Suppl. 3. Evaluation metrics, baseline implementation, and statistical analysis.

1. Evaluation Metrics

We conducted stage-wise assessments on the Gold Standard dataset to quantify precision preservation and recall recovery. We matched predicted ROR IDs against manually validated ground-truth labels and categorized the results using the confusion matrix criteria:

- True Positive (TP): A valid institution exists, and the model links it to the correct ROR ID.
- False Positive (FP): The model assigns an ROR ID to an affiliation string that does not correspond to a valid ROR-registered institution, or assigns an incorrect ID to a valid institution.
- False Negative (FN): A valid institution exists, but the model fails to assign an ROR ID.
- True Negative (TN): The affiliation string does not correspond to any valid ROR-registered institution, and the model correctly abstains.

We computed standard performance metrics with a primary emphasis on precision to control for false positives. The equations are as follows:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Baseline Method: Generic spaCy NER

To assess the value of domain adaptation, we implemented a baseline using an off-the-shelf spaCy Named Entity Recognition (NER) model. This approach represents a generic NLP strategy lacking domain-specific knowledge of scholarly affiliations.

spaCy is a widely used NLP tool for automatically identifying and classifying named entities—such as persons, organizations, and geopolitical entities—in unstructured text. As an open-source NLP library extensively adopted in both industrial and academic settings, spaCy provides pretrained language models such as en_core_web_sm, which are trained on large-scale general-domain corpora (e.g., news articles and web text). These models perform tokenization, part-of-speech (POS) tagging, and dependency parsing, and subsequently predict entity types including PERSON, ORG (organization), and GPE (geographic and political entity). We focus on the model's ability to identify organizational entities within PubMed affiliation strings.

The evaluation followed a two-stage baseline procedure: generic entity recognition using spaCy NER pipeline, followed by naive string-based linking to the ROR corpus.

2.1. Entity Recognition

We used the en_core_web_sm model of spaCy, trained on general-domain corpora, to identify organizational entities (ORG type in spaCy). We processed PubMed affiliation strings using nlp.pipe with

a batch size of 1,000 for efficient inference. We retained only entities labeled as ORG and applied minimal post-processing to remove trailing punctuation.

2.2. Naive Linking

To link extracted mentions to the ROR corpus, we used spaCy's Matcher in a strict surface-form configuration. We converted ROR canonical names into token patterns using the LOWER attribute for case insensitivity. We enabled FULL_SPAN_ONLY to require complete token alignment and ALLOW_TRAIL_PUNCT to tolerate punctuation artifacts. This baseline excluded geographic or contextual validation.

3. Additional Statistical Analysis Results

3-1.Performance Comparison by McNema's Test

Table S3.1

Comparison	**Chi-squared value**	***p*-value**	**Accuracy gain with 95% CI**	**Win/loss odd ratio**
Stage2 vs. Baseline	19629.76	< 0.001***	0.45 [0.44, 0.45]	34.55
Stage2 vs. Stage1	1600.19	< 0.001***	0.04 [0.03, 0.04]	12.65
Stage2 vs. Fuzzy	8931.00	< 0.001***	0.21 [0.20, 0.21]	25.40

3-2. Bootstrap Confidence Interval of F1-score

We used the percentile bootstrap method with 2,000 repetitions. Since testing the increase in the Stage 2 model's F1-score required three sets of tests, we adopted the Bonferroni correction and adjusted the significance level to 1.7%. All CIs do not include zero, indicating the statistical significance of the difference..

Table S3.2

F1-score	**Coef.**	**[Lower Bound, Upper Bound] (98.3% CI)**
Stage2	.95	[.94 .95]
Baseline	.58	[.58 .59]
Stage1	.92	[.91 .93]
Fuzzy	.80	[.79 .81]
Stage2 vs. Baseline	.37	[.36 .38]
Stage2 vs. Stage1	.03	[.02 .03]
Stage2 vs. Fuzzy	.14	[.14 .15]