RWD62

Development and Validation of an Automated Cancer Stage Classifier for Real-World Oncology Data Mapped to the OMOP Common Data Model

Fabienne Ver Donck[1], Dries Hens[1], Clara L. Oeste[1], Annelies Verbiest[2]

[1]LynxCare Clinical Informatics, Leuven, Belgium; [2]University Hospital Antwerp, Edegem, Belgium; on behalf of the FAIR-ICI project authors

BACKGROUND & AIMS

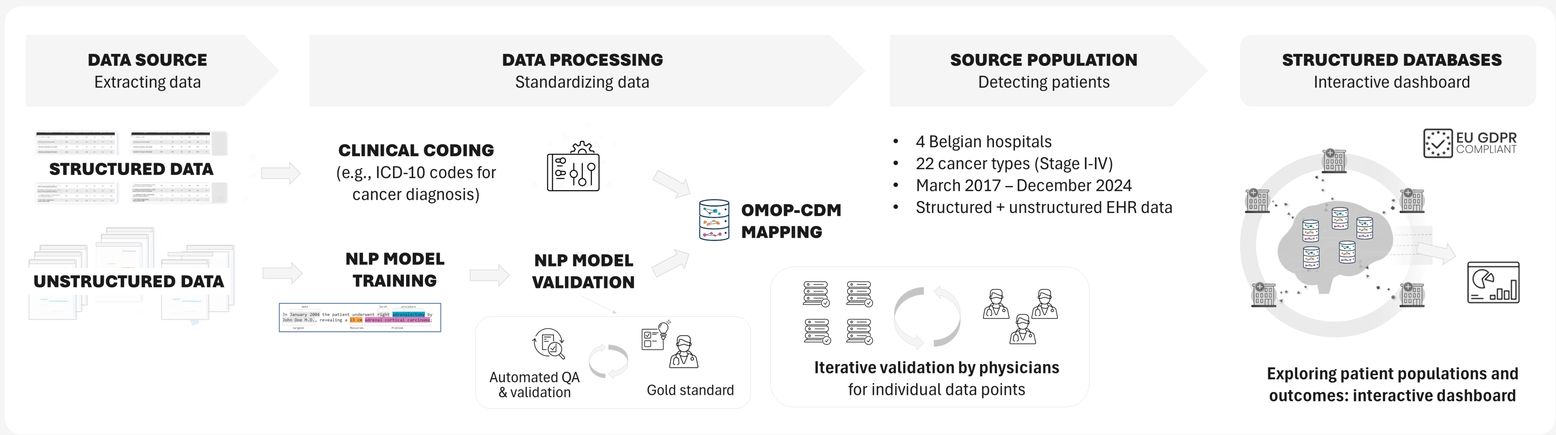
- **Cancer stage** is critical for oncology RWE but is often **missing/inconsistent in EHRs**.
- Heterogeneous recording across sites/tumor types limits **comparability and reproducibility.**
- **Mapping to OMOP CDM** enables harmonized inputs from structured fields and NLP-extracted text.

Aims:

- Develop and validate an **automated, rule-based cancer stage classifier** covering **>20 cancer types.**
- Integrate **structured TNM and NLP-derived** evidence to maximize completeness.
- Encode **UICC 8th edition** staging per cancer type with **pathological > clinical** precedence and inference from **partial TNM data sources.**

METHODS

- **Population:** n=3,231 cancer patients initiating ICI (2017–2024) at 4 Belgian hospitals.
- **Inputs:** structured TNM, structured metastatic entries, and free text (NLP) from EHRs.
- **Metastasis definitions:** metastatic (Stage IV or documented metastasis) vs non-metastatic (Stage I–III, no metastasis).
- **Validation:** iterative refinement with oncologist review across tumor types.
- **Missing data–tolerant TNM:** stage resolves with incomplete T/N/M when the missing part does not affect staging (e.g., any T, N3M0 ⇒ Stage III in breast cancer).



RESULTS

Figure 1. Staging of the ICI-treated population

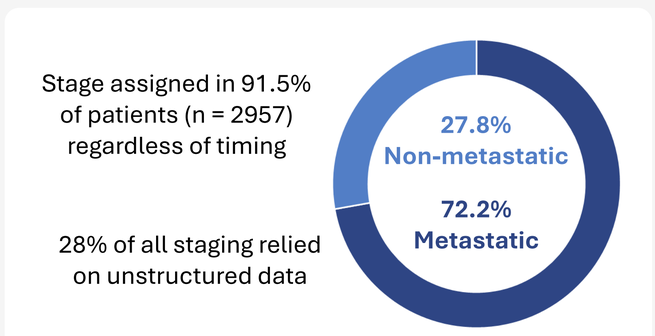


Figure 2. Data sources to derive staging

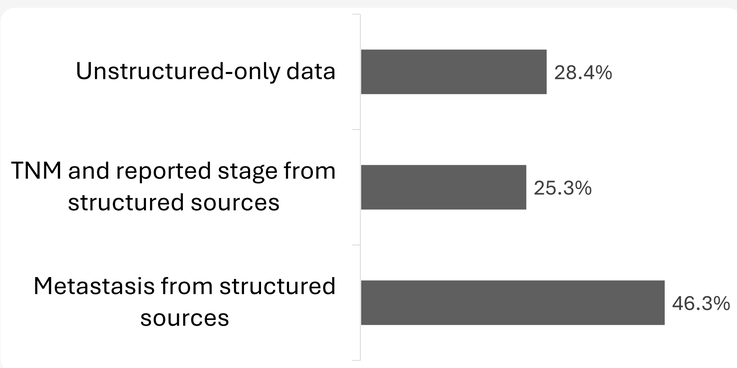


Figure 3. Truncated TNM availability and distribution (dashboard view). Counts for clinical (cT, cN, cM) and pathological (pT, pN, pM) components, plus source-unspecified T/N/M. Categories are truncated to 0–4 with an unknown category. Bars show patient counts.

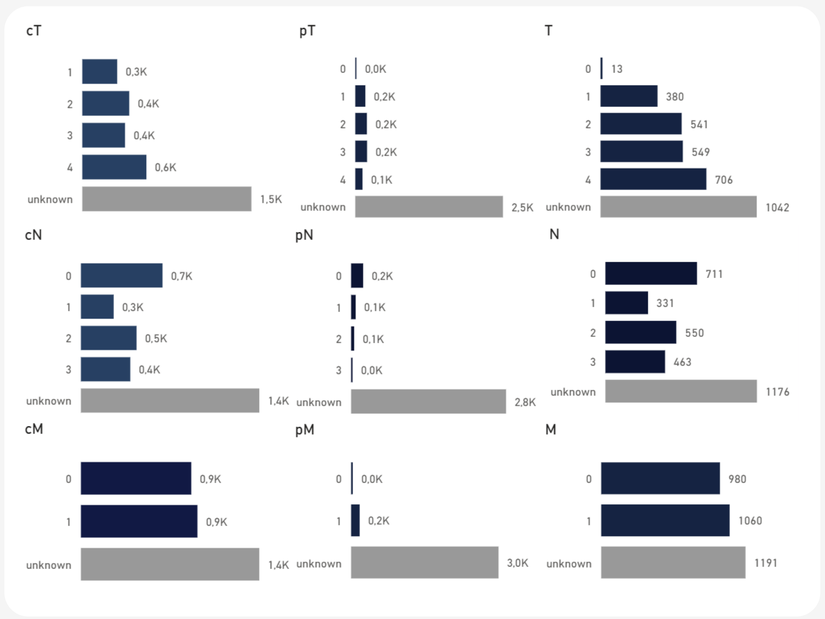
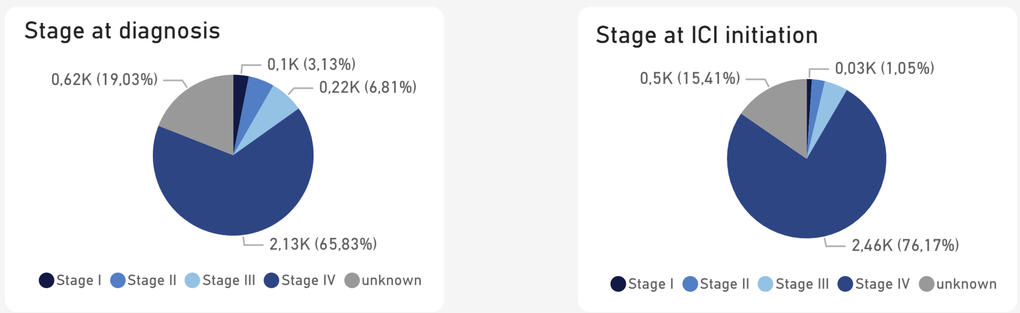


Figure 4. Stage distributions at two time points (dashboard view). Pie charts showing the proportion of stage I-IV and unknown for (left) stage **at diagnosis** and (right) stage **at ICI initiation.**



Classifier governance

- **Priority:** Metastasis → Structured staging → NLP staging.
- **Diagnosis stage:**
  - Metastasis ≤60 d post-diagnosis or before.
  - Else: Structured staging → NLP staging.
- **Stage at ICI initiation:**
  - Metastasis pre-ICI initiation.
  - Else: highest stage from structured staging, NLP staging within −120 to +60 d of ICI initiation.
  - Metastasis ≤60 d post-ICI initiation.
- **Traceability:** Each stage assignment retains provenance tags (structured/NLP/mixed; inferred vs explicit).

CONCLUSIONS

- A **scalable, interpretable classifier** can standardize stage assignment across **>20 cancers**.
- Combining structured + NLP-derived inputs within OMOP CDM markedly **improves completeness and reproducibility.**
- Code and logic rules will be made publicly available, supporting **reuse across sites**.

Implications:

- Facilitates **harmonized staging for federated networks** and multi-center RWE.
- **Reduces bias from missing/incomplete TNM** and enhances study generalizability.

Acknowledgements: We'd like to thank the lead physicians and their teams (Dr. Verbiest at UZA, Prof. Vulsteke at AZMM, Prof. Debruyne at AZ Groeninge, Prof. Aspeslagh at UZ Brussel) participating in the FAIR-ICI project for their contributions, collaboration, and clinical validation support.

Contact: Fabienne Ver Donck, LynxCare Clinical Informatics, Leuven, Belgium.
fabienne.verdonck@lynx.care