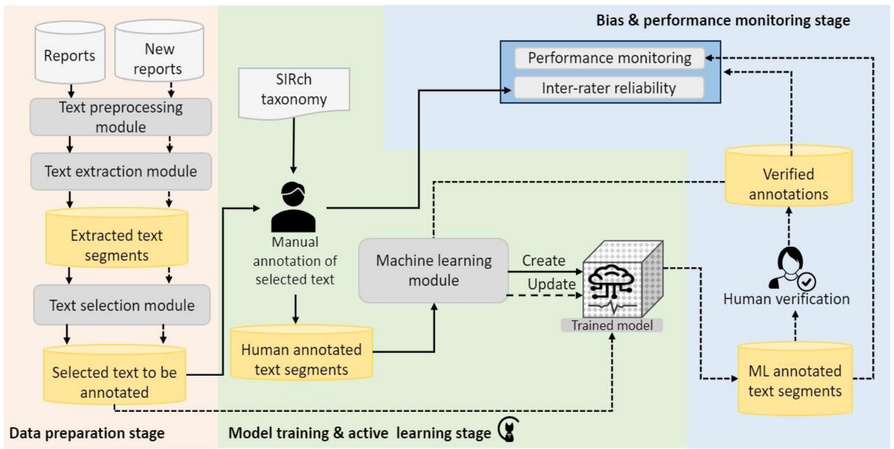
Unveiling Disparities in Maternity Care: AI-powered Analysis of Maternity Incident Investigation Reports

Georgina Cosma[1*], Mohit Kumar Singh[1], Patrick Waterson[2], and Gyuchan Thomas Jun[2] and Jonathan Back[3]

1 Department of Computer Science, School of Science, Loughborough University, Loughborough, United Kingdom
2 School of Design and Creative Arts, Loughborough University, Loughborough, United Kingdom
3 Health Services Safety Investigations Body (HSSIB), United Kingdom

***Contact: Prof. Georgina Cosma**
Email: g.cosma@lboro.ac.uk

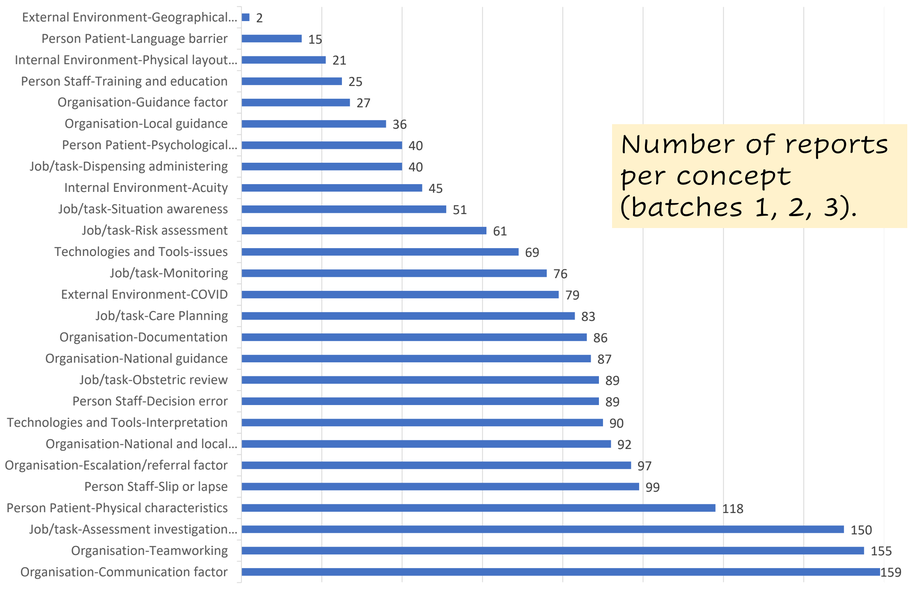
I-SIRch tool: concept annotation [1]



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)		

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

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



Number of reports per concept (batches 1, 2, 3).

	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

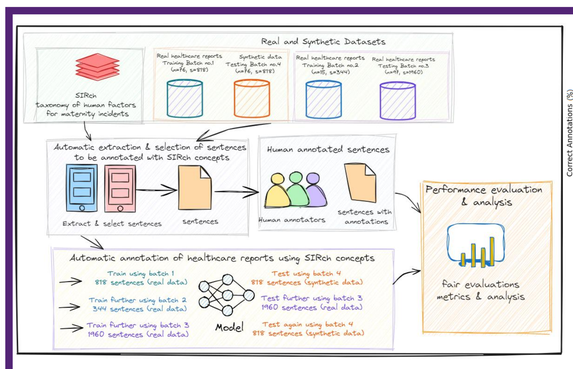
Avg: Average; SD: Standard deviation.

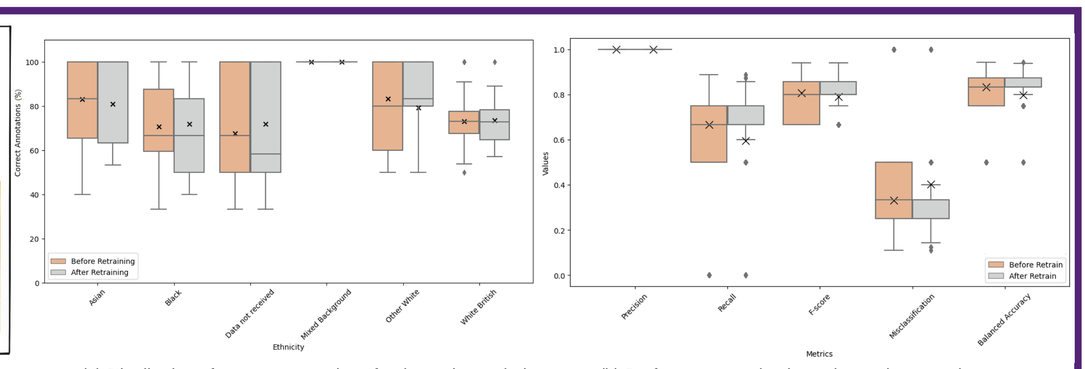
Results of Test A and Test C can be compared because both tests were conducted on Batch 4.

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]

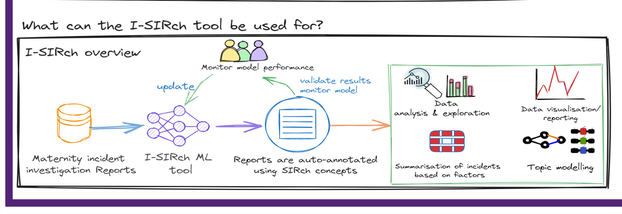
Test C Performance of I-SIRch when tested on synthetic data (batch 4) after training on batches 1, 2 & 3.

I-SIRch tool development & usage [1]





(a) Distribution of correct annotations for the various ethnic groups (b) 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).

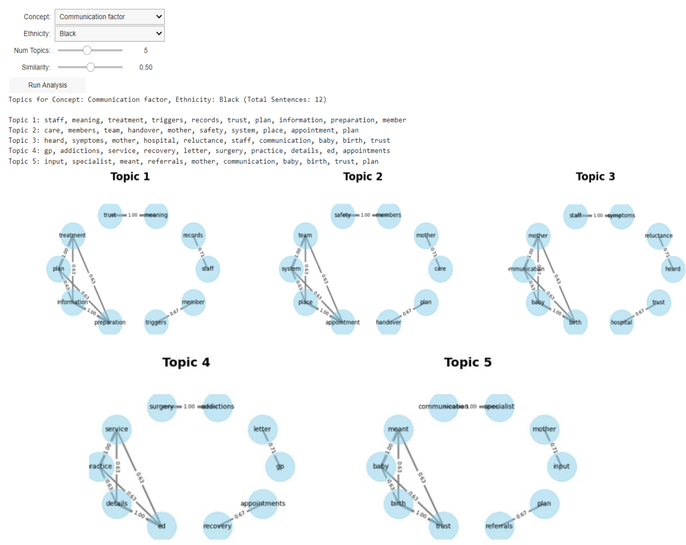


Thorough evaluations to ensure model equity during the concept annotation process.

Topic modelling [2]

Table 3: Sentence counts by ethnicity and topic. Up to 5 topics per group were generated. Duplicated topics were removed, hence some groups have < 5 topics.

Ethnicity	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Total sentences
Asian	22	24	14	9	10	79
Black	30	21	15	7	4	77
Data not received	24	14	10	3	0	51
Mixed Background	9	2	0	0	0	11
Other White	21	11	4	1	1	38
White British	132	121	115	179	140	687



Communication factor

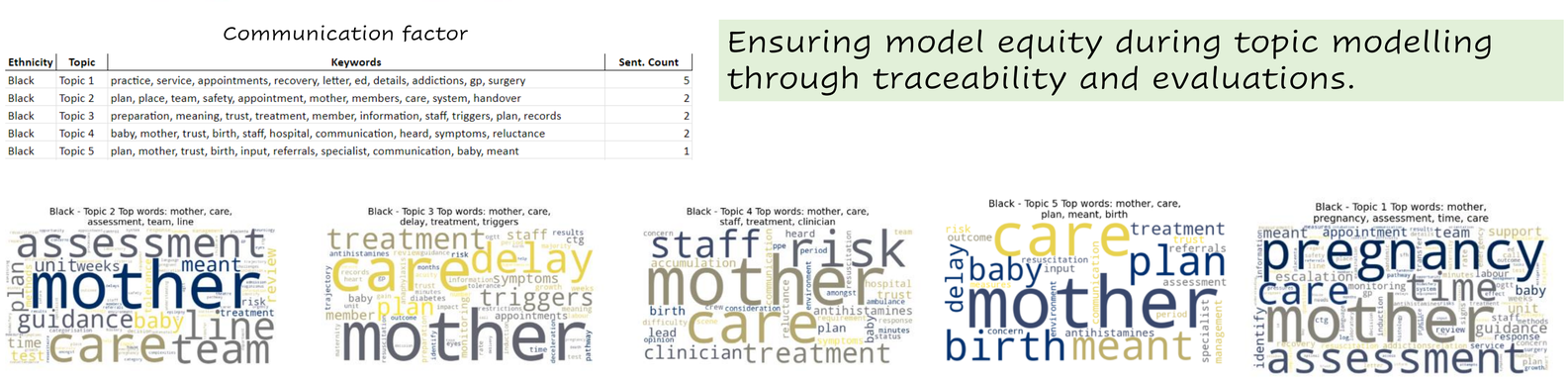
Ethnicity	Topic	Keywords	Sent. Count
Black	Topic 1	practice, service, appointments, recovery, letter, ed, details, addictions, gp, surgery	5
Black	Topic 2	plan, place, team, safety, appointment, mother, members, care, system, handover	2
Black	Topic 3	preparation, meaning, trust, treatment, member, information, staff, triggers, plan, records	2
Black	Topic 4	baby, mother, trust, birth, staff, hospital, communication, heard, symptoms, reluctance	2
Black	Topic 5	plan, mother, trust, birth, input, referrals, specialist, communication, baby, meant	1

A combination of offline and online methods were developed and utilised to ensure data protection whilst enabling advanced analysis, with offline processing for sensitive data and online processing for non-sensitive data using the `Claude 3 Opus' language model.

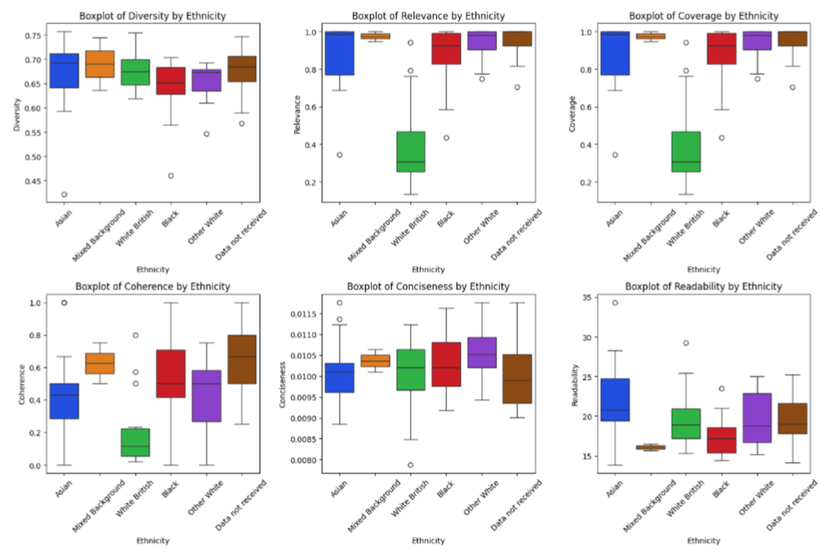
Topics around the sentences of the Black ethnic group

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

Ensuring model equity during topic modelling through traceability and evaluations.



Summarisation [3]



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

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.

Table 3: Average and std. values of metrics by ethnicity when using BART

Metric	Asian	Black	DNR	MB	OW	WB
Diversity	0.67 ± 0.08	0.64 ± 0.07	0.67 ± 0.06	0.69 ± 0.08	0.65 ± 0.04	0.68 ± 0.04
Relevance	0.89 ± 0.18	0.87 ± 0.17	0.95 ± 0.09	0.97 ± 0.04	0.93 ± 0.10	0.40 ± 0.22
Coverage	0.89 ± 0.18	0.87 ± 0.17	0.95 ± 0.09	0.97 ± 0.04	0.93 ± 0.10	0.40 ± 0.22
Coherence	0.46 ± 0.32	0.54 ± 0.29	0.65 ± 0.26	0.63 ± 0.18	0.42 ± 0.27	0.18 ± 0.19
Conciseness	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
Readability	21.54 ± 5.12	17.53 ± 2.54	19.23 ± 3.37	16.08 ± 0.59	19.54 ± 3.63	19.62 ± 3.58

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

The work was jointly funded by The Health Foundation and the NHS AI Lab at the NHS Transformation Directorate, and supported by the National Institute for Health Research. The project is entitled "I-SIRch - Using Artificial Intelligence to Improve the Investigation of Factors Contributing to Adverse Maternity Incidents involving Black Mothers and Families" AI_HI200006. The authors would like to acknowledge MNSI for their feedback on the paper.

Our publications:
1. 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
2. Unveiling Disparities in Maternity Care: A Topic Modelling Approach to Analysing Maternity Incident Investigation Reports, AIiH 2024, LNCS, 2024
3. Intelligent Multi-Document Summarisation for Extracting Insights on Racial Inequalities from Maternity Incident Investigation Reports, AIiH 2024, LNCS, 2024