Embedding-Based Density Sampling
for Efficient Quality Assurance in OMOP

Background & Methods

- Traditional validation is **annotation-heavy**
- Clinical text is highly redundant - enabling **semantic clustering**

We built a human-in-the-loop **validation framework** combining:

- **Embedding-based density sampling**
- **Label propagation**

This reduces **annotation effort** while maintaining **robust quality metrics**[1,2]

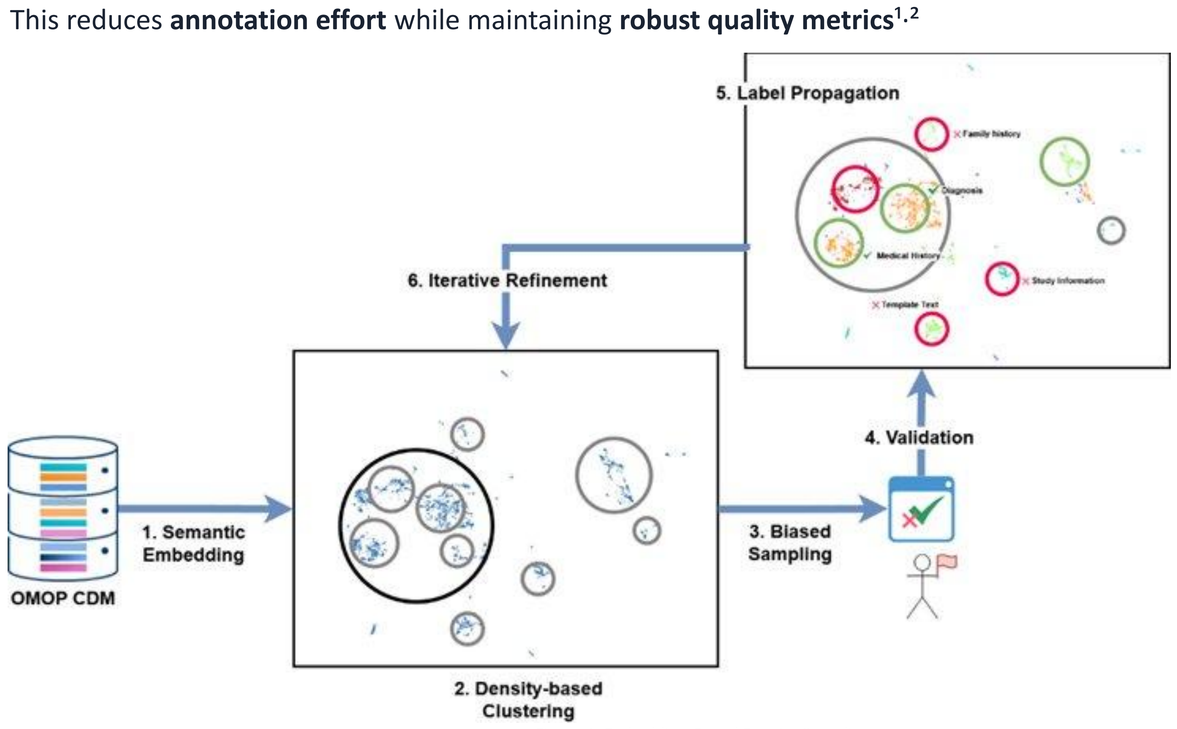


Figure 1. Embedding-Based Density Sampling

Corpus: 31,924 Dutch clinical notes

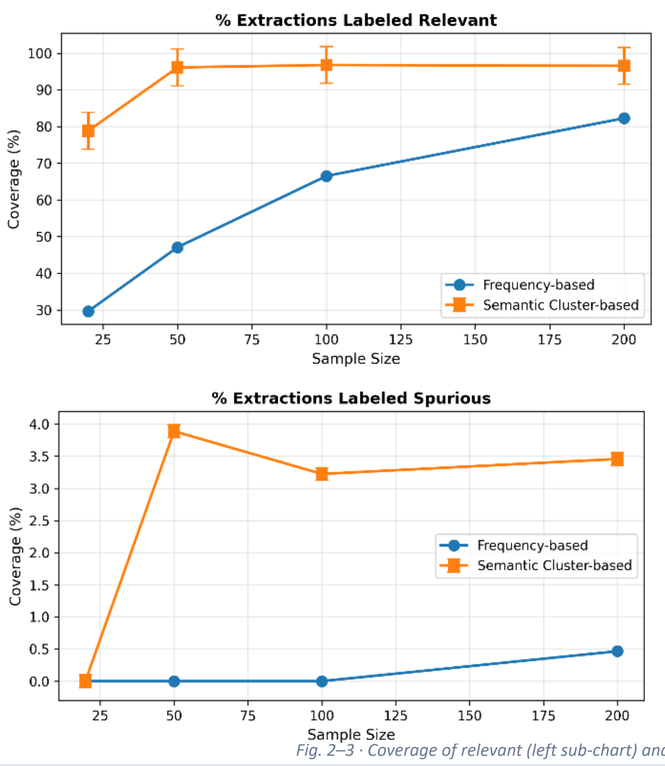
"HER2 amplification" datapoint	"Radiotherapy" datapoint
3,008 mentions 592 contexts	48,672 concepts 15,903 contexts

1. **Semantic embedding**
   surrounding context → 384-dim dense vector space
2. **Density-based clustering**
   HDBSCAN applied in the semantic vector space
3. **Cluster-diverse sampling**
   compared against frequency-based baseline
4. **Human-in-the-loop validation**
   reviewer labels relevant vs spurious
5. **Label propagation**
   to similar unlabeled concepts within clusters
6. **Iterative refinement**
   impure clusters re-sampled if budget allows

Results

1. **Clusters were mostly clean** and representative; one sample usually validated the whole group.
2. **Noise text fell into distinct clusters**, easy to reject at a glance
3. Single labels propagated across clusters, **cutting review load**
4. **Cluster-based sampling gave far higher coverage than frequency sampling** (38%/97% vs 12%/65% at 100 concepts).

HER2 gene amplification



Radiotherapy

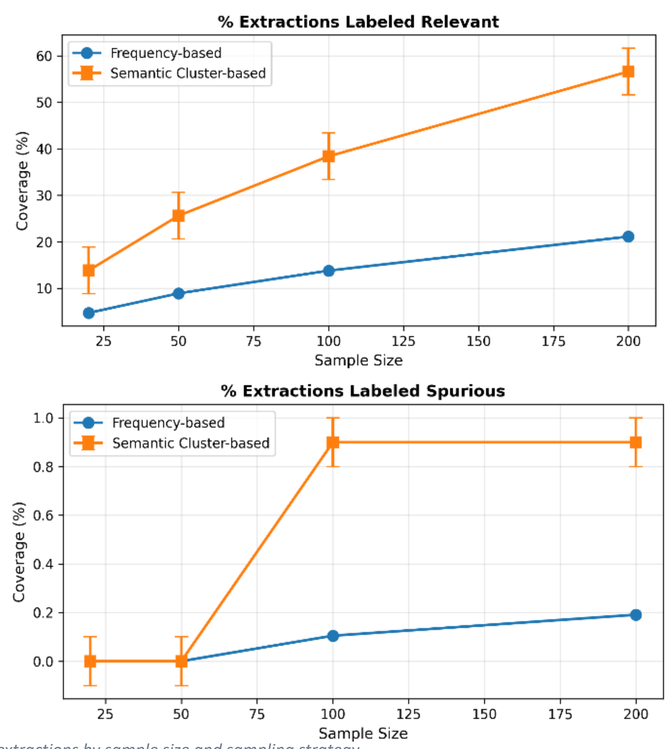


Fig. 2–3 · Coverage of relevant (left sub-chart) and spurious (right sub-chart) extractions by sample size and sampling strategy.

Annotated extraction snippets

Datapoint	Snippet (Dutch clinical note)	Valid?
HER2	Voor borstca: ER: + PR: + Neu: 1+	✓
HER2	FISH LSI HER2/CEP17: geen amplificatie	✓
HER2	OSE2101 vaccine: CEA, HER2-neu, MAGE-2…	✗
HER2	CRP 34, WBC 19.55, Neu 17.34k (blood count)	✗
Radiotherapy	- Externe radiotherapie (pancranieel)	✓
Radiotherapy	Na 2 cycli chemoradiotherapie: CT-evaluatie	✓
Radiotherapy	RT: onverdacht adenoom (rectal exam note)	✗
Radiotherapy	RT: normaal prostaatkapsel (prostate note)	✗

Conclusions

1. Embedding-based density sampling offers major efficiency gains.
2. Natural clustering in clinical text allows fewer annotations for solid validation.
3. Embeddings work in multilingual contexts of European OMOP data networks.
4. Lower validation burden helps accelerate OMOP adoption.

References 1. Dasgupta S, Hsu D. Hierarchical sampling for active learning. ICML 2008. 2. Sterckx L et al. Active learning and semantic clustering for noise reduction in distant supervision. NIPS 2014.

OHDSI

Lucas Sterckx · Narges Farokhshad · Clara L. Oeste
LynxCare, Leuven, Belgium.

LYNXCARE