# ArcMap – A new tool to accelerate real-world data standardisation at scale

Joseph Cronin PhD[1], Lawrence Adams[2], Keiran Tait[1], Janie Baxter[1], Robert Dürichen PhD[1] [1]: Arcturis Data, Building One, Oxford Technology Park, Technology Drive, Kidlington, OX5 1GN UK. [2]Artificial Intelligence Centre for Value Based Healthcare, London, United Kingdom, SE1 7EU

## Introduction

- Real-world data (RWD) is primarily collected for patient care, not research, leading to inconsistencies in data quality and structure across healthcare providers.
- Variability in data collection and management – even within the same Electronic Patient Record (EPR) systems – hinders large-scale RWE studies.
- Data standardisation is essential for enabling meaningful analysis across institutions, requiring both schema alignment and concept mapping.
- Manual mapping of medical concepts to standardised vocabularies (e.g., SNOMED, LOINC) is time-consuming and relies heavily on clinical expertise.
- Natural language processing (NLP) currently being explored in relation to standardizing clinical data[1,2] and mapping medical concepts to structures ontologies

Example scenario:

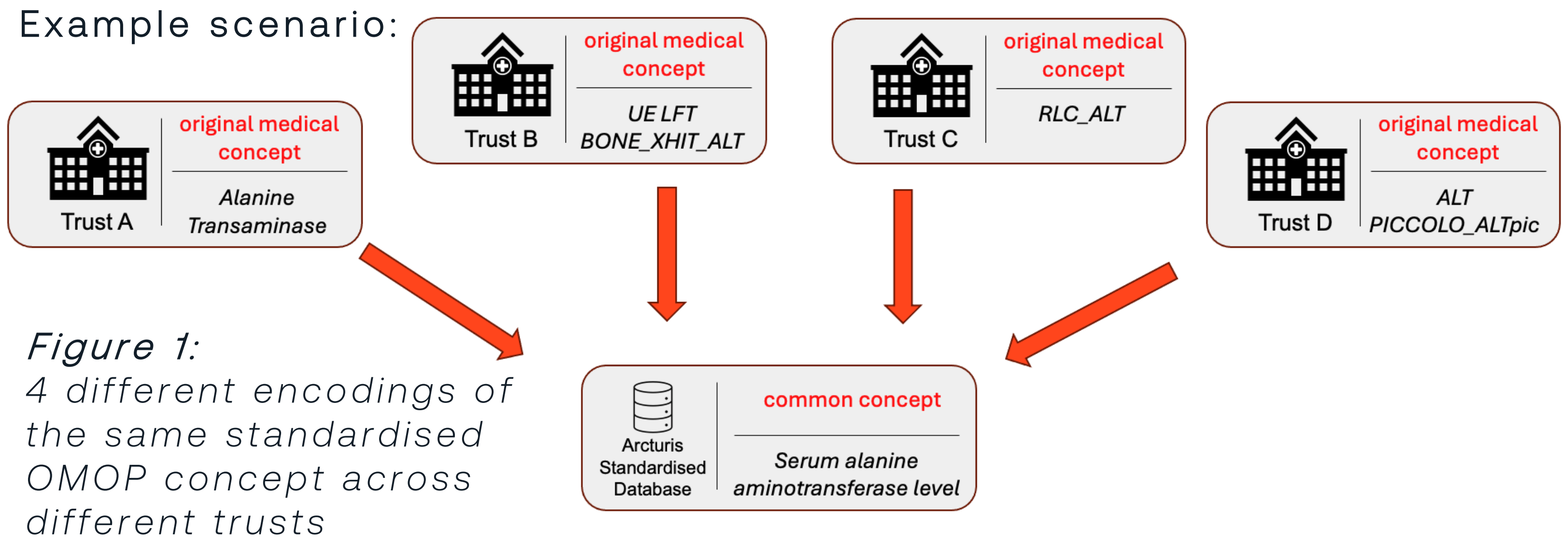


*Figure 1:* *4 different encodings of the same standardised OMOP concept across different trusts*

## Methods

- Base Model: ArcMap is built on the BioLORD3 model, fine-tuned using trust-specific medical concept encodings to enhance domain relevance.
- Training Dataset: Over 14,000 medical concepts from 4 NHS trusts were manually mapped to OMOP standard concepts (SNOMED[4] for lab tests, DM+D[5] for medications) by clinical experts.
- Training Strategy: Used triplet loss with one positive and 50 negative samples per source concept; data augmentation included word order shuffling to increase training diversity.
- Embedding Optimization: During training, embeddings of source concepts were moved closer to correct mappings and further from incorrect ones using cosine similarity.
- Evaluation Setup: Simulated onboarding of a new trust by training on three trusts and testing on the fourth; performance measured by top-1 and top-5 accuracy of predicted OMOP concepts.

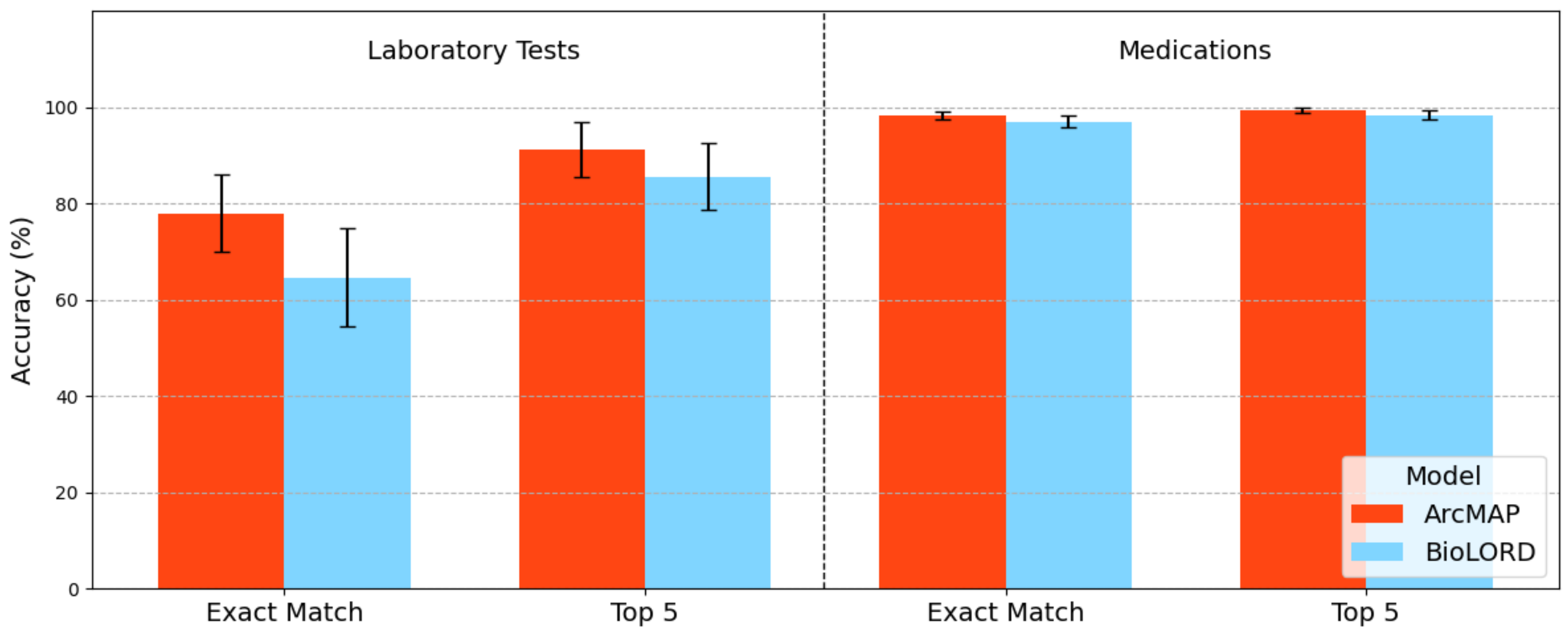
## Objectives

- Introduce **ArcMap**, a machine learning-based tool designed to streamline the medical concept standardisation process.
- **Reduce the manual burden** on clinical experts by providing automated suggestions for concept mappings.
- **Improve the efficiency and scalability** of data standardisation for RWE studies.
- **Enable continuous learning** within the tool to enhance accuracy and adaptability to new data sources over time.

## Results

- High Accuracy for Medications: ArcMap achieved an exact match accuracy of 98.20% for medication names, slightly outperforming BioLORD (96.99%).
- Superior Performance on Lab Tests: For laboratory test names, ArcMap showed a clear improvement with a top-5 accuracy of 91.07%, compared to 85.59% for BioLORD.
- Higher Variance in Labs arises from higher incidence of abbreviations and “codes” in lab tests compared to medication names. (e.g. UE LFT vs. paracetamol). Lab tests also varied more between trusts.
- Generalisation Across Trusts: Results were averaged across 4 NHS trusts, demonstrating ArcMap’s ability to generalise to unseen data sources.
- Improved Standardisation: ArcMap enables clinical experts to find the correct standardised concept in over 90% of cases from a shortlist of five, significantly accelerating the data standardisation process.

*Figure 2:* *Average prediction accuracies for ArcMAP and BioLORD models across both data modalities. Error bars show standard deviation across each NHS trust*



## Conclusion

- ArcMap significantly advances the data standardisation process by automating the mapping of medical concepts to standard vocabularies, reducing reliance on manual expert review.
- Scalability and adaptability: ArcMap generalises well across different NHS trusts and improves over time through continuous learning, making it suitable for large-scale, multi-institutional RWD projects.
- Regulatory readiness: The tool ensures traceability of mappings, supporting its use in regulatory-grade real-world evidence (RWE) studies.
- Impact on healthcare research: By accelerating and improving data standardisation, ArcMap enables more robust and timely insights from RWD, ultimately contributing to better patient outcomes and evidence-based decision-making.

## Outlook

- **Implement continuous learning pipeline** to iteratively improve model accuracy
- Extend approach to **further data modalities**

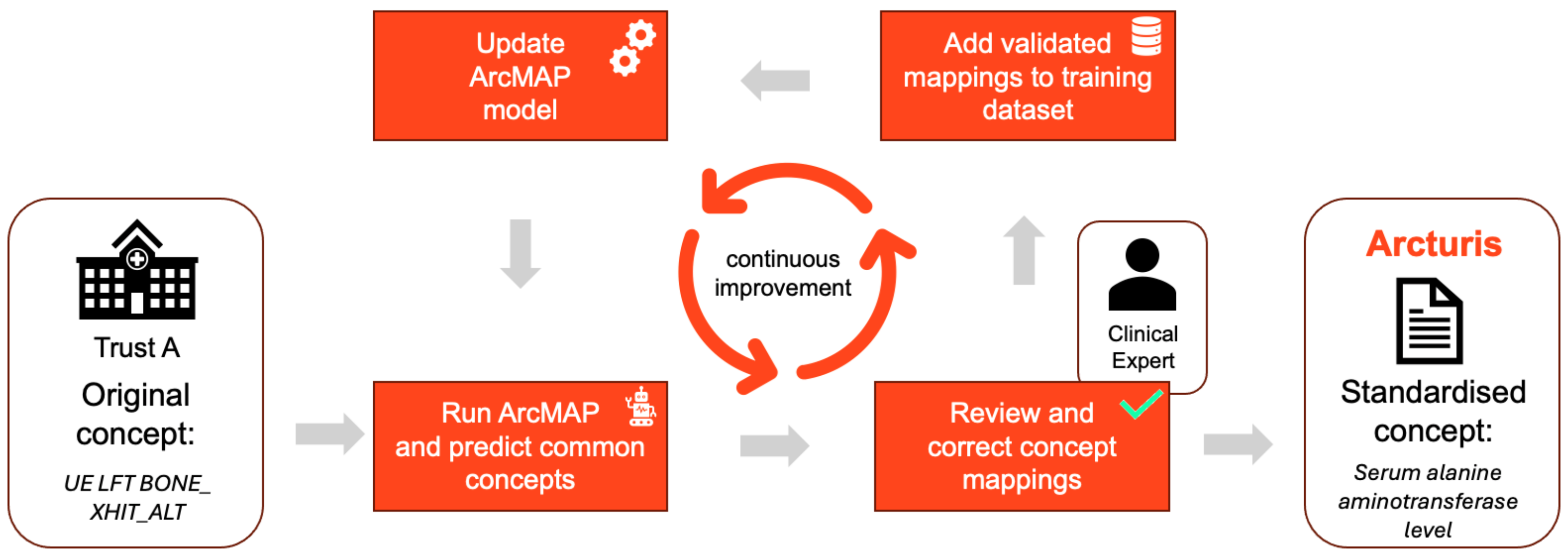

Trust A
Original concept:
UE LFT BONE_ XHIT_ALT
Run ArcMAP and predict common concepts
Review and correct concept mappings
Clinical Expert
Add validated mappings to training dataset
Update ArcMAP model
continuous improvement
Arcturis
Standardised concept:
Serum alanine aminotransferase level

*Figure 3:* *Schematic showing proposed continuous learning process*

## Acknowledgements & References

This study uses anonymised clinical data collected by NHS centres as part of routine care. The contributing centres to this study were Oxford University Hospitals, Chelsea and Westminster Hospital, Milton Keynes University Hospital and Hampshire Hospitals NHS Foundation Trusts. The study was submitted to and approved by the Research Oversight Subcommittee of the Arcturis Real-World Data Network research database under NHS Research Ethics Committee (REC) approval 24/YH/0164.

1. Kersloot, M.G., van Putten, F.J.P., Abu-Hanna, A. et al. Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. J Biomed Semant 11, 14 (2020). https://doi.org/10.1186/s13326-020-00231-z
2. Hier et al. “Efficient Standardization of Clinical Notes using Large Language Models”, arXiv: 2501.00644, 2024
3. Remy et al.: “BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights”, arXiv: 2311.16075, 2023
4. SNOMED CT codes, NHS England, https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct
5. Dictionary of medicines and devices (DM+D), NHS England, https://digital.nhs.uk/services/terminology-and-classifications/dm-d