Suppl. 1. Methods for constructing the ROR corpus, PubMed sample, and gold standard

1. ROR Corpus Construction and Normalization

1.1. Data Source and Schema

We sourced institutional reference data from the official Research Organization Registry (ROR) distribution (v2.0) and converted the CSV files into JavaScript Object Notation (JSON) format. To preserve multilingual characters and UTF-8 encoded special characters, we applied UTF-8-sig encoding during the conversion process. The resulting dataset contains 163,963 institutional entries with 29 variables.

We extracted canonical names, aliases, acronyms, and geographic attributes to build a domain-specific rule set for the spaCy EntityRuler. This rule set was uploaded to spaCy's blank pipeline, which was then configured as a domain-adapted entity-linking environment. This environment integrated name normalization, acronym expansion and geographic metadata (city, state, country) to optimize matching precision. The pipeline was able to link PubMed affiliation strings to ROR IDs.

The corpus schema is indexed by ror_id (Table S1).

Table S1.1. ROR Corpus Schema

Field	Description	Example
id	Unique ROR Identifier (URL format)	https://ror.org/05a28rw58
name	Institution's Standard English Name	Seoul National University
country	Country Name and ISO Code	{"country_name": "South Korea", "country_code": "KR"}
region	Geographic Information	{"city": "Seoul", "country": "South Korea"}
location	Lat-Long Coordinates	{"lat": 37.459882, "lng": 126.951905...}
links	Official Website	["http://www.snu.ac.kr"]
types	Institution Type	["Education"]
external_ids	External Identifier Mappings	{"ISNI": {"preferred": "0000 0001 2281 955X"}}
aliases	Abbreviations/Alternatives	["SNU"]
labels	Multilingual Names	[{"label": "서울대학교", "iso639": "ko"}]

1.2. Corpus Refinement

We optimized the corpus for matching performance through three iterative *refinement phases*:

Phase 1: Field Consolidation and Normalization

We consolidated key textual fields from the ROR dataset—specifically names, aliases, labels, and acronyms—into a unified names field. Similarly, we combined geographical attributes (city, state1, state2, and country name) to construct standardized region and country fields.

We normalized all textual fields by lowercasing characters, removing language codes (e.g., "en") and parenthetical content, trimming redundant whitespace, and applying Unicode NFKD normalization to remove diacritics.

Preliminary tests indicated that while this improved consistency, certain valid but generic acronyms created semantic ambiguity. For example, the institution "AND" (ROR ID: https://ror.org/05k7ts633) frequently triggered false positives in PubMed affiliation strings.

Phase 2: Stopword and Ambiguity Filtering

To suppress erroneous matches due to semantically non-discriminative strings, we applied three filters to the names field:

1. *Length Filter*: We removed acronyms with fewer than three characters long to eliminate low-specificity tokens.
2. *Pattern Filter*: We excluded strings beginning with numbers or symbols to remove tokens and data artifacts generated during data transformation.
3. *Stopword Filter*: We eliminated generic stopwords (length greater than 3) lacking discriminative value.

This process pruned non-informative tokens from the matching patterns without removing the associated ROR IDs.

Phase 3: Manual Enrichment and ROR Variant Generation

Phase 2 filtering reduced mismatched cases. However, further reductions were possible by addressing real-world variations missed by strict exact matching. To address such mismatches we introduced two refinement strategies:

1. *Manual Alias Enrichment*: Some institutions appeared as affiliations in PubMed frequently, but they were not found in the ROR registry. We manually mapped 18 high-frequency institutions, sub-schools (e.g., "Harvard Medical School") and healthcare providers that appeared in PubMed but lacked explicit entries or aliases in the ROR registry.
2. *ROR Variant Generation*: Many ROR canonical names contained commas or hyphens while the same institutions often omitted them in PubMed affiliation texts. This omission leads to greater false negatives. We generated character-free variants for canonical names containing commas or hyphens and stored them as variants of the canonical ROR entries. For example, we mapped the ROR canonical form "University of California, San Diego" to the common PubMed variant "University of California San Diego".

We exported these refined patterns as the spaCy's EntityRuler pattern files for the Stage 1 exact matching model.

2. PubMed Study Sample

We targeted Digital Health-related publications indexed in PubMed between 2014 and 2024, a period

selected for improved metadata consistency, especially in the author-affiliation structure. The search strategy combined Medical Subject Headings (MeSH) and title/abstract keywords related to digital health including “Digital Health,” “Telemedicine,” “Artificial Intelligence,” “Machine Learning,” “Electronic Health Records,” “Wearable Devices,” and “Precision Medicine.” We applied the “Humans”[MeSH] filter to exclude non-human research.

We retrieved 482,086 PMID records using the NCBI E-utilities API (last accessed December 20, 2025) in Extensible Markup Language (XML) format to preserve full metadata structure. From the retrieved XML records, we extracted 4,062,669 raw author - affiliation strings related to Digital Health.

3. Gold Standard Benchmark Construction

To evaluate the models, we constructed a Gold Standard dataset by randomly sampling 4,000 PMIDs from the 482,086 PubMed records published between 2014 and 2024, and then extracting 46,378 author-affiliation strings from the 4,000 PMID records.

Annotation Protocol
Four trained annotators with one adjudicator labeled the dataset using a two-pass process:

1. Pass 1 (Initial Labeling): Each annotator processed a unique subset of 1,000 PMIDs. Using the ROR registry, institutional websites, the GRID archive, and Wikidata, they mapped strings to canonical ROR IDs. Ambiguous cases were labeled “Abstain”.
2. Pass 2 (Cross-Review): Annotators cross-reviewed peer subsets to ensure every string received independent dual validation.
3. Adjudication: Disagreements were escalated to an expert adjudicator. Irreconcilable cases were labeled “Unlinkable” to avoid speculative matches.

The ROR IDs serve as the ground truth for institutional identification. To ensure accurate identifier selection, the canonical ROR institution names were referenced as needed.

Table S1.2. Gold Standard Dataset Schema

Field	**Description**
PMID	Unique identifier of the PubMed article
Original affiliation	Raw affiliation as appeared in PubMed
ROR name	Canonical ROR institution name
ROR ID	Linked ROR identifier (or *Unlinkable*)
Annotation status	Final decision after adjudication