PharmaSUG China 2026 - Paper SA-168

A Traceable and Reproducible Framework for Oncology Efficacy Data Review Using IMWG Criteria: A BCMA Case Study

Yun Jiang, Yingying Deng, Manman Lu, and Longhui Li, Pfizer Inc.

ABSTRACT

Oncology efficacy review often requires complex algorithmic assessment across multiple clinical data sources. In BCMA studies, efficacy review based on International Myeloma Working Group criteria requires integration of laboratory, bone marrow, imaging, and clinical assessment data. Because response outcomes are derived through multiple rule-based steps, manual review can be time-consuming and difficult to standardize.

This paper presents a traceable and reproducible framework for IMWG-based efficacy data review. The framework was developed in R within Positron and supports SDTM-to-ADaM derivation through modular, parameter-level steps. Each step generates intermediate outputs, issue reports, and links to the final response result. A dashboard supports study-level monitoring, subject-level review, parameter-level traceability, and issue investigation, while an AI assistant helps users understand algorithm logic, derivation decisions, and response assessment evidence.

The framework converts complex response assessment into a transparent and reproducible workflow. Parameter-level outputs allow programmers to trace how individual assessments contribute to endpoint results, while automated issue reports and interactive review tools improve consistency, efficiency, and interpretability. This approach provides a practical model for applying automation and AI-assisted review to clinical analyses with complex endpoint derivations.

INTRODUCTION

B-cell maturation antigen (BCMA) has become an important therapeutic target in multiple myeloma and is the focus of many recent clinical studies [1]. Demonstrating treatment efficacy is a key objective of these studies and relies on accurate assessment of patient response.

Response assessment in multiple myeloma is commonly based on International Myeloma Working Group (IMWG) criteria [2,3]. IMWG response categories are derived from multiple assessment components, including serum protein electrophoresis, urine protein electrophoresis, serum free light chain measurements, bone marrow evaluations, and imaging assessments. These components must be evaluated together to determine a patient's overall response status.

Unlike directly collected endpoints, IMWG efficacy results are generated through a series of derivation rules across multiple data sources and time points. Understanding how individual assessments contribute to a final response outcome can therefore be challenging. Reviewers often need to trace derived results back to supporting evidence and investigate discrepancies across assessment components.

To improve transparency and review efficiency, there is a need for approaches that provide clear links between source data, intermediate assessments, and final efficacy endpoints. This paper presents a framework designed to support traceable and reproducible IMWG-based efficacy data review using BCMA studies as a case study.

METHODOLOGY

IMWG RESPONSE ASSESSMENT HIERARCHY

Figure 1 illustrates the hierarchy of assessment components used in International Myeloma Working Group (IMWG) response determination [3]. Response categories are derived through the integration of multiple assessment components, including serum protein electrophoresis, urine protein electrophoresis, serum free light chain measurements, bone marrow evaluations, imaging assessments, and clinical observations. These components contribute to intermediate assessments, which are subsequently used to derive overall response outcomes.

Because response outcomes depend on multiple assessment components and their interrelationships, review requires visibility into both intermediate assessments and the supporting source data.

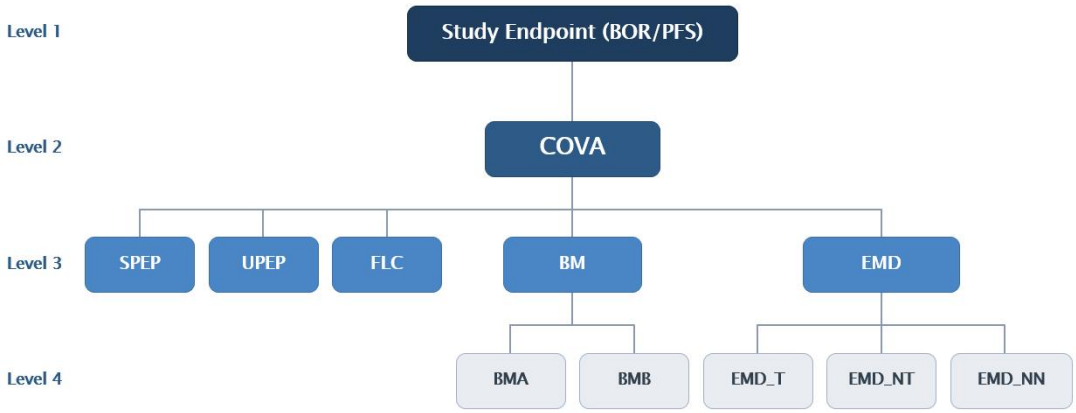


Figure 1. Hierarchy of IMWG Response Assessment Components

Abbreviations: *BOR = best overall response; PFS = progression-free survival; COVA = Clinical Overall Response Assessment; SPEP = serum protein electrophoresis; UPEP = urine protein electrophoresis; FLC = serum free light chain; BM = bone marrow; BMA = bone marrow aspirate; BMB = bone marrow biopsy; EMD = extramedullary disease; EMD_T = target extramedullary disease; EMD_NT = non-target extramedullary disease; EMD_NN = non-target non-extramedullary disease or new lesion.*

FRAMEWORK DESIGN

Figure 2 illustrates how derivation outputs, review artifacts, dashboard visualizations, and AI-assisted review are generated from a common traceable workflow.

Source data are transformed into SDTM and ADaM datasets, where component responses and endpoint results are derived for downstream review. In parallel, the framework generates review materials, including summary outputs, SDTM-to-ADaM listings, and automated issue reports. These outputs provide a common foundation for dashboard-based review and AI-assisted investigation.

The framework was developed in R within the Positron environment [4,5] and supports derivation from Study Data Tabulation Model (SDTM) to Analysis Data Model (ADaM) datasets [6,7]. Synthetic data were used for all examples and demonstrations presented in this paper.

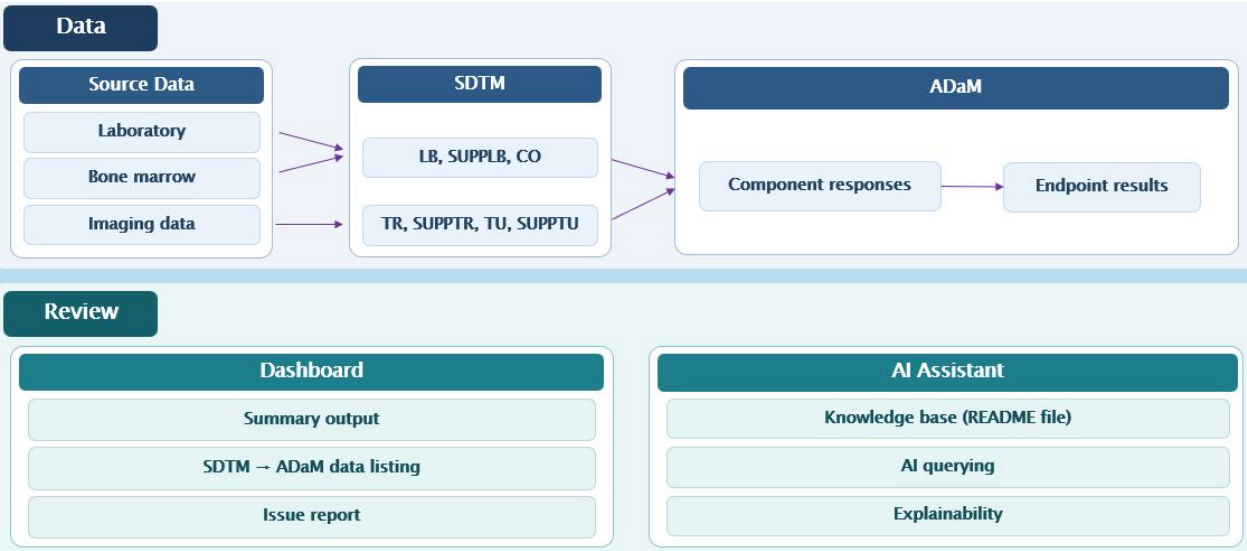


Figure 2. Architecture of the IMWG Efficacy Review Framework

DASHBOARD-BASED REVIEW

An interactive dashboard was developed as the primary review interface for framework-generated outputs. The dashboard was designed to provide organized access to study-level summaries, subject-level response histories, component-level assessments, and automated review findings. Representative dashboard capabilities are described in the Results section.

AI-ASSISTED REVIEW

An AI assistant with chatbot-style interaction was integrated into the framework to support interpretation and validation of IMWG response derivations. The AI assistant is supported by a structured knowledge base containing assessment definitions, derivation algorithms, input variables, and response determination rules.

Users can submit natural language queries to understand derivation logic, investigate supporting evidence, and perform rule-based review of calculated results. Representative examples of AI-assisted explanation and QC validation are presented in the Results section.

RESULTS

DASHBOARD REVIEW CAPABILITIES

The dashboard serves as the primary review interface for framework-generated outputs and consolidates efficacy results, assessment data, and review findings within a single environment. The dashboard is organized into four modules: Study Overview, Subject Explorer, Parameter Review, and QC & Issue Center.

- The **Study Overview** module summarizes key study metrics, including analysis population counts, best overall response distributions, and progression-free survival event summaries. These metrics provide reviewers with an overall assessment of endpoint results and help identify areas requiring further investigation.
- The **Subject Explorer** module supports subject-level review by presenting response histories and clinical overall response assessment timelines. This view allows reviewers to examine how longitudinal assessments contribute to endpoint derivations for individual subjects.
- The **Parameter Review** module provides standardized outputs for individual IMWG assessment components. Reviewers can examine component-level results and supporting evidence used in response determination, enabling investigation of unexpected classifications and derivation outcomes.
- The **QC & Issue Center** module presents automated review findings generated during derivation and validation. Identified discrepancies can be traced to the corresponding subjects, visits, and assessment components, supporting focused investigation of potential data or derivation issues.

Together, these modules support review activities ranging from study-level endpoint evaluation to detailed investigation of subject-level assessments and derivation-related issues.

AI-ASSISTED INTERPRETATION AND VALIDATION

The framework incorporates an AI assistant supported by a structured knowledge base maintained as markdown README files. Each README serves as a documented definition of a response algorithm and includes the algorithm objective, required input variables, baseline definition, and rule hierarchy used during derivation. This standardized documentation enables the AI assistant to retrieve study-specific algorithm logic and use it as the basis for interpretation and validation of derived results.

To support consistent documentation across assessment components, all README files follow a common structure. Key sections include the response algorithm title, overview, input variables, baseline definition, and ordered response rules. This design helps guide the AI assistant’s responses using documented derivation logic rather than relying only on general-purpose AI reasoning.

Table 1. Representative Structure of a Response Algorithm README

README Section	Purpose
Response Algorithm Title	Identifies the response algorithm or component documented by the file.
Overview	Summarizes the objective of the algorithm and its role in response derivation.
Input Variables	Defines the variables required to evaluate the response rules.
Baseline Definition	Documents how baseline values are identified and used in response assessment.
Rule Hierarchy	Specifies the ordered decision logic applied during derivation (e.g., PD, CR, VGPR, PR, MR).

Abbreviations used in Figures 3 and 4: PD = progressive disease; CR = complete response; VGPR = very good partial response; PR = partial response; MR = minimal response. SMPRESP = derived SPEP response.

Figure 3 demonstrates an AI-assisted interpretation use case. The user submits the prompt: "Use the spep_response.md as the algorithm definition. Why is subject DMY001 Visit 2 MR?" The AI assistant retrieves the spep_response.md documentation, identifies the relevant assessment variables for the specified subject and visit, and evaluates the response according to the documented rule hierarchy. The response explains how the observed parameter values satisfy the criteria for a Minimal Response classification and provides the supporting evidence used in the decision process. This capability helps reviewers understand not only the assigned response category, but also the derivation pathway that led to the result.

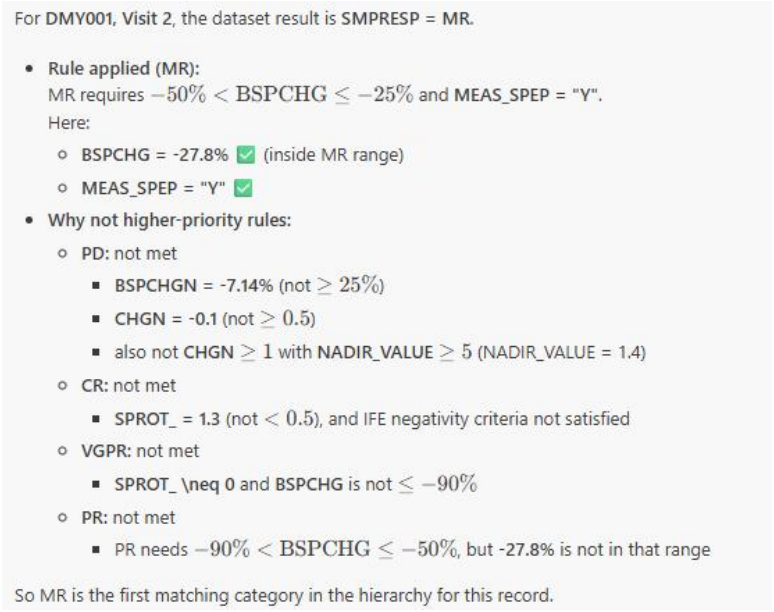
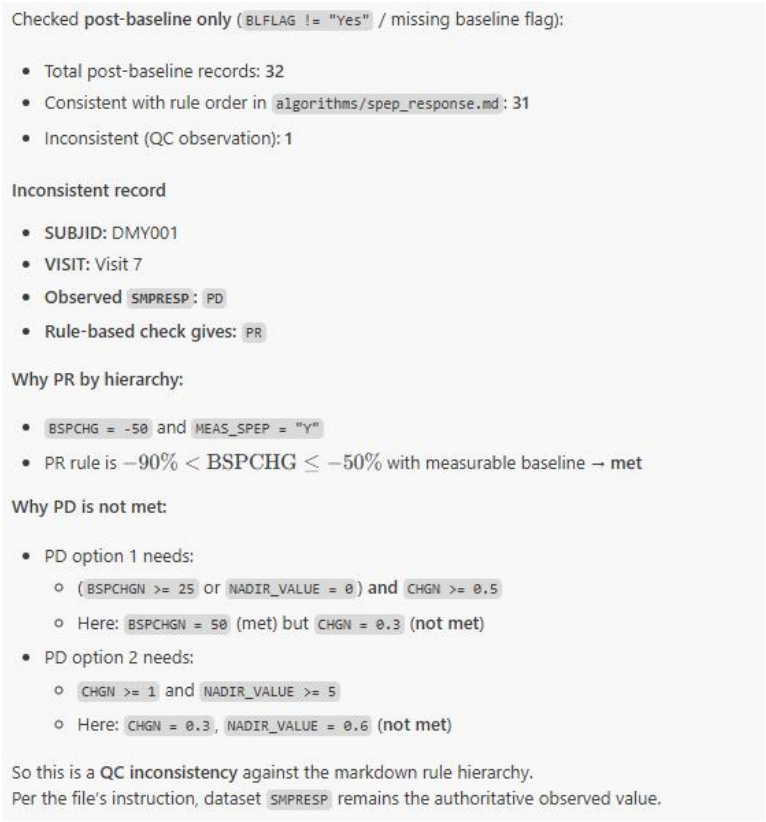


Figure 3. AI-Assisted Response Explanation Example

Figure 4 demonstrates an AI-assisted validation use case. The user submits the prompt: "Based on spep_response.md, please recalculate the SMPRESP results and check any inconsistent values and explain them. Note: only check post baseline record." Using the documented algorithm as the reference specification, the AI assistant recalculates the expected response results and compares them with the reported SMPRESP values. When inconsistencies are identified, the AI assistant highlights the affected

records and explains the discrepancy according to the documented response rules. This capability provides a documentation-driven review mechanism that can assist reviewers in identifying potential derivation or data-related issues for further investigation.



Checked **post-baseline only** (`BLFLAG != "Yes"` / missing baseline flag):

- Total post-baseline records: **32**
- Consistent with rule order in `algorithms/spep_response.md`: **31**
- Inconsistent (QC observation): **1**

Inconsistent record

- **SUBJID:** DMY001
- **VISIT:** Visit 7
- **Observed** `SMPRESP`: `PD`
- **Rule-based check gives:** `PR`

Why PR by hierarchy:

- `BSPCHG = -50` and `MEAS_SPEP = "Y"`
- PR rule is $-90\% < \text{BSPCHG} \leq -50\%$ with measurable baseline → **met**

Why PD is not met:

- PD option 1 needs:
  - (`BSPCHGN >= 25` or `NADIR_VALUE = 0`) and `CHGN >= 0.5`
  - Here: `BSPCHGN = 50` (met) but `CHGN = 0.3` (**not met**)
- PD option 2 needs:
  - `CHGN >= 1` and `NADIR_VALUE >= 5`
  - Here: `CHGN = 0.3`, `NADIR_VALUE = 0.6` (**not met**)

So this is a **QC inconsistency** against the markdown rule hierarchy.
Per the file's instruction, dataset `SMPRESP` remains the authoritative observed value.

Figure 4. AI-Assisted QC Validation Example

Together, these examples demonstrate how structured README documentation can support both response interpretation and rule-based validation. By combining documented algorithms with framework-generated outputs, the AI assistant provides a consistent and traceable approach for reviewing IMWG response derivations.

CONCLUSION

IMWG-based response assessment requires the integration of multiple assessment components and derivation rules, making efficacy review challenging in BCMA studies [1,3]. Although final endpoint results are essential for analysis, review activities often require understanding how individual assessments contribute to response determinations and investigating discrepancies across multiple data sources.

This paper presented an IMWG efficacy review framework that combines component-level derivation outputs, automated issue reporting, dashboard-based review, and AI-assisted interpretation within a unified workflow. By preserving intermediate assessment results and their relationships to derived endpoints, the framework enables reviewers to examine response evidence, investigate derivation pathways, and perform targeted review of potential data or programming issues.

The dashboard provides a centralized environment for study-level, subject-level, and parameter-level review, while the AI assistant leverages structured README-based algorithm documentation to support response explanation and rule-based validation. Together, these capabilities improve accessibility of complex derivation logic and facilitate more efficient review of IMWG response assessments.

Although demonstrated using BCMA studies and IMWG criteria, the underlying approach can be adapted to other clinical analyses that involve complex endpoint derivations, multiple assessment components,

and traceability requirements. By combining standardized derivation outputs, structured review workflows, and documentation-driven AI assistance, the proposed framework provides a practical model for improving transparency, reproducibility, and efficiency in clinical efficacy data review.

REFERENCES

[1] Tai, Y.-T., and K. C. Anderson. 2019. B cell maturation antigen (BCMA)-based immunotherapy for multiple myeloma. Expert Opinion on Biological Therapy 19(11): 1143-1156. doi:10.1080/14712598.2019.1641196.

[2] Rajkumar, S. V., M. A. Dimopoulos, A. Palumbo, J. Bladé, G. Merlini, M.-V. Mateos, S. Kumar, J. Hillengass, E. Kastritis, P. Richardson, O. Landgren, B. Paiva, A. Dispenzieri, B. Weiss, X. LeLeu, S. Zweegman, S. Lonial, L. Rosinol, E. Zamagni, S. Jagannath, O. Sezer, S. Y. Kristinsson, J. Caers, S. Z. Usmani, J. J. Lahuerta, H. E. Johnsen, M. Beksac, M. Cavo, H. Goldschmidt, E. Terpos, R. A. Kyle, K. C. Anderson, B. G. M. Durie, and J. F. San Miguel. 2014. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. The Lancet Oncology 15(12): e538-e548. doi:10.1016/S1470-2045(14)70442-5.

[3] Kumar, S., B. Paiva, K. C. Anderson, B. Durie, O. Landgren, P. Moreau, N. Munshi, S. Lonial, J. Bladé, M.-V. Mateos, M. Dimopoulos, E. Kastritis, M. Boccadoro, R. Orlowski, H. Goldschmidt, A. Spencer, J. Hou, W. J. Chng, S. Z. Usmani, E. Zamagni, K. Shimizu, S. Jagannath, H. E. Johnsen, E. Terpos, A. Reiman, R. A. Kyle, P. Sonneveld, P. G. Richardson, P. McCarthy, H. Ludwig, W. Chen, M. Cavo, J.-L. Harousseau, S. Lentzsch, J. Hillengass, A. Palumbo, A. Orfao, S. V. Rajkumar, J. San Miguel, and H. Avet-Loiseau. 2016. International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. The Lancet Oncology 17(8): e328-e346. doi:10.1016/S1470-2045(16)30206-6.

[4] R Core Team. 2025. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

[5] Posit Software, PBC. 2026. Positron. Available at: https://positron.posit.co/. Accessed July 26, 2026.

[6] Clinical Data Interchange Standards Consortium (CDISC). 2021. Study Data Tabulation Model Implementation Guide Version 3.4. Austin, TX: CDISC. Available at: https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-4.

[7] Clinical Data Interchange Standards Consortium (CDISC). 2021. Analysis Data Model Implementation Guide Version 1.3. Austin, TX: CDISC. Available at: https://www.cdisc.org/standards/foundational/adam/adamig-v1-3.

CONTACT INFORMATION

Comments and questions are welcome. Contact the author at:

Name: Yun Jiang
Enterprise: Pfizer (China) Research and Development Co.,Ltd.
Email: yun.jiang2@pfizer.com