Journal of the Society for Clinical Data Management

Breen TE, Neal AU. The Endpoints Dataset: A Quality Control Method for Review and Analysis of Critical Efficacy Endpoints Data. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 1, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.174

DESIGN MANUSCRIPT

The Endpoints Dataset: A Quality Control Method for Review and Analysis of Critical Efficacy Endpoints Data

Timothy E. Breen and Adelai U. Neal

Introduction: Clinical trial research is increasingly utilizing risk-based quality management systems. The identification and management of critical data elements is a crucial component of these systems. A quality control method is required to meet the needs of risk-based quality management systems with respect to critical data elements.
Objective: A quality control method, the endpoints dataset, is described to ensure critical efficacy endpoints data are identified and managed to guarantee fitness for purpose and support good decision making.
Methods: The endpoints dataset consists of four components: demographics, disposition, endpoints, and analysis. The structure of the four components are described in detail. A hypothetical endpoints dataset based on a randomized oncology clinical trial is provided to illustrate the detailed description of the endpoints dataset (supplemental material).
Results: The endpoints dataset compiles all required data to review and analyze the primary and secondary objectives of a clinical trial. Efficacy endpoints data are derived from clinical trial and external data for review and to support analysis. Analysis data is derived to directly support biostatistical analysis. All data is formatted in one record per subject.
Conclusions: The generation of the endpoints dataset requires clinical data management and biostatistical teams to understand and to agree on critical efficacy endpoints and analysis. Biostatistical analysis of the primary and many secondary endpoints can be carried out using only the endpoints dataset. This quality control method can be used for any type of clinical trial. The endpoints dataset is compared to the Analysis Data Model (CDISC ADaM).

Keywords: Data Quality; Quality Control Method; Critical Data Elements; Efficacy Endpoints

Background

Industry and regulatory agencies have emphasized the use of a risk-based quality management system in clinical trial research.[1,2] Fundamental components of a quality system are the identification and management of critical data elements that ensure "the reliability of trial results".[3] Critical efficacy endpoints are a special subset of critical data elements that "are chosen to assess drug effects".[4] These efficacy data endpoints are used to evaluate the primary and secondary trial objectives, and to support good decision making. Efficacy endpoints data are located across the clinical trial database and even in external data sources. For example, an oncology clinical trial with a primary objective of progression-free survival compared between two groups with high and low values of a selected analyte will require endpoints data from multiple clinical trial database tables and external lab data. Ensuring all endpoints data are complete, accurate, valid, and consistent for analysis demands an understanding of critical efficacy endpoints and of the exact format of the data for analysis. A quality control (QC) method is proposed herein to ensure critical efficacy endpoints data are fit for purpose and are appropriate for analysis prior to publication or conversion for regulatory submission. This method is not meant to replace the Clinical Data Interchange Standards Consortium (CDISC) standards, especially the Analysis Data Model (ADaM) submission standard. The example of an oncology trial as described above will be used to illustrate this QC method. This method, however, could be applied to any type of clinical trial.

Methods

The endpoints dataset has four data components: demographics, disposition, endpoints, and analysis. The purpose of the endpoints dataset is threefold. First, all data relevant to analysis of primary and secondary objectives are compiled in one record per study subject. This format for endpoints data makes review more effective and efficient. Second, the metadata and data for derived endpoints are provided to support the outcome for each endpoint. Third, the analysis metadata and

Hoosier Cancer Research Network, Indianapolis, Indiana, US
Corresponding author: Timothy E. Breen (tbreen@hoosiercancer.org)

data are derived from the endpoints data. Any efficacy endpoint and the resulting analysis data can therefore be traced through a single record in the endpoints dataset. The generation of an endpoints dataset will be illustrated using a hypothetical oncology randomized open label clinical trial with a primary objective of progression-free survival (PFS). The secondary objectives will be best overall response and overall survival (OS). The time to event endpoints, PFS and OS, are defined as in the US Food and Drug Administration (FDA) Guidance[5] and will be analyzed using Kaplan-Meier analysis.[6]

The demographics data consist of the clinical site, subject ID, patient demographics data, stratification factors, and arm assignments (supplemental material). Clinical site is included to provide for analysis of data to identify site-specific effects. Demographics data allows for description of the study sample and arms by age, sex, race, ethnicity, etc. In addition, the demographics data can be used to analyze for effects associated with demographic factors. Stratification factors are used to analyze the balance between arms; arm assignments provide for the assessment of balance, and for analysis by arm. In the hypothetical trial example, there are two stratification factors: cancer stage, and histology; and two arms: control (A), and treatment (B). Additional analysis factors could be added to this component, such as a blood chemistry level or tissue level of a selected analyte. All data copied directly from data collected must have the same variable name, variable format, and values as the data collected to ensure traceability.

The disposition data component contains the enrollment date, randomization date, first treatment date, last treatment date, off treatment and off study information, date of death, and last visit date (supplemental material). Randomization date is the start date used to compute time to event for Kaplan-Meier analysis.[5,6] Also, the data from this component are used to validate disposition and endpoints data. Off treatment and off study data include the first and last treatment dates, off study date, and reasons for off treatment and off study. The text data to further explain the reasons for off treatment and off study are also included. The off treatment date, off study date, date of death, and last visit date are also used to determine PFS and OS status and dates as defined by the FDA Guidance on cancer trial endpoints.[5] First and last treatment dates should be determined from actual treatment data and not from reported off treatment dates. Actual first and last treatment dates are more accurate than treatment dates inferred from disposition data. Again, all data copied directly from data collected must have the same variable name, variable format and values as the data collected to ensure traceability.

The endpoints data component of the endpoints dataset contains the dates and outcomes that will determine the efficacy endpoints (supplemental material). PFS and best overall response will be determined and coded by the response evaluation criteria in solid tumors – RECIST v1.1.[7] Objective response for PFS and best overall response are determined by calculating the change in tumor burden at each disease evaluation. The objective response is then coded as a complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD). The endpoints component will contain data for each subject with the date and objective response for PFS and best overall response. The PFS date is usually the date from the most current disease evaluation or date of progression. The best overall response date is the date of the best response from the start of study treatment. The dates and objective responses must follow the definitions in the study protocol. To ensure traceability, variable names for the dates and responses copied from the data collected must have the same variable name, variable format and values as in the data collected. Derived variables, such as the best overall response date, must be described in the metadata and can be maintained across protocols to promote standardization of the endpoints dataset. In the hypothetical trial, the PFS date and response variables are named RCSTDAT and RCSTRESP and the best overall response variable names are BRCRPDAT and BRCRESP. The DAT component of the date variable names conforms to the CDISC Clinical Data Acquisition Standards Harmonization (CDASH) standard for dates.[8]

The final component of the endpoints dataset is the analysis component. This component consists of derived data for the Kaplan-Meier analysis of the efficacy endpoints. The two time-to-event analyses in the hypothetical trial are PFS and OS. PFS is defined as the time from randomization until disease progression or death from any cause. If a subject has not progressed or died, the status of the subject is censored at the last disease evaluation. OS is defined as the time from randomization to death from any cause. If the subject has not died, the status is censored at the last contact with the subject. Censoring is a concept of the time-to-event analysis and is required for the Kaplan-Meier analysis. These definitions should be followed based on the protocol. Each subject will therefore have a PFS date and an OS date in addition to a censoring variable for each analysis. Variable names must be assigned for these dates and censoring data. In the hypothetical trial, the date variable names are PFSDAT, OSDAT and the censoring variables names are PFSSTAT and OSSTAT. These variable names and formats must be described in the metadata. The censoring variables are coded 0 – the event has not occurred, or 1 – the event has occurred. The values and coding definitions must be included in the metadata. Finally, Kaplan-Meier analysis requires the actual time to event, which is calculated in months. The time unit for months used by most biostatisticians is calculated by the formula 365.25/12 which equals 30.4375 days. The time to event for PFS is calculated by the formula (PFSDAT-Randomization Date)/30.4375. Likewise, the OS time to event is calculated by the formula (OSDAT-Randomization Date)/30.4375. If the endpoints dataset is developed as a spreadsheet, these times can be calculated in the spreadsheet. The variable names for the PFS and OS times to event are PFSTIM and OSTIM respectively. The TIM component of these variable names is also compatible with the CDASH standard.[8] The time to event variable names, variable formats, variable coding, and derivation

equations must be described in the metadata. The variable names, formats, coding and derivation equations can be maintained across protocols to promote standardization of the endpoints dataset.

The metadata can be compiled in a document, such as a data dictionary. The data dictionary will contain the source-collected data and any derived data. The format of the data dictionary should include the data table name of the collected data as specified in the clinical trial database as well as the variable format, the name of the code list for the variable, and any data derivation descriptions. The data dictionary will also include, in a separate component, code lists and data derivations that are referenced in the list of data tables. The hypothetical endpoints dataset includes code lists and data derivations as would appear in a data dictionary.

One final column should be added to the endpoints dataset following the time-to-event and censoring variables: a comments column. There are always unusual outcomes for a few subjects. For example, a patient withdraws consent after the first treatment is administered. This subject will have a first treatment date, a last treatment date, off treatment and off study information, but no post-baseline disease evaluations or resulting efficacy endpoints data. A comment that explains the circumstances will provide everyone involved with the endpoints data and the analysis as to why this data is missing.

Finally, the endpoints dataset should be complied and validated by separate individuals. For example, one clinical data manager can compile the endpoints dataset and another clinical data manager can validate it. The independent validation of the endpoints dataset is absolutely crucial as this is the dataset that will be used by the biostatistician for analysis and ultimately be the basis for decision making. The endpoints dataset can be compiled manually, or by programming, or by a combination of programming and manual methods. A spreadsheet provides flexibility to compile and validate the endpoints dataset even if the initial dataset is generated through programming. Once validated, the spreadsheet can easily be converted to either SAS or R-programming dataset formats.

Results

The endpoints dataset described for a hypothetical oncology clinical trial provides all the relevant efficacy endpoints data for review and analysis for each subject in one record. First, the compilation of all efficacy endpoints in one record per subject focuses clinical data management and biostatistical analysis staff on the fitness for purpose of the efficacy endpoints data. Second, the derived efficacy endpoints data for PFS and best overall response are determined for each subject and can be reviewed with respect to disposition and other data. For example, a subject with an off treatment reason of disease progression should have a PD coded for the PFS RECIST response (RCSTRESP) on the appropriate disease evaluation date (RCSTDAT). The endpoints component can also be reviewed with respect to other data, including reported disease response, and survival status. Third, the data required for Kaplan-Meier analysis can be derived from the disposition and endpoints data and is available for analysis. A biostatistician can execute the Kaplan-Meier analyses directly from the PFSTIM, PFSSTAT, OSTIM, and OSSTAT variables. Consistent use of variable names across endpoints datasets can make analysis of trials across the organization very efficient. In addition, the biostatistician can also efficiently review the data for fitness for purpose.

The most important result of the generation of an endpoints dataset, as illustrated with the hypothetical trial, is the integration of all information related to efficacy evaluation. This integration of information includes metadata. The experimental design, the protocol objectives, the definition of endpoints, the appropriate data elements, and the derivation of analysis data must be understood to generate the endpoints dataset. The clinical data management and biostatistical teams must have a unified understanding of the origins and derivation of the critical efficacy endpoints data.

The endpoints dataset method has been used in multiple published clinical trials.[9,10,11]

Discussion

Analysis datasets have been used for regulatory submissions for many years.[12] The current standard for regulatory data submission to the FDA is the CDISC Analysis Data Model (ADaM).[12] This standard meets the needs of the FDA and of industry for regulatory submissions for approval. The ADaM standard is primarily used by the clinical trial statistical staff and FDA regulatory staff at the time of submission. The endpoints dataset is proposed for use by clinical data management staff and clinical trial statistical staff to ensure critical efficacy endpoints data are fit for purpose and support good decision making. While the ADaM standard is a detailed and well designed standard for submission, the endpoints dataset is offered as a QC method for collection, management, and biostatistical analysis for publication or preparation of ADaM datasets. The design of the endpoints dataset can form the foundation for the creation of ADaM datasets. The endpoints dataset has many features in common with the ADaM subject-level analysis dataset (ADSL) and the basic data structure (BDS). The ADSL has one record per subject, contains subject-level population flags, and treatment, demographics, randomization, subgrouping, and important timing variables. The BDS contains the data related to the statistical analyses. The endpoints dataset combines these two ADaM data structures into one dataset for clinical data management and statistical analysis.

Conclusions

The endpoints dataset provides a QC method to compile all required efficacy data and derived efficacy endpoints data. The process of generating an endpoints dataset demands a total understanding of the clinical trial design, and agreement by clinical data management, and biostatistics. In addition, data that supports other analyses, such as demographics, stratification factors, and correlatives are included for review and analysis. This QC method serves multiple purposes and increases the reliability of efficacy

results for decision making. While the endpoints dataset generation was illustrated with an oncology clinical trial, any type of clinical trial could apply this QC method.

Additional File

The additional file for this article can be found as follows:

- **Supplemental Material.** Hypothetical Endpoints Dataset. DOI: https://doi.org/10.47912/jscdm.174.s1

Competing Interests

The authors have no competing interests to declare.

References

1. **The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2), Section 5.0. Published 9 Nov 2016, accessed 11 Jan 2021. https://www.ich.org/page/efficacy-guidelines.
2. **European Medicines Agency (EMA).** Reflection Paper on Risk Based Quality Management in Clinical Trials. Published 18 Nov 2013, accessed 21 Jan 2021. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-risk-based-quality-management-clinical-trials_en.pdf.
3. **The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2), Section 5.0.1. Published 9 Nov 2016, accessed 11 Jan 2021. https://www.ich.org/page/efficacy-guidelines.
4. **The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** General Considerations for Clinical Studies E8(R1) Draft version, Section 5.1.4. Published 8 May 2019, accessed 11 Jan 2021. https://www.ich.org/page/efficacy-guidelines.
5. U.S. Food and Drug Administration Guidance Documents. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics Guidance for Industry. Published December 2018, accessed 21 Jan 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics.
6. **Kaplan EL, Meier P.** Nonparametric Estimation from Incomplete Observations. *J AM STAT ASSOC*. 1958; 53(282): 457–481. DOI: https://doi.org/10.1080/01621459.1958.10501452
7. **Eisenhauer EA, Therasse P, Bogaerts J,** et al. New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guidance (version 1.1). *Eur J Cancer*. 2009; 45: 228–247. DOI: https://doi.org/10.1016/j.ejca.2008.10.026
8. **Clinical Data Interchange Standards Consortium (CDISC).** Clinical Data Acquisition Standards Harmonization Implementation Guide for Human Clinical Trials Version 2.1 (Final). Published 1 Nov 2019, accessed 21 Jan 2021. https://www.cdisc.org/standards/foundational/cdash.
9. **Radovich M, Jiang G, Bradley H,** et al. Association of Circulating Tumor DNA and Circulating Tumor Cells after Neoadjuvant Chemotherapy with Disease Recurrence in Patients with Triple-Negative Breast Cancer PrePlanned Secondary Analysis of the BRE12–158 Randomized Clinical Trial. *JAMA Oncol*. 2020; 6(9): 1410–1415. DOI: https://doi.org/10.1001/jamaoncol.2020.2295
10. **Durm GA, Jabbour SK, Althouse SK,** et al. A Phase 2 Trial of Consolidation Pembrolizumab following Concurrent Chemoradiation for Patients with Unresectable Stage III Non-Small Cell Lung Cancer: Hoosier Cancer Research Network LUN 14–179. *Cancer*. 2020; 126(19): 4353–4361. DOI: https://doi.org/10.1002/cncr.33083
11. **Dudek AZ, Liu LC, Gupta S,** et al. Phase Ib/II Clinical Trial of Pembrolizumab with Bevacizumab for Metastatic Renal Cell Carcinoma: BTCRC-GU14-003. *J. Clin. Oncol*. 2020; 38: 1138–1145. DOI: https://doi.org/10.1200/JCO.19.02394
12. **Clinical Data Interchange Standards Consortium (CDISC).** Analysis Data Model (ADaM) Version 2.1 (Final). Published 17 Dec 2009, accessed 22 Apr 2022. https://www.cdisc.org/standards/foundational/adam.

How to cite this article: Breen TE, Neal AU. The Endpoints Dataset: A Quality Control Method for Review and Analysis of Critical Efficacy Endpoints Data. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 1, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.174

Submitted: 29 March 2022 **Accepted:** 05 May 2023 **Published:** 29 August 2023

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS