Merging and Validation of Cancer Registry Data using AI

ZuVaKI

WORKSHOP 4: IT tools and novel AI approaches for cancer registration

Philipp Leppert
IDG Institut für digitale Gesundheitsdaten RLP gGmbH

13.11.2023

Gefördert durch:

Bundesministerium für Gesundheit

aufgrund eines Beschlusses des Deutschen Bundestages

European Commission

Project Members

Krebsregister Rheinland-Pfalz
im Institut für digitale Gesundheitsdaten

Deutsches Kinderkrebsregister

kkn
klinisches krebsregister niedersachsen

JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Hessisches Krebsregister

Krebsregister Baden-Württemberg

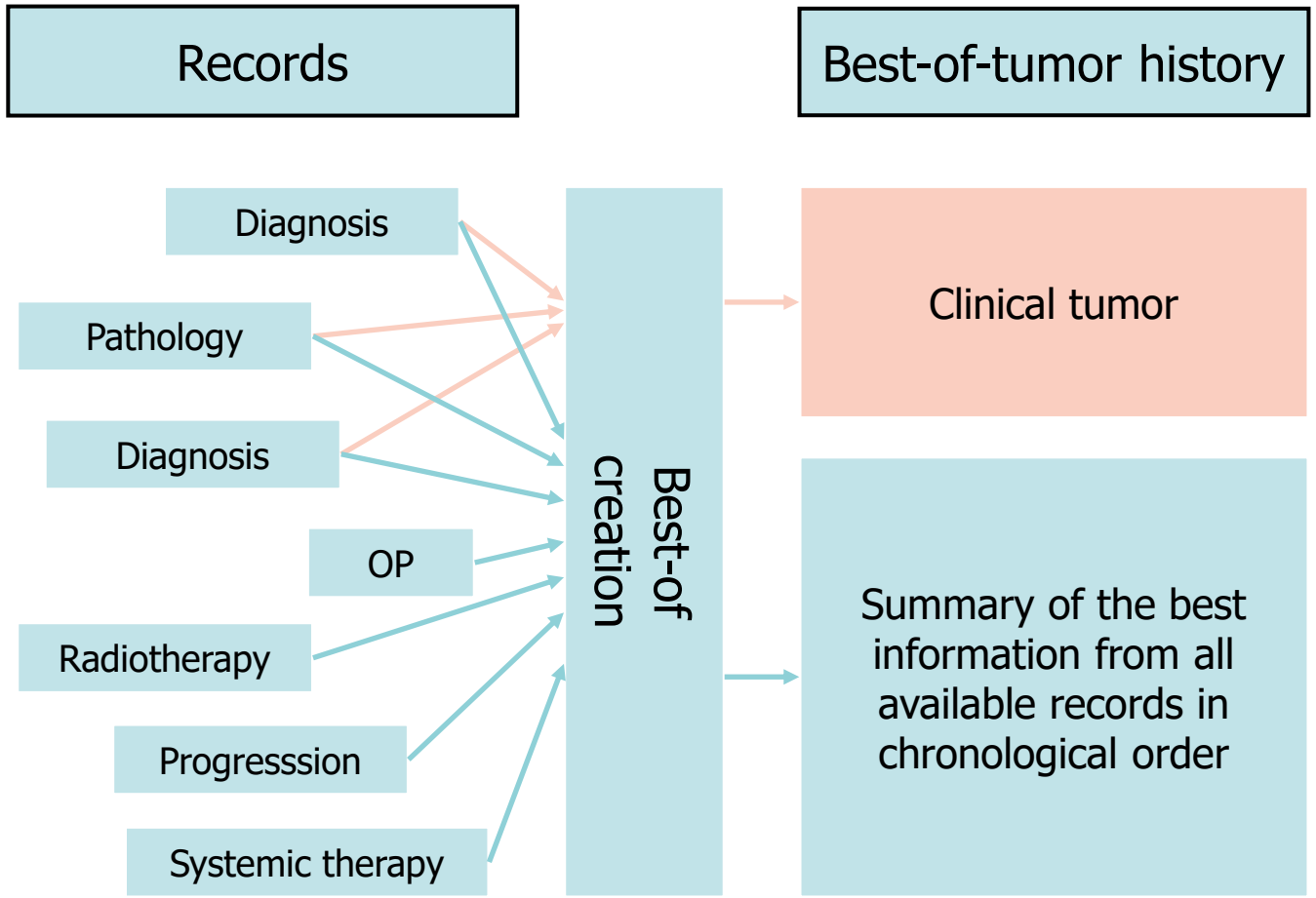
Project Scope

Fields of investigation

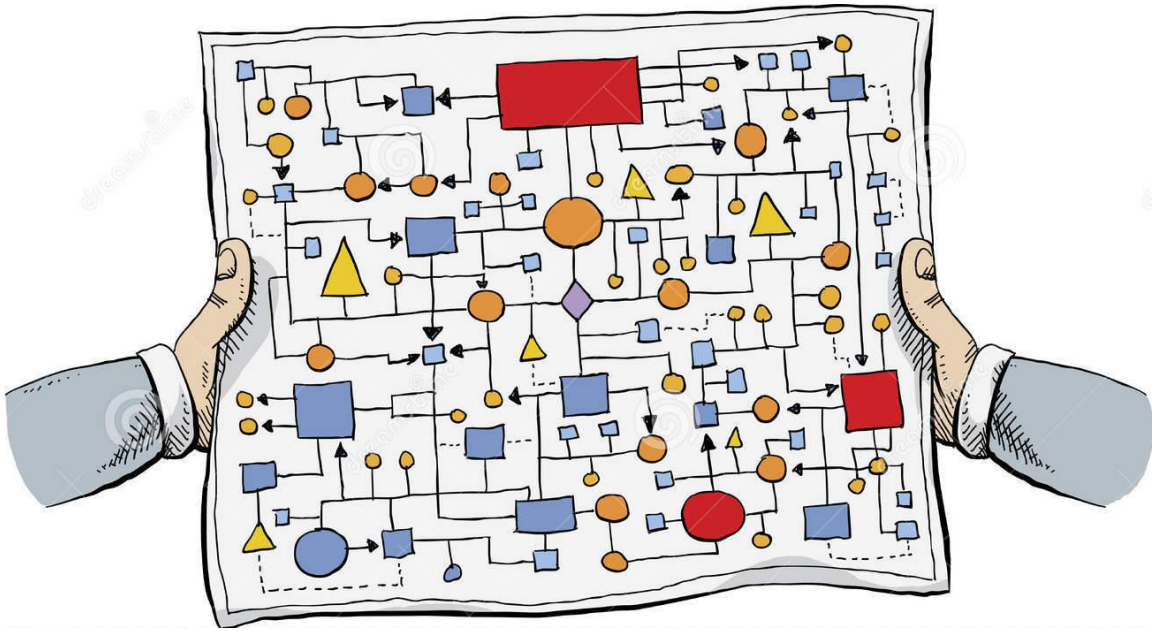
- Identification of conspicuous records in cancer registry data (anomaly detection)
- Merging of possibly contradictory information on tumour diseases into a best-of tumour history (record fusion)

- **Further information (only in German)**
  - https://zuvaki.de
  - Project profile @BMG

Best-of Creation (Record Fusion)



- in part already automated, but mostly done manually



https://www.dreamstime.com/stock-illustration-complicated-flowchart-cartoon-two-arms-holding-tangled-flow-chart-paper-image77231017

Best-of Creation via AI

Goal

- Harmonising information from different records into a valid best-of tumour history through AI procedures

How?

- Supervised learning
  - e.g. Random forest

Labels?

- Manually created best-of datasets

Approaching Quality Assurance

	Targeted quality assurance (queries / plausibility checks)	**Untargeted quality assurance** (data driven)
Advantages	• Find only actual quality problems • Type of quality problem known	• Find quality problems without restrictions • Find quality problems with complex interrelationships
Limitations	• Only find quality problems we are looking for • Difficult to map complex relationships	• Not all cases are quality problems • Type of quality problem unknown

Quality Assurance via AI (Anomaly Detection)

Goal

- Finding unusual records through anomaly detection procedures at the domain of:
  - records, tumours, patients

How?

- Unsupervised learning
- Categorical data
  - e.g. Autoencoder, Frequent Pattern Based Outlier Detection

Area of application

- After Record import? After record processing? In between?

Challenges (so far)

Data from state cancer registries

- Different IT infrastructure and data management
- Different legal foundation

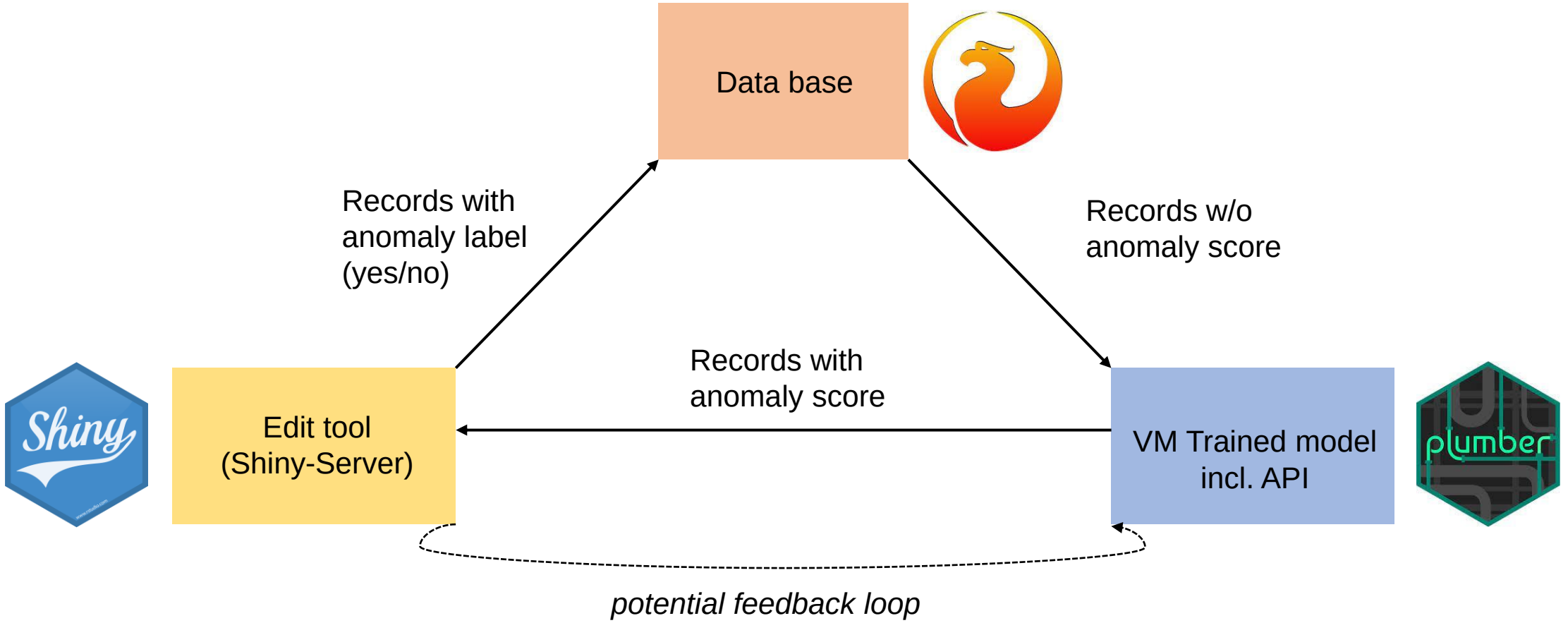
German Childhood Cancer Registry (DKKR)

- Different data basis than the other state cancer registries

Domain knowledge for validation of AI procedures required

- Especially in the context of anomaly detection (unsupervised learning)

Hands-on: Open-Source Validation Approach



Read More

- Röchner, P., Rothlauf, F.
  Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries.
  BMC Med Res Methodol 23, 125 (2023).
  https://doi.org/10.1186/s12874-023-01946-0

or via web browser: tiny.cc/zuvaki

www.encr.eu