Application of an AI-Powered Terminology Management Solution (TMS) in the Real-World Data (RWD) FAIRification process

Alena Fedarovich, Brad Farrell, Alex Kadhim, Tracy Ballinger, Rob Beetel, Oleg Stroganov, Leonya Ivanov,
Vishnu Govindaraj, Candace Ruff, Shahraz Niwaz
Rancho Biosciences, LLC

Abstract

There has been an increased interest in the use of real-world data (RWD) and real-world evidence (RWE) to facilitate drug discovery, development, and regulatory decision making. Utilization of RWD as a promising tool to answer key questions in the areas of clinical pharmacology and translational science is limited due to challenges posed by quality issues and integration from various sources. Leveraging extensive curation expertise, Rancho Biosciences has developed Terminology Management Solution (TMS), a user-friendly tool designed to support scientists by simplifying the curation process. TMS uses the power of AI to scan and annotate large text datasets, aligning them with over 50 biopharma and biomedical standards. TMS supports both public and custom ontologies, ensuring comprehensive data interoperability. It simplifies data alignment processes with a lightweight, pre-configured solution accessible via an intuitive UI or robust API suite, enhancing research efficiency and data accuracy. As the ultimate solution for precise terminology mapping, TMS empowers data management and the FAIRification process, making data Findable, Accessible, Interoperable, and Reusable. This is especially critical for RWD integration and transformation into a comprehensive data product ready for downstream analyses or submission to regulatory agencies.

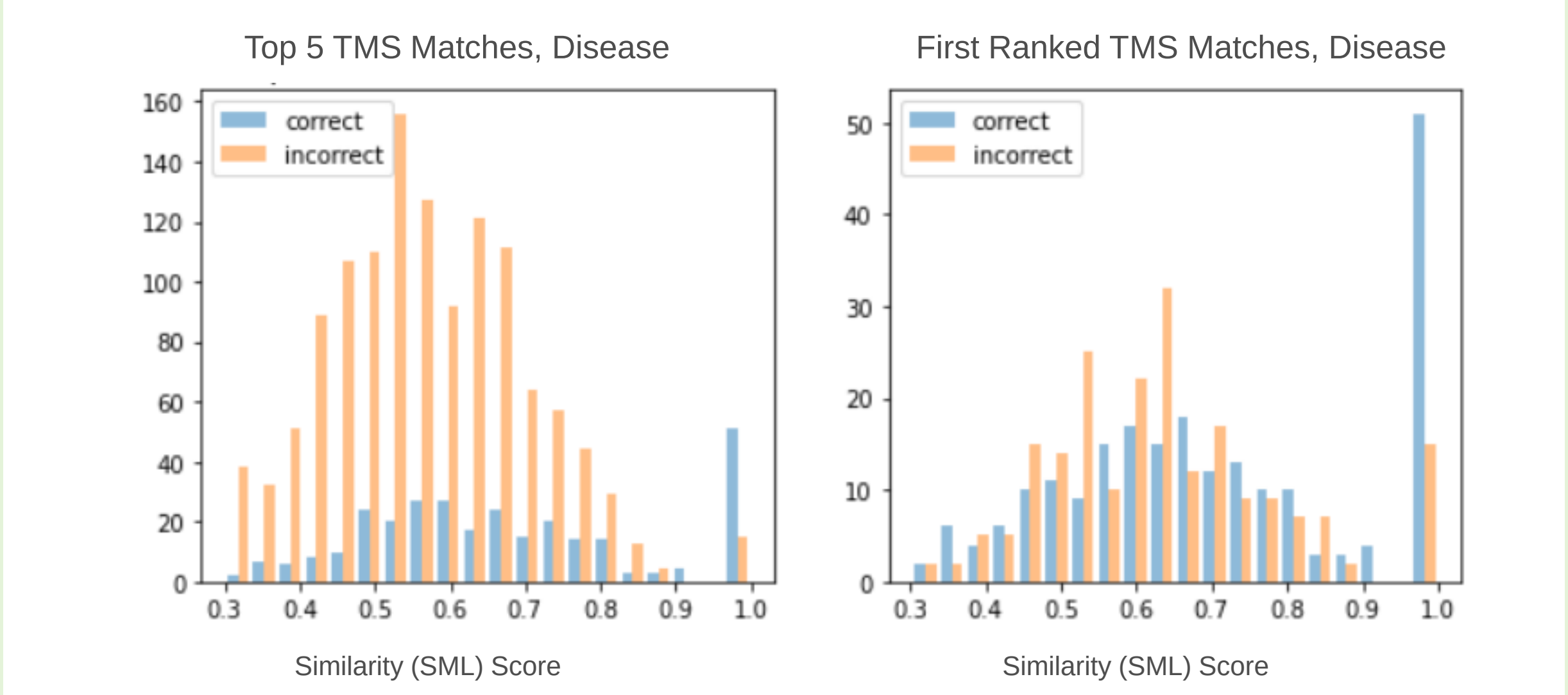
TMS utilizes both AI-assisted semantic (AI) and phonetic (Fuzzy) algorithms making it a one-stop-shop for data harmonization, alignment, and mapping. We rigorously evaluated TMS against existing commercial tools across essential tasks such as term harmonization, ontology mapping, and data extraction from unstructured sources. To provide a comprehensive assessment of each tool's potential in streamlining the curation process, the evaluation was focused on accuracy and efficiency, usability, support, and customization.

We also explored how the use of AI in combination with Fuzzy enriches the outcomes of the terminology mapping upon TMS incorporation into a RWD harmonization and integration pipeline.

The results demonstrated the robust capabilities of TMS, particularly its superiority in precision and recall compared to other evaluated tools. TMS excelled in accurately mapping a vast array of terms to respective ontologies and displayed a potential for substantial timesaving in manual curation processes. Use of AI in combination with Fuzzy enriches mapping outcomes of the RWD harmonization and integration pipeline. These highlight TMS' role as a pivotal asset in RWD/RWE curation, promising a significant leap forward in the accuracy and efficiency of data harmonization efforts.

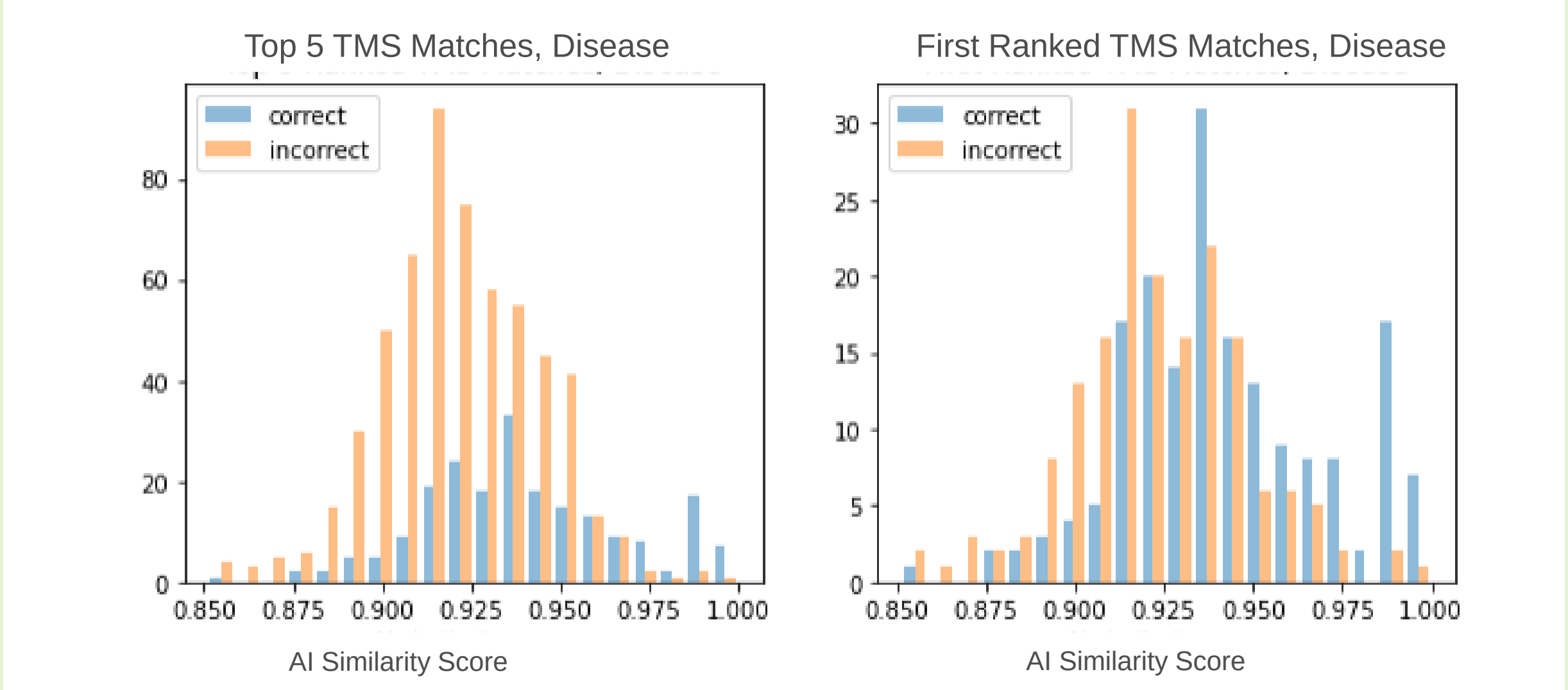
TMS: Mapping Algorithm Comparison

Fuzzy mapping algorithm enables fast phonetic mapping and provides similarity score outputs.



Similarity (SML) scores are shown for correct and incorrect matches of disease terms: the top 5 ranking matches are shown on the left, and only the best matches are shown on the right.

AI-assisted semantic (AI) term mapping involves linking and translating terms between different vocabularies or databases. This process facilitates the understanding and integration of diverse data sources by establishing equivalences or relationships between terms with similar meanings or concepts.

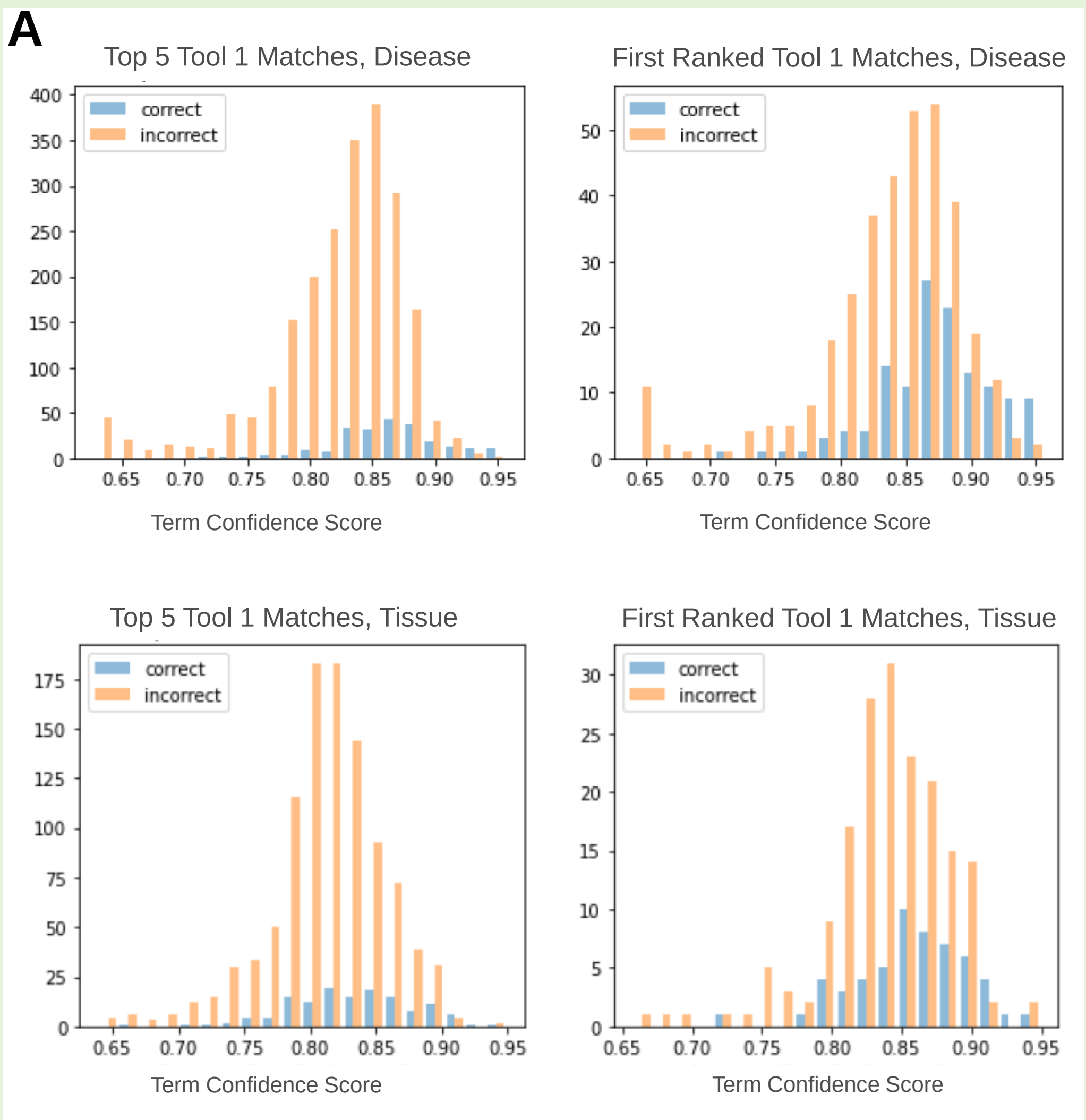


AI cosine similarity scores are displayed for correct and incorrect matches of disease terms: the top 5 ranking matches are shown on the left, and only the best matches are shown on the right.

The AI cosine similarity score demonstrates better separation between correct and incorrect mappings of 476 RWD disease terms to DOID compared to TMS Fuzzy SML score or commercial Tool 1 confidence scores.

TMS Compared to Other Automated Ontology Mapping Tools

TMS Fuzzy and AI algorithms vs. Commercial Tool 1



B

Algorithm	Tool 1	TMS Fuzzy	TMS AI
Disease Rank 1	132 (38%)	219 (46%)	179 (38%)
Disease Top 5	225 (47%)	296 (62%)	207 (43%)
Tissue Rank 1	55 (24%)	137 (59%)	123 (53%)
Tissue Top 5	134 (58%)	171 (74%)	142 (61%)

(A) Tool 1 confidence scores for correct and incorrect matches of disease and tissue terms are shown: The top 5 ranking matches are shown on the left, and only the best matches are shown on the right.

(B) Accuracy of Rancho TMS with two scoring algorithms is shown. Each tool was tested against a set of 476 RWD disease terms mapped to DOID and 232 RWD tissue terms mapped to Uberon. The top five ranked results were returned for each tool, and the number of correct matches was counted both within the top match and among the top five matches. TMS Fuzzy method scoring performs the best, but TMS AI methods also perform well, with up to 74% and 61% of terms correctly mapped, respectively.

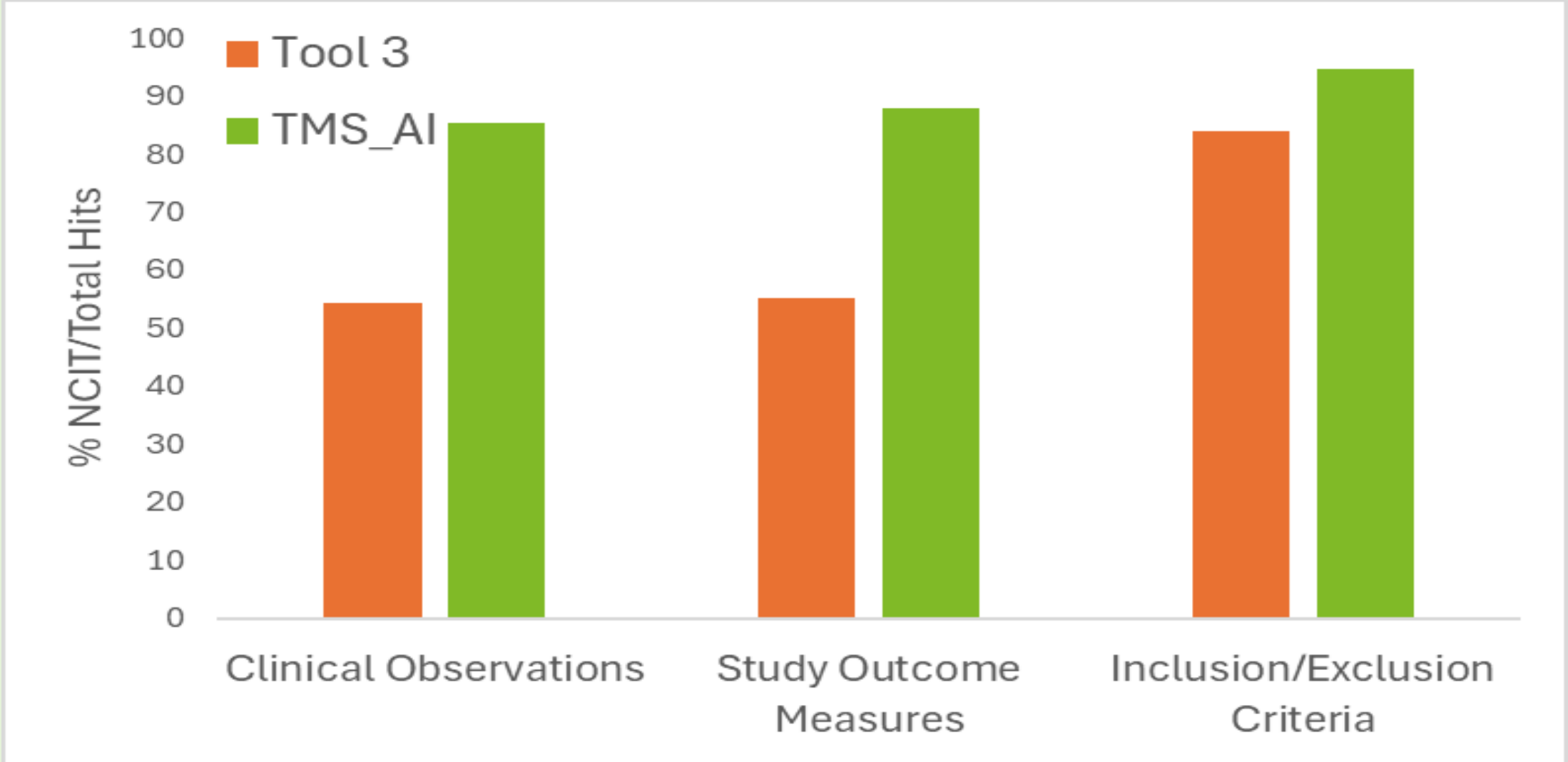
TMS Fuzzy vs. Commercial Tool 2

770 terms were run against DOID using both TMS Fuzzy and commercial Tool 2. TMS performed slightly better on the recall. Although precision was the same for both tools.

Automatic matching results review	Tool 2	TMS Fuzzy
True positive (TP)	491	513
False positive (FP)	219	234
False negative (FN)	42	5
Not matched terms by either algorithm	18	18
All	770	770
Recall (TP/(TP+FN))	0.92	0.99
Precision (TP/(TP + FP))	0.69	0.69

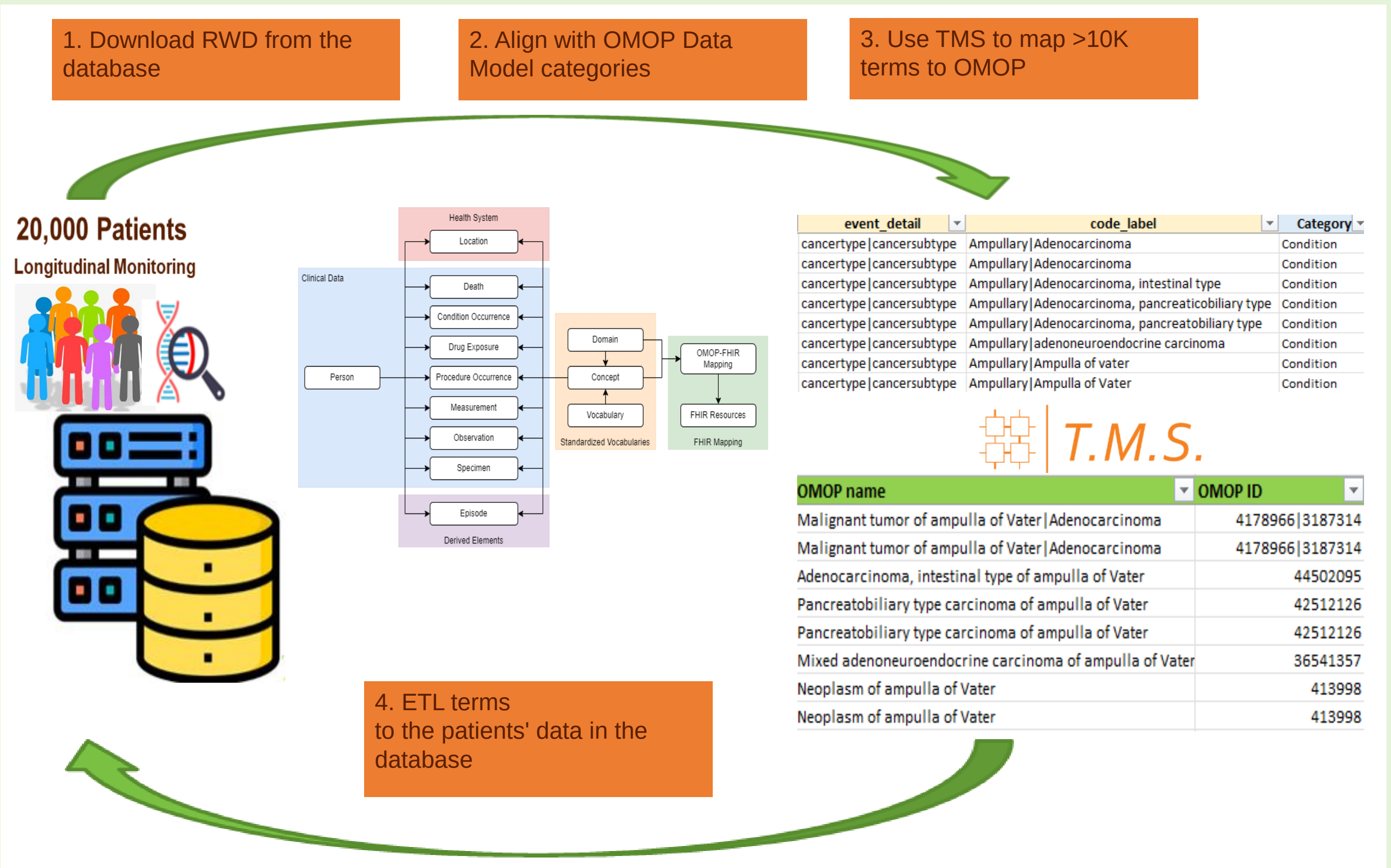
TMS AI vs. Commercial Tool 3

A test was performed to determine if TMS AI could find more NCIT hits than Tool 3. Terms from three protocol sections were run through Tool 3 pipeline and TMS AI. In all test cases, TMS outperformed Tool 3, finding between 10.9 to 32.6 percentage point increase in NCIT matching terms.



TMS Use Case in the Real-World Data (RWD) FAIRification process

Rancho used an OMOP-based data model to harmonize the unstructured RWD coming from multiple sources, vendors and systems. TMS was incorporated in the pipeline developed to align RWD from de-identified patients with CDM categories, clean and standardize terms, and prepare them for database ingestion and ETL terms to the patients' data in an internal database. Both TMS Fuzzy and AI algorithms were used to map >10,000 terms to OMOP CDM condition and observation categories, followed by a manual QC performed by experienced data integrity specialists. A pipeline, including TMS for data curation and ETL for data ingestion was delivered to a client, saving ~500 hours of curation and QC time.


1. Download RWD from the database

2. Align with OMOP Data Model categories

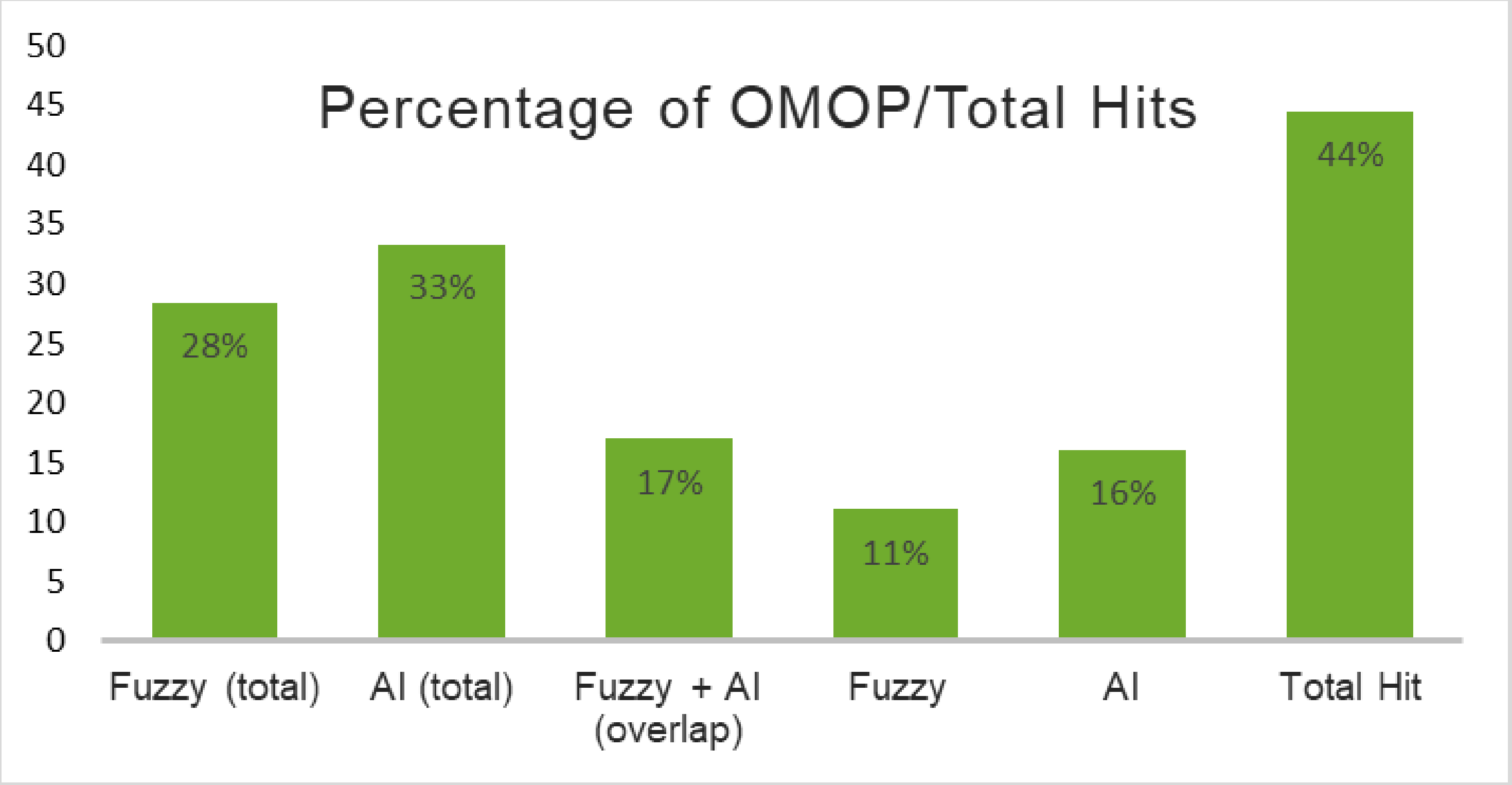
3. Use TMS to map >10K terms to OMOP

4. ETL terms to the patients' data in the database

event_detail	code_label	Category
cancertype|cancersubtype	Ampullary|Adenocarcinoma	Condition
cancertype|cancersubtype	Ampullary|Adenocarcinoma	Condition
cancertype|cancersubtype	Ampullary|Adenocarcinoma, intestinal type	Condition
cancertype|cancersubtype	Ampullary|Adenocarcinoma, pancreaticobiliary type	Condition
cancertype|cancersubtype	Ampullary|Adenocarcinoma, pancreatobiliary type	Condition
cancertype|cancersubtype	Ampullary|adenoneuroendocrine carcinoma	Condition
cancertype|cancersubtype	Ampullary|Ampulla of vater	Condition
cancertype|cancersubtype	Ampullary|Ampulla of Vater	Condition

OMOP name	OMOP ID
Malignant tumor of ampulla of Vater|Adenocarcinoma	4178966|3187314
Malignant tumor of ampulla of Vater|Adenocarcinoma	4178966|3187314
Adenocarcinoma, intestinal type of ampulla of Vater	44502095
Pancreatobiliary type carcinoma of ampulla of Vater	42512126
Pancreatobiliary type carcinoma of ampulla of Vater	42512126
Mixed adenoneuroendocrine carcinoma of ampulla of Vater	36541357
Neoplasm of ampulla of Vater	413998
Neoplasm of ampulla of Vater	413998

To explore how the use of AI/Fuzzy enriches the outcomes of the terminology mapping a subset of randomly selected RWD disease terms was mapped to OMOP conditions using both algorithms. 17% of the correctly mapped by Fuzzy and AI terms overlapped, 11% and 16% additional terms were mapped correctly by Fuzzy and AI respectively resulting in 14 percentage point increase.



Contact Us

Rancho bio sciences