Three different data extraction methods lead to different data outputs from the same general practice

Merriman M, Marchesi A, Rabbi M, Luzuriaga J, Liu C (AIHW), and Aigner, K (CHN)

Introduction

Primary health care data is a known information gap for effective population health monitoring, research, policy, and planning (AIHW, 2025). A key challenge is the complexity and diversity of the primary care data ecosystem in a privatised and unregulated health information sector.

Electronic health records of regular clients are recorded within the practice management system (PMS) of a general practice for clinical use and decision support. An AIHW survey of 31 PHNs in April 2025 showed that over 95% of the 6,095 practices that submitted Practice Incentives Program Quality Improvement (PIPQI) data used Best Practice (71.9%) or Medical Director software (25.1%) and there are at least 15 other software products in operation. Three different extraction tools extract de-identified data from the PMS for a variety of service planning and population health analysis reasons: CAT4, POLAR, and Primary Sense.

Diverse software tools and methods for capturing general practice data can cause inconsistencies in the outputs generated, with new tools or versions potentially leading to data inaccuracies and loss of context (Canaway et al., 2022, AIHW, 2024).

Capital Health Network (CHN) and AIHW collaborated to compare aggregated outputs generated by the same practice using 3 different data extraction methods.

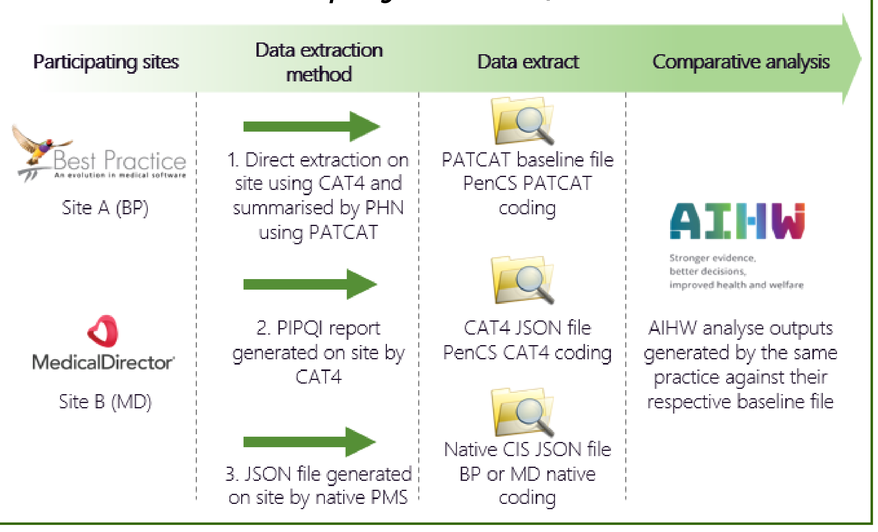
Methods

CHN engaged two general practices to participate in this project, one used 'Best Practice' and the other used 'Medical Director Clinical' as their PMS. The two sites each used three different data extraction methods whilst controlling for the same client population, reference period, and date of extraction (*Figure 1*):

- Method 1 - The practice used PenCS CAT4 to extract data and then the PHN used PATCAT to summarise data (referred to as 'PATCAT baseline file')
- Method 2 - The practice used PenCS CAT4 to extract data (referred to as 'CAT4 JSON file')
- Method 3 – The practice used its native PMS (either BP or MD) to extract data (referred to as 'native CIS JSON file')

Data extracts were based on the specification of the PIP Eligible Dataset (DoHAC, 2020). This dataset captures the recording of service event information in 10 Quality Improvement Measures (QIMs) across 6,095 practices for over 20 million regular clients that visited a practice 3 or more times in the 2 years prior to data extraction.

Figure 1: Data extraction pathways and project workflow



Results

The distribution of regular clients aged 15 years and over was analysed for each extraction method, using the practice's PATCAT baseline as the benchmark (*Figure 2*). Comparing the distribution of clients across age groups showed that the PATCAT extract (Method 1) and CAT4 JSON extract (Method 2) were identical for the BP practice only – The other extraction methods had a different number and distribution of regular clients for the BP practice (Method 3) and MD practice (Methods 1, 2, and 3). The native CIS generated JSON files (Method 3) had the largest difference in regular clients compared to the PATCAT baseline (Method 1) for both the BP and MD practice.

When examining the results for specific QIMs (*Figure 3*), the PATCAT baseline (Method 1) had different results across most measures compared to the CAT4 JSON (Method 2) and native CIS JSON files (Method 3), which was expected given the differences in the number of regular clients that were captured across extracts. When comparing the PATCAT extracts with CAT4 JSON extracts, the BP practice had the same result for 14 out of 19 QIM subgroups, whereas the MD practice only matched results for 3 out of 19 QIM subgroups. Each practice's native CIS generated JSON file had different results to the PATCAT baseline across all indicators. These results indicate that different coding has been applied to generate each aggregate extract.

Figure 2: Regular client (15+ years) counts by age group for BP and MD sites

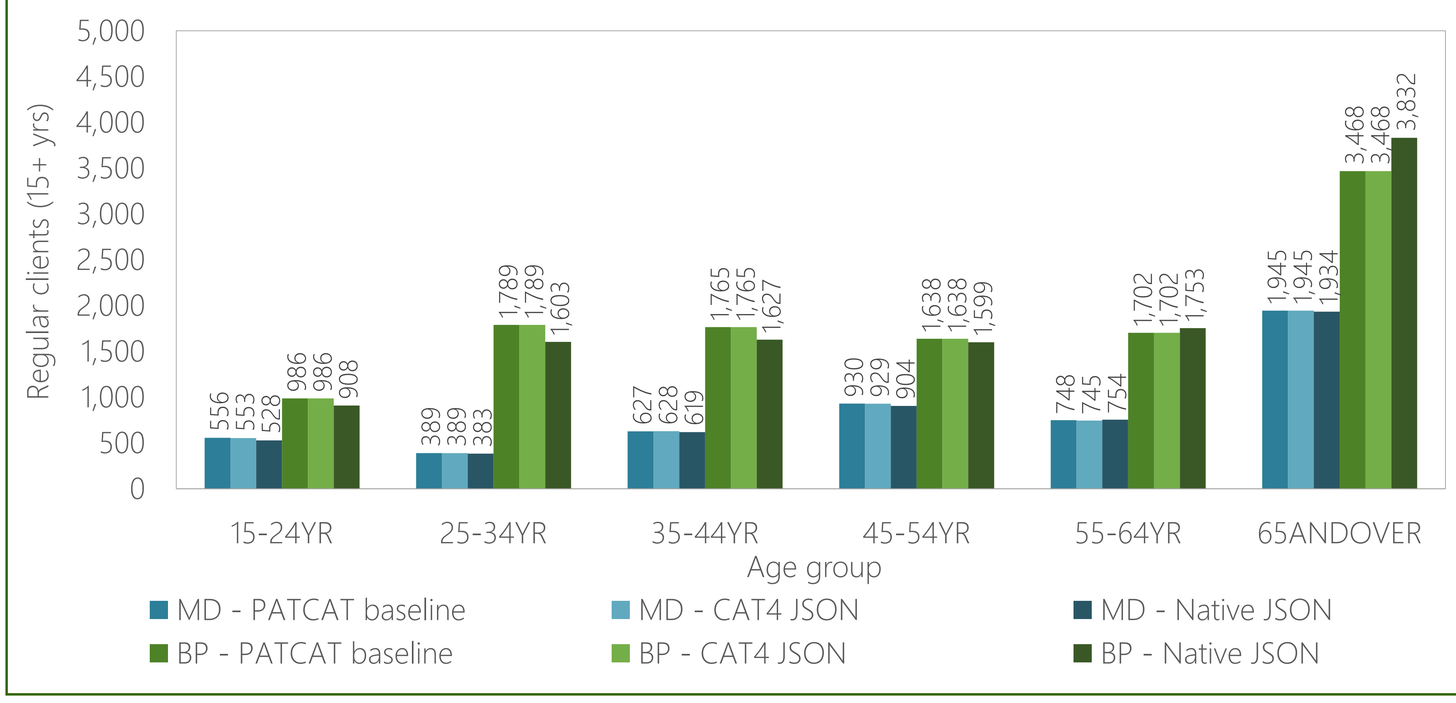


Figure 3: Data extract comparisons by practice

Comparison	Best Practice Comparison to PATCAT baseline		Medical Director Comparison to PATCAT baseline	
	JSON-CAT4	JSON-BP	JSON-CAT4	JSON-MD
QIM1 – T1DM	✓	×	✓	×
QIM1 – T2DM	✓	×	×	×
QIM1 – Undefined DM	✓	×	✓	×
QIM2a – Recorded	✓	×	×	×
QIM2b – Current smoker	✓	×	×	×
QIM2b – Ex smoker	✓	×	×	×
QIM2b – Never smoked	✓	×	×	×
QIM3a – Recorded	×	×	×	×
QIM3b – Underweight	×	×	×	×
QIM3b – Healthy	×	×	×	×
QIM3b – Overweight	×	×	×	×
QIM3b – Obese	×	×	×	×
QIM4	✓	×	×	×
QIM5	✓	×	×	×
QIM6	✓	×	✓	×
QIM7	✓	×	×	×
QIM8	✓	×	×	×
QIM9	✓	×	×	×
QIM10	✓	×	×	×

Discussion

These results show that the same practice produces inconsistent data outputs if using different extraction methods, even when controlling for the same client population, reference period, and date of extraction. This means that each software product for extracting general practice data is likely to produce inconsistent results due to differences in how the proprietary coding logic has been written and applied to interpret a technical specification.

Data quality testing at the point of extraction and data submission needs to be an ongoing and embedded process as each new software type or update has the potential to generate different results. It remains a challenge to be able to review and critique the proprietary coding, case definitions and mapping tables of software providers.

Given the highly diverse software environment, establishing processes to monitor software types and versions enables the isolation of files with sufficient data quality to be included for reporting. The development of agreed data and quality standards at all stages of the clinical and research data use lifecycle will greatly improve the benefits of secondary use primary care data (Canaway et al., 2022).

References

AIHW. (2024). Practice Incentives Program Quality Improvement Measures: annual data update 2023–24. https://www.aihw.gov.au/reports/primary-health-care/pipqi-measures-2023-24/contents/technical-notes/interpreting-pipqi-data

AIHW. (2025). Primary Health Care Data Development. https://www.aihw.gov.au/reports-data/health-welfare-services/primary-health-care/primary-health-care-data-development

Canaway, R., Boyle, D., Manski-Nankervis, J.-A. and Gray, K. (2022). Identifying primary care datasets and perspectives on their secondary use: a survey of Australian data users and custodians. BMC Medical Informatics and Decision Making, 22(1). doi: https://doi.org/10.1186/s12911-022-01830-9

DoHAC. (2020). Practice Incentives Program quality improvement measures – Technical specifications. https://www.health.gov.au/resources/publications/practice-incentives-program-quality-improvement-measures-technical-specifications?language=en