Development and Implementation of a Dynamic Framework for Assessment and Improvement of Registries' Data Quality

RWD60

Mai Duong, PhD,[1] Nahila Justo, PhD, MBA, MPhil,[2,3] Noami Berfeld, MSc,[1] Charlotte Pettersson, MSc,[2] Hai Nguyen, PhD,[1] Kristina Kastreva, PhD,[4] Vitaliy Matyushenko, MSc,[5] Mgr. Lenka Mokrá,[6] Farjana Ali, BSc,[7] Neil Bennett, BSc,[7] Emma Watson, BSc,[7] Dino Masic, PhD,[7] Annie Poll, PhD[7]

[1]Thermo Fisher Scientific, London, UK; [2]Thermo Fisher Scientific, Stockholm, Sweden; [3]Karolinska Institute, Stockholm, Sweden; [4]University Hospital "Alexandrovska", Medical University Sofia, Sofia, Bulgaria; [5]Ukrainian SMA Registry, Kharkiv, Ukraine; [6]Institut biostatistiky a analýz, s.r.o.Brno, Czech Republic; [7]TREAT-NMD Services Ltd, Newcastle Upon Tyne, UK

Background

- According to the EMA Guideline on registry-based studies[1], quality management should be continuous and regularly assessed throughout the registry's lifetime. Uncertainties in data quality can undermine confidence in the evidence generated. The EMA data quality framework[1] provides guidelines for ensuring data quality for regulatory decision-making, applicable to any data source. A current long-term study using multiple SMA registries offers an ideal opportunity to adopt EMA guidance, which can be applied to registries in other therapeutic areas.

Objectives

- **To develop a dynamic framework for assessing and improving registry data reliability, extensiveness, and coherence and to implement it across six TREAT-NMD registries (BNMDR [Belgium], Neuromuscular Disease Registries in Bulgaria, Georgia, and Latvia, ReaDy [Czechia] and Ukrainian SMA Registry) to support a study in SMA.**

Methods

- A data quality framework and improvement process (DQF&IP) was developed based on EMA guidance and Findable, Accessible, Interoperable, and Reusable (FAIR[2,3]) principles, and implemented across registries taking part in an ongoing study. This framework includes different data-quality dimensions and subdimensions, as well as their characterization and related metrics to make reliable assessments.
- The DQF&IP (**Figure 1**) is composed of metadata (variable definitions, coding, formats, dimensions, quality metrics) and processes, which include periodic quality assessment, issue investigation, root-cause analysis, and improvement planning, which is implemented iteratively in a continuous feedback loop.

Figure 1. Data quality framework and improvement process

Metadata

Check item	Variable	Variable Definitions	Coding/ Categorization	Formats	Quality Dimension	Quality Sub-dimension	Quality Metrics
1	Variable 1				Extensiveness		
2	Variable 1				Coherence		
3	Variable 1				Reliability		

Processes

Periodic quality assessment	Issue investigation via source data verification	Issue resolutions	Root cause investigation	Improvement plan
• By member registries • By TREAT-NMD • By Thermo Fisher Scientific	• By TREAT-NMD • By member registries	• Correct errors • Collect missing data	• e.g., lack of stop rules for implausible values, mapping/ transformation errors	e.g., adding stopping rules Fix data mapping/ transformation issues

Feedback loop processes

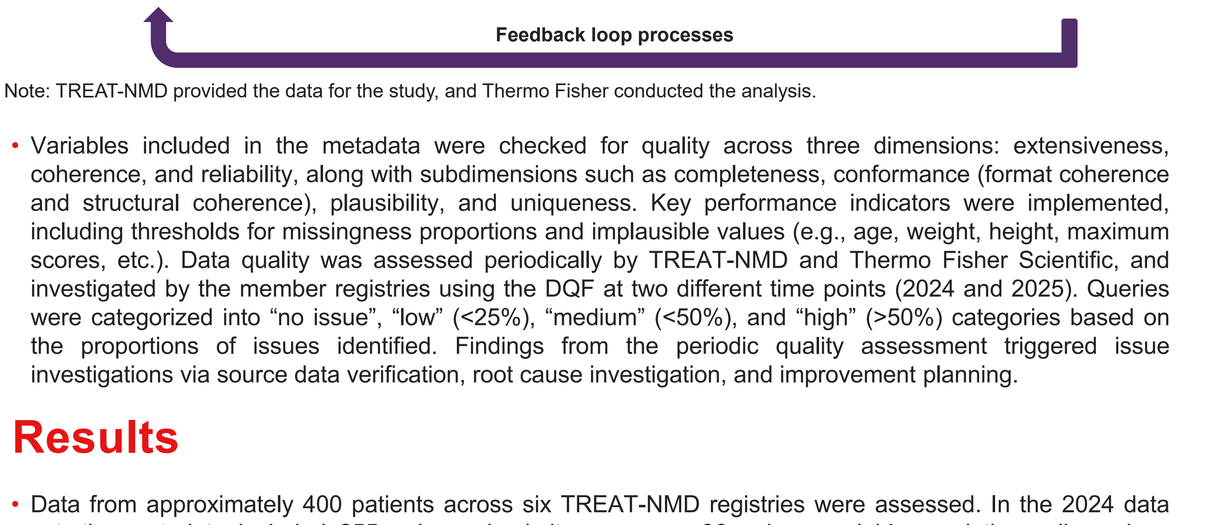
Note: TREAT-NMD provided the data for the study, and Thermo Fisher conducted the analysis.

- Variables included in the metadata were checked for quality across three dimensions: extensiveness, coherence, and reliability, along with subdimensions such as completeness, conformance (format coherence and structural coherence), plausibility, and uniqueness. Key performance indicators were implemented, including thresholds for missingness proportions and implausible values (e.g., age, weight, height, maximum scores, etc.). Data quality was assessed periodically by TREAT-NMD and Thermo Fisher Scientific, and investigated by the member registries using the DQF at two different time points (2024 and 2025). Queries were categorized into "no issue", "low" (<25%), "medium" (<50%), and "high" (>50%) categories based on the proportions of issues identified. Findings from the periodic quality assessment triggered issue investigations via source data verification, root cause investigation, and improvement planning.

Results

- Data from approximately 400 patients across six TREAT-NMD registries were assessed. In the 2024 data cut, the metadata included 255 unique check items across 96 unique variables and three dimensions (coherence: 36.1%, extensiveness: 43.5%, reliability: 20.4%). In 2025, the list of unique check items was expanded, the metadata increased to 476 check items across 153 unique variables and three dimensions (coherence: 23.1%, extensiveness: 33.8%, reliability: 43.1%).
- Across both 2024 and 2025, most cases fell into the "no issue" category (**Figure 2**). When errors occurred, they were mostly categorized as "low", with few cases classified as "medium" or "high". In the reliability dimension, most cases reported no issues. The extensiveness dimension showed relatively more spread across categories. While "no issue" remained common, there were noticeable "low" and "medium" cases, and even some "high" cases, specifically in 2025. For coherence, most cases were recorded as "no issue," with very few "low", "medium", or "high" cases. High frequencies were mostly related to missingness.

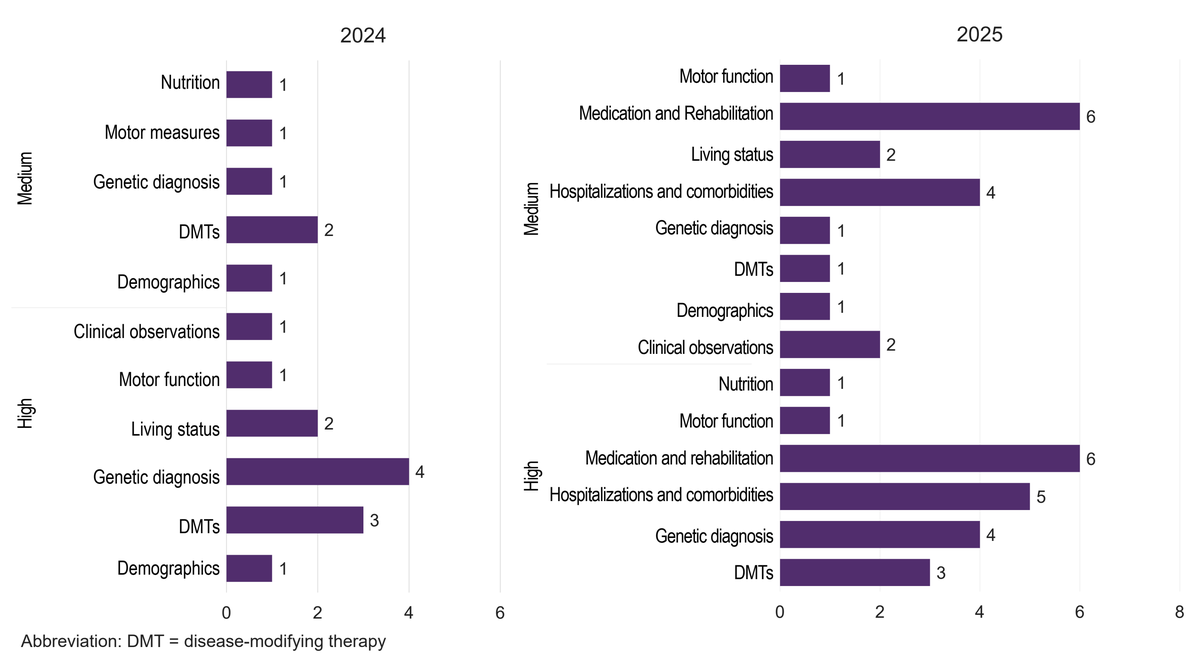
Figure 2. Distribution and number of unique data checks by dimensions and issue categorization



Results (cont.)

- In the TREAT-NMD case report form (CRF), items are organized into modules. Among the "medium" category in 2024, the most common queries related to the DMTs module, while the "high" category had the highest frequency of queries in the genetic diagnosis module **(Figure 3).** In 2025, medication and rehabilitation had the highest frequency in both the "medium" and "high" categories.

Figure 3. Distribution of data quality issues by group of variables



Abbreviation: DMT = disease-modifying therapy

- Among the queries identified over the last few years, **Table 1** below provides insight into the top key issues, their root causes, and the improvement plans that have been discussed or are currently being discussed.

Table 1. Key issues, root causes, and improvement plans for identified queries

Key Issue	Root Cause	Improvement Plan
The variable "motor ability observed in clinic" appeared missing for a three registries, raising concern about outcome validity	Two registries only report results for those observed in clinic; one indicated that all information from a certain period should be marked as being observed in clinic	Ongoing discussions with the registries with the aim to update the dataset to ensure completeness and avoid future flagging
Conflicting information between "status" variables (e.g., current ventilation) and date/episode variables (start and end date)	Delay in updating status variable despite start/ end dates being recorded	Where possible, registries will update the 'status' variable at the same time when recording dates
Absence of timely updates for variables that require regular updating	Ongoing status (e.g., symptomatic status) and dates not consistently recorded unless there is a change; registries typically update only when changes occur	Where possible, registries to record information even if the status hasn't changed

Limitations

- The missingness examined as part of the quality-control process was based on the absence of expected information indicated by other variables. It was not possible to determine if the entire entry was missing (e.g., a treatment was given, however, neither the treatment name nor date was recorded).
- Additionally, while beneficial, the registry may retrospectively update data previously entered, including between data cuts. Consequently, data-quality issues that were not present in one data cut may be newly introduced and identified in subsequent data cuts.
- Most registries map their data to the TREAT-NMD conceptual disease model. While developing the DQF&IP, it was observed that this mapping process removed the dependence between variables, making it challenging to quantify true missingness.

Conclusions

- **The prospective data collection in clinical registries allows for adaptability and quality improvement throughout the data lifecycle. Our study's DQF&IP highlights the importance of having a clear and dynamic framework to continuously assess and enhance the quality of registry data. This supports research sustainability and benefits the patient community. Additionally, this flexible framework can be adapted to registry data for other indications.**

References

1. EMA. Guideline on registry-based studies 2021. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf
2. EMA. Data Quality Framework for EU medicines regulation 2023. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf
3. EMA. Good Practice Guide for the use of the HMA-EMA: Catalogues of real-world data sources and studies. 2025. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-hma-ema-catalogues-real-world-data-sources-studies_en.pdf

Disclosures

MD, **NJ**, **NB**, **CP**, and **HN** are employees of PPD™ Evidera™ Real-World Data & Scientific Solutions, Thermo Fisher Scientific. The work presented in this poster is part of a larger study sponsored by F. Hoffmann-La Roche AG. However, the opinions expressed in this poster do not necessarily represent the views of the study Sponsor.

Acknowledgments

Editorial and graphic design support was provided by Caroline Cole and Richard Leason of Thermo Fisher Scientific.