AI-CARE ETL PIPELINE: FOUNDATION FOR NATIONAL AND INTERNATIONAL AI APPLICATIONS

MARKUS SAUERBERG, SEBASTIAN GERMER, CHRISTIANE RUDOLPH, HENRIK KUSCHE, AND THE AI-CARE WORKING GROUP

GMDS Annual Meeting 2025

Bundesministerium für Gesundheit

DLR Projektträger

AI-CARE

HAMBURGISCHES KREBSREGISTER

9th of September 2025 | Jena

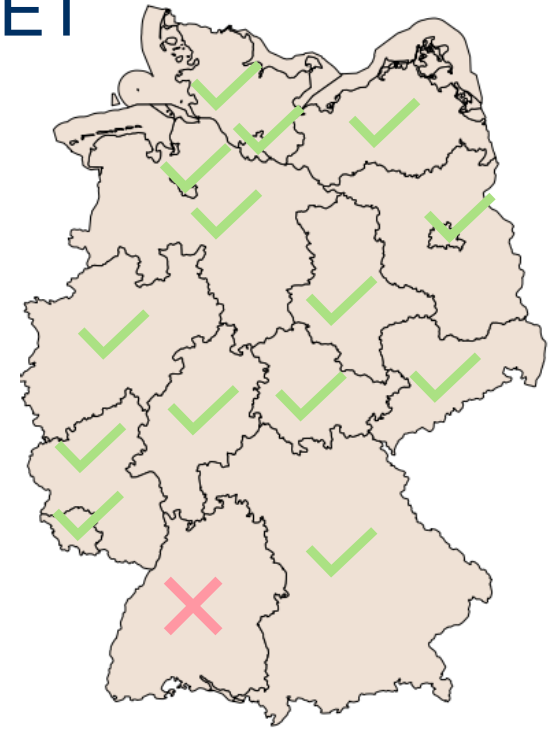
Hamburg

AI-SUPPORTED, CARE-ORIENTED USE OF CANCER REGISTRY DATA – AI-CARE

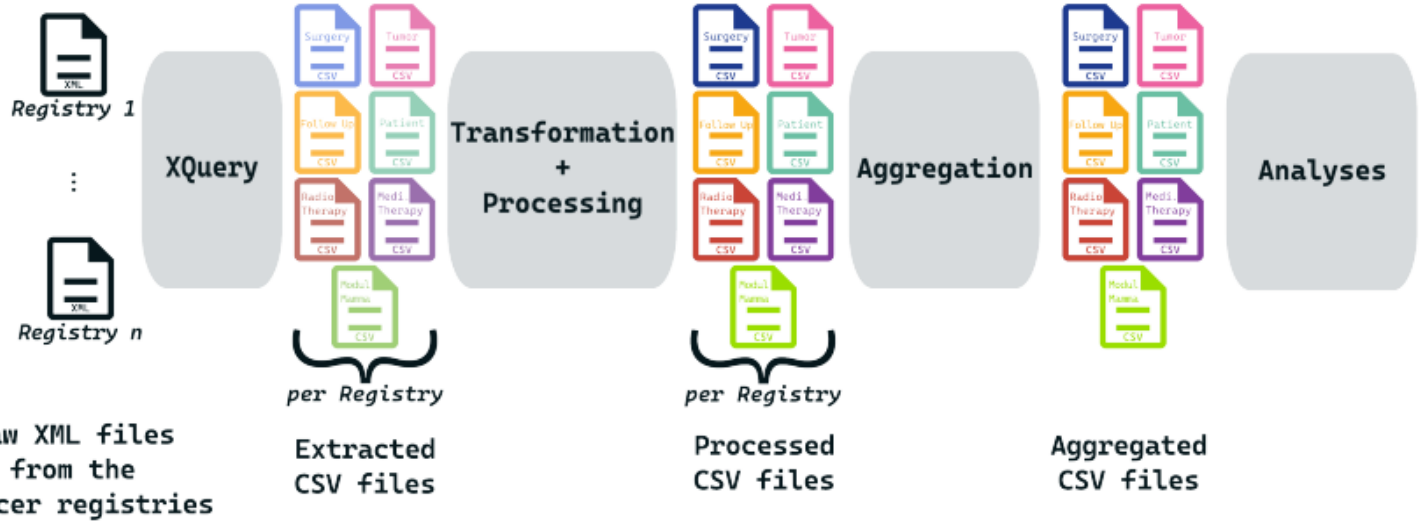
- In collaboration with 12 project partners, cancer registry data is being collected, harmonized, and analyzed
- The project is part of the research initiative *Krebsregisterdaten zusammenführen und intelligent nutzen: Innovationsprojekte für Künstliche Intelligenz*, funded by the Federal Ministry of Health
- Project duration: 01.09.2022 to 31.08.2025

THE ETL-PIPELINE REFINES STATE-SPECIFIC DATASETS AND INTEGRATES THEM INTO A HARMONIZED, NATIONWIDE DATASET

- Data obtained from 14 cancer registries in Germany (oBDS format)
- Data refers to the period 2014 to 2024
- Focuses on four cancer entities (lung, thyroid, breast, and non-Hodgkin lymphoma)



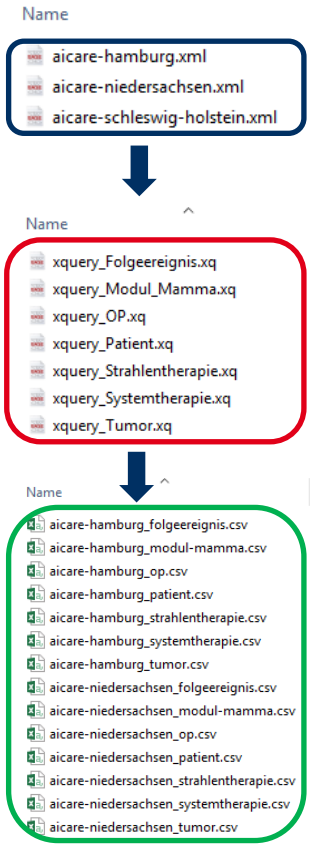
Example transformation:
1A ⟷ 1a for TNM
C34.19 ⟷ C43.1 for ICD10
15154 ➡ 15082 for Inzidenzort

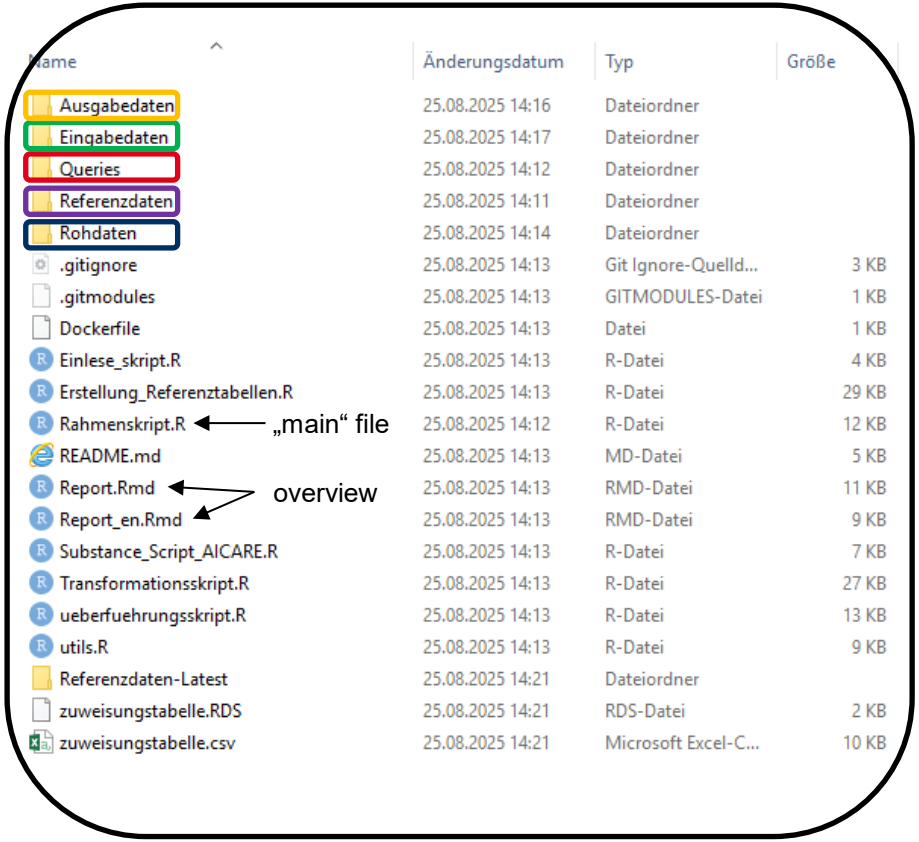


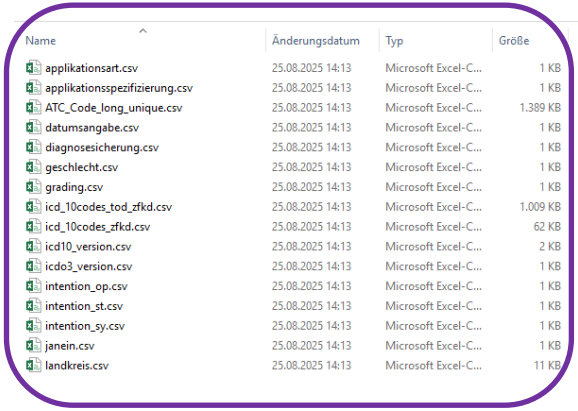
Interface: „ZfKD Lieferdatensatz“

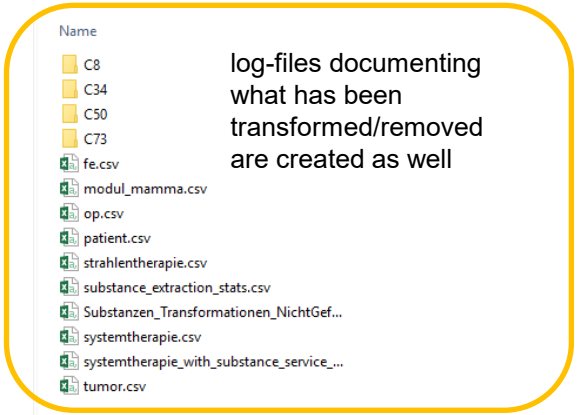
Matching with reference tables

“Ready-to-analyze” dataset that includes state-specific cancer registry data


Name
aicare-hamburg.xml
aicare-niedersachsen.xml
aicare-schleswig-holstein.xml
Name
xquery_Folgeereignis.xq
xquery_Modul_Mamma.xq
xquery_OP.xq
xquery_Patient.xq
xquery_Strahlentherapie.xq
xquery_Systemtherapie.xq
xquery_Tumor.xq
Name
aicare-hamburg_folgeereignis.csv
aicare-hamburg_modul-mamma.csv
aicare-hamburg_op.csv
aicare-hamburg_patient.csv
aicare-hamburg_strahlentherapie.csv
aicare-hamburg_systemtherapie.csv
aicare-hamburg_tumor.csv
aicare-niedersachsen_folgeereignis.csv
aicare-niedersachsen_modul-mamma.csv
aicare-niedersachsen_op.csv
aicare-niedersachsen_patient.csv
aicare-niedersachsen_strahlentherapie.csv
aicare-niedersachsen_systemtherapie.csv
aicare-niedersachsen_tumor.csv


Name
Änderungsdatum
Typ
Größe
Ausgabedaten 25.08.2025 14:16 Dateiordner
Eingabedaten 25.08.2025 14:17 Dateiordner
Queries 25.08.2025 14:12 Dateiordner
Referenzdaten 25.08.2025 14:11 Dateiordner
Rohdaten 25.08.2025 14:14 Dateiordner
.gitignore 25.08.2025 14:13 Git Ignore-Quelld... 3 KB
.gitmodules 25.08.2025 14:13 GITMODULES-Datei 1 KB
Dockerfile 25.08.2025 14:13 Datei 1 KB
Einlese_skript.R 25.08.2025 14:13 R-Datei 4 KB
Erstellung_Referenztabellen.R 25.08.2025 14:13 R-Datei 29 KB
Rahmenskript.R 25.08.2025 14:12 R-Datei 12 KB
README.md 25.08.2025 14:13 MD-Datei 5 KB
Report.Rmd 25.08.2025 14:13 RMD-Datei 11 KB
Report_en.Rmd 25.08.2025 14:13 RMD-Datei 9 KB
Substance_Script_AICARE.R 25.08.2025 14:13 R-Datei 7 KB
Transformationsskript.R 25.08.2025 14:13 R-Datei 27 KB
ueberfuehrungsskript.R 25.08.2025 14:13 R-Datei 13 KB
utils.R 25.08.2025 14:13 R-Datei 9 KB
Referenzdaten-Latest 25.08.2025 14:21 Dateiordner
zuweisungstabelle.RDS 25.08.2025 14:21 RDS-Datei 2 KB
zuweisungstabelle.csv 25.08.2025 14:21 Microsoft Excel-C... 10 KB
„main“ file
overview


Name
Änderungsdatum
Typ
Größe
applikationsart.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
applikationsspezifizierung.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
ATC_Code_long_unique.csv 25.08.2025 14:13 Microsoft Excel-C... 1.389 KB
datumsangabe.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
diagnosesicherung.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
geschlecht.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
grading.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
icd_10codes_tod_zfkd.csv 25.08.2025 14:13 Microsoft Excel-C... 1.009 KB
icd_10codes_zfkd.csv 25.08.2025 14:13 Microsoft Excel-C... 62 KB
icd10_version.csv 25.08.2025 14:13 Microsoft Excel-C... 2 KB
icdo3_version.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
intention_op.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
intention_st.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
intention_sy.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
janein.csv 25.08.2025 14:13 Microsoft Excel-C... 1 KB
landkreis.csv 25.08.2025 14:13 Microsoft Excel-C... 11 KB


Name
C8
C34
C50
C73
fe.csv
modul_mamma.csv
op.csv
patient.csv
strahlentherapie.csv
substance_extraction_stats.csv
Substanzen_Transformationen_NichtGef...
systemtherapie.csv
systemtherapie_with_substance_service_...
tumor.csv
log-files documenting what has been transformed/removed are created as well

Statistics on Output Data

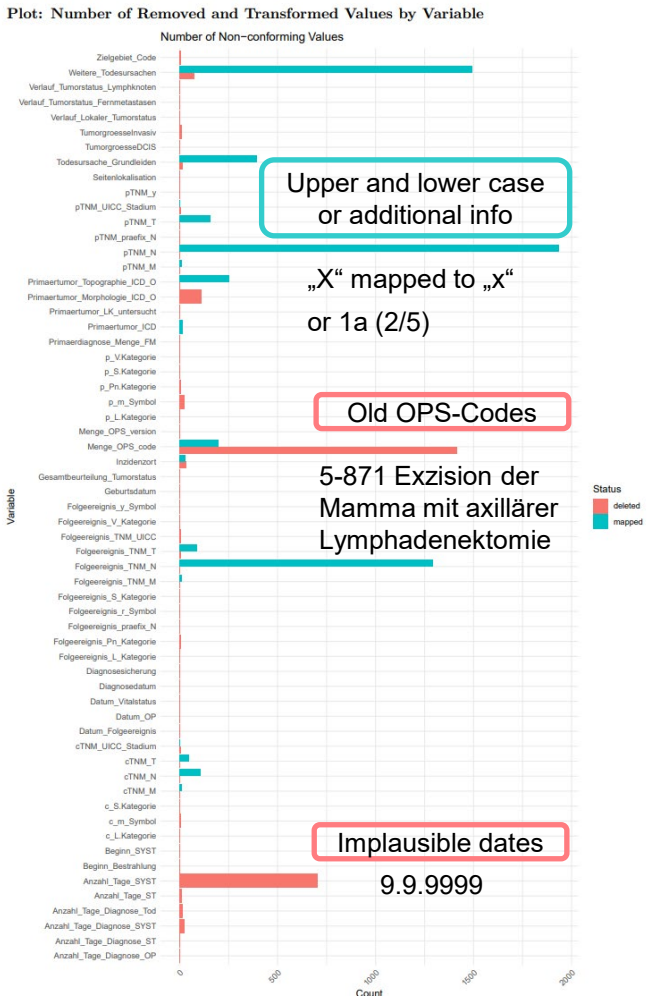
Table: Number of Cases by Module and Entity

Module	Total	Lung Cancer	Breast Cancer	Thyroid Cancer	Non-Hodgkin Lymphoma
fe	1518227	467974	847553	46659	155832
modul_mamma	388817	25	388594	0	168
op	597293	115720	428505	31925	21098
patient	1055118	395382	504320	40022	124908
strahlentherapie	539415	169401	332137	19335	18514
systemtherapie	873565	288537	455217	1870	127895
tumor	1092215	403806	520891	40337	127099

Over 1 million patients and tumors

Total_Cases	Total_Removals	Total_Transformations	Unique_Cases	Unique_Removals	Unique_Transformations
372835	147713	225122	8636	2584	6052

More transformations than removals



PIPELINE WILL BE PUBLISHED SOON

- ETL-Pipeline produces a „ready for analysis“ dataset that has been used, for example, to conduct survival analysis for lung cancer patients
  - Germer, S., Rudolph, C., Katalinic, A., Rath, N., Rausch, K., & Handels, H. (2025). Lung Cancer Survival Estimation Using Data from Seven German Cancer Registries. Studies in Health Technology and Informatics, 327, 457-461
- Pipeline can be extended to more entities such as prostate cancer and colon cancer
- Substances and protocols are reported as free text, making harmonization more challenging
  - R and Python scripts using fuzzy matching lead to promising results and are included in the ETL-Pipeline → “spacy-matching” package on pypi.org

Universitätsklinikum Hamburg-Eppendorf

GMDS 2025

AI-CARE ETL Pipeline: Foundation for National and International AI Applications

Jasmin Carus[1] und Mareile Beernink[2]
[1] Universitäres Cancer Center Hamburg, UKE
[2] Institut für Angewandte Medizininformatik, UKE

Universitäres Cancer Center HAMBURG

gmds Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V.

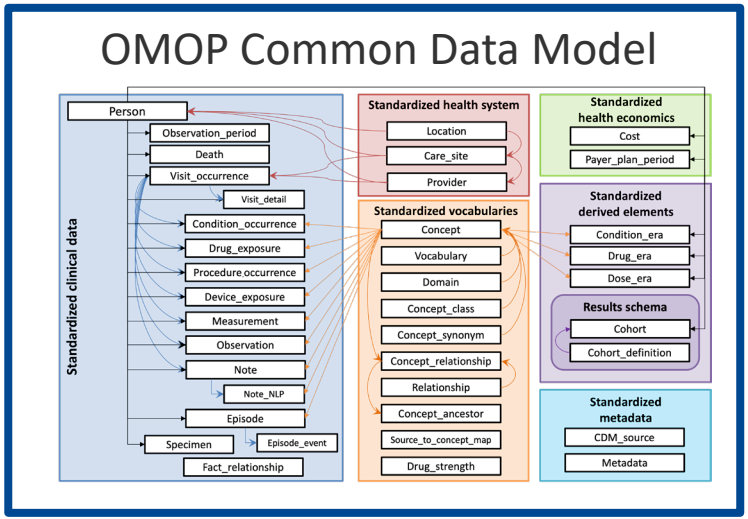
Universitätsklinikum Hamburg-Eppendorf

… from National to International …

Privacy Protection

Integrity

Availability

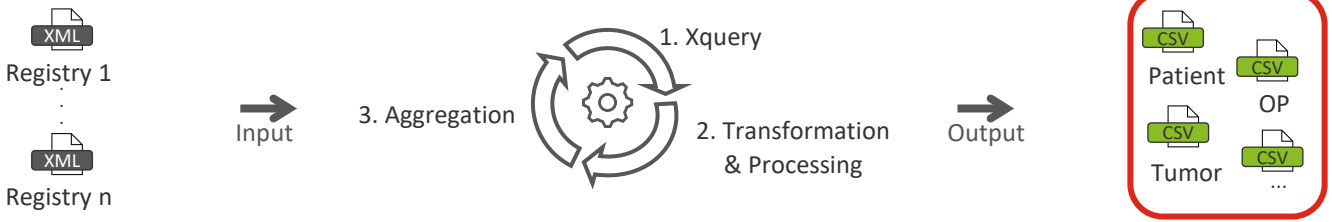


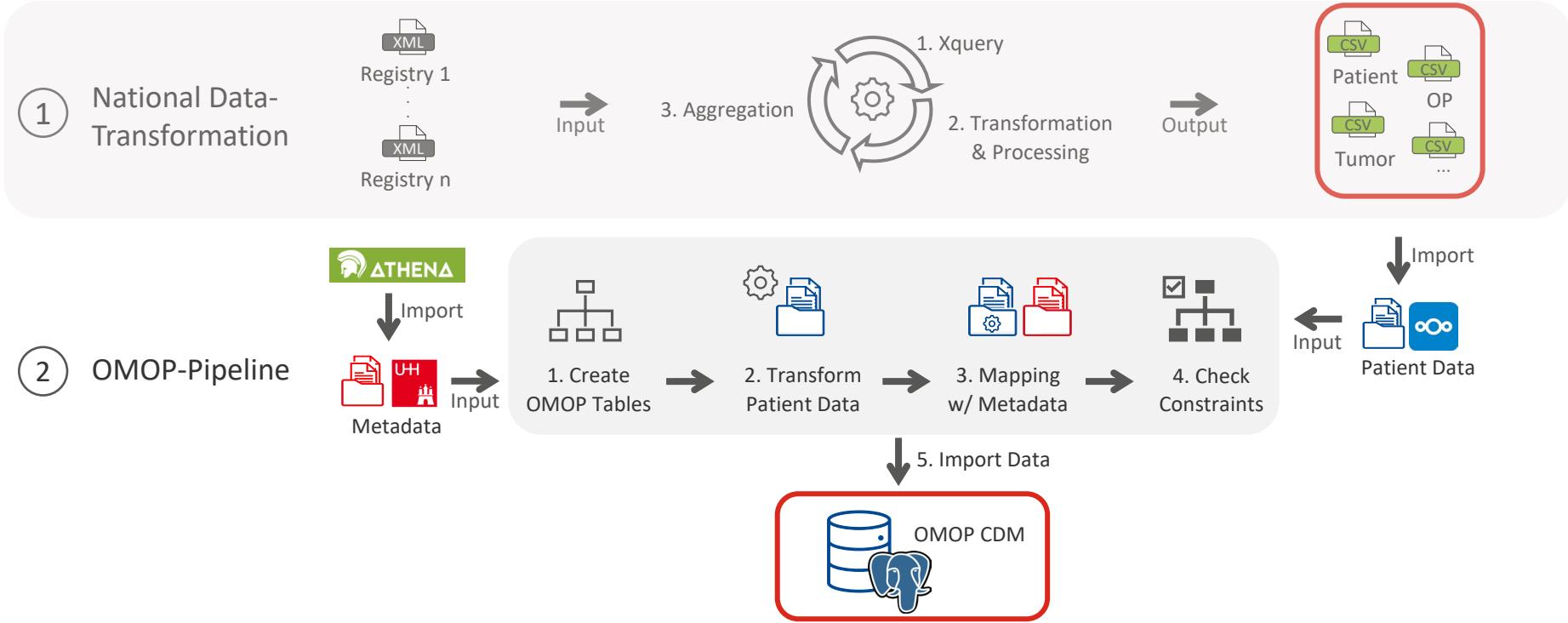
Big Data

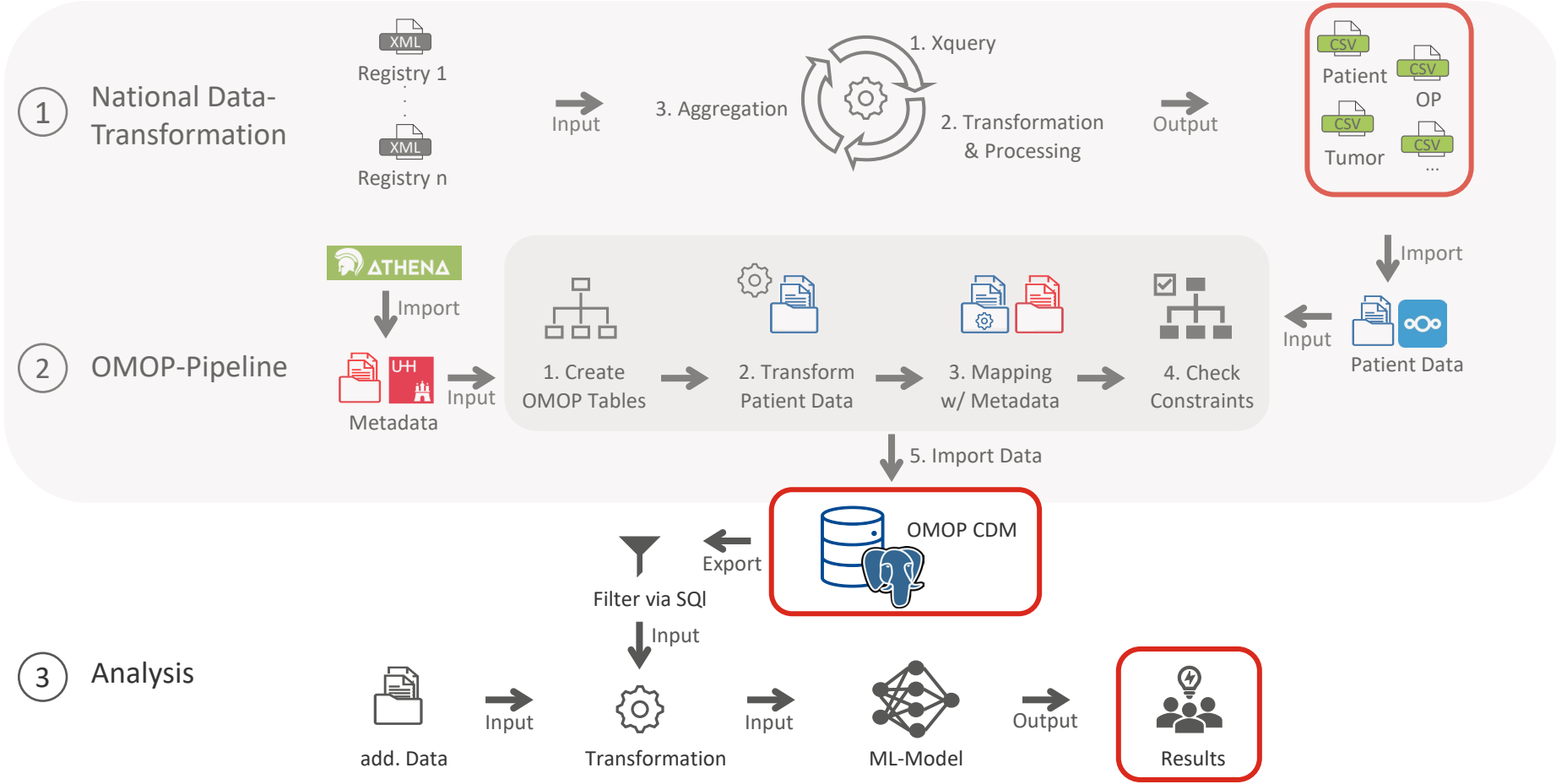
Rare Disease

New Technology

1 National Data-Transformation



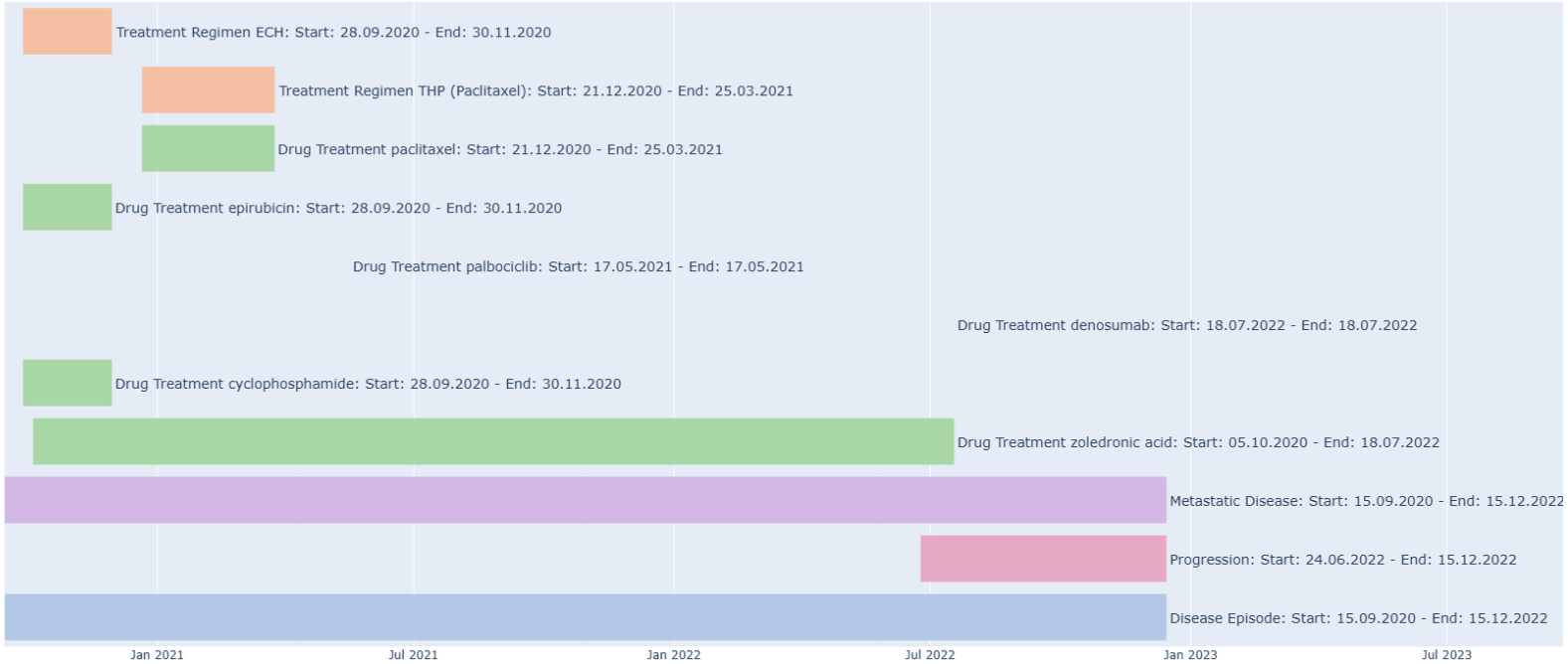

1
National Data-Transformation
XML
Registry 1
XML
Registry n
Input
3. Aggregation
1. Xquery
2. Transformation & Processing
Output
CSV
Patient
CSV
OP
CSV
Tumor
CSV
...
2
OMOP-Pipeline
ATHENA
Import
Metadata
Input
1. Create OMOP Tables
2. Transform Patient Data
3. Mapping w/ Metadata
4. Check Constraints
Import
Input
Patient Data
5. Import Data
OMOP CDM


1
National Data-Transformation
XML
Registry 1
XML
Registry n
Input
3. Aggregation
1. Xquery
2. Transformation & Processing
Output
CSV
Patient
CSV
OP
CSV
Tumor
CSV
...
2
OMOP-Pipeline
ATHENA
Import
UH
Metadata
Input
1. Create OMOP Tables
2. Transform Patient Data
3. Mapping w/ Metadata
4. Check Constraints
Input
Import
Patient Data
5. Import Data
OMOP CDM
Export
Filter via SQl
Input
3
Analysis
add. Data
Input
Transformation
Input
ML-Model
Output
Results

Patient 	1,055,118	1,055,178	≈99,9%

Patient	1,055,118	1,055,178	≈99,9%
Condition	2,143,812	SNOMED CT	ICD-O, ICD-10
Drugs	1,157,220	Guidelines for ATC classification and DDD assignment 2025 (NIPH)	
Procedure	3,014,164	SNOMED CT	OPS 2025 Systematisches Verzeichnis
Measurement & Observation	7,100,837	CANCER MODIFIER VOCABULARY, NAACCR	ICD-O, SNOMED CT

… Exploring **patient journey**



Episode

… Exploring **patient journey**

… **Patient-centered** AI-analysis:

- Patient prognosis: *Risk of recurrence, side effects, or survival probability*
- Digital Twin: *simulation of different treatment strategies*
- Next Best Action Recommendation: *personalized treatment strategies*

Episode

… Exploring cohort **journey**

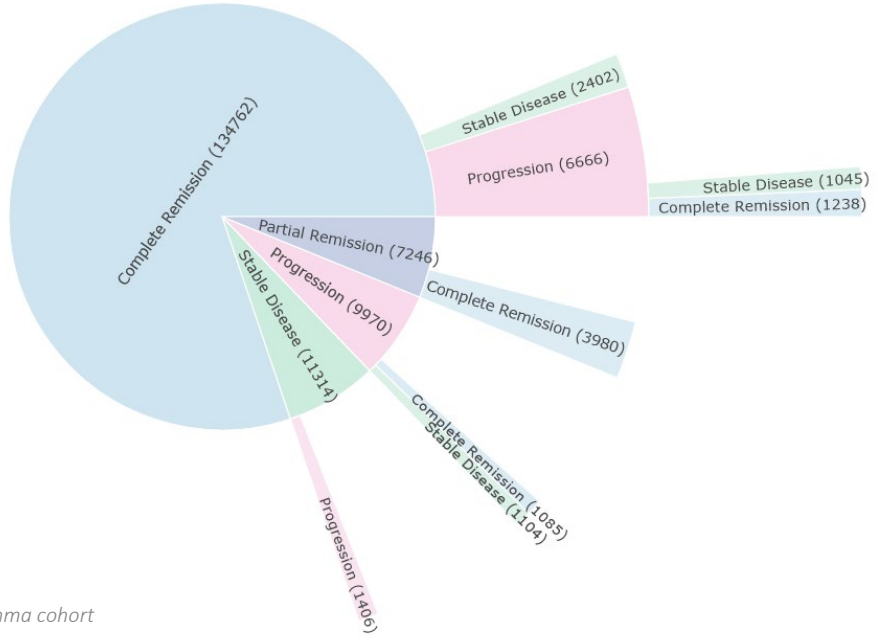


Fig1: Disease course of Mamma cohort after initial treatment

Cohort insights

Disease Dynamic

… Exploring cohort **journey**

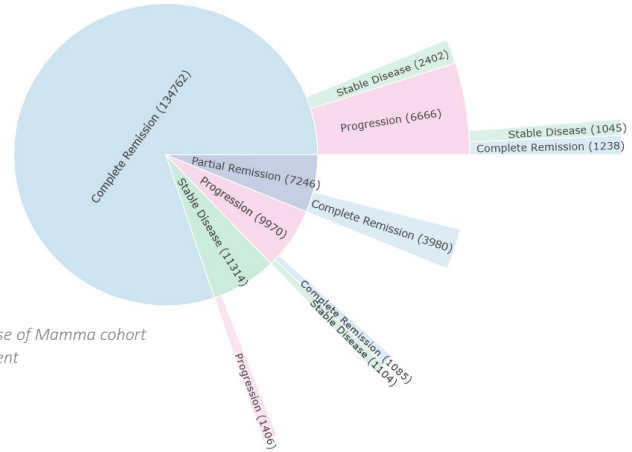


Fig1: Disease course of Mamma cohort after nitial treatment

… **Cohort-centered** AI-analysis:

Cohort insights

- Survival Analysis / Time-to-event Prediction: *Time to progression, recurrence, or death (overall survival, progression-free survival*)
- Clustering: *Identification of oncological subgroups*
- Tracteory Mining: *patterns in disease trajectories or treatment sequences*
- Treatment Response Prediction: *How likely is a cohort to respond to a specific treatment?*

Disease Dynamic

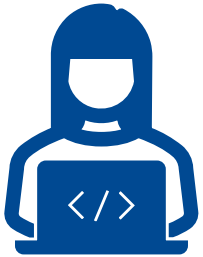
Regimen

Next Steps…

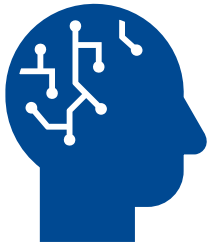
Collaboration

Validation

Publication

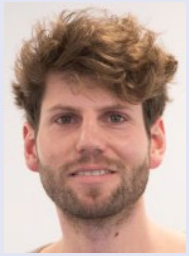


Development



AI-Analysis

Thank you for your Attention!
Questions?



Dr. Markus Sauerberg

markus.sauerberg@bwfgb.hamburg.de

Behörde für Wissenschaft, Forschung, Gleichstellung und Bezirke, Freie und Hansestadt Hamburg

Jasmin Carus

j.carus@uke.de

Universitäres Cancer Center Hamburg, UKE

Mareile Beernink

m.beernink@uke.de

Institut für Angewandte Medizininformatik, UKE