For the transformation to the OMOP common data model with the SwitchBox, it is needed to transform the source data to a defined dataset first. For the first module, the data dictionary from the MSDA Core Dataset was used to prepare the transformation process. For a better handling of source data, the categories are needed to be transformed to single CSV files (representing each category of the Core Dataset). The single CSV files structures, as they are needed as input files for the SwitchBox, are illustrated below. In general, the variables, their specified names and the value options are described in the latest version of the Core Dataset and must be used.

Important: Every column as described below must be included in the CSV even if it is not collected in the data source!

Note: For all dates, if dd or MM is not allowed/not collected, 15 is used for missing days and 06 for missing months.

For some categories there have been changes to the data dictionary to ensure a better processability. This will be clearly explained per file / category, if applicable.

1. Demographic Data -> patient.csv

- For "race_ethnicity" there are no values specified since this is a very locally dependent variable. Use text (strings) for the value.
- The "smoking_count" was introduced as a single variable (column) instead of being a value for "smoking". If a patient is a "current_smoker", this variable should have a number of cigarettes smoked per day (int).

patient.csv										
patient_id	**date_visit**	**date_birth**	**sex**	**residence**	**race_ethnicity**	**education**	**employment**	**smoking**	**smoking_count**	**ms_family**
0	2017-12-30	1989-07-14	male	residence_QE	*Ethnicity*	isced_3	student	former_smoker		no
1	2017-12-17	1991-11-04	male	residence_AJ		isced_3	parental_leave	former_smoker		no
2	2018-06-08	1952-01-23	female	residence_CC	*Race*	isced_5	retired	never_smoked		no
3	2022-04-21	1965-12-24	male	residence_AM	*Ethnicity*	isced_2	med_retired_other	current_smoker	15	yes
4	2017-12-11	2005-01-06	male	residence_NY		isced_3	student	current_smoker	4	no
5	2019-08-25	1992-02-16	female	residence_NT		isced_4	employed	never_smoked		unknown
…	…	…	…	…	…	…	…	…	…	…

2. Disease-specific Information: Disease History -> disease_history.csv

No changes from the data dictionary.

disease_history.csv					
patient_id	**date_visit**	**date_diagnosis**	**date_onset**	**csf_olib**	**ms_course**
87	2020-08-28	2014-11-14	2014-10-02	yes	rrms
79	2022-10-01	2007-07-12	2006-07-15	unknown	spms
47	2019-04-04	2012-04-06	2012-02-23	yes	ppms
39	2022-09-28	2006-10-07	2006-01-22	unknown	rrms
24	2021-10-20	2007-03-14	2005-05-20	no	cis
17	2020-05-27	1999-03-01	1998-12-16	unknown	spms
…	…	…	…	…	…

3. Disease-specific Information: Disease Status -> disease_status.csv

- Depending on the type of registry/cohort, usually only one variable is filled in for "edss_score" and "pdss_score". Same applies to "ms_status_clin" or "ms_status_pat".
- To make the linkage with the laterality easier, the nine-hole peg test (ninehpt) was split up in "ninehpt_right" and "ninehpt_left". To prevent coding issues, the variable was renamed to "ninehpt_xx" (formerly 9HPT). The value is currently the average of both measurement times per hand.
- The symptoms ("current_sympt"), along with "sever_symp" and "treat_symp" were extracted into a separate CSV file (see below) to enable a dedicated data collection on several symptoms with their assessment during one visit.

disease_status.csv										
patient_id	date_visit	ms_status_clin	ms_status_pat	edss_score	pdss_score	t25fw	ninehpt_right	ninehpt_left	vib_sense	sdmt
76	2017-07-17	statusC_stable	statusP_better	1.5		7.2	23.3	26.7	vib_decreased_right	32
97	2022-08-14	statusC_act_progr	statusP_worse	6.5		22.1	33.4	35.8	vib_normal_right	26
74	2022-04-30		statusP_worse2		4	16.4	28.3	31.6	vib_normal_right	59
93	2016-02-29	statusC_act_progr	statusP_worse2	7.0		25.7	34.6	35.3	vib_normal_right	15
27	2019-06-21		statusP_better		3	9.1	26.7	24.8	vib_normal_right	48
22	2022-03-20	statusC_stable	statusP_better	5.5		19.8	31.9	38.6	vib_decreased_right	56
...	...	...	...	...	...	...	...	...	...	...

3.1 Current symptoms and their severity and treatment -> symptom.csv

- The rating scale for symptom severity was adapted to a scale from 0 to 6 (0: not affected at all, 1: very mild limitation/ I make minor adjustments, 2: mild limitation/ I make frequent adjustments, 3: moderate limitation/ I reduced my daily activities, 4: severe limitation/ I gave up some activities, 5: very severe limitation/ I'm unable to do many daily activities, 6: total limitation/ I'm unable to do most daily activities) to match the SymptoMScreen.
- For every symptom during a visit, a new row is populated for the concerning patient_id and date_visit.

symptom.csv				
patient_id	date_visit	current_symptom	sever_symp	treat_symp
34	2018-04-23	symp_fatigue	2	no
70	2018-11-12	symp_bladder	1	no
94	2020-11-28	symp_cognition	6	yes

94	2020-11-28	symp_vision	1	no	
53	2022-03-20	symp_sensory	5	yes	
53	2022-03-20	symp_vision	3	yes	
...	...	...	...	...	

4. Disease Specific Information: Relapses -> relapses.csv

No changes from the data dictionary.

relapses.csv					
patient_id	date_visit	relapse	date_relapse	relapse_treat	relapse_recovery
14	2019-09-19	no			
58	2017-04-11	unknown			
34	2021-01-10	yes	2020-12-28	no	unknown
3	2019-05-30	yes	2019-05-10	yes	partial_recovery
5	2016-11-12	no			
99	2019-04-24	yes	2019-04-19	no	compl_recovery
...	...	...	...	...	...

5. Para-clinical Investigations: MRI -> mri.csv

No changes from the data dictionary.

mri.csv							
patient_id	date_visit	mri	mri_region	mri_date	mri_gd_les	mri_new_les_T1	mri_new_les_T2
83	2016-12-11	no					
38	2022-02-17	yes	mri_myelon	2022-01-15	2	1	1
98	2014-11-01	no					
21	2021-02-12	yes	mri_myelon	2020-12-13	1	0	1
83	2017-04-24	yes	mri_myelon	2017-03-13	1	2	1
85	2016-05-21	yes	mri_brain	2016-04-23	0	1	0
...	...	...	...	...	...	...	...

6. Comorbidities -> comorbidities.csv

- For every comorbidity during a visit or in the patient's trajectory, a new row is populated for the concerning patient_id and date_visit.
- If comorbidities are collected that don't fit the specified values from the Core Dataset, "com_other" should be used. Ideally, at least a body system is then added to the comorbidity to ensure some knowledge transfer.

comorbidities.csv				
patient_id	**date_visit**	**comorbidity**	**com_type**	**com_system**
37	2018-02-05	no	com_epilepsy	
37	2018-02-05	yes	com_other	sys_cardio
35	2015-04-04	no	com_chronic_lung	
41	2019-01-01	yes	com_arrhythmia	
53	2019-04-01	yes	com_autoimmune	
53	2020-04-01	yes		sys_skel
…	…	…	…	…

7. Disease-modifying Treatment (DMT): for current and past treatments -> dmt.csv

No changes from the data dictionary.

dmt.csv						
patient_id	**date_visit**	**dmt_status**	**dmt_type**	**dmt_start**	**dmt_stop**	**dmt_stop_reas**
26	2016-04-11	dmt_naive				
81	2019-12-01	no	ponesimod	2017-05-25	2018-06-05	stop_patient_choice
41	2022-06-01	no	dimethyl_fumarate	2016-11-09	2021-11-12	stop_intolerance
99	2021-07-14	yes	ocrelizumab	2021-04-11		
23	2022-05-01	yes	ublituximab	2021-11-23		
66	2021-01-03	no	interferon	2018-06-21	2020-04-09	stop_contraindication
…	…	…	…	…	…	…

8. Non-pharmaceutical Treatments -> npt.csv

No changes from the data dictionary.

npt.csv				
patient_id	**date_visit**	**np_treat_type**	**np_treat_start**	**np_treat_stop**
19	2020-07-29	rehabilitation	2017-01-26	2017-04-26
93	2015-03-12	np_other	2013-07-21	2014-02-22
57	2021-08-06	speech_therapy	2020-03-09	
45	2018-01-22	physiotherapy	2017-10-15	
70	2016-11-22	rehabilitation	2016-06-09	2016-07-09
74	2022-07-14	speech_therapy	2021-06-11	2022-04-10
…	…	…	…	…