106)
Week 16	113	[106, 120)
Week 18	127	[120, 134)
Week 20	141	[134, 148)
Week 22	155	[148, 162)
Week 24	169	[162, 183)
Week 28	197	[183, 211)
Week 32	225	[211, 239)
Week 36	253	[239, 267)
Week 40	281	[267, 295)
Week 44	309	[295, 323)
Week 48	337	[323, 379)
Week 60	421	[379, 463)
Week 72	505	[463, 547)
Week 84	589	[547,631)
Week 96	673	[631, 715)
Week 108	757	[715, 799)
Week 120	841	[799, 883)
Week 132	925	[883, 967)
Week 144	1009	[967, 1051)
Week 156	1093	[1051, 1100)

Table 35 Analysis visit windows for Urinalysis through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 24	169	[2, 253)
Week 48	337	[253, 421)
Week 72	505	[421, 589)
Week 96	673	[589, 883)
Week 156	1093	[883, 1100)

Table 36 Analysis visit windows for Biomarkers (including NfL and Ig levels) through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 12	85	[2, 127)
Week 24	169	[127, 253)
Week 48	337	[253, 421)
Week 72	505	[421, 589)
Week 96	673	[589,757)
Week 120	841	[757,925)
Week 144	1009	[925,1051)
Week 156	1093	[1051,1100)

Table 37 Analysis visit windows for PD/Biomarkers measured as part of the PD substudy (BTKO, cellular function and cytokines) through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 2	15	[2, 22)
Week 4	29	[22, 43)
Week 8	57	[43, 71)
Week 12	85	[71, 127)
Week 24	169	[127, 172)

Table 38 Analysis visit windows for HBV DNA through Week 156

Analysis visit	Nominal analysis visit day	Analysis visit window (days)
Baseline	1	≤1
Week 4	29	[2, 43)
Week 8	57	[43, 71)
Week 12	85	[71, 127)
Week 24	169	[127, 211)
Week 36	253	[211, 295)
Week 48	337	[295, 379)
Week 60	421	[379, 463)
Week 72	505	[463, 547)
Week 84	589	[547, 631)
Week 96	673	[631, 757)
Week 120	841	[757, 967)
Week 156	1093	[967, 1100)

18.5 Tier 1 AEs and AESIs

A reference spreadsheet containing the list of AE terms to programmatically flag participants with Tier 1 AEs and AESIs will be shared by the Safety Team and regularly updated as per MedDRA version updates.

The latest version of the file at time of analysis will be used.

18.6 Laboratory Parameters to be Summarized in the TLFs

Table 39 Laboratory Parameters to be Summarized in the TLFs

	Names of Clinical Safety Laboratory Evaluations in Protocol	If gradable parameter, corresponding evaluation names in NCI-CTCAE 5.0	Worst on treatment value based on normal range
Biochemistry	Albumin	Hypoalbuminemia	LOW
	Aspartate aminotransferase	Aspartate aminotransferase increased	HIGH
	Alanine aminotransferase	Alanine aminotransferase increased	HIGH
	Alkaline phosphatase	Alkaline phosphatase increased	HIGH
	Gamma-Glutamyl-transferase	GGT increased	HIGH
	Lactate dehydrogenase	Blood lactate dehydrogenase increased	HIGH
	Bilirubin (total)	Blood bilirubin increased	HIGH
	Protein (total)		LOW
	Creatinine	Creatinine increased	HIGH
	eGFR	Chronic kidney disease	LOW
	Amylase	Serum amylase increased	HIGH
	Lipase	Lipase increased	HIGH
	Blood urea nitrogen		HIGH
	Glucose	Hypoglycemia	LOW
	Sodium	Hypernatremia	HIGH
	Sodium	Hyponatremia	LOW
	Potassium	Hyperkalemia	HIGH
	Potassium	Hypokalemia	LOW
	Chloride		NA
	Calcium	Hypercalcemia*	HIGH
	Calcium	Hypocalcemia*	LOW
	Magnesium	Hypermagnesemia	HIGH
	Magnesium	Hypomagnesemia	LOW
	Phosphate		LOW
	Total Carbon Dioxide	Blood bicarbonate decreased	LOW
Hematology	Hematocrit		LOW/HIGH
	Hemoglobin	Hemoglobin increased	HIGH
	Hemoglobin	Anemia	LOW
	Mean corpuscular volume (MCV)		NA
	Mean corpuscular hemoglobin (MCH)		NA
	Mean corpuscular hemoglobin concentration (MCHC)		NA

	Names of Clinical Safety Laboratory Evaluations in Protocol	If gradable parameter, corresponding evaluation names in NCI-CTCAE 5.0	Worst on treatment value based on normal range
	Red blood cell count		NA
	Reticulocyte count		NA
	Platelet count	Platelet count decreased	LOW
	White blood cell count	Leukocytosis	HIGH
	White blood cell count	White blood cell decreased	LOW
	White blood cell differentials and absolute counts: Basophils[$]		NA
	White blood cell differentials and absolute counts: Eosinophils[$]	Eosinophilia	HIGH
	White blood cell differentials and absolute counts: Lymphocytes[$]	Lymphocyte count increased	HIGH
	White blood cell differentials and absolute counts: Lymphocytes[$]	Lymphocyte count decreased	LOW
	White blood cell differentials and absolute counts: Monocytes[$]		NA
	White blood cell differentials and absolute counts: Neutrophils[$]	Neutrophil count decreased	LOW
Urinalysis	pH		NA
	Nitrite		NA
	Protein		NA
	Blood		NA
	Glucose		NA
	Ketones bodies		NA
	Urobilinogen		NA
	Bilirubin		NA
	Leukocyte esterase by dipstick		NA
	Specific gravity		NA
	Microscopic examination (if blood or protein is abnormal)		NA
	βhCG (women only)		NA
Coagulation	International normalized ratio		NA
	Partial thromboplastin time		NA
Reflex Testing for HBV DNA	HBV DNA PCR		HIGH
	Antinuclear antibody, antismooth muscle antibody, antibody to liver kidney microsomes		NA
	Alkaline phosphatase, Albumin		NA

	Names of Clinical Safety Laboratory Evaluations in Protocol	If gradable parameter, corresponding evaluation names in NCI-CTCAE 5.0	Worst on treatment value based on normal range
Hepatic/Autoimmune Panel (to be performed in the event of elevated LFTs)	Anti-HAV IgM, HBsAg, anti-HBc, anti- HBsAg, anti-HCV, anti-HEV IgG and IgM, anti-VCA IgG and IgM, anti-EA IgG, anti-EBNA IgG, anti-CMV IgG and IgM, EBV PCR, and CMV PCR		NA
	Ferritin/ Transferrin saturation		NA
	Fibrinogen, ESR, hsCRP		NA
	Focused Genetic Testing (e.g High Iron Fe)		NA

* Corrected calcium (mmol/L) to be used in CTCAE grading derived as 'measured total calcium (mmol/L) + 0.02 (40 – serum albumin [g/L]'.

$ For WBC differential counts (total neutrophil [including bands], lymphocyte, monocyte, eosinophil, and basophil counts), the absolute value will be used when reported. When only percentages are available (this is mainly important for neutrophils and lymphocytes, because the CTCAE grading is based on the absolute counts), the absolute value is derived as follows: Derived differential absolute count = (WBC count) * (Differential %value / 100).

If the range for the differential absolute count is not available (only range for value in % is available) then Grade 1 will be attributed to as follows:

- Lymphocyte count decreased:
  - derived absolute count does not meet Grade 2-4 criteria, and
  - % value < % LLN value, and
  - derived absolute count ≥ 800/mm3
- Neutrophil count decreased:
  - derived absolute count does not meet Grade 2-4 criteria, and
  - % value < % LLN value, and
  - derived absolute count ≥ 1500/mm3

In NCI-CTCAE v5.0 the following parameters are not gradable at Baseline as grading implies dependencies to Baseline (e.g. ALT Grade 1: >ULN - 3.0 x ULN if Baseline was normal; 1.5 - 3.0 x Baseline if Baseline was abnormal):

- Eosinophilia (added term, Grade 1 depends on Baseline)
- Alanine aminotransferase increased
- Alkaline phosphatase increased
- Aspartate aminotransferase increased
- Blood total bilirubin increased
- Creatinine increased
- GGT increased
- INR increased (dependent on Baseline also in v4.03)

In addition, in v5.0 the term 'Hyperglycemia' does not depend on glucose laboratory measurements anymore and is thus excluded from laboratory analyses and part of AE reporting only.

For the term 'Hemoglobin increased' the dependency to Baseline is removed in v5.0. 'INR increased' is dependent on Baseline values and intake of anticoagulants in both versions.

18.7 Handling of laboratory results with modifiers

Some laboratory parameters results are presented with modifiers in SDTM.LB (<xx, >xx, <=xx, >=xx). In order e.g. to use them in summary statistics outputs, laboratory modifiers will be handled as follows:

- When the modifier is equal to "<" the derived laboratory value will be equal to the original value minus a quantity. The quantity will be equal to:
  - 1 if the original value is an integer
  - 0.1 if the original value has one significant digit after the decimal place
  - 0.01 if the original value has 2 significant digits after the decimal place
  - Etc…

As the derived value must always be positive, there is an exception for original values equal to 1, 0.1, 0.01, and so on. For such cases, the quantity will be equal to the quantity defined above divided by 10.

Examples:

1. If the original value is equal to "<15", then the corresponding numeric value for summary statistics will be equal to 14
2. If the original value is equal to "<0.5", then the corresponding numeric value for summary statistics will be equal to 0.4
3. If the original value is equal to "<0.001", then the corresponding numeric value for summary statistics will be equal to 0.0009

- When the modifier is equal to ">" the derived laboratory value will be equal to the original value plus a quantity. The quantity will be equal to:
  - 1 if the original value is an integer
  - 0.1 if the original value has 1 significant digit after the decimal place
  - 0.01 if the original value has 2 significant digits after the decimal place
  - Etc…

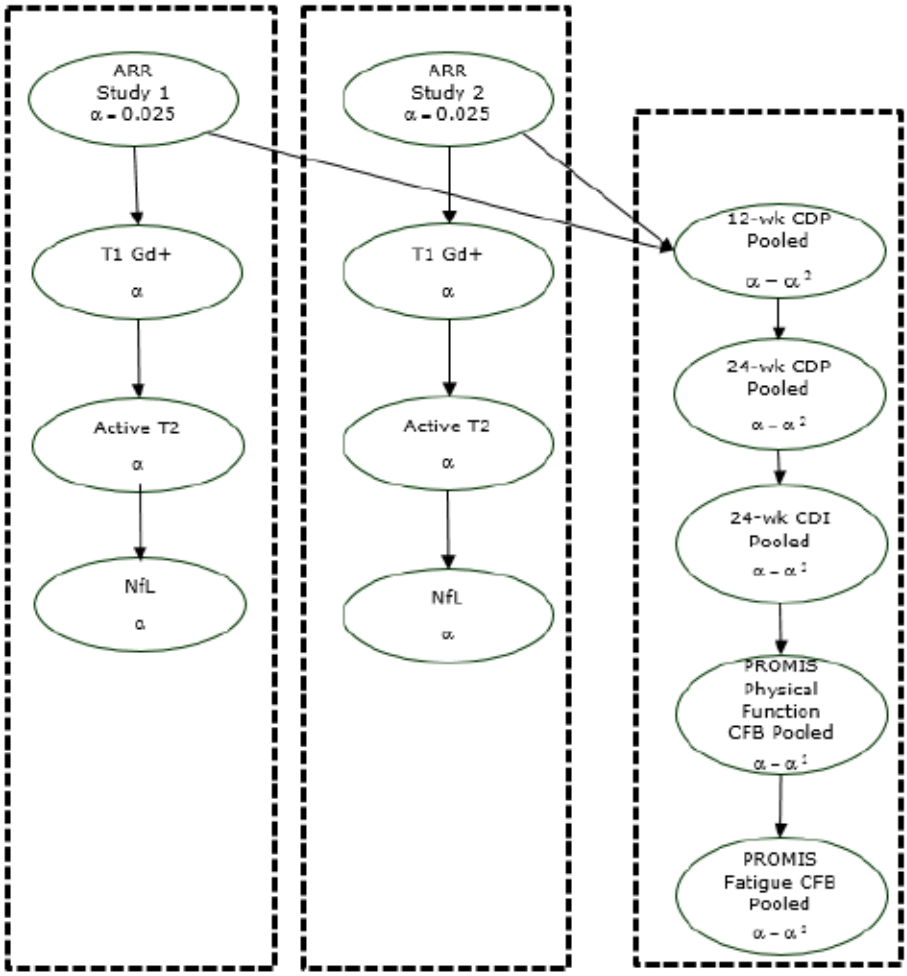
Examples:

1. If the original value is equal to ">20", then the corresponding numeric value for summary statistics will be equal to 21
2. If the original value is equal to ">8.5", then the corresponding numeric value for summary statistics will be equal to 8.6
3. If the original value is equal to ">1.045", then the corresponding numeric value for summary statistics will be equal to 1.046

Exceptions:

1. For Cytomegalovirus IgM Antibody, Epstein-Barr Capsid IgM Antibody, Epstein-Barr Early D Antigen IgG Ab, Epstein-Barr Nuclear Antigen 1 IgG Ab, when the valus is "<yy" then corresponding numeric value for summary statistics will be equal to "yy/2".
2. Hepatitis B Virus DNA Quant PCR : If the original value is equal to "<10", then the corresponding numeric value for summary statistics will be equal to 5 (for other original values, the general rule above applies)

18.8 Multiplicity Graph



ARR = Annualized Relapse Rate, CDP = Confirmed Disability Progression, CFB = change from Baseline, Gd^+ = gadolinium positive, NfL = Neurofilament Light Chain, PROMIS = Patient Reported Outcomes Measurement Information System, T1 and T2 = type of Magnetic Resonance Image, wk = Week.

18.9 Statistical Methodology

The document detailing statistical methodology and procedures is found in *ctp-ms200527-0080-PPD -stat-method-v1-0.docx*.

18.10 SAP for IDMC/SMC

The IDMC SAP is found in ctp-ms200527-0080-PPD -sap-idmc-v7-0.docx.

BREEZE location:

MS200527_0080-Statistical Analysis Plan-IDMC SAP V7.0/Document ID: VV-CLIN-333164

MS200527_0080-Statistical Analysis Plan-IDMC SAP Shells V7.0/Document ID: VV-CLIN-333165

Signature Page for VV-CLIN-171569 v11.0

Approval Task	PPD Nonclinical 21-Nov-2023 15:25:06 GMT+0000

Approval Task	PPD Clinical 21-Nov-2023 15:44:37 GMT+0000

Signature Page for VV-CLIN-171569 v11.0