Validating the accuracy of automated data extraction for an obstetric anaesthetic dataset: a single centre quality improvement project

Aoife Kelly[1]; Wayne Henry[2]; Gary Stocks[2]; Chris Mullington[1,2]
1. Imperial College London, 2. Imperial College Healthcare NHS Trust

Imperial College London

Imperial College Healthcare NHS Trust

Introduction

- Secondary use of patient data is vital for service development, quality improvement and research.[1]
- The transition of the NHS from paper to electronic patient records is making automated extraction of large data sets increasingly possible.
- To ensure data quality, it is vital that validation of the data extraction process itself is performed before the dataset is collected.[2]
- The Neonatal Outcome and Epidural Hyperthermia (NOEH) study requires automated extraction of a dataset from >100,000 electronic delivery records.
- This single centre quality improvement project aimed to validate the NOEH dataset extraction process.

Methods

- Six NOEH dataset parameters had not been automatically extracted before and therefore required validation.
- The extraction process was validated using a cyclical iterative method (Figure 1).
- Samples of each parameter were extracted from between 25 and 100 randomly selected electronic delivery records.
- Automated extraction was performed by a senior clinical analyst (WH).
- The extraction process was deemed acceptable when parameters were extracted with >95% accuracy.

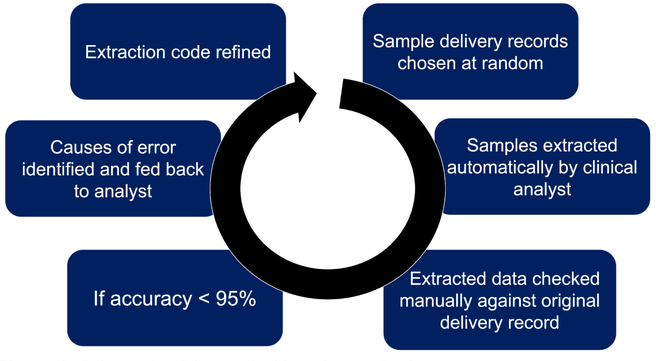


Figure 1: Schematic of the cyclical iterative method

Results

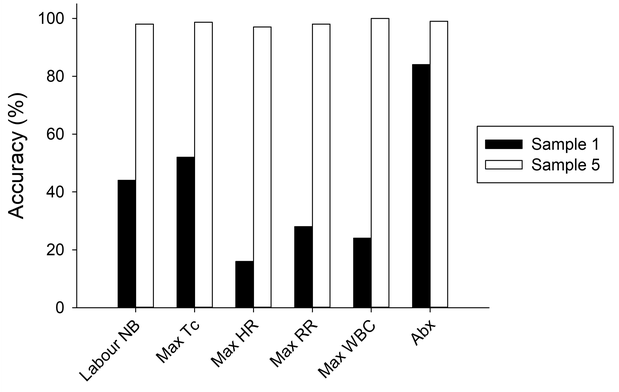


Figure 2: Extraction accuracy in the first and fifth (last) samples of the six novel NOEH study dataset parameters. NB = neuraxial blockade.

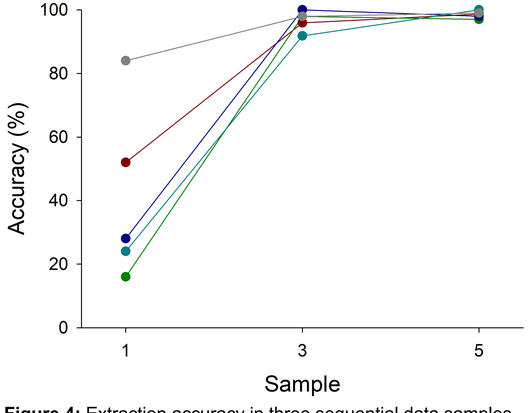


Figure 4: Extraction accuracy in three sequential data samples of five novel NOEH study dataset parameters.

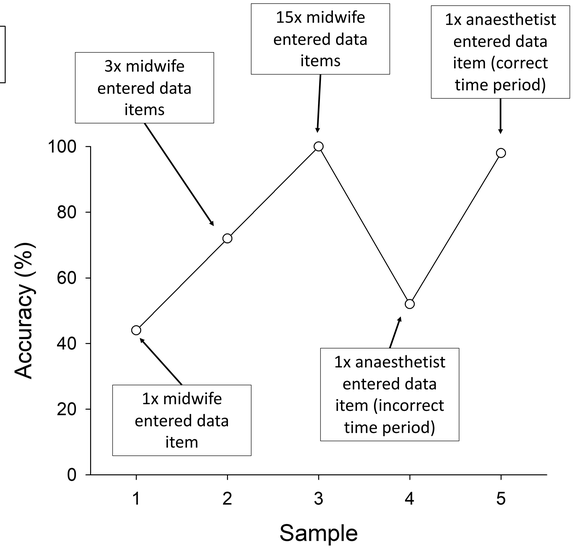
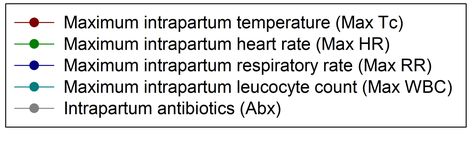


Figure 3: Labour neuraxial blockade extraction accuracy in five sequential data samples. Annotations describe the source of the extracted data.



Results

- In the first cycle 62/150 (41%) items were extracted correctly and no parameters were extracted with >95% accuracy.
- Common errors were extraction from incorrect time periods and extraction of incorrect data items.
- After five cycles, 571/580 (98%) items were extracted correctly and all parameters were extracted with >95% accuracy.

Conclusions

- We are confident that the NOEH dataset is an accurate representation of the original electronic delivery records.
- However, without robust analysis and refinement of the data extraction process, the NOEH dataset would have been inaccurate.
- The cyclical iterative approach proved an effective method of ensuring automated data extraction accuracy.

Future Implications

- The parameters validated in this project will be available for use in future projects.
- It is likely that similar levels of inaccuracy are present in other automatically extracted NHS datasets, although the sources of inaccuracy may vary.[2]
- It is possible that the accuracy of national obstetric anaesthesia datasets, such as the NHS digital maternity-services-dataset, contain similar levels of inaccuracy.[3]
- Robust validation of automated data extraction processes at a local level is vital to ensure the quality of both local and national datasets.

References

1. NHS. Information Governance. hps://www.england.nhs.uk/ig/about/
2. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. *Int J Med Inform* 2016; **94:** 271-4.
3. NHS Digital. Maternity Services Data Set. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/maternity-services-data-set