PharmaSUG 2025 - Paper RW-154

Addressing early challenges in RWD data standardization for analysis and reporting of RWE studies

Li Ma, Bo Zheng, and Xingshu Zhu, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

The increasing reliance on Real-World data (RWD) in medical research necessitates efficient methods for data transformation to utilize modular standardized codes. The fragmented non-standard format of RWD makes it inefficient and complicated to generate reproducible Real-World Evidence (RWE). This paper explores a conversion process of RWD into a Basic Data Structure (BDS) format similar to Analysis Data Model (ADaM). It aims to explore a methodology of identifying common data elements in RWE data and mapping them to an ADaM-like BDS specifications format to facilitate use of existing standard code for RWE analysis. These efforts lay the groundwork for reproducible generation and analysis of RWE.

INTRODUCTION

Currently, there is no universally accepted standard for real-world evidence (RWE) data analysis that is specifically equivalent to those developed by the Clinical Data Interchange Standards Consortium (CDISC). However, well-defined frameworks, guidelines, and initiatives, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), have emerged to facilitate the standardization of RWE data collection, analysis, and reporting. The adoption of OMOP-CDM has been steadily increasing globally, with approximately 12% of electronic medical records (EMRs) worldwide converted to this format by 2022, encompassing data from 453 databases that account for more than 928 million unique patient records across 41 countries [1]. While these are noteworthy achievements in the effort to standardize RWE data, one practical limitation of OMOP-CDM is that it is intended for usage with existing large-scale observational data sources, such as electronic health records (EHRs) or administrative claims databases. OMOP-CDM does not always adequately accommodate other less common and smaller-scale examples of real-world data sources, such as those gathered from medical encounters, including patient registries, surveys, questionnaires, and primary data collection studies (PDCS).

Many small-scale real-world data sources, such as proprietary EHRs, claims databases, and data from medical encounters in PDCS, often do not have standardized formats comparable to SDTM or ADaM used in randomized trials. This gap in standardization presents several challenges:

- Integrating and analyzing multi-source data becomes inefficient and error-prone without a consistent structure.
- Identifying and resolving data issues is difficult without predefined validation rules.
- Lack of standards hinders documenting transformations and sharing reproducible analyses.
- Standard clinical trial tools and processes are not easily adaptable.

Standardizing a Real-World Evidence (RWE) analysis dataset at an early stage is crucial in addressing data quality issues, ensuring reproducible results, and avoiding redundant efforts in data cleaning and management. Establishing data standards facilitates quality, reproducibility, transparency, and the efficient utilization of existing analytical resources for RWE. This paper outlines our approach to converting fragmented, non-standard real-world data into an ADaM-like Basic Data Structure format used in our PDCS. This method enhances data consistency, readability, and organization, facilitating easier comparison and integration of data from various real-world sources.

The diagram in Figure 1, below, depicts the flow of the approach in data preparation; the bottom one is the proposed workflow for RWE PDCS project using the randomized controlled trial (RCT) general pathway as a reference:

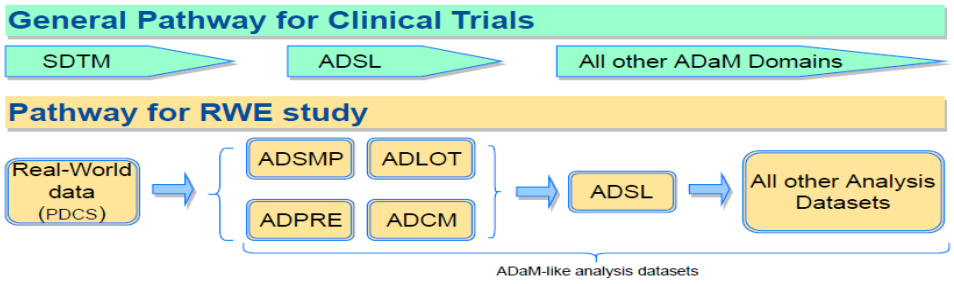

Figure 1

Please note that the analysis datasets derived from real-world data in this paper are prefixed with 'AD' (e.g., ADSMP), showing that they are meant for analysis datasets of nonrandomized and noninterventional studies rather than conforming to any existing ADaM standard analysis datasets.

TYPE OF RAW RWD DATA SOURCES

Figure 2[2] illustrates the most common types of raw Real-World data sources, and there are various solutions available for handling RWD in the market. Most of the data is available in large health databases for research, either in raw form or in a structured format like OMOP, which is standardized but not yet at the submission-ready level of CDISC standards. Numerous publications address the refinement of OMOP-CDM and other structured health databases to ensure alignment with CDISC standards. These established resources should be consulted when performing data management steps to convert formatted RWE data that already has a basic structure into a refined form that meets the stricter requirements of CDISC standards.

Figure 2

This paper focuses specifically on managing raw RWE data for primary data collection studies that are typically gathered using data collection forms and manual steps to aggregate the data. There are also increasing cases where PDCS RWE data is integrated with clinical trial data to conduct a variety of research such as health economics and outcomes research, pharmacovigilance, and post-marketing surveillance studies. Establishing a standard approach for these studies is essential because the data collected for PDCS are often sent in for statistical analysis in various file formats that lack a standard structure, consistent variable names, or uniform data ingestion protocols. Additionally, the raw RWE data received typically has less vigorous data management and may contain inconsistencies in variable naming, mixed column formatting, mismatched headers, and missing or unknown values. Creating a standardized basic data structure helps to address these core data wrangling and processing issues, enhancing the overall reliability and comparability of the data as well as the reproducibility of the statistical analysis results. Finally, if the RWE data does need to be analyzed alongside clinical trial data, it's important to have it in a basic structured format to reliably analyze. This data structure can standardize

formats, variable names, and coding schemes, aiding the integration of RWE with clinical trial data for accurate comparisons and robust analyses.

This is the general flow of the primary data collection process before it arrives in the hands of the analysis programmers:

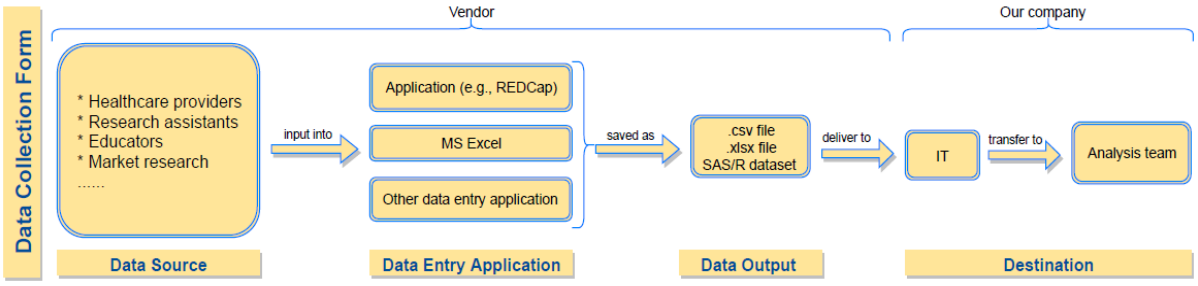

Figure 3

Initially, collected data is manually entered into the REDCap system, MS Excel, or other data entry applications, and then the entered data are exported from REDCap into formats such as .xlsx or other compatible formats before being delivered to analysis team for analysis and reporting.

As shown in the workflow in Figure 1, we bypass the step of converting RWD to SDTM before creating the analysis datasets. In other words, we directly map the raw RWD collected from primary data to an ADaM-like Basic Data Structure (BDS) format. We've implemented this approach across multiple PDCS projects and therapeutic areas and have continued to build standards and tools to enhance its capabilities. For example, we mapped health surveillance system RWE data into an ADaM-like BDS and were able to analyze and create refined summary datasets and results with our existing clinical trial standard reporting macros. Next, the refined RWE dataset and a subset of clinical trial data were used to compare results across studies to detect potential underreporting and other irregularities.

Taking a direct route from RWE to ADaM offers many short-term efficiency advantages since it helps to streamline the workflows for many RWE studies which will not be submitted to the FDA or another regulatory agency and significantly speeds up the analysis dataset development process. The potential downsides include more challenges to ensure data traceability and the lack of development towards a true end to end RWE to CDISC standards conversion process. Despite these challenges, the lessons learned from developing this approach over time also gave us the insight and a starting point for developing additional processes towards the final goal, a true end to end RWE to SDTM to ADaM process to prepare for future RWE regulatory agency submission projects.

MAPPING PROCESS

The primary data we have collected is mostly organized at the subject level, with all relevant information for each subject contained within a single record. This structure simplifies data management and makes it easier to work with the data directly, provided the number of columns is limited. However, if the number of columns exceeds 50, browsing the data becomes more challenging. In non-laboratory sample data files (classified as 'clinical' files which has detailed information about the subject such as demographics, timepoints, treatment history, and other endpoint variables), the number of columns mostly varies between 80 and 200, depending on the project. These clinical files are typically accompanied by supplemental laboratory results that could have detailed disease, biomarker, microbiology, laboratory, and other measures of interest.

In one of our larger primary data collection studies, the clinical file comprises approximately 200 columns (partial column names are provided in the APPENDIX). Each subject record includes multiple pre-treatment therapies, systemic treatments documented across three lines of therapy, progression information spanning three lines, best response details over three lines, and treatment discontinuation information also extending up to three lines, among other details. Imagine the inconvenience and eye

strain caused by navigating this extensive data in SAS or other platforms. The sheer volume of information can make it challenging to locate specific details, leading to potential oversights and dramatically increasing the time required to manage and analyze the data effectively. This complexity not only hampers productivity but may also affect the accuracy of the findings, highlighting the need for more streamlined or normalized data management solutions.

CLINICAL DATA EXAMPLE

Below is a partial snapshot of the analysis dataset ADPRE, which is created from subject-level raw clinical data, transforming all variables from a single row to a vertical display. The data depicted here and on following pages is mock data and not actual data values from any patient.

SUBJID	PARAMN	PARAM	STARTDT	STOPDT	PROGDT	BESTRESP	DCSREAS
AAA	1	Neoadjuvant treatment	04JAN2014	07FEB2013	.	Not assessable/not evaluable/unknown	Toxicity
AAA	2	Adjuvant treatment	05MAY2014	07AUG2014	20JAN2016	Complete response/complete remission	Completed therapy
AAA	3	Radiation 1	.	.	.		
AAA	4	Radiation 2	.	.	.		
AAA	5	Metastasectomy 1	04JAN2016	.	.		
AAA	6	Metastasectomy 2	.	.	.		
AAA	7	Metastasectomy 3	.	.	.		
AAA	8	Primary surgery	07MAR2014	.	.		
BBB	1	Neoadjuvant treatment	07MAR2014	07APR2014	.	Partial response/partial remission	Completed therapy
BBB	2	Adjuvant treatment	05JUL2014	07NOV2014	21DEC2015	Complete response/complete remission	Completed therapy
BBB	3	Radiation 1	07MAR2014	07APR2014	.		
BBB	4	Radiation 2	04JAN2018	07JAN2018	.		
BBB	5	Metastasectomy 1	.	.	.		
BBB	6	Metastasectomy 2	.	.	.		
BBB	7	Metastasectomy 3	.	.	.		
BBB	8	Primary surgery	04JUN2014	.	.		

Figure 4 (Clinical data example)

ADPRE is the analysis dataset that focuses on assessing treatments administered prior to the initiation of any line of therapy:

- In Figure 4, the analysis time points for all measurements (start day, end day, progression date, best response) are presented across seven columns in addition to SUBJID.
- This improves the organization the raw data received into an analysis-ready state versus storing all measurements into a single consolidated long row.

By adopting the PARAM concept from ADaM, we can effectively identify and categorize various pre-treatments measured before systemic therapy, enhancing data analysis and reporting.

Similarly, the line of treatment measurements can be placed into the ADLOT dataset, which reflects the subject's treatment journey and is applicable only after systemic therapy has commenced:

- In Figure 5, the sequence of therapies received is recategorized under the variable PARAM.
- The new PARAM variable organizes the corresponding regimen name, start-and-stop dates, best response, progression date, and discontinuation date for each line of therapy.

SUBJID	PARAMN	PARAM	REGIME	STARTDT	STOPDT	TRTDY	PROGDT	BESTRESP	DCSREAS
AAA	1	Systemic therapy - 1st line	FOLFIRI/EGFR	10JUL2016	16MAY2017	310	16JUL2017	Stable disease	Completed therapy
AAA	2	Systemic therapy - 2nd line	5-FU or capecitabine with VEGF or EGFR	09AUG2017	15FEB2018	190	15APR2018	Stable disease	patient refusal
AAA	3	Systemic therapy - 3rd line	FOLFIRI/EGFR	09JUN2018	16OCT2018	129	16DEC2018	Stable disease	Patient refusal
BBB	1	Systemic therapy - 1st line	FOLFIRI	10DEC2015	16DEC2017	737	16DEC2017	Stable disease	Completed therapy
BBB	2	Systemic therapy - 2nd line		.	.	.	.	.	.
BBB	3	Systemic therapy - 3rd line		.	.	.	.	.	.

Figure 5 (Clinical data example)

LABORATORY DATA EXAMPLE

We have also had laboratory data file paired with the corresponding clinical file that includes biomarker samples and related data. The analysis focuses on HER2 and PD-L1 as biomarkers of interest, aiming to

explore their expression and co-expression within a specific disease context. The laboratory data is structured as a denormalized file, where all variables are presented in a single long row. However, many laboratory-related details stored in this long row format can be condensed in the ADSMP dataset (Figure 6) using only two variables: LBSSEQ and LBRSLT. In the BDS-like format, each individual column in the laboratory data file is transformed into LBSSEQ in the ADSMP dataset. This approach enhances the readability of the data.

SUBJID	LBTESTCD	LBTEST	LBSSEQ	LBSSEQN	LBRSLT	BODYSI	LBDT	SMPSPDT	SDSECTDT
AAA	HER2	Human epidermal growth factor receptor 2	Average Of Green Signals	1	.	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Average Of Red Signals	2	.	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Homogenous Dark Basolateral Pattern	3	ABSENT	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	IHC Assay Specific Comment	4	BLUSH WITH BASAL STAINING PRESENT	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	IHC Failure Code	5		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	IHC Staining Artifacts	6	NONE	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	IHC Staining Pattern	7	FOCAL STAINING	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	ISH Assay Specific Comment	8		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	ISH Failure Code	9		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	ISH Ratio Of Sum Of Signals	10	.	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	ISH Staining Artifacts	11		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	ISH Status	12		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Maximum Of Green Signals	13	.	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Maximum Of Red Signals	14	.	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Overall Status	15	NEGATIVE	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	HER2	Human epidermal growth factor receptor 2	Staining Intensity	16	0	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Approximate Number of Viable Tumor Cells	1	>=100	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Assay Specific Comments	2		Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Combined Positive Score	3	0	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Evaluable	4	EVALUABLE	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Expression Level 01	5	CPS <1	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Expression Level 10	6	CPS <10	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Mononuclear Inflammatory Cells Present	7	YES	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Mononuclear Inflammatory Density Score	8	0	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Negative Control Cell Line	9	PASS	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Negative Control Reagent Acceptable	10	YES	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Negative Control Tissue	11	PASS	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Non Specific Background Acceptabble	12	YES	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Positive Control Cell Line	13	PASS	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Positive Control Tissue	14	PASS	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Stainatt	15	INITIAL	Primary	15FEB2014	05SEP2023	31AUG2023
AAA	PDL1	Programmed death ligand 1	Tumor Proportion Score	16	0	Primary	15FEB2014	05SEP2023	31AUG2023

Figure 6 (Lab data example)

ADSL RWE DATASET EXAMPLE

The ADSL dataset generated for RWE is very similar to the ADaM standard version and includes:

- Basic demographic information, disposition status, treatment details, subgrouping and stratification variables, other relevant information, and population flag.
- Derived flag variables to indicate the completion of each prior treatment (under PARAM in Figure 4) and the completion of each line of therapy (under PARAM in Figure 5).
- Endpoint biomarker information such as the PD-L1 and HER2 results are sourced from the ADSMP dataset.

Consequently, all key subject-level information is consolidated in the ADSL dataset, while detailed information can be found in the ADPRE, ADLOT, ADSMP, and other relevant datasets. While it may be tempting to include all possible details in the ADSL dataset, it should primarily focus on:

- the patient identifiers from the mapped raw data.
- the key derived variables necessary to generate the tables, listings, and figures (TLFs).

Additional AD datasets can be created as needed to house detailed information and provide additional layers of traceability for the mapped or derived variables. Existing ADaM standard variable and dataset naming conventions from the CDISC implementation guide should be respected whenever possible and ensures that the final analysis datasets are as compliant as possible. Adhering to these standards in RWE allows for the use of existing analysis and reporting (A&R) macros and standards from clinical trials programming.

RADIAFL1	RADIAFL2	METSE1FL	METSE2FL	METSE3FL	PRIMSGFL	TRT01	TRT02FL	TRT03F	PDL1C1P	HER2P	PDL1C1PN	PDL1C2P	PDL1C2PN
0	.	1	0	.	1	.	.	.	Positive	Negative	2	Negative	1
1	0	1	0	.	1	.	.	.	Negative	Negative	1	Negative	1
0	.	1	1	1	1	.	.	.	Positive	Positive	2	Negative	1
0	.	1	1	0	1	.	.	.	Negative	Negative	1	Negative	1
1	0	0	.	.	1	1	0	.	Negative	Not Evaluable	1	Negative	1
1	0	1	0	.	1	.	.	.	Negative	Negative	1	Negative	1
0	.	1	1	0	1	1	1	1	Positive	Negative	2	Negative	1
1	0	1	1	0	1	1	1	1	Positive	Negative	2	Negative	1
1	0	1	0	.	0	1	1	0	Positive	Negative	2	Negative	1
0	.	1	0	.	1	.	.	.	Positive	Negative	2	Negative	1
0	.	1	1	0	1	1	1	1	Positive	Negative	2	Negative	1
1	0	1	1	0	0	1	1	0	Negative	Negative	1	Negative	1
0	.	1	1	1	1	1	0	.	Positive	Negative	2	Negative	1
1	0	1	1	1	1	1	0	.	Positive	Negative	2	Positive	2
1	0	1	0	.	1	.	.	.	Negative	Negative	1	Negative	1
0	.	0	.	.	1	1	0	.	Positive	Negative	2	Negative	1
0	.	1	0	.	1	1	1	0	Positive	Negative	2	Negative	1
1	1	1	1	1	1	1	0	.	Positive	Negative	2	Negative	1

Figure 7 (Clinical data example)

The creation of an ADaM-like dataset, such as ADSL for a real-world PDCS project can rely on other ADaM-like datasets, including ADPRE, and ADLOT (Illustrated in the bottom section of Figure 1). In this context, these datasets can be developed independently of the ADSL-like dataset. This contrasts with the creation of analysis datasets in Randomized Controlled Trials (RCTs), where the ADSL dataset is typically generated first.

In both cases, the ADSL dataset serves as a foundational dataset that consolidates key subject-level information essential for subsequent analyses (refer to the upper section of Figure 1):

- The ADSL-like BDS formatted dataset mirrors the previous subject-level dataset by keeping one subject per row.
- And using flag indicators to connect to detailed information in other datasets like ADLB, ADPRE, and ADLOT.

With this flexible approach our RWE reporting macros, which were originally developed for denormalized real-world PDCS data, can also be used with these ADaM-like datasets. This, combined with clinical trial programming standards, creates a comprehensive analysis and reporting package to generate the TLFs deliverable.

CONCLUSION

The increasing reliance on RWE PDCS in medical research underscores the need for efficient data transformation methods to utilize modular standardized codes. This paper presented a comprehensive methodology for converting fragmented, non-standard RWD into an ADaM-like Basic Data Structure (BDS) format:

- By identifying common data elements in RWE data and mapping them to ADaM-like BDS specifications, we can leverage existing standardized code from clinical trials to enhance the RWE analysis and reporting process.
- The development of custom ADaM-like datasets such as ADPRE and ADLOT facilitated the integration of various data sources and types common in primary data collection studies and patient-reported outcomes.

Our approach also addressed several key challenges in managing and analyzing RWD, including inconsistent data structures, variable names, and data ingestion protocols. By leveraging the principles and structures of defined ADaM datasets, we ensure that the refined RWE data is organized, consistent, and easier to work with. Jumping directly from raw RWE data to ADaM-like and bypassing the SDTM mapping portion gave us more flexibility to gain efficiency advantages for study completion while also

creating a framework for future RWE to SDTM to ADaM conversions for any studies that will need to be submitted to a regulatory agency or for cross-study comparison of results and findings.

In conclusion, our proposed standardized approach for transforming raw PDCS data into an ADaM-like basic data structure lays a solid foundation for reproducible and efficient RWE analysis, particularly in cases where popular RWE standards like OMOP-CDM may not be readily or efficiently adaptable due to the nature of the data collected. Moving forward, we aim to continue refining and enhancing this methodology to ultimately develop a comprehensive end-to-end process for converting raw RWE data files into a CDISC-compliant format that includes the SDTM step. This will help to fully harness the potential of RWE research as a valuable supplement to traditional randomized controlled trials.

REFERENCES

1. Seamless EMR data access: Integrated governance, digital health and the OMOP-CDM | BMJ Health & Care Informatics
   https://informatics.bmj.com/content/31/1/e100953

2. Real World Data
   https://www.cdisc.org/standards/real-world-data

ACKNOWLEDGMENTS

The authors would like to thank their management teams as well real-world evidence and epidemiology colleagues at Merck & Co., Inc., Upper Gwynedd, PA, USA, for their advice on this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Li Ma
Merck & Co. Inc.
351 N Sumneytown Pike, North Wales, PA 19454, USA
li.ma2@merck.com

Xingshu Zhu
Merck & Co. Inc.
351 N Sumneytown Pike, North Wales, PA 19454, USA
xingshu_zhu@merck.com

Bo Zheng
Merck & Co. Inc.
351 N Sumneytown Pike, North Wales, PA 19454, USA
bo.zheng1@merck.com

TRADEMARK

PC SAS version 9.4 (TS1M4) and Microsoft 365 MSO were used in this paper.

APPENDIX

Partial names of columns or variables are provided in the example for the PDCS raw clinical file within the paper:

VARIABLE NAME	VARIABLE LABEL
SUBJID	Subject ID
AGEDX	Age at initial diagnosis of colorectal cancer (CRC)
AGEDXINI	Age at initial diagnosis
BESTRESPADJUV	Best response adjuvant
BESTRESPNEOADJUV	Best response neoadjuvant
BESTRESP1L	Best response 1L
BESTRESP2L	Best response 2L
BESTRESP3L	Best response 3L
METSECDT1	Date of Metastasectomy 1 (1st) – metastatic tumor removal procedure
METSECDT2	Date of Metastasectomy 2 (2nd) – metastatic tumor removal procedure
METSECDT3	Date of Metastasectomy 3 (3rd) – metastatic tumor removal procedure
DATEPROG1L	Date of documented progression 1L
DATEPROG2L	Date of documented progression 2L
DATEPROG3L	Date of documented progression 3L
DATEPROGADJUV	Date of progression/recurrence after adjuvant
DATEPROGNEOADJUV	Date of progression/recurrence after neoadjuvant
DATEPROGRADIAT1	Date progression/recurrence radiation 1
DATEPROGRADIAT2	Date progression/recurrence radiation 2
DATEPROGMETSECTMY1	Date progression/recurrence metastasectomy 1
DATEPROGMETSECTMY2	Date progression/recurrence metastasectomy 2
DATEPROGMETSECTMY3	Date progression/recurrence metastasectomy 3
DATEPROGSURG	Date of progression/recurrence after surgery
DATEPROGFREE1L	Date known without progression 1L
DATEPROGFREE2L	Date known without progression 2L
DATEPROGFREE3L	Date known without progression 3L
DCT01RSN	Reason for discontinuing - 1L
DCT02RSN	Reason for discontinuing - 2L
DCT03RSN	Reason for discontinuing - 3L
DCTADJRSN	Reason for discontinuing adjuvant
DCTNADRSN	Reason for discontinuing neoadjuvant
ECOG	ECOG performance status at time of initial diagnosis
ELIGFL	Patient meets all eligibility criteria
LALIVEDT	Last date known alive
OTHDISCONT1L	Other reason discontinuing 1L
OTHDISCONT2L	Other reason discontinuing 2L
OTHDISCONT3L	Other reason discontinuing 3L
OTHDISCONTADJUV	Other reason discontinuing adjuvant
OTHDISCONTNEOADJUV	Other reason discontinuing neoadju
OTHMALIGN	Other malignancies prior to CRC DX (except non-melanoma skin cancers)

PCRRSLT	PCR results
PROG1L	Progression 1L
PROG2L	Progression 2L
PROG3L	Progression 3L
PROGAJFL	Flag - prog/recur after adjuvant treatment
RADIATDOSE1	Radiation dose 1
RADIATDOSE2	Radiation dose 2
RADIOSDT1	Radiation start date 1
RADIOSDT2	Radiation start date 2
RADIOEDT1	Radiation stop date 1
RADIOEDT2	Radiation stop date 2
RADIAFL1	Radiation therapy 1
RADIAFL2	Radiation therapy 2
REGIM1L	Regime 1L treatment
REGIM2L	Regime 2L treatment
REGIM3L	Regime 3L treatment
SMPCOMP	If sample complete
SMPID1	Sample ID 1
SMPID2	Sample ID 2
SIDEDCERN	Sidedness of a cancer site location
TR01SDT	Start date of systemic therapy - 1L
TR02SDT	Start date of systemic therapy - 2L
TR03SDT	Start date of systemic therapy - 3L
ADJUVSDT	start date adjuvant
NADJSDT	start date neoadjuvant
TR01EDT	Last date of systemic therapy (1L)
TR02EDT	Last date of systemic therapy (2L)
TR03EDT	Last date of systemic therapy (3L)
ADJUVEDT	stop date adjuvant
NADJEDT	stop date neoadjuvant
PROGA1DT	Date of progression after ablation 1
PROGA2DT	Date of progression after ablation 2
PROGA3DT	Date of progression after ablation 3
PROGA4DT	Date of progression after ablation 4