

Harmonising Scotland's Cancer Data

Quantifying concordance across national datasets

Aims:

- Understand why data conflicts arise when CIP datasets overlap
- Create metadata to assist CIP users

Objectives

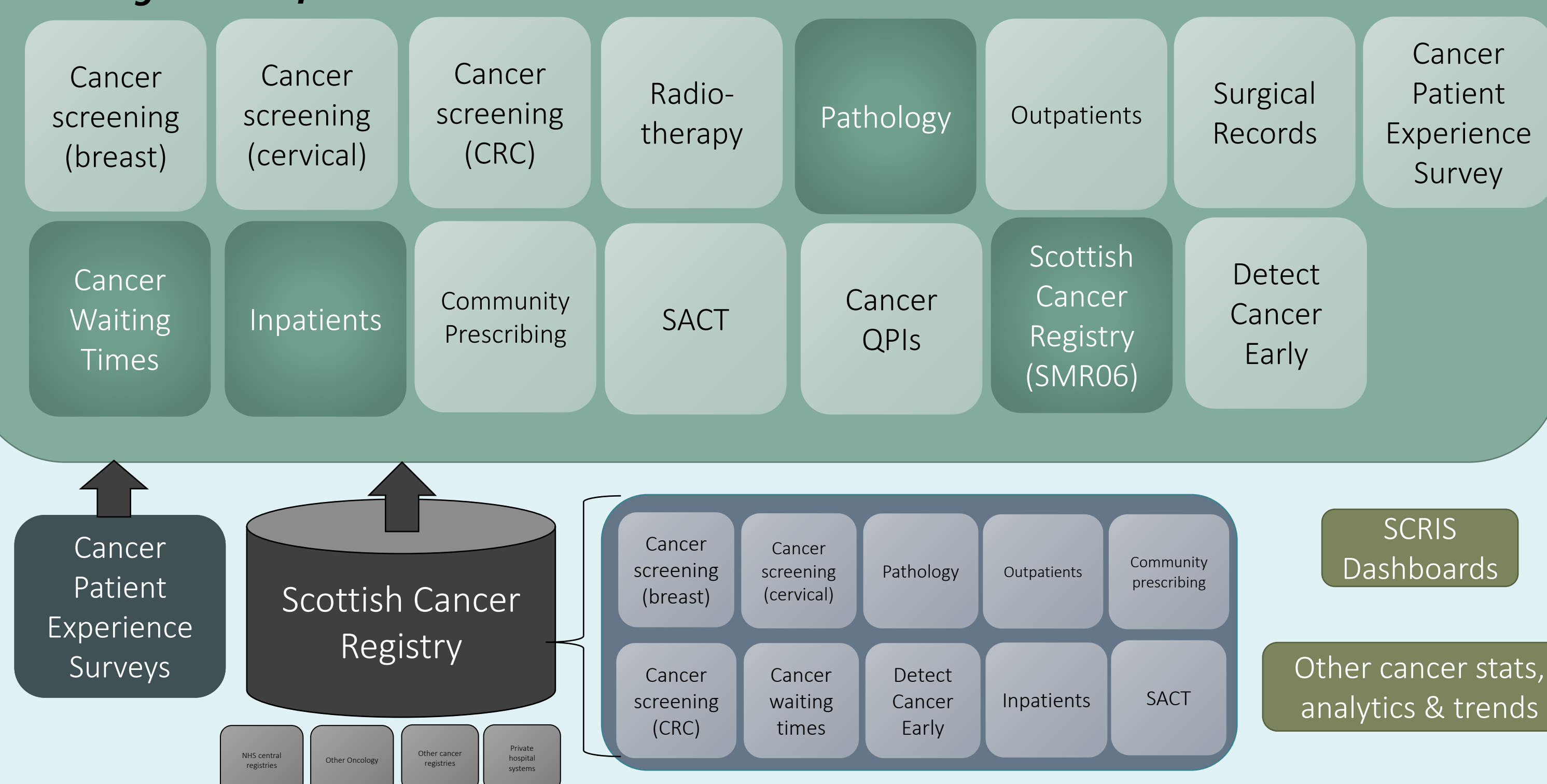
- Quantify discordance across datasets
- Demonstrate discordance on Scottish Cancer Registry and QPI datasets
- Design a reproducible protocol that can be expanded across CIP

Conclusion

- Data linkage within the CIP will highlight discordance between datasets
- Clinicians & Data Scientists must collaborate to ensure users can access accurate data.
- Data harmonisation using standardised protocols should be developed for use on all CIP datasets

Figure 1: the Cancer Intelligence Platform

CANCER INTELLIGENCE PLATFORM (CIP)



Implications for Data Science within the Scottish Cancer Intelligence Platform

- Automated harmonisation
- Rapid production of metadata for users
- Triaging for validation
- Scalable protocol for remaining CIP datasets

Developing and Testing the Harmonisation Protocol

1. Review data dictionaries

Dictionaries for Cancer Registry and QPI datasets reviewed

2. Identify overlapping domains

Fourteen domains recording similar but differently coded data were identified

3. Define 'agreement' vs 'conflict'

A set of logical rules were devised to discern concordance and discordance between datasets. These are stored as JSON files

4. Create linked datasets

We identified 34,744 cases of a first episode of colorectal cancer in the cancer registry which had a matching record in the QPI dataset.

5. Construct logic engines.

JSON Logic and the Validate package (R) were used to 'test' the matched records against the rules to determine concordance between datasets.

6. Generate summary statistics

Logical 'ruleset' stored in JSON files

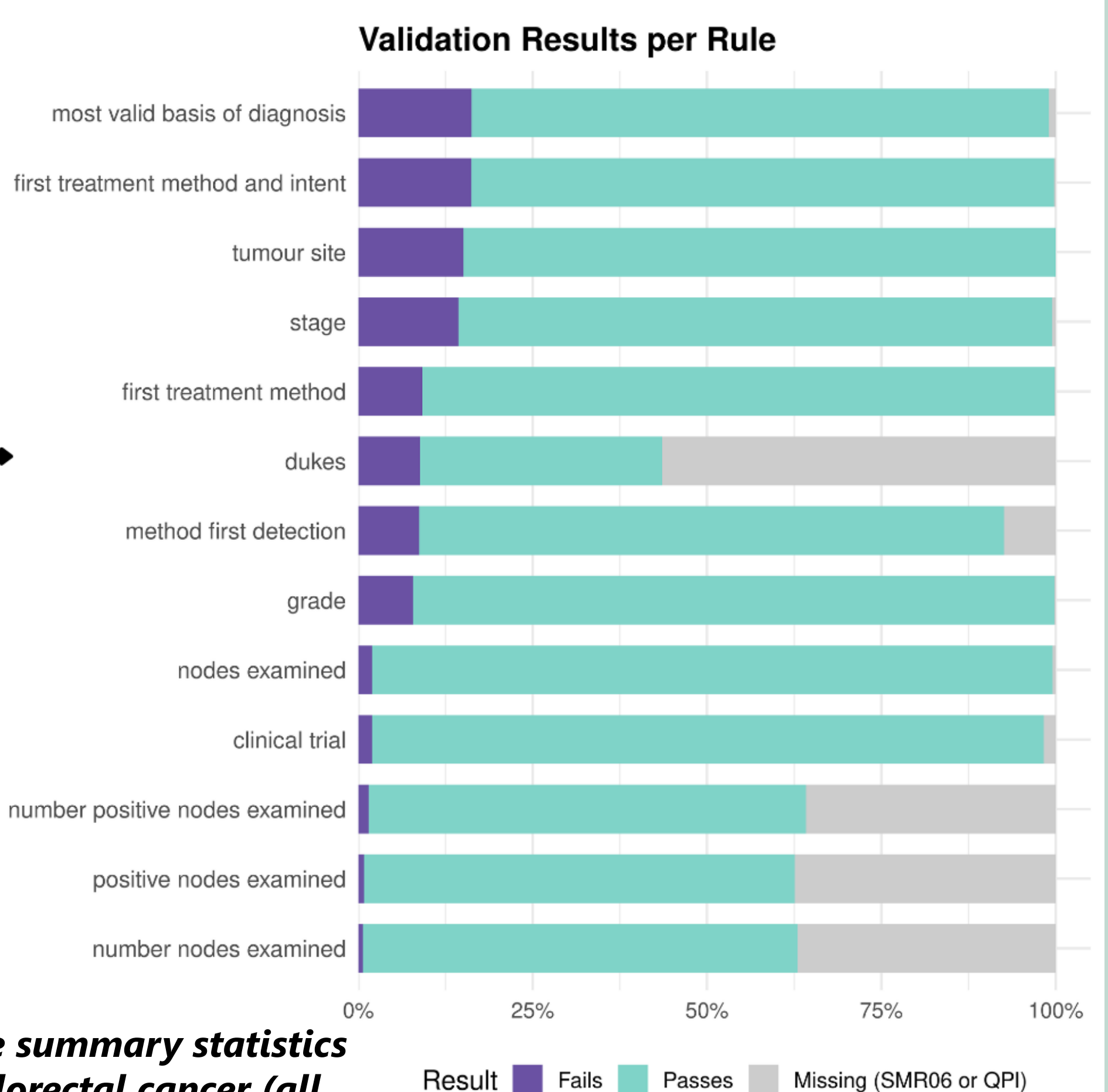
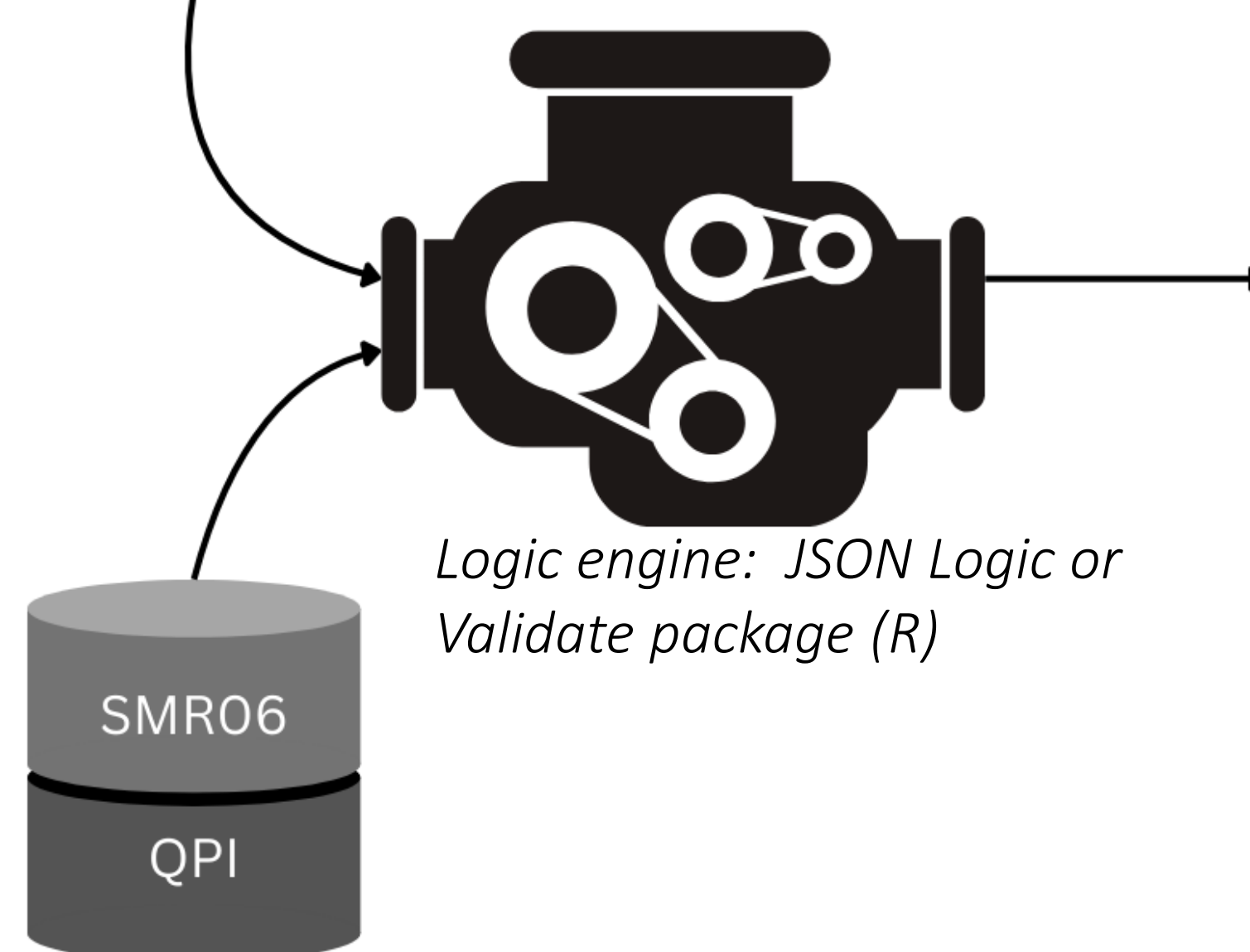


Figure 2: using the harmonisation protocol to produce summary statistics of concordance between SMR06 and QPI records of colorectal cancer (all domains)

Example: 'Mode of Referral'

A closer look at this domain shows:

- 83.2% overall agreement: Of 34,744 matched records, 28,903 were concordant on mode of referral
- Of the 5,841 discordant records, the majority of discordance arose from missing data in the QPI
- For all cases of missing data in QPI, a valid 'non missing' code was present in SMR06, highlighting the potential for data supplementation

% Agreement of SMR06 with QPI: Mode of Referral

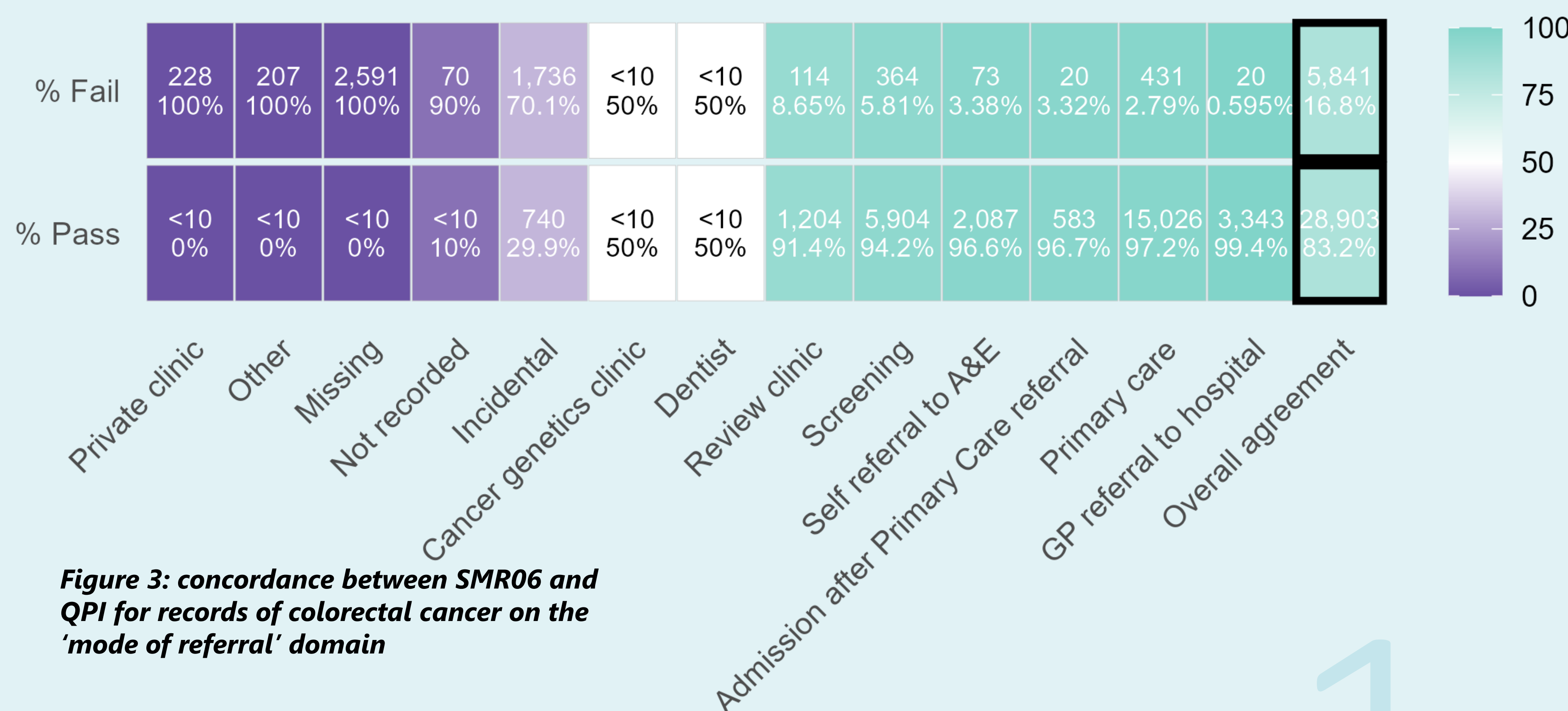


Figure 3: concordance between SMR06 and QPI for records of colorectal cancer on the 'mode of referral' domain