Intelligent Multi-Document Summarisation for Extracting Insights on Racial Inequalities from Maternity Incident Investigation Reports

Prof. Georgina Cosma

Prof. Patrick Waterson

Dr Mohit Singh
(Research Associate in AI)

Dr Jonathan Back

HSSIB

Prof. Thomas Jun

Contact: Georgina Cosma
g.cosma@lboro.ac.uk

Loughborough University

Loughborough University

HSIB
HEALTHCARE SAFETY INVESTIGATION BRANCH

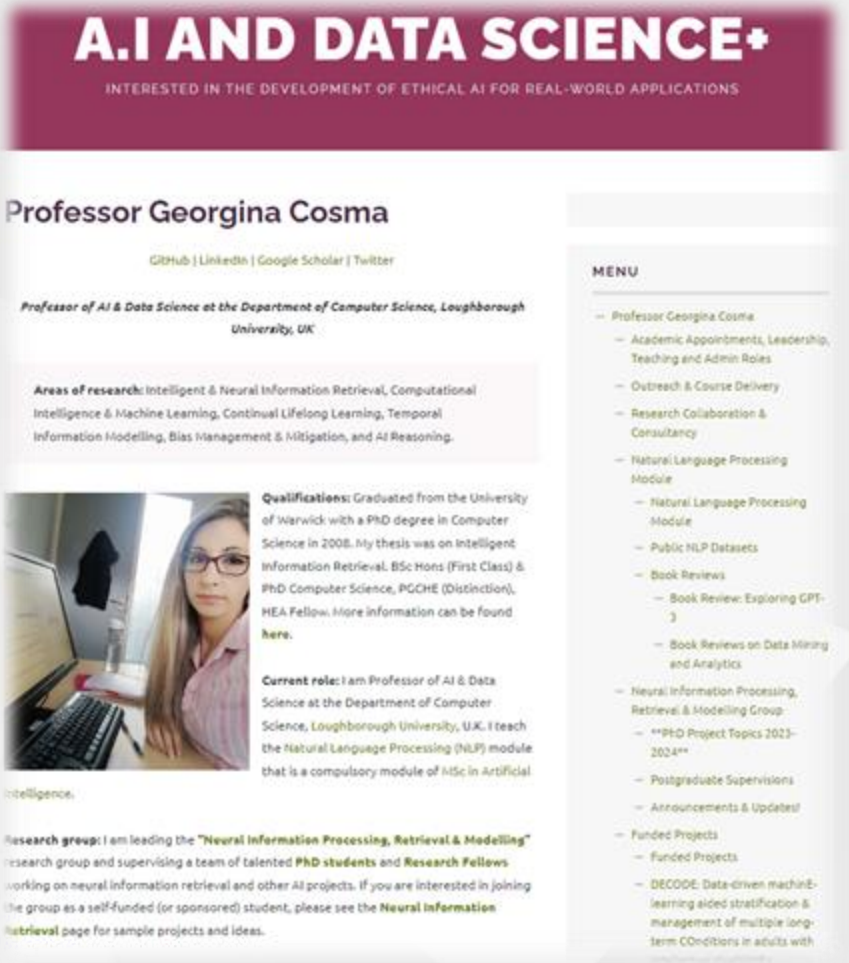
#InspiringWinners since 1909

Georgina Cosma

Professor of AI and Data Science

Department of Computer Science
School of Science

Loughborough University



Email: g.cosma@lboro.ac.uk
Twitter: @gcosma1
Website: https://datascienceplus.blog/
Univ.: https://www.lboro.ac.uk/departments/compsci/staff/georgina-cosma/

The maternal death rate in 2020-22 was 13.41 deaths per 100,000 maternities. This is significantly higher than the maternal death rate of 8.79 deaths per 100,000 maternities reported in the previous complete three year period (2017-19).

Maternal death rate for women from Black ethnic backgrounds has decreased slightly from the rate in 2019-21 but Black women remain three times more likely to die compared to White women.

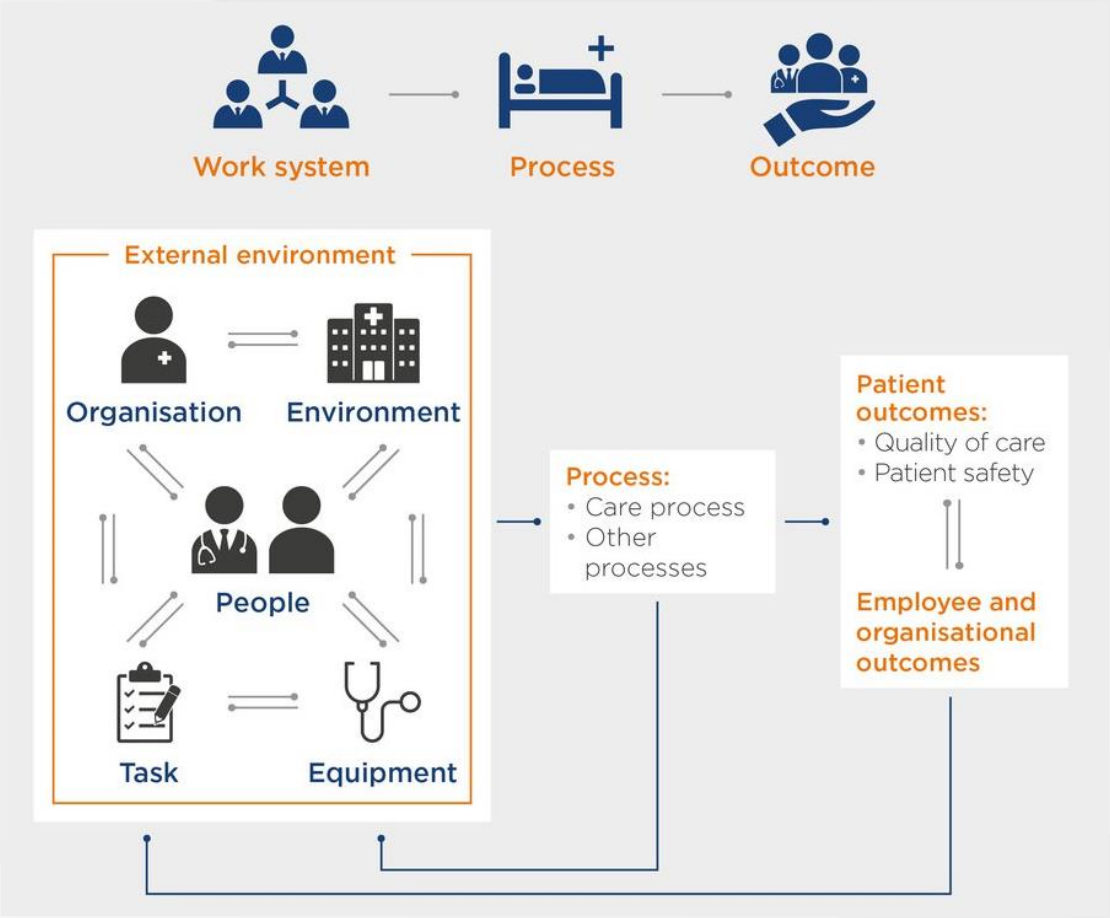
The maternal death rate for women from Asian ethnic backgrounds remains two times higher than that of White women.

About the project

- HSIB was established in 2017 to improve patient safety through independent investigations
- Transformed in 2023 into MNSI (Maternity and Newborn Safety Investigations) and HSSIB (Health Services Safety Investigations Body)
- Produced reports with investigation findings and recommendations

Project aim: To develop a reliable and unbiased machine learning based tool to **extract** and **analyse** intelligence from maternity investigation reports provided by HSSIB/MNSI, using human factor concepts specifically designed for maternity investigations.

What are Human Factors?



Systems Engineering Initiative for Patient Safety (SEIPS) adapted from Holden et al (2013)

Source: Investigation report- Never events: analysis of HSIB's national investigations

Supplementary File SF1. SIRch taxonomy of human factors for maternity investigations

1. External Environment
   - (a) Policy factor
   - (b) Societal factor
   - (c) Economic factor
   - (d) COVID ✓
   - (e) Geographical factor (e.g. Location of patient)
2. Internal Environment
   - (a) Physical layout and Environment
   - (b) Acuity (e.g., capacity of the maternity unit as a whole)
   - (c) Availability (e.g., operating theatres)
   - (d) Time of day (e.g., night working or day of the week)
3. Organisation
   - (a) Team culture factor (e.g., patient safety culture)
   - (b) Incentive factor (e.g., performance evaluation)
   - (c) Teamworking
   - (d) Communication factor
     - i. Between staff
     - ii. Between staff and patient (verbal)
   - (e) Documentation
   - (f) Escalation/referral factor (including fresh eyes reviews)
   - (g) National and/or local guidance
   - (h) Language barrier
4. Jobs/Task
   - (a) Assessment, investigation, testing, screening (e.g., holistic review)
   - (b) Care planning
   - (c) Dispensing, administering
   - (d) Monitoring
   - (e) Risk assessment
   - (f) Situation awareness (e.g., loss of helicopter view)
   - (g) Obstetric review
5. Technologies and Tools
   - (a) Issues
   - (b) Interpretation (e.g., CTG)
6. Person
   - (a) Patient (characteristics and performance)
     - i. Characteristics
       - A. Physical characteristics
       - B. Psychological characteristics (e.g., stress, mental health)
       - C. Language competence (English)
       - D. Disability (e.g., hearing problems)
       - E. Training and education (e.g., attendance at ante-natal classes)
       - F. Record of attendance (e.g., failure to attend antenatal classes)
     - ii. Performance
       - A. Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)
       - B. Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)
       - C. Intentional rule breaking (deliberately do something different from rules)
   - (b) Staff (characteristics and performance)
     - i. Characteristics
       - A. Physical characteristics
       - B. Psychological characteristics (e.g., stress, mental health)
       - C. Language competence (English)
       - D. Disability (e.g., hearing problems)
       - E. Training and education (e.g., attendance at ante-natal classes)
       - F. Record of attendance (e.g., failure to attend antenatal classes)
     - ii. Performance
       - A. Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)
       - B. Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)
       - C. Intentional rule breaking (deliberately do something different from rules)

SIRch taxonomy

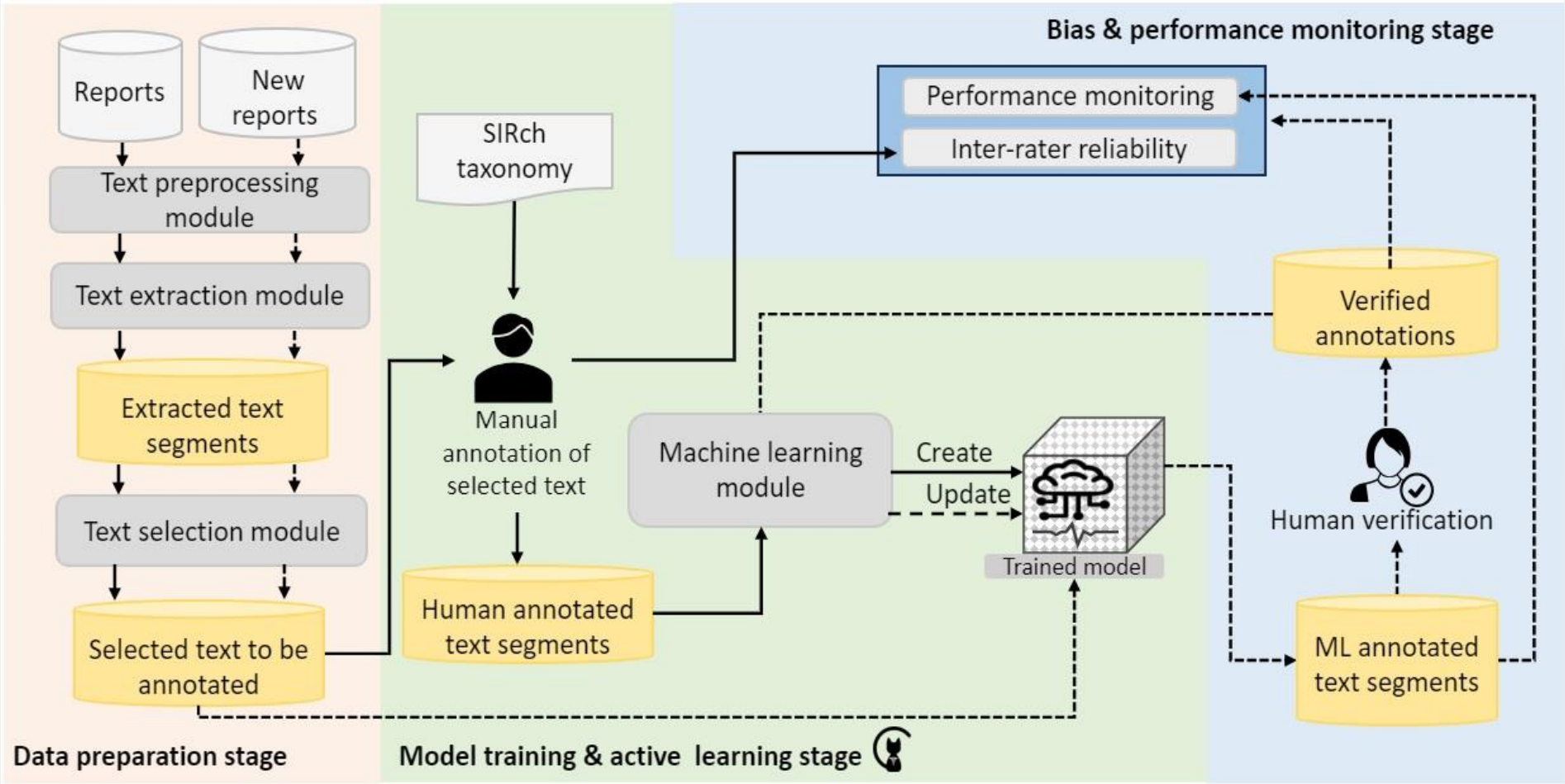
Example

"The Trust's clinical practice guideline relating to antenatal care sets the criteria for mothers who have uncomplicated pregnancies to have their care provided by a midwife throughout."

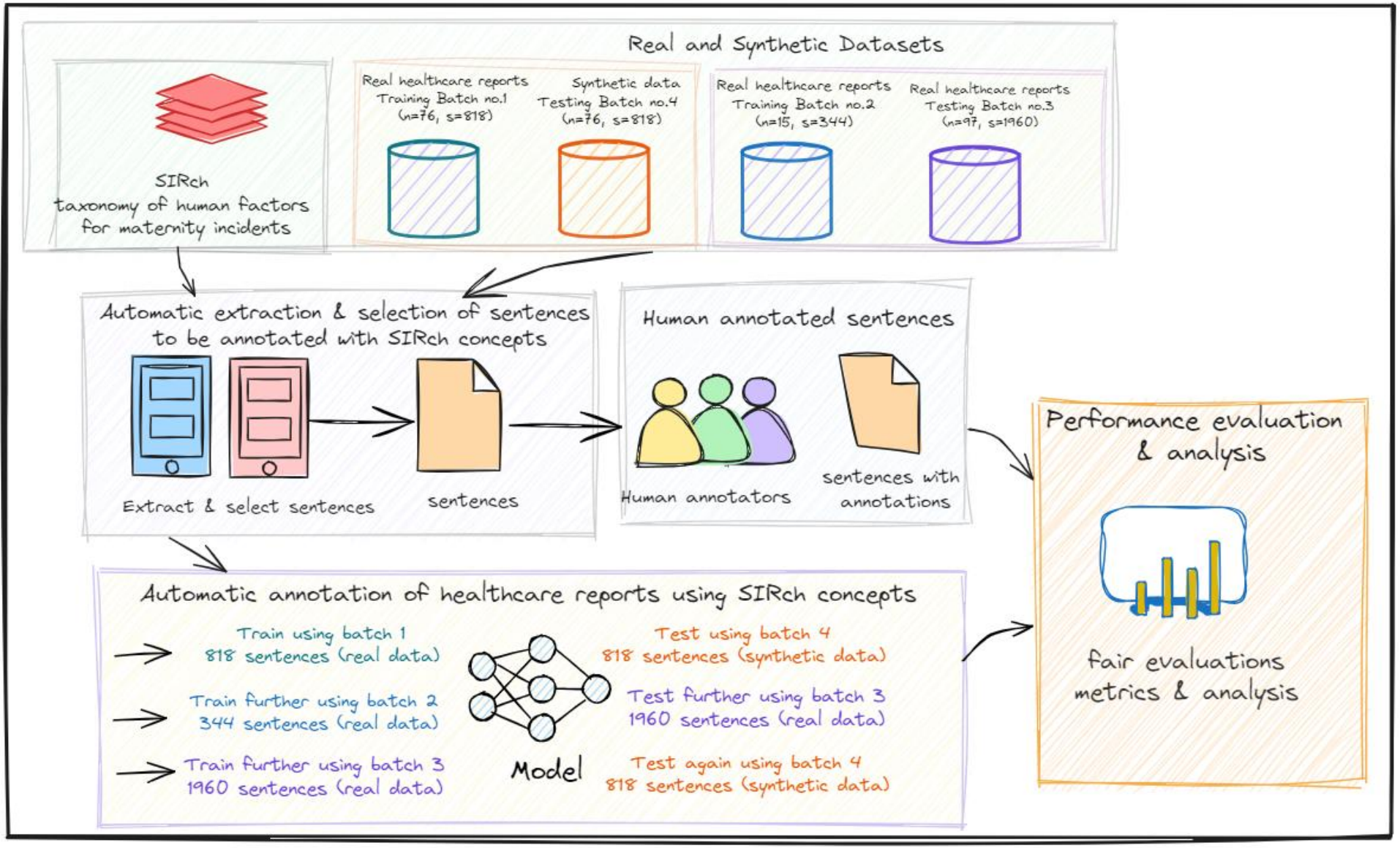
[guideline, local and national guideline] [care, job/task/Care planning] …

Annotated sentence

I-SIRch tool


Bias & performance monitoring stage
Performance monitoring
Inter-rater reliability
Reports
New reports
SIRch taxonomy
Text preprocessing module
Text extraction module
Extracted text segments
Text selection module
Selected text to be annotated
Manual annotation of selected text
Human annotated text segments
Machine learning module
Create
Update
Trained model
Verified annotations
Human verification
ML annotated text segments
Data preparation stage
Model training & active learning stage

I-SIRch tool: Training, testing and development



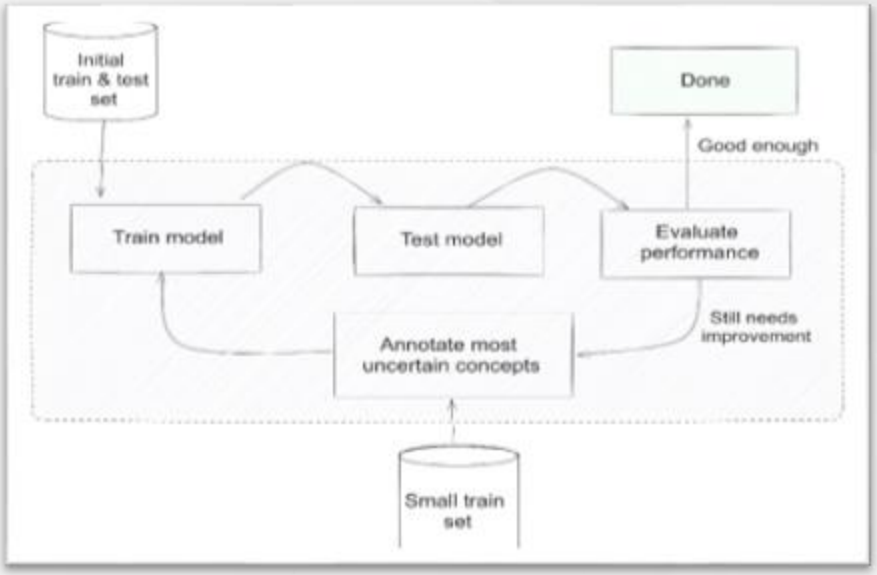
Datasets

No. of reports	no. of concepts	no. of sentences
Batch 1 (n=76 real)	970	818
Batch 2 (n=15 real)	452	344
Batch 3 (n=97 real)	2644	1960
Batch 4 (n=76 synthetic)	970	818
Reports per year (excluding batch 4): 2019 (n=4), 2020(n=115), 2021 (n=42), 2022 (n=27)		

Supplementary Table S6. Performance evaluation for each test. The results of Test A and Test C can be compared because both tests were conducted on Batch 4. Test B was conducted on Batch 3 and hence cannot be directly compared to the results of Test A and Test C.

	Test with real data		Tests with synthetic data			
	Test B		Test A		Test C	
	Avg	SD	Avg	SD	Avg	SD
Precision	0.87	0.34	1.00	0.00	1.00	0.00
Recall	0.93	0.18	0.60	0.23	0.67	0.15
F-score	0.96	0.10	0.79	0.08	0.81	0.08
Misclassification	0.19	0.35	0.40	0.23	0.33	0.15
Accuracy	0.81	0.35	0.60	0.23	0.67	0.15
Balanced Accuracy	0.90	0.18	0.80	0.11	0.83	0.08
Avg: Average; SD: Standard deviation.						

I-SIRch tool: evaluations



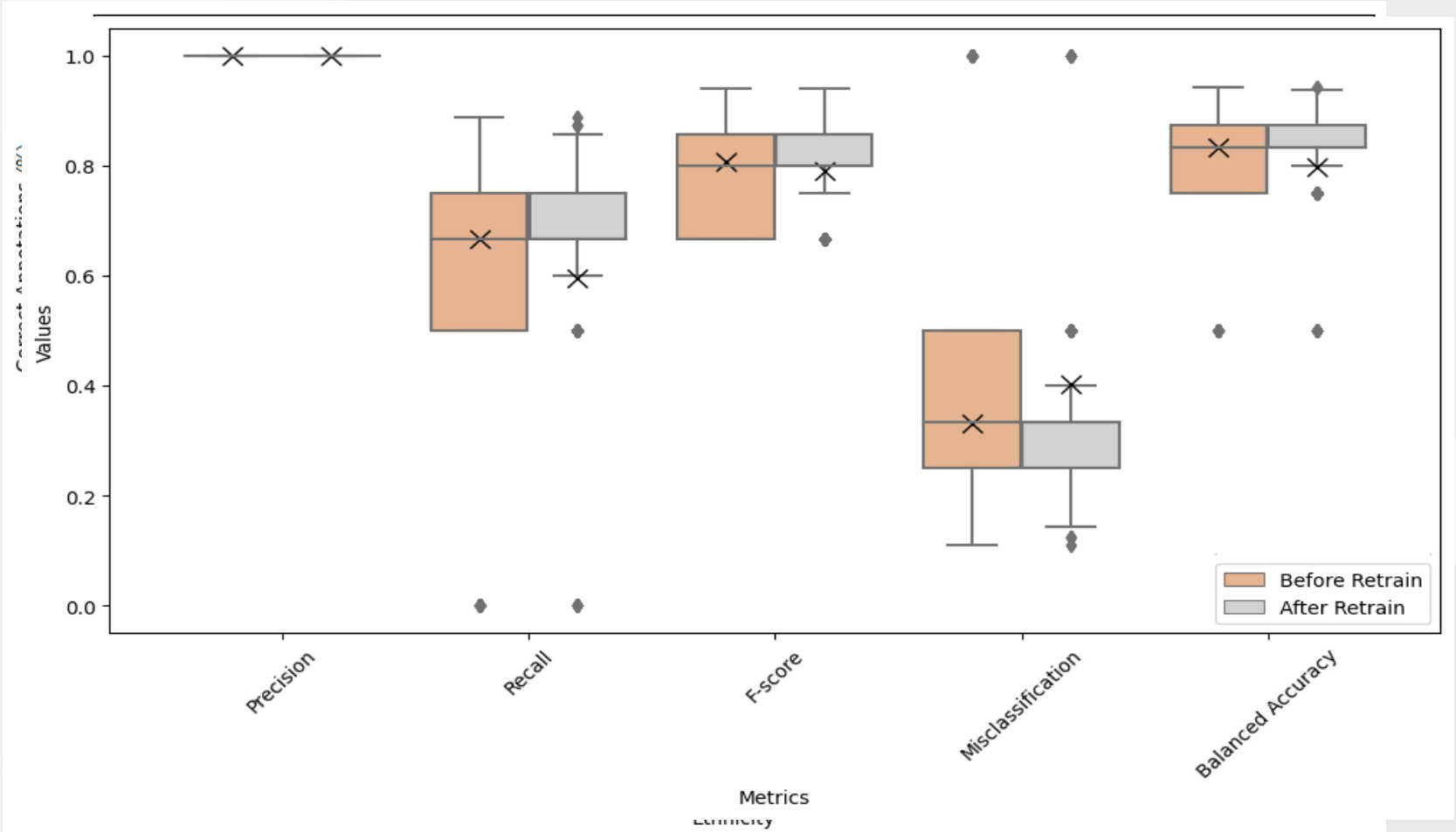
Ethnic group	No. of reports	No. of concepts across the reports	Average no. of concepts per report
Asian	6	87	15
Black	7	81	12
Data not received	4	55	14
Mixed Background	1	13	13
White British	52	688	14
White Other	6	46	8
Total:	76	970	Average: 13

Table 2. Test C results. Performance of I-SIRch when tested on the synthetic data (batch 4) after training on Batch 1 and retaining on Batches 2 and 3. Table shows results of I-SIRch performance across ethnic groups, showing number of annotations, mean values, standard deviations (SD), and 95% Confidence Intervals (CI).

Ethnic group	Metric	Mean ± SD	95% CI
Asian (n = 87)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.68 ± 0.14	[0.65, 0.71]
	F-score	0.81 ± 0.09	[0.79, 0.83]
	Misclassification	0.32 ± 0.14	[0.29, 0.35]
	Balanced Accuracy	0.84 ± 0.07	[0.83, 0.85]
Black (n = 81)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.65 ± 0.17	[0.61, 0.69]
	F-score	0.80 ± 0.07	[0.78, 0.82]
	Misclassification	0.35 ± 0.17	[0.31, 0.39]
	Balanced Accuracy	0.82 ± 0.09	[0.80, 0.84]
Data not received (n = 55)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.65 ± 0.21	[0.59, 0.71]
	F-score	0.82 ± 0.08	[0.80, 0.84]
	Misclassification	0.35 ± 0.21	[0.29, 0.41]
	Balanced Accuracy	0.82 ± 0.10	[0.79, 0.85]
Mixed Background (n = 13)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.69 ± 0.12	[0.62, 0.76]
	F-score	0.81 ± 0.09	[0.76, 0.86]
	Misclassification	0.31 ± 0.12	[0.24, 0.38]
	Balanced Accuracy	0.85 ± 0.06	[0.81, 0.89]
Other White (n = 46)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.64 ± 0.19	[0.58, 0.70]
	F-score	0.80 ± 0.09	[0.77, 0.83]
	Misclassification	0.36 ± 0.19	[0.30, 0.42]
	Balanced Accuracy	0.82 ± 0.09	[0.79, 0.85]
White British (n = 688)	Precision	1.00 ± 0.00	[1.00, 1.00]
	Recall	0.67 ± 0.15	[0.66, 0.68]
	F-score	0.81 ± 0.08	[0.80, 0.82]
	Misclassification	0.33 ± 0.15	[0.32, 0.34]
	Balanced Accuracy	0.84 ± 0.07	[0.83, 0.85]

Statistics about Batch 1 provided by HSIB (ethnicity was only known for Batch 1)

I-SIRch tool: evaluations



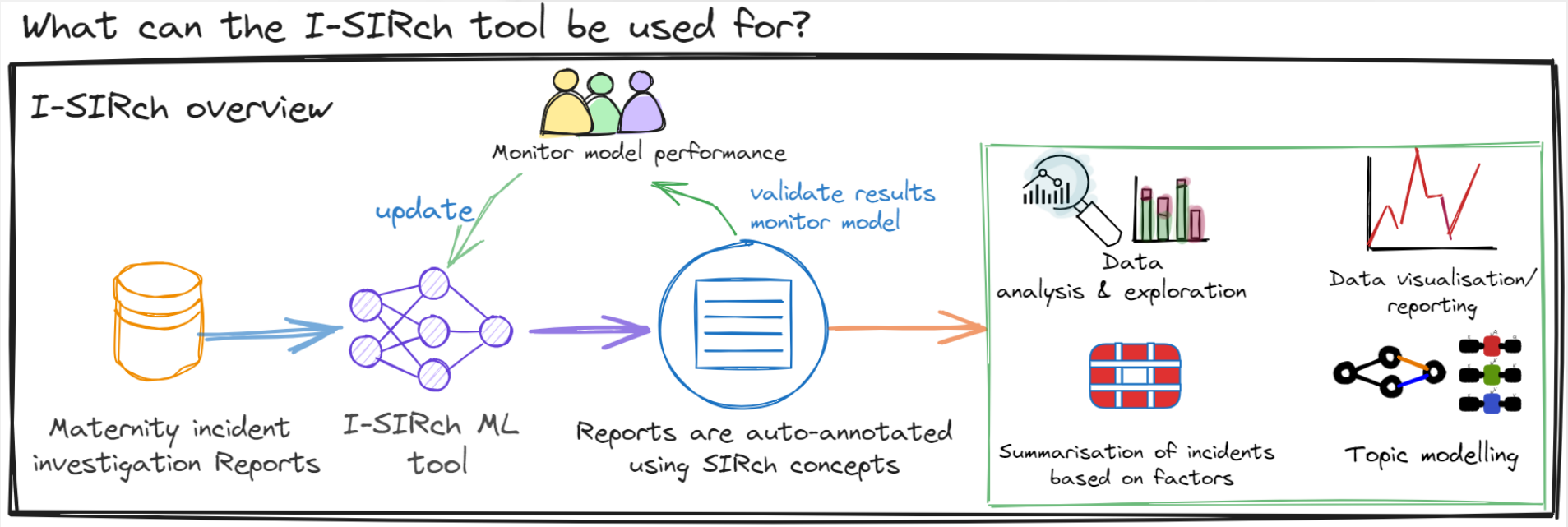
Performance evaluation using various metrics. Orange bars show the test results of Test A, when the model is trained on Batch 1 (real data) and tested on Batch 4 (synthetic data). Grey bars show the test results of Test C, when the model is trained on real data batches (Batches 1, 2, and 3) and tested on synthetic data (Batch 4).

Distribution of correct annotations for the various ethnic groups

Thorough evaluations to ensure model equity during the concept annotation process

Number of reports containing per concept. These were utilised for training the model (batches 1, 2, 3).

I-SIRch tool



Our publications:

- I-SIRch: AI-Powered Concept Annotation Tool For Equitable Extraction And Analysis Of Safety Insights From Maternity Investigations, Int. Journal of Population Data Science, 2024

- Unveiling Disparities in Maternity Care: A **Topic Modelling** Approach to Analysing Maternity Incident Investigation Reports, AIiH 2024, LNCS, 2024

- Intelligent **Multi-Document Summarisation** for Extracting Insights on Racial Inequalities from Maternity Incident Investigation Reports, AIiH 2024, LNCS, 2024

I-SIRch tool for summarisation

188 anonymised maternity investigation reports annotated with 27 SIRch human factors concepts.

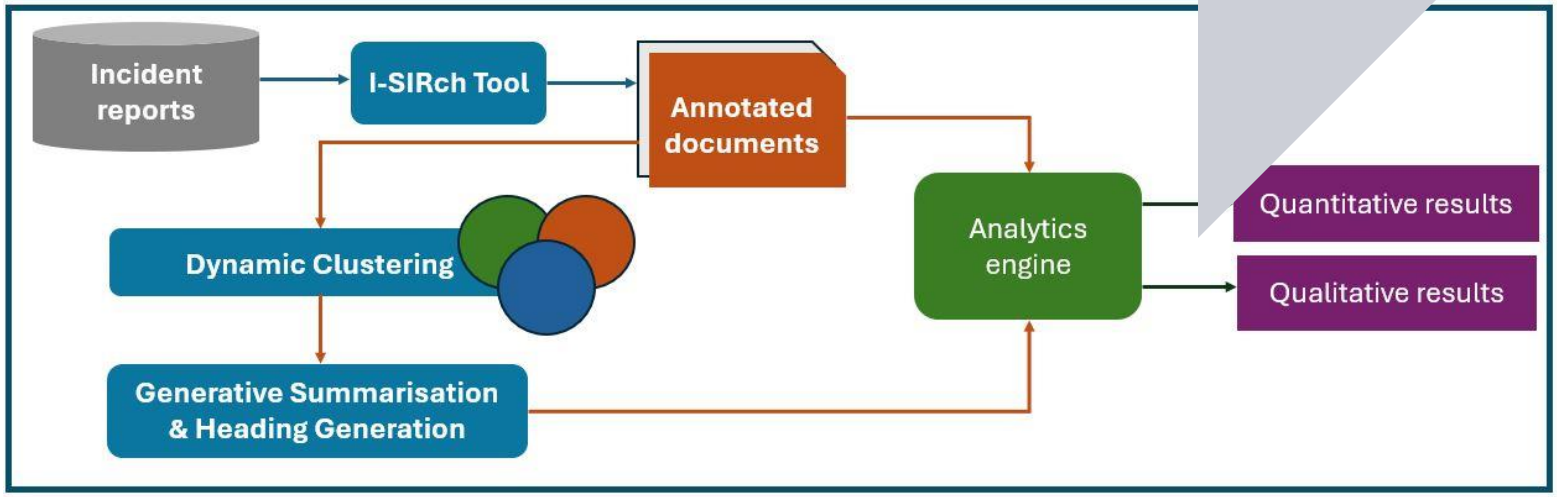
I-SIRch:CS groups the annotated sentences into clusters using sentence embeddings and k-means clustering, maintaining traceability via file and sentence IDs.

Summaries generated for each cluster using offline state-of-the-art abstractive summarisation models (BART, DistilBART, T5).

Evaluations using metrics assessing summary quality attributes.

Generated summaries are linked back to the original file and sentence IDs, ensuring traceability and allowing for verification of the summarised in-formation.

Results demonstrate BART's strengths in creating informative and concise summaries.



Designed to facilitate the aggregation and analysis of safety incident reports while ensuring traceability throughout the process.

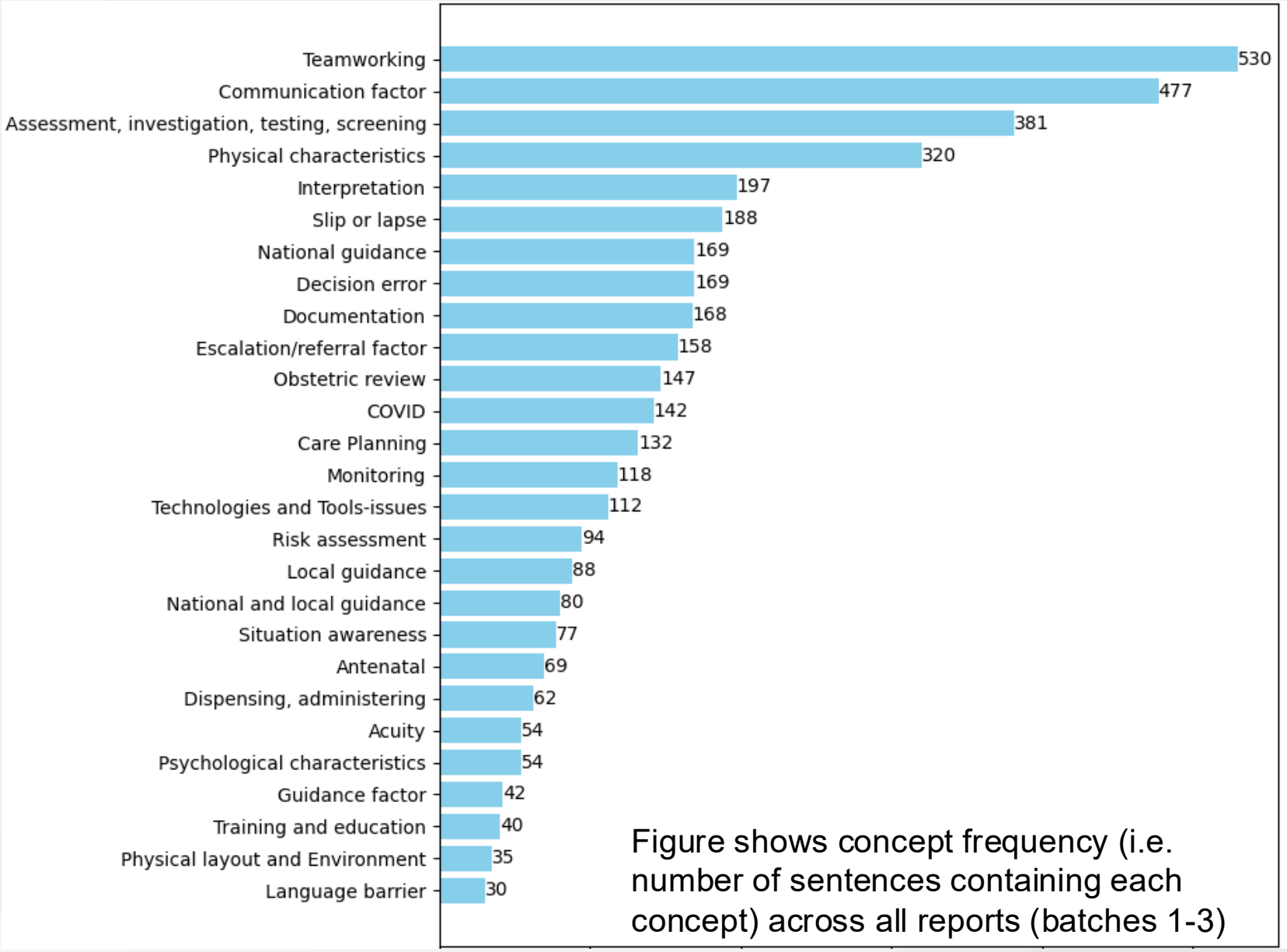
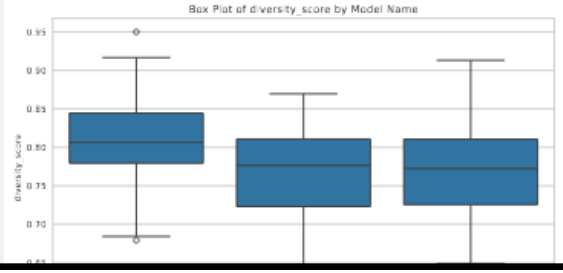
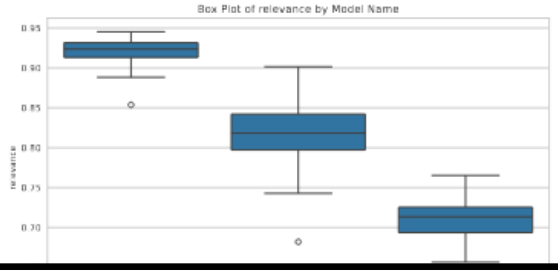


Figure shows concept frequency (i.e. number of sentences containing each concept) across all reports (batches 1-3)





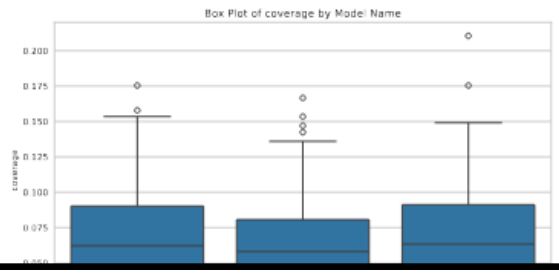
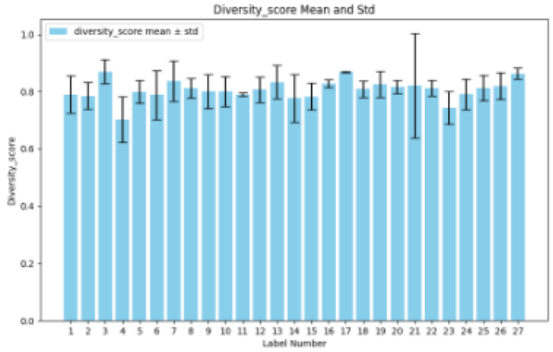


Table 2: Evaluation metrics for summarisation models

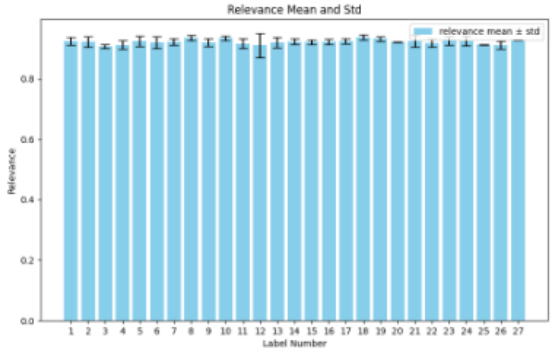
Metric	BART	DistilBART	T5-small
Diversity Score	0.806 ± 0.058	0.770 ± 0.057	0.771 ± 0.059
Relevance	0.922 ± 0.015	0.709 ± 0.028	0.818 ± 0.035
Coverage	0.070 ± 0.034	0.072 ± 0.034	0.064 ± 0.032
Coherence	0.794 ± 0.022	0.674 ± 0.038	0.670 ± 0.030
Conciseness	0.021 ± 0.004	0.019 ± 0.003	0.022 ± 0.004
Readability	189.148 ± 5.863	186.896 ± 6.253	190.288 ± 5.799

(d) Coherence boxplot (e) Conciseness boxplot (f) Readability boxplot

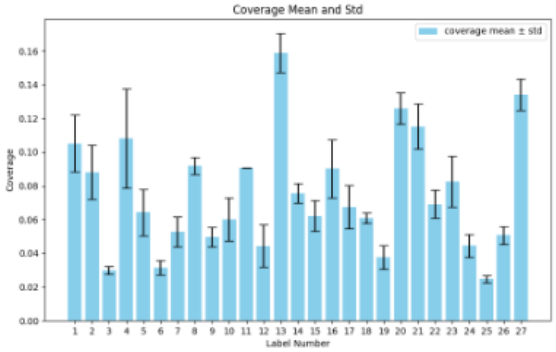
Fig. 2: Overall comparison of evaluation metrics.



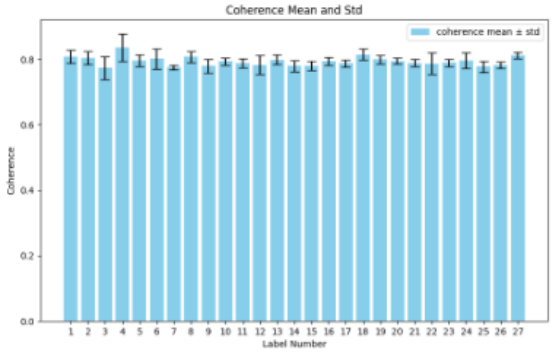
(a) Diversity Score



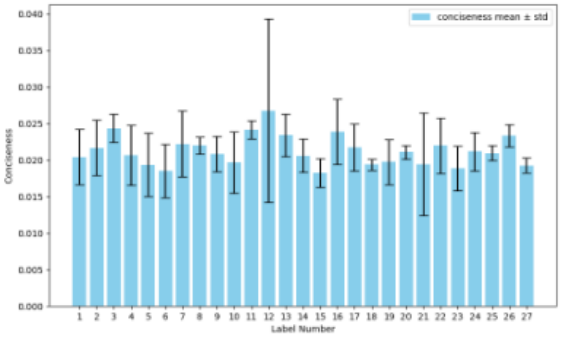
(b) Relevance



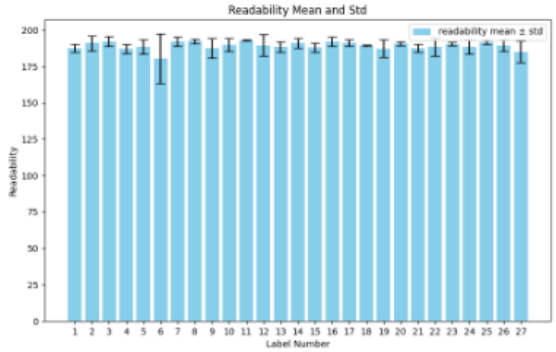
(c) Coverage



(d) Coherence



(e) Conciseness



(f) Readability

Fig. 3: Summary of model evaluation metrics for BART.

Performance across each concept

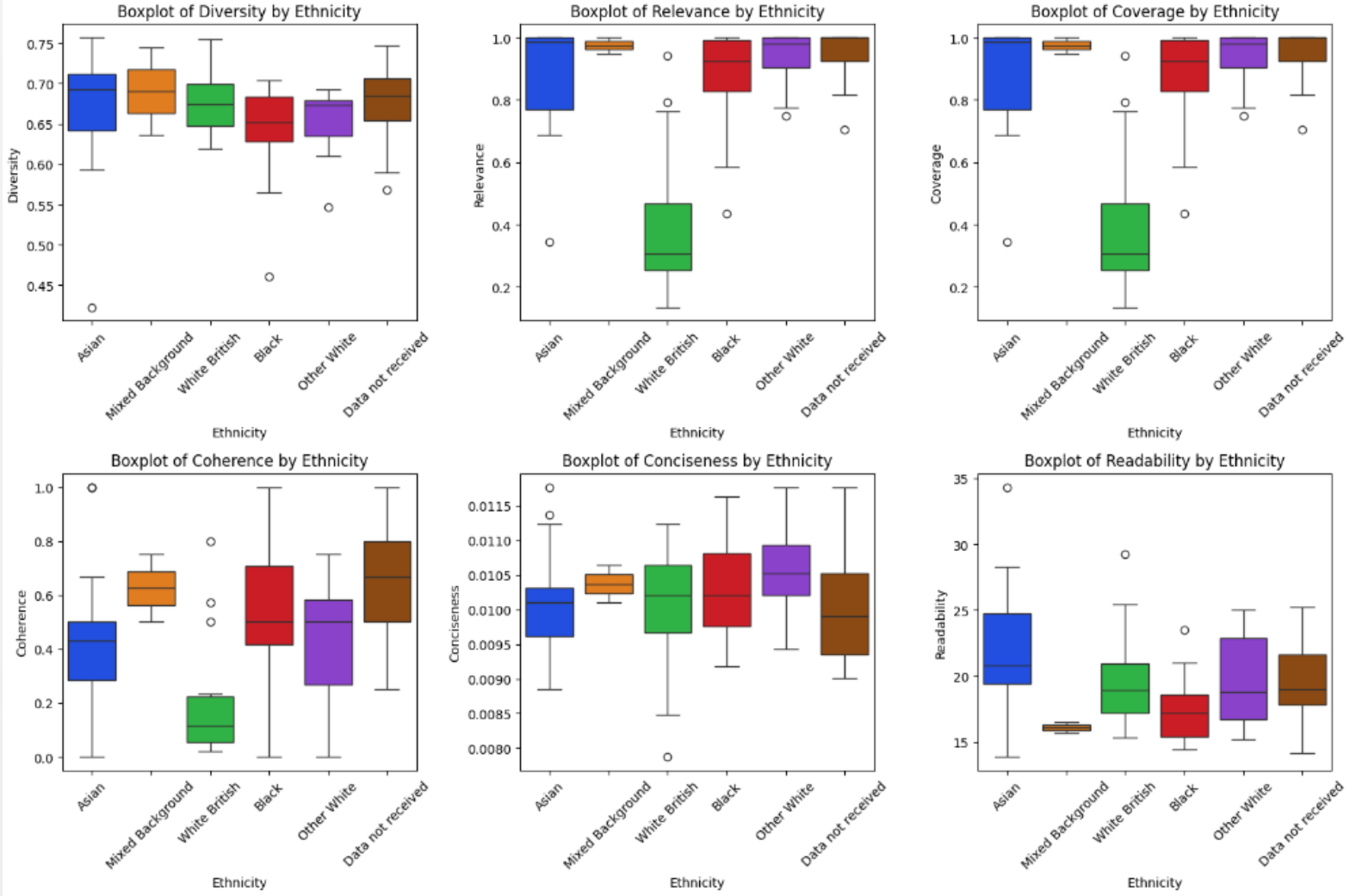


Fig. 4: Results per ethnic group when using BART

Ethnicity	Sentence Count
Asian	87
Black	81
Mixed Backg.	13
Other White	46
White British	688
Grand Total	**915**

I-SIRch tool for summarisation

Table 4: Sample summaries from multiple reports for two ethnic groups. Ethnicities cannot be disclosed. Each summary is generated from multiple reports.

Concept	Summary
Communication	**Staff not heard.** Staff voiced their concerns about this decision to the senior obstetrician. They were left feeling that their concerns had not been heard. There was no formal debriefing afterwards, which staff would have valued. The opportunity to share reflections and learning was not completed with all staff involved. The incremental delays caused by finding and allocating staff to the concurrent theatre cases, and communication breakdown within the team further impacted on the DDI.
Acuity	**Reviews with seniors did not occur.** The ambulance Trust was experiencing high volumes of 999 calls at the time of a Mother's call. Due to the high acuity on the labour ward the initial decisions were not discussed with the senior clinician. A senior face to face review did not occur until 16:05 hours, 2 hours and 50 minutes after the initial recognition of abnormalities of the Baby's heart rate. The abnormal CTG trace from the IOL suite was not reviewed by the senior obstetrician.

Ethical risks of abstractive summarisation

Risk of information hallucination and bias amplification

Risk of inadequate control over content

Risk in processing sensitive data

Sample size and variability

I-SIRch tool for topic modelling

3.2 Topics around the sentences of the Black ethnic group

Healthcare Processes and Assessments: The Black ethnic group contributed the most sentences (29) to this topic, which covers various healthcare processes and assessments. This suggests a significant focus on the effectiveness and efficiency of these processes. Concepts such as "Assessment, investigation, testing, screening", "Care Planning", and "Risk assessment" indicate a focus on the quality and comprehensiveness of patient assessments and care planning. Keywords like "pathway", "assessment", "care", "plan", and "review" relate to the various stages and components of healthcare processes.

Patient Care and Management: The group contributed 24 sentences to this topic, highlighting their concerns regarding patient care and management. This reflects an emphasis on ensuring appropriate and effective care for patients. Concepts such as "Escalation/referral factor", "Psychological characteristics", and "Teamworking" suggest a focus on the various aspects of patient care, including

Key Human Factors Identified

- **Organisation-Teamworking** in 155 reports, highlighting the critical role of effective collaboration in maternity care.
- **Organisation-Communication** in 159 reports, underscoring the importance of clear information exchange in healthcare settings.
- **Assessment, investigation, testing, and screening** in 150 reports, indicating potential gaps in patient evaluation processes.
- **Patient physical characteristics** in 118 reports, suggesting the significance of individual patient factors in care outcomes.
- **Interpretation of technologies and tools** (e.g., CTG) in 90 reports, pointing to challenges in using and understanding medical equipment.
- **Staff-related factors such as slips/lapses** in 99 reports and decision errors (89 reports) were prominent, revealing human performance concerns.
- **COVID-19 in 79 reports**, demonstrating the pandemic's significant effect on maternity care.
- **Organisational factors like documentation** (86 reports) and **escalation/referral** (97 reports) were common, suggesting systemic challenges.
- **National and local guidance issues** in 92 reports, indicating potential problems with policy implementation or clarity.

Patient and Public Involvement

Limitations

- Lack of Patient and Public Involvement (PPI) in assessing the framework's outputs
- PPI is crucial to ensure solutions are relevant to patient needs and experiences, potentially improving summaries and real-world applicability
- PPI can also increase the model's transparency and trustworthiness

Future development goals

- Integrating PPI feedback to enhance the framework's effectiveness and contributions to patient safety and care quality

Conclusion

- The I-SIRch:CS framework automates analysis, modelling and summarisation of textual data in maternity incident reports
- This holds significant potential for uncovering critical insights and contributing factors to preventable harm

Future work will focus on:

- Further enhancing traceability by providing clear links between summaries and original reports for easy verification
- Expanding further on explainable AI techniques for summarisation
- Extend the study and use findings to influence policy on maternal deaths

This project is independent research funded by NHSX and The Health Foundation and it is managed by the National Institute for Health Research (AI_HI200006). The views expressed in this publication are those of the author(s) and not necessarily those of the NHSX, The Health Foundation, National Institute for Health Research, or the Department of Health and Social Care.