

PhUSE US Connect 2020

Data Wrangling – Combining and Restructuring QS related ADaM Datasets for the purpose of pathways predictive modeling.

Stanley Brill

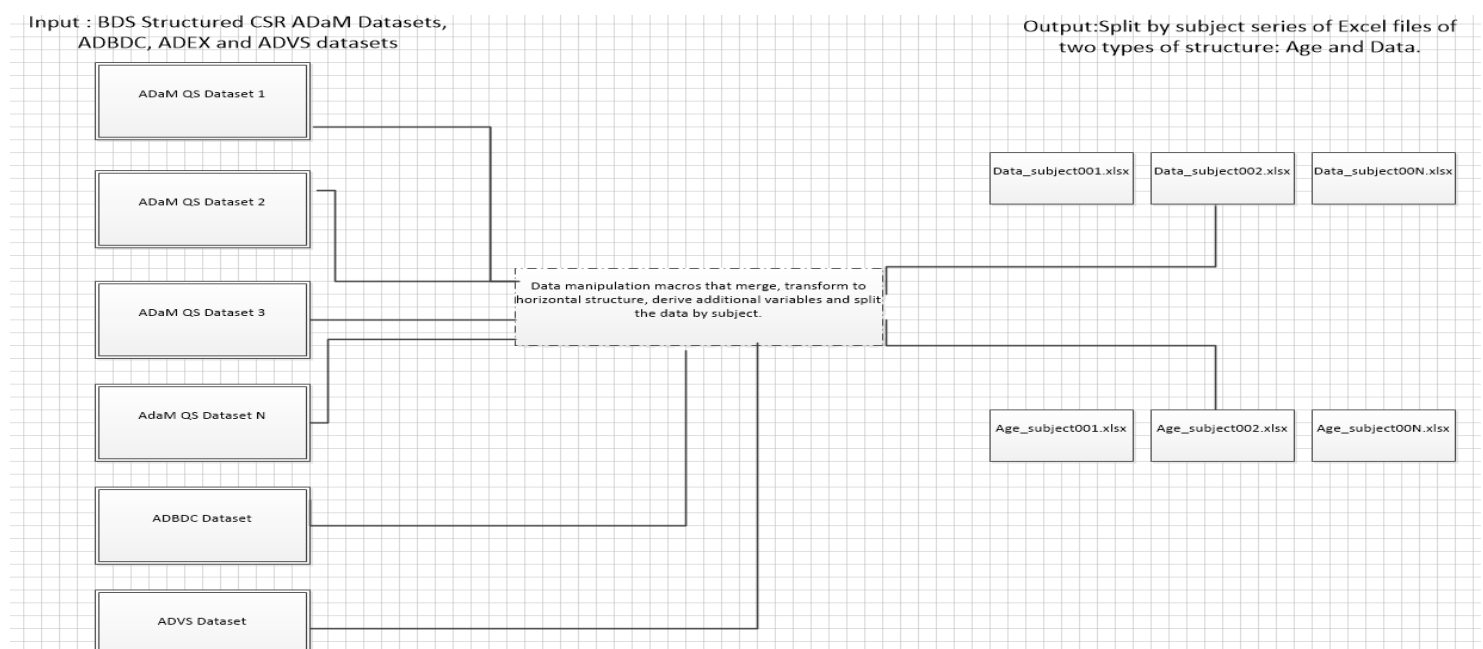
Janssen Pharmaceuticals - Johnson & Johnson Pharmaceutical Research and Development, Titusville NJ, U.S.A.

Abstract

The implementation of Data Science analytical capabilities in the realm of clinical trials of the pharmaceutical research industry is paramount to attain innovation and sharpen technological edge. Data Science is an emerging discipline in which quantitative and analytical approaches, processes, and tools are developed and integrated to extract knowledge and insights from increasingly large and complex sets of clinical data. This includes utilizing complex techniques like predictive modeling (machine Learning, Deep Learning) and Artificial Intelligence. Often, statistical programmers who work on analyzing clinical trials may be asked to collaborate with their data scientists' peers to provide a tailored clinical data structure that would fit the requirement to be used as an input for the analytical tools. This abstract talks about a use case in the-neuroscience therapeutic area where integrated data from the AdaM QS related domains was created predicting a relapse in Major Depressive disorder (MDD) subjects on an investigational drug.

Introduction.

The idea is to use analysis datasets (Specifically, AdaM efficacy datasets in the BDS structure), containing certain necessary scales. For example: MADRS (Montgomery–Åsberg Depression Rating Scale) , CGIS (The Clinical Global Impression Scale), EQ5D (standardized instrument for measuring generic health status), SDS (Sheehan Disability Scale), PHQ9 (The Patient Health Questionnaire), BPICSS (Bladder Pain / Interstitial Cystitis Symptom Score). The data is then transposed by subject ID and assessment date and additional variables are derived for analysis. The horizontal datasets are then adjusted for two types : age of assessments and actual value.



Process flow.

The initial input BDS structured datasets may look like those below, snapshots taken for one subject as an example.

Dataset BPICSS (partial view for one subject).

	USUBJID	ADT	ADTM	ADY	AVISIT	APHASE	APHASEN	PARAM	PARAMCD	AVAL
1	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Urinated Because of Bladder Pain	BPIC0101	0
2	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Need to Urinate After Urinating	BPIC0102	0
3	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Urinate to Avoid Worse Pain	BPIC0103	0
4	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 How Often, Pressure in Bladder	BPIC0104	0
5	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 How Often, Pain in Bladder	BPIC0105	0
6	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Bother, urinary frequency, day	BPIC0106	0
7	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Bother, urinary frequency, night	BPIC0107	0
8	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 Worst Bladder Pain	BPIC0108	0
9	ABC001	19SEP2017	19SEP2017:09:07:54		1 Baseline(OP)	OPTIMIZATION		2 BPIC-SS Total Score	BPIC0109	0
10	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Urinated Because of Bladder Pain	BPIC0101	0
11	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Need to Urinate After Urinating	BPIC0102	0
12	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Urinate to Avoid Worse Pain	BPIC0103	0
13	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 How Often, Pressure in Bladder	BPIC0104	0
14	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 How Often, Pain in Bladder	BPIC0105	0
15	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Bother, urinary frequency, day	BPIC0106	0
16	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Bother, urinary frequency, night	BPIC0107	0
17	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 Worst Bladder Pain	BPIC0108	0
18	ABC001	27SEP2017	27SEP2017:08:38:04		9 Week 1(OP)	OPTIMIZATION		2 BPIC-SS Total Score	BPIC0109	0

Dataset PHQ9 (partial view for one subject).

	USUBJID	ADT	ADTM	ADY	AVISIT	APHASE	APHASEN	PARAM	PARAMCD	AVAL
1	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Little Interest/Pleasure in Things	PHQ0101	1
2	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Feeling Down Depressed or Hopeless	PHQ0102	1
3	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Trouble Falling or Staying Asleep	PHQ0103	2
4	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Feeling Tired or Little Energy	PHQ0104	2
5	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Poor Appetite or Overeating	PHQ0105	0
6	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Feeling Bad About Yourself	PHQ0106	1
7	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Trouble Concentrating on Things	PHQ0107	1
8	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Moving Slowly or Fidgety/Restless	PHQ0108	0
9	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Thoughts You Be Better Off Dead	PHQ0109	0
10	ABC001	03OCT2017	03OCT2017:07:55:10		15 Week 2(OP)	OPTIMIZATION		2 Total Score	PHQ0111	8
11	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Little Interest/Pleasure in Things	PHQ0101	1
12	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Feeling Down Depressed or Hopeless	PHQ0102	1
13	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Trouble Falling or Staying Asleep	PHQ0103	2
14	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Feeling Tired or Little Energy	PHQ0104	2
15	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Poor Appetite or Overeating	PHQ0105	0
16	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Feeling Bad About Yourself	PHQ0106	0
17	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Trouble Concentrating on Things	PHQ0107	1
18	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Moving Slowly or Fidgety/Restless	PHQ0108	0
19	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Thoughts You Be Better Off Dead	PHQ0109	0
20	ABC001	16OCT2017	16OCT2017:07:58:46		28 Week 4(OP)	OPTIMIZATION		2 Total Score	PHQ0111	7

Dataset MADRS (partial view for one subject).

	USUBJID	ADT	ADY	AVISIT	APHASE	APHASEN	PARAM	PARAMCD	AVAL
1	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Reported Sadness	MADR0201	0
2	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Apparent Sadness	MADR0202	0
3	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Inner Tension	MADR0203	0
4	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Reduced Sleep	MADR0204	4
5	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Reduced Appetite	MADR0205	0
6	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Concentration Difficulties	MADR0206	3
7	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Lassitude	MADR0207	3
8	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Inability to Feel	MADR0208	3
9	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Pessimistic Thoughts	MADR0209	1
10	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 MADR02-Suicidal Thoughts	MADR0210	0
11	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 Derived MADRS Total Score	MADR0212	14
12	ABC001	19SEP2017		1 Baseline(OP)	OPTIMIZATION		2 Improvement in Total MADRS Score >= 50%	MADR0213	1
13	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Reported Sadness	MADR0201	3
14	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Apparent Sadness	MADR0202	2
15	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Inner Tension	MADR0203	3
16	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Reduced Sleep	MADR0204	3
17	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Reduced Appetite	MADR0205	0
18	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Concentration Difficulties	MADR0206	4
19	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Lassitude	MADR0207	5
20	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Inability to Feel	MADR0208	4
21	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Pessimistic Thoughts	MADR0209	2
22	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 MADR02-Suicidal Thoughts	MADR0210	0
23	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 Derived MADRS Total Score	MADR0212	26
24	ABC001	02OCT2017		14 Week 2(OP)	OPTIMIZATION		2 Improvement in Total MADRS Score >= 50%	MADR0213	0

Corresponding Baseline Information:

BPICSS

	USUBJID	ADT	ADY	APHASE	APHASEN	PARAMCD	BASE_AVAL
1	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0101	0
2	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0102	0
3	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0103	0
4	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0104	0
5	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0105	0
6	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0106	0
7	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0107	0
8	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0108	0
9	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0109	0
10	ABC001	21AUG2017	-29	INDUCTION		1 BPIC0110	0

PHQ9

	USUBJID	ADT	ADY	APHASE	APHASEN	PARAMCD	BASE_AVAL
1	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0101	3
2	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0102	3
3	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0103	3
4	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0104	1
5	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0105	3
6	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0106	3
7	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0107	2
8	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0108	2
9	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0109	0
10	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0111	20
11	ABC001	21AUG2017	-29	INDUCTION		1 PHQ0112	20

MADRS

	USUBJID	ADT	ADY	APHASE	APHASEN	PARAMCD	BASE_AVAL
1	ABC001	21AUG2017	-29	INDUCTION		1 MADR0201	5
2	ABC001	21AUG2017	-29	INDUCTION		1 MADR0202	4
3	ABC001	21AUG2017	-29	INDUCTION		1 MADR0203	3
4	ABC001	21AUG2017	-29	INDUCTION		1 MADR0204	4
5	ABC001	21AUG2017	-29	INDUCTION		1 MADR0205	4
6	ABC001	21AUG2017	-29	INDUCTION		1 MADR0206	5
7	ABC001	21AUG2017	-29	INDUCTION		1 MADR0207	5
8	ABC001	21AUG2017	-29	INDUCTION		1 MADR0208	5
9	ABC001	21AUG2017	-29	INDUCTION		1 MADR0209	4
10	ABC001	21AUG2017	-29	INDUCTION		1 MADR0210	0
11	ABC001	21AUG2017	-29	INDUCTION		1 MADR0211	39
12	ABC001	21AUG2017	-29	INDUCTION		1 MADR0212	39

In the next step the datasets are transposed by subject ID and date and time of the assessment to obtain horizontal structure needed.

BPICSS transposed:

	USUBJID	ADT	ADTM	AVISIT	ADY	A_BPIC0101	A_BPIC0102	A_BPIC0103	A_BPIC0104	A_BPIC0105	A_BPIC0106	A_BPIC0107	A_BPIC0108	A_BPIC0109
1	ABC001	19SEP2017	19SEP2017:09:07:54	Baseline(OP)	1	0	0	0	0	0	0	0	0	0
2	ABC001	27SEP2017	27SEP2017:08:38:04	Week 1(OP)	9	0	0	0	0	0	0	0	0	0
3	ABC001	16OCT2017	16OCT2017:08:03:15	Week 4(OP)	28	0	0	0	0	0	0	0	0	0
4	ABC001	14NOV2017	14NOV2017:16:21:13	Week 8(OP)	57	0	0	0	0	0	0	0	0	0
5	ABC001	08DEC2017	08DEC2017:13:57:40	Week 12(OP)	81	0	0	0	0	0	0	0	0	0
6	ABC001	08JAN2018	08JAN2018:15:40:49	Week 4(MA)	112	0	0	0	0	0	0	0	0	0
7	ABC001	25JAN2018	25JAN2018:17:04:30	Week 8(MA)	129	0	0	0	0	0	0	0	0	0

PHQ9 transposed:

	USUBJID	ADT	ADTM	AVISIT	ADY	A_PHQ0101	A_PHQ0102	A_PHQ0103	A_PHQ0104	A_PHQ0105	A_PHQ0106	A_PHQ0107	A_PHQ0108	A_PHQ0109	A_PHQ0111
1	ABC001	03OCT2017	03OCT2017:07:55:10	Week 2(OP)	15	1	1	2	2	0	1	1	0	0	8
2	ABC001	16OCT2017	16OCT2017:07:58:46	Week 4(OP)	28	1	1	2	2	0	0	1	0	0	7
3	ABC001	31OCT2017	31OCT2017:07:34:22	Week 6(OP)	43	0	0	3	2	0	0	0	0	0	5
4	ABC001	14NOV2017	14NOV2017:16:16:40	Week 8(OP)	57	0	0	2	1	0	0	0	0	0	3
5	ABC001	28NOV2017	28NOV2017:15:58:36	Week 10(OP)	71	1	1	1	2	0	0	1	0	0	6
6	ABC001	08DEC2017	08DEC2017:13:54:37	Week 12(OP)	81	0	0	2	1	0	0	2	0	0	5
7	ABC001	27DEC2017	27DEC2017:15:38:55	Week 2(MA)	100	0	0	0	1	0	0	0	0	0	1
8	ABC001	08JAN2018	08JAN2018:15:37:35	Week 4(MA)	112	0	0	1	1	0	0	0	0	0	2
9	ABC001	22JAN2018	22JAN2018:15:36:15	Week 6(MA)	126	1	1	1	1	0	1	1	0	0	6

MADRS transposed:

	USUBJID	ADT	AVISIT	ADY	A_MADR0201	A_MADR0202	A_MADR0203	A_MADR0204	A_MADR0205	A_MADR0206	A_MADR0207	A_MADR0208	A_MADR0209	A_MADR0210	A_MADR0212	A_MADR0213
1	ABC001	19SEP2017	Baseline(OP)	1	0	0	0	4	0	3	3	3	1	0	14	1
2	ABC001	02OCT2017	Week 2(OP)	14	3	2	3	3	0	4	5	4	2	0	26	0
3	ABC001	10OCT2017	Week 3(OP)	22	0	0	0	3	0	3	3	3	3	0	15	1
4	ABC001	15OCT2017	Week 4(OP)	27	1	0	0	5	0	3	2	4	1	0	16	1
5	ABC001	22OCT2017	Week 5(OP)	34	0	0	0	5	0	3	2	2	1	0	13	1
6	ABC001	30OCT2017	Week 6(OP)	42	0	0	0	3	0	2	3	3	0	0	11	1
7	ABC001	06NOV2017	Week 7(OP)	49	0	0	0	3	0	2	3	3	1	0	12	1
8	ABC001	13NOV2017	Week 8(OP)	56	1	0	3	3	0	4	4	2	3	0	20	0
9	ABC001	20NOV2017	Week 9(OP)	63	3	3	3	2	0	4	5	3	3	0	26	0
10	ABC001	27NOV2017	Week 10(OP)	70	0	0	1	1	0	4	3	2	2	0	13	1
11	ABC001	03DEC2017	Week 11(OP)	76	1	0	0	1	0	3	2	1	2	0	10	1
12	ABC001	07DEC2017	Week 12(OP)	80	1	0	0	4	0	2	2	3	1	0	13	1
13	ABC001	17DEC2017	Week 1(MA)	90	1	0	0	4	0	1	1	2	0	0	9	1
14	ABC001	26DEC2017	Week 3(MA)	99	0	0	1	2	0	0	2	2	0	0	7	1
15	ABC001	07JAN2018	Week 4(MA)	111	0	0	0	0	0	1	2	1	0	0	4	1
16	ABC001	14JAN2018	Week 5(MA)	118	1	0	0	1	0	3	2	1	1	0	9	1
17	ABC001	21JAN2018	Week 6(MA)	125	3	3	3	3	0	2	5	2	3	0	24	0

In the next phase, we split the datasets by unique subject ID, QS efficacy domain and a single scale of assessment within this domain, while creating two different structures: Age and LOCF.

From the original transposed data ADBPICSS dataset:

	USUBJID	ADT	A_BPIC0101	A_BPIC0102	A_BPIC0103	A_BPIC0104	A_BPIC0105	A_BPIC0106	A_BPIC0107	A_BPIC0108	A_BPIC0109
1	ABC001	19SEP2017	0	0	0	0	0	0	0	0	0
2	ABC001	27SEP2017	0	0	0	0	0	0	0	0	0
3	ABC001	16OCT2017	0	0	0	0	0	0	0	0	0
4	ABC001	14NOV2017	0	0	0	0	0	0	0	0	0
5	ABC001	08DEC2017	0	0	0	0	0	0	0	0	0
6	ABC001	08JAN2018	0	0	0	0	0	0	0	0	0
7	ABC001	25JAN2018	0	0	0	0	0	0	0	0	0

We are creating separate datasets of AGE structure: This dataset represents the age of assessment. Here in the example you see that they are split by subject and individual scale assessment from the transposed dataset, and that the dates are padded, meaning that records are inserted between ADT=19SEP207 and ADT = 27SEP2017 to represent all dates. The original values of the assessments are replaced “0” and incremented by 1 until you reach the next assessment which is set back to 0 again. This process applies for all split datasets across domains and individual scales for all subjects.

(Partial example of AGE structure split datasets for one subject for BPICSS domain)

	USUBJID	ADT	A_BPIC0101
1	ABC001	19SEP2017	0
2	ABC001	20SEP2017	1
3	ABC001	21SEP2017	2
4	ABC001	22SEP2017	3
5	ABC001	23SEP2017	4
6	ABC001	24SEP2017	5
7	ABC001	25SEP2017	6
8	ABC001	26SEP2017	7
9	ABC001	27SEP2017	0
10	ABC001	28SEP2017	1
11	ABC001	29SEP2017	2
12	ABC001	30SEP2017	3
13	ABC001	01OCT2017	4
14	ABC001	02OCT2017	5
15	ABC001	03OCT2017	6
16	ABC001	04OCT2017	7
17	ABC001	05OCT2017	8
18	ABC001	06OCT2017	9
19	ABC001	07OCT2017	10
20	ABC001	08OCT2017	11
21	ABC001	09OCT2017	12
22	ABC001	10OCT2017	13
23	ABC001	11OCT2017	14
24	ABC001	12OCT2017	15
25	ABC001	13OCT2017	16
26	ABC001	14OCT2017	17
27	ABC001	15OCT2017	18

	USUBJID	ADT	A_BPIC0102
1	ABC001	19SEP2017	0
2	ABC001	20SEP2017	1
3	ABC001	21SEP2017	2
4	ABC001	22SEP2017	3
5	ABC001	23SEP2017	4
6	ABC001	24SEP2017	5
7	ABC001	25SEP2017	6
8	ABC001	26SEP2017	7
9	ABC001	27SEP2017	0
10	ABC001	28SEP2017	1
11	ABC001	29SEP2017	2
12	ABC001	30SEP2017	3
13	ABC001	01OCT2017	4
14	ABC001	02OCT2017	5
15	ABC001	03OCT2017	6
16	ABC001	04OCT2017	7
17	ABC001	05OCT2017	8
18	ABC001	06OCT2017	9
19	ABC001	07OCT2017	10
20	ABC001	08OCT2017	11
21	ABC001	09OCT2017	12
22	ABC001	10OCT2017	13
23	ABC001	11OCT2017	14
24	ABC001	12OCT2017	15
25	ABC001	13OCT2017	16
26	ABC001	14OCT2017	17
27	ABC001	15OCT2017	18

Similarly, LOCF structure split datasets are created with padded dates, the only difference is that the original values of transposed scales assessments are preserved and retained for padded dates using LOCF principle

	USUBJID	ADT	A_BPIC0101		USUBJID	ADT	A_BPIC0102
1	ABC001	19SEP2017	0	1	ABC001	19SEP2017	0
2	ABC001	20SEP2017	0	2	ABC001	20SEP2017	0
3	ABC001	21SEP2017	0	3	ABC001	21SEP2017	0
4	ABC001	22SEP2017	0	4	ABC001	22SEP2017	0
5	ABC001	23SEP2017	0	5	ABC001	23SEP2017	0
6	ABC001	24SEP2017	0	6	ABC001	24SEP2017	0
7	ABC001	25SEP2017	0	7	ABC001	25SEP2017	0
8	ABC001	26SEP2017	0	8	ABC001	26SEP2017	0
9	ABC001	27SEP2017	0	9	ABC001	27SEP2017	0
10	ABC001	28SEP2017	0	10	ABC001	28SEP2017	0
11	ABC001	29SEP2017	0	11	ABC001	29SEP2017	0
12	ABC001	30SEP2017	0	12	ABC001	30SEP2017	0
13	ABC001	01OCT2017	0	13	ABC001	01OCT2017	0
14	ABC001	02OCT2017	0	14	ABC001	02OCT2017	0
15	ABC001	03OCT2017	0	15	ABC001	03OCT2017	0
16	ABC001	04OCT2017	0	16	ABC001	04OCT2017	0
17	ABC001	05OCT2017	0	17	ABC001	05OCT2017	0
18	ABC001	06OCT2017	0	18	ABC001	06OCT2017	0
19	ABC001	07OCT2017	0	19	ABC001	07OCT2017	0
20	ABC001	08OCT2017	0	20	ABC001	08OCT2017	0
21	ABC001	09OCT2017	0	21	ABC001	09OCT2017	0
22	ABC001	10OCT2017	0	22	ABC001	10OCT2017	0
23	ABC001	11OCT2017	0	23	ABC001	11OCT2017	0
24	ABC001	12OCT2017	0	24	ABC001	12OCT2017	0
25	ABC001	13OCT2017	0	25	ABC001	13OCT2017	0
26	ABC001	14OCT2017	0	26	ABC001	14OCT2017	0
27	ABC001	15OCT2017	0	27	ABC001	15OCT2017	0
28	ABC001	16OCT2017	0	28	ABC001	16OCT2017	0

In the final stage, all the information is merged horizontally together with the auxiliary domains such as vitals , demographics , additional derived variables such as AGE_MDD , REL_ASSESS , BASE_ASSESS. Two separate structures of LOCF and AGE are preserved and separated Excel files by subject are generated:

(Partial view of the final excel file of one subject AGE structure)

1	Unique Subject Identifier	Analysis Date	ADY	Age	Sex	Race	Dose Administered	DOSE_ADMIN_FLAG	Analysis Visit	BASE_DIABP	BASE_PULSE	BASE_RESP	BASE_SYSBP
2	ABC001	19-Sep-17	1	38	M	WHITE			0 Baseline(OP)	75	60	14	109
3	ABC001	20-Sep-17	2	38	M	WHITE			0 Baseline(OP)	75	60	14	109
4	ABC001	21-Sep-17	3	38	M	WHITE			0 Baseline(OP)	75	60	14	109
5	ABC001	22-Sep-17	4	38	M	WHITE			0 Baseline(OP)	75	60	14	109
6	ABC001	23-Sep-17	5	38	M	WHITE			0 Baseline(OP)	75	60	14	109
7	ABC001	24-Sep-17	6	38	M	WHITE			0 Baseline(OP)	75	60	14	109
8	ABC001	25-Sep-17	7	38	M	WHITE			0 Baseline(OP)	75	60	14	109
9	ABC001	26-Sep-17	8	38	M	WHITE			0 Baseline(OP)	75	60	14	109
10	ABC001	27-Sep-17	9	38	M	WHITE	9/27/2017 8:40		1 Baseline(OP)	75	60	14	109
11	ABC001	28-Sep-17	10	38	M	WHITE			0 Baseline(OP)	75	60	14	109
12	ABC001	29-Sep-17	11	38	M	WHITE			0 Baseline(OP)	75	60	14	109
13	ABC001	30-Sep-17	12	38	M	WHITE			0 Baseline(OP)	75	60	14	109
14	ABC001	1-Oct-17	13	38	M	WHITE			0 Baseline(OP)	75	60	14	109
15	ABC001	2-Oct-17	14	38	M	WHITE			0 Week 2(OP)	75	60	14	109
16	ABC001	3-Oct-17	15	38	M	WHITE	10/3/2017 8:25		1 Week 2(OP)	75	60	14	109
17	ABC001	4-Oct-17	16	38	M	WHITE			0 Week 2(OP)	75	60	14	109
18	ABC001	5-Oct-17	17	38	M	WHITE			0 Week 2(OP)	75	60	14	109
19	ABC001	6-Oct-17	18	38	M	WHITE			0 Week 2(OP)	75	60	14	109
20	ABC001	7-Oct-17	19	38	M	WHITE			0 Week 2(OP)	75	60	14	109
21	ABC001	8-Oct-17	20	38	M	WHITE			0 Week 2(OP)	75	60	14	109
22	ABC001	9-Oct-17	21	38	M	WHITE			0 Week 2(OP)	75	60	14	109
23	ABC001	10-Oct-17	22	38	M	WHITE			0 Week 2(OP)	75	60	14	109
24	ABC001	11-Oct-17	23	38	M	WHITE	10/11/2017 9:05		1 Week 2(OP)	75	60	14	109
25	ABC001	12-Oct-17	24	38	M	WHITE			0 Week 2(OP)	75	60	14	109
26	ABC001	13-Oct-17	25	38	M	WHITE			0 Week 2(OP)	75	60	14	109
27	ABC001	14-Oct-17	26	38	M	WHITE			0 Week 2(OP)	75	60	14	109
28	ABC001	15-Oct-17	27	38	M	WHITE			0 Week 2(OP)	75	60	14	109
29	ABC001	16-Oct-17	28	38	M	WHITE	10/16/2017 9:00		1 Week 4(OP)	75	60	14	109

Continued on the next page:

[illegible][illegible]

These final datasets are sent to the data scientist for use in a modeling software to determine risk factors for certain diseases. Data science is an emerging field for the traditional statistical programmers. It involves mostly data manipulation or data mining for data modelers and forecasters. Data Science is about combining data and using intuition and experimentation to obtain insights on how to conduct smarter trials in the future. Data science methods dealing with large segments of structured and unstructured clinical datasets may offer us prescriptive hints to understand ongoing trends and offer actionable solutions.

