LYNXCARE

RWD188

A HUMAN-IN-THE-LOOP VALIDATION INTERFACE FOR SCALABLE QUALITY ASSURANCE OF OMOP DATABASES

Lucas Sterckx, PhD, Narges Farokhshad, MSc, Stephanie Vandeput, MSc, Siddharth Agarwal, MSc, Dries Hens, MD

LynxCare Clinical Informatics, Leuven, Belgium

OBJECTIVES

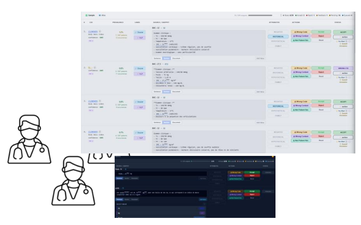
NLP quality assurance remains resource-intensive. Each processing run can introduce errors, and review effort remains constant even as extraction quality improves. Error-hunting rarely quantifies overall review effort or accuracy. We developed an **expert-validator interface** with intelligent sampling to **reduce validation time** and produce **statistically sound quality metrics** for OMOP-compliant databases.

METHODS



1 Select samples

One datapoint at a time, starting with text shared by the most patients. Identical text with the same extracted term is validated once and never sampled again. A semantic map lets reviewers label a cluster of similar snippets in one action.



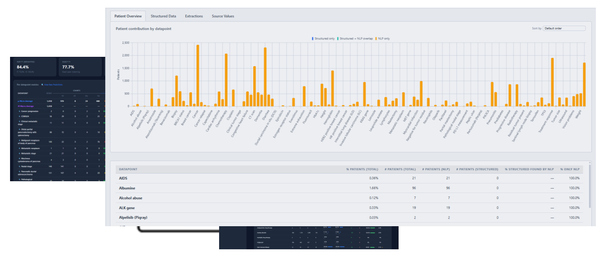
2 Expert review

Experts can accept, reject, or flag a wrong concept code, wrong meaning in context or no relevance to the patient. They can then fix the code and negated, historical, hypothetical and experiencer attributes. Past decisions on the same code and term are pre-filled to confirm.



3 Expand context

A context toggle widens the view from sentence to list item, section or full document. Each decision is stored with the context level it was made at.



4 Store and monitor

All decisions, accepted ones included, enter one annotation store. Dashboards show acceptance and edit rates, error types and coverage per datapoint, NLP version and reviewer, plus F1, attribute and coding accuracy across NLP builds.

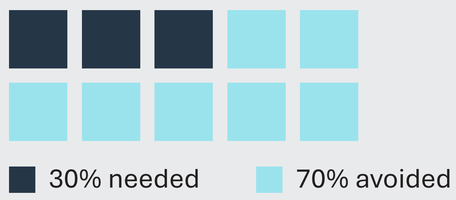
RESULTS

5-fold
higher review throughput

Prior generic dashboard
16
Expert-validator interface
86
Concepts reviewed per hour

The prior dashboard supported error flagging only.

70%
fewer annotations
for representative coverage



Identical text is validated once, text shared by the most patients is reviewed first, and validated extractions are never sampled again.

ONCOLOGY, SITE B
One hospital · 59 datapoints

2,595,001	713,356	6,565	2,094
NLP extractions	unique text samples	patients	expert decisions

Most-shared text first

≈10× more extractions covered per expert decision

Reviewed samples	≈37
Average sample	3.6

Bars: extractions per unique text sample

Reviewing 0.3% of unique samples covered 3.0% of all extractions. Within single datapoints, reviewing 2.1–6.2% of unique samples covered 20.0–60.0% of extractions.

Acceptance rate per datapoint

95.9% of 2,094 decisions accepted (lung cancer and staging)

Staging 9 T/N/M-type datapoints	≥98.4%
Disease course 3 datapoints	≥94.1%
Generic “Cancer” 1 datapoint	80.4%
Lung-cancer histology 4 related datapoints	71.0%

Validate per site

Site A Site B

84.7% accepted at Site A (lung cancer, 3,293 decisions)

	Site A	Site B
Tumor stage	77.4%	99.0%
Lung-cancer histology	97–100%	71.0%

Site A errors concentrate in genomic biomarkers; clinical and pathological T/N/M reach 94–100% at both sites.

Acceptance rate: share of reviewed extractions that were accepted. Review starts with the most frequent text, so acceptance rates are not estimates of overall precision. One decision can count toward several related datapoints. Site B datapoints shown each have at least 100 decisions.

VALIDATE ONCE. REUSE ACROSS RUNS.

Validated extractions are never re-sampled, so expert time goes only to unreviewed extractions. Logging correct extractions as well as errors limits bias toward edge cases and builds a reusable benchmark. Metrics demonstrate progress.

CONCLUSIONS

A purpose-built interface, intelligent sampling and approval logging make quality assurance measurable and scalable. Continuous monitoring becomes feasible for resource-constrained OMOP implementations, supporting reliable observational research across multilingual data networks.

Contact: Siddharth Agarwal, LynxCare Clinical Informatics, Leuven, Belgium. info@lynx.care