Enhancing Biomedical Data FAIRification by Streamlining Data Harmonization with a Terminology Management Solution

Rancho Biosciences, LLC, PO Box 7208, Rancho Santa Fe, CA 92067

Alena Fedarovich, Andrey Kalinin, Aish Pathak, Vishnu Govindaraj, Leonya Ivanov, Candace Ruff

Background

Effective data management is the cornerstone of reliable data insights. Streamlining critical tasks such as data integration, harmonization, and standardization are key to ensuring your data is analysis ready and downstream results are rooted in a solid data foundation. Originally developed as an internal tool to automate client terminology mapping projects, our Terminology Management Solution (TMS) has evolved over the past decade into a robust automated solution, shaped by scientific expertise, practical use, and real-world feedback in diverse biomedical contexts.

TMS effectively integrates terminology from different domains into cohesive datasets significantly reducing errors and the time required for curation tasks. TMS can reduce manual curation time by 50% or more depending on selected ontology and complexity of raw data. Several use cases, including the FAIRification of real-world data (RWD), highlight TMS' significant value in biomedical data curation.

Recent TMS tool enhancements mark a significant leap towards improving accuracy and efficiency of data harmonization used for drug discovery and development, understanding of disease mechanisms, and other life sciences applications.

A Decade of Innovation

2015-2018

The Foundation

- Implemented fuzzy algorithm.
- Developed Excel plugin for data curation with ontology-based data harmonization.

2019-2022

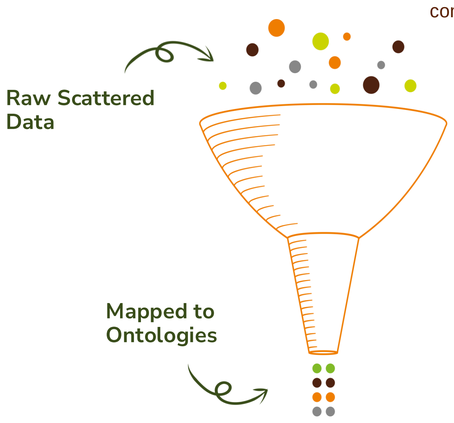
Advancement

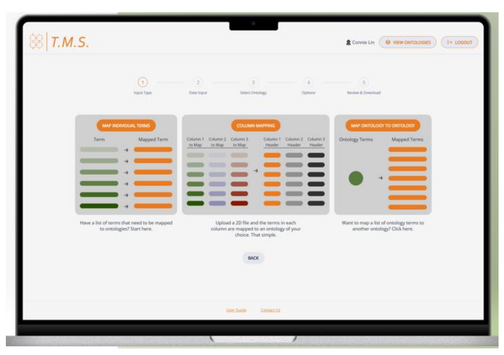
- Developed data curation interfaces with ontology-based controlled vocabs.
- Performed benchmarking and comparison to other mapping tools.

2023+

Breakthrough

- Introduced AI algorithm for AI Powered precision
- Launched to market.





TMS Supported Ontologies and Key Features

Enhanced with advanced AI-assisted mapping algorithms, a user-friendly interface, and an extensive collection of over 40 pre-loaded public biomedical ontologies, TMS enables flexible and accurate ontology mapping driving fast and reliable results. Additionally, TMS offers versatile capabilities such as user-defined mapping rules, custom ontology ingestion, and browsing of ontological trees.

1. Automated Mapping

- TMS uses semantic AI and fuzzy phonetic algorithms
- Auto-detects spelling variations & synonyms for accurate mapping
- Enables quick corrections & custom term additions

2. Supported Ontologies

Anatomy: BTO, CELLO, CL, CLO, FMA, UBERON

Assay: BAO, CHMO, OBI

Disease/Phenotype: DOID, HP, ICDO3, ICD10CM, ICD11, MONDO, OGMS, OMIM, ORDO

Drug: ATC, CHEBI, DRON, RXNORM, VO

Gene/Protein: GO, HGNC, PR

Multi-Modal: CRISP, EFO, MESH, NCIT, OBA, OMOP, SNOMEDCT, UMLS

Other: AFO, CDISC, EDAM, LOINC, PATO, QUDT, RS, UO

3. Simple to Access

- TMS has an intuitive Web Interface - Annotate terms and manage ontologies with spreadsheet-like ease
- Powerful API - Seamlessly integrate into existing data pipelines to automate mapping

Term Standardization for AI Benchmarking Use Case

To help assess target gene safety profiles from various sources an AI-based algorithm was developed by a large pharma. The client needed to validate AI safety predictions, but manually generating and mapping over 200 liability terms to ontologies was time-consuming. A scalable, automated solution was required to ensure reliable benchmarking and minimize manual effort.

We used TMS to automate the mapping of terms annotated from different sources, enabling a consistent 'apples-to-apples' comparison with the AI algorithm. This saved over 100 hours by eliminating manual ontology mapping

01 Annotate
Extracted liability terms (disease, condition, anatomical region) for 10 target genes from publications.

02 MedDRA Map
Used TMS to map liability terms to MedDRA for defining disease and anatomical categories for TLA.

03 MeSH Map
Used TMS to map liability terms to MeSH.

04 AI Benchmark
Used categorized liabilities to validate the AI algorithm's gene safety predictions compared to human-produced TLAs.

journal	abstract	Liability (raw)
Gastroenterology	The effect of endoge	colon carcinoma
Clinical and experi	1. The dose-respons	colon cancer
British journal of pl	The present study in	colon cancer
Journal of cancer re	Polyamines are esse	colonic carcinoma
International journ	Gastrin has been sug	colon cancer

Gene	Disease Category (MedDRA)	Anatomy Category (MeSH)	Liability
CCKBR	Colorectal neoplasms malignant	Digestive System	Colon cancer
CCKBR	Colorectal neoplasms malignant	Digestive System	Colorectal cancer
CCKBR	Gastrointestinal mucosal dystrophies	Digestive System	Hypergastrinemia
CCKBR	Gastrointestinal mucosal dystrophies	Digestive System	Hypersecretion gastric
CCKBR	Gastrointestinal neoplasms malignant	Digestive System	Gastric cancer

Rancho biosciences

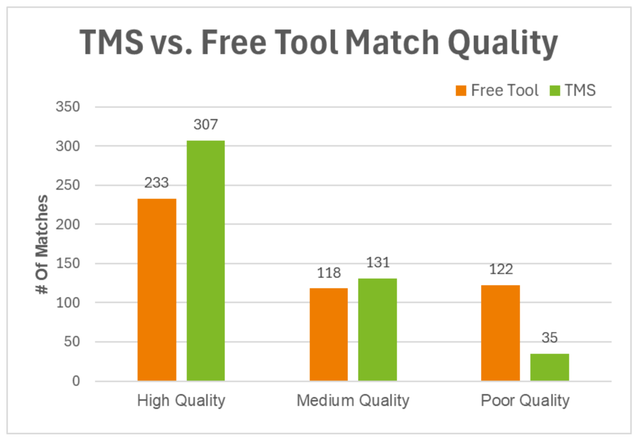
Download Poster

TMS performance quality

A thorough evaluation of TMS' potential to streamline data curation processes was conducted by benchmarking it to existing commercial and free tools. In comparison with a free terminology mapping tool using a list of patient conditions from doctor's notes as input, TMS demonstrated over 30% increase in number of high-quality hits and a threefold decrease in number of poor-quality hits. The assessment highlights TMS' strong performance, particularly in terms of precision and accuracy.

(A) To evaluate TMS' performance on real-world data, 473 patient conditions from doctor notes were mapped to OMOP Conditions by either a free terminology mapping tool or by TMS. With input term examples like "breast invasive ductal carcinoma", "INVASIVE MAMMARY/DUCTAL", "her2+er+prt", and "IDC left", the task was very challenging. The free tool failed to produce a relevant match in over 25% of cases, while TMS, empowered by AI-assisted algorithm, failed in only about 7% of cases. Moreover, TMS produced over 30% more high-quality hits.

(B) Examples where the free tool produced a poor-quality hit, and TMS produced a high-quality hit.



A: Number of hits produced by each tool in each category

All matches were manually categorized into:
- 'High quality' (exact or very close match)
- 'Medium quality' (relevant, but with some details missing)
- 'Poor quality' (completely or mostly irrelevant).

B: Mapping examples with tool-generated match scores to the right of the respective matches.

Input Term	Free Tool (FT) Mapping	FT Score	TMS Mapping	TMS Score
Ampullary cancer pancreatico-biliary subtype	Biliary acute pancreatitis	0.41	Pancreatobiliary type carcinoma of ampulla of Vater	0.93
Appendiceal Goblet cell adeno	Appendiceal colic	0.52	Goblet cell carcinoid of appendix	0.91
Appendiceal Mucinous/Signet Ring	Appendiceal colic	0.41	Signet ring cell carcinoma of appendix	0.9
c lung mets	Uremic lung	0.60	Metastatic lung carcinoma	0.88
DCIS cancer	Malignant neoplastic disease	0.50	Ductal carcinoma in situ of breast	0.88
Invasive Small Cell Carino	Invasive hydatidiform mole	0.62	Small cell carcinoma	0.91
lung high grade adenoca	Primary high grade serous adenocarcinoma of ovary	0.56	Adenoma of lung	0.86
Lung NSCCC	Multiple symmetrical lipomatosis	0.32	Non-small cell lung cancer	0.88
Melanoma- Left posterior leg	Strain of muscle of left posterior lower leg	0.49	Melanoma in situ of left lower limb	0.94
Melanoma Right Axilla	Laceration of right axilla	0.68	Malignant melanoma of axilla	0.96
Merkel cell Unknown primary site	Malignant tumor of unknown origin	0.65	Merkel cell carcinoma of unknown primary site	0.96
squamous cell, poorly differerntiated	Poorly differentiated chordoma	0.57	Squamous cell carcinoma	0.89

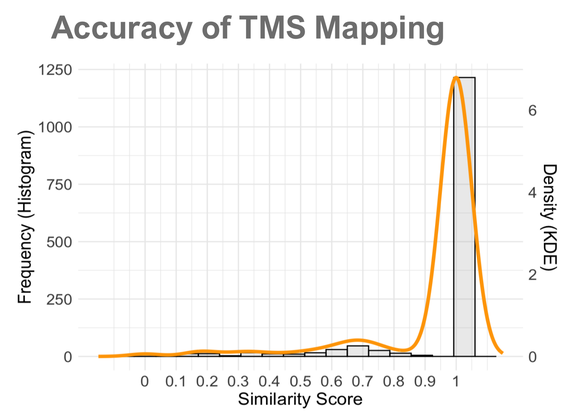
RWD Terminology Mapping Use Cases

I. With a list of 1,875 raw unique generic drug names related to 45.5K drug names from a RWD dataset as input TMS achieved 65% of high-quality matches to OMOP drugs.

A: Examples of drug names mapped to OMOP drug ingredients using TMS

Unique generic drug names were mapped to their corresponding ingredients using phonetic mapping algorithm which provides similarity score outputs ranging from **0 to 1**, with "1" indicating an exact match or recognized synonym in the OMOP ontology. The tool allows selection of the most relevant ingredient by choosing the highest-scoring term. This TMS mapping was followed up with manual QC which confirmed the accuracy of the mapped ingredients to their corresponding generic drug names and 45.6k unique drug names.

Drug Name	Generic Drug Name	TMS Mapped Ingredient	Similarity Score
ziagen 300 mg tablet	abacavir sulfate	abacavir	1
abelcet 5 mg/ml susp	amphotericin b lipid complex	amphotericin B	1
carbidopa-levodopa-enta 200 mg	carbidopa/levodopa/entacapone	carbidopa / entacapone / levodopa Oral Gel	0.75
fenofibric 135mg dr capsules	fenofibric acid (choline)	choline fenofibrate	1
butalbital/aspirin/caffeine 325mg cap	butalbital/aspirin/caffeine	aspirin 325 MG / butalbital 50 MG / caffeine 40 MG Oral Capsule	0.9
dexamethasone 0.1% eye drop	dexamethasone sodium phosphate	dexamethasone	1



B: Distribution of similarity scores across generic drug names from TMS mapping

The histogram (gray) represents the frequency of similarity scores (left y-axis) across the RWD dataset, and the kernel density estimation (KDE) curve (orange) provides a smoothed probability distribution (right y-axis) of the similarity scores. A clear peak at similarity score **1** indicates accurate mappings where the tool identified exact matches.

II. After genetic testing, patient data is returned from multiple sources, vendors and systems (LIMS, EPIC, etc.). This clinical and genetic data requires aggregation, harmonization and standardization for specific use-case analyses and for sharing with agencies and pharmaceutical companies. Clinical data from 20,000 de-identified patients was standardized by Rancho using TMS. Approximately 10,500 terms were mapped to OMOP CDM categories for database ingestion. A pipeline integrating TMS for clinical data curation and ETL for clinical/genetic data ingestion was delivered to a client, saving more than 500 hours of curation and QC time.