I-SIRch: AI-powered concept annotation tool for equitable extraction and analysis of safety insights from maternity investigations

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

[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
[†] These authors contributed equally to this work

Abstract

Background

Maternity care is a complex system involving treatments and interactions between patients, healthcare providers, and the care environment. To enhance patient safety and outcomes, it is crucial to understand the human factors (e.g. individuals' decisions, local facilities) influencing healthcare. However, most current tools for analysing healthcare data focus only on biomedical concepts (e.g. health conditions, procedures and tests), overlooking the importance of human factors.

Methods

We developed a new approach called I-SIRch, using artificial intelligence to automatically identify and label human factors concepts in maternity investigation reports describing adverse maternity incidents produced by England's Healthcare Safety Investigation Branch (HSIB). These incident investigation reports aim to identify opportunities for learning and improving maternal safety across the entire healthcare system. Unlike existing clinical annotation tools that extract solely biomedical insights, I-SIRch is uniquely designed to capture the socio-technical dimensions of patient safety incidents. This innovation enables a more comprehensive analysis of the complex systemic issues underlying adverse events in maternity care, providing insights that were previously difficult to obtain at scale. Importantly, I-SIRch employs a hybrid approach, incorporating human expertise to validate and refine the AI-generated annotations, ensuring the highest quality of analysis.

Findings

I-SIRch was trained using real data and tested on both real and synthetic data to evaluate its performance in identifying human factors concepts. When applied to real reports, the model achieved a high level of accuracy, correctly identifying relevant concepts in 90% of the sentences from 97 reports (Balanced Accuracy of $90\% \pm 18\%$ (Recall $93\% \pm 18\%$, Precision $87\% \pm 34\%$, F-score $96\% \pm 10\%$). Applying I-SIRch to analyse these reports revealed that certain human factors disproportionately affected mothers from different ethnic groups. In particular, gaps in risk assessment were more prevalent for minority mothers, whilst communication issues were common across all groups but potentially more for minorities.

Interpretation

Our work demonstrates the potential of using automated tools to identify human factors concepts in maternity incident investigation reports, rather than focusing solely on biomedical concepts. This approach opens up new possibilities for understanding the complex interplay between social, technical and organisational factors influencing maternal safety and population health outcomes. By taking a more comprehensive view of maternal healthcare delivery, we can develop targeted interventions to address disparities and improve maternal outcomes. Targeted interventions to address these disparities could include culturally sensitive risk assessment protocols, enhanced language support, and specialised training for healthcare providers on recognising and mitigating biases. These findings highlight the need for tailored approaches to improve equitable care delivery and outcomes in maternity services. The I-SIRch framework thus represents a significant advancement in our ability to extract actionable intelligence from healthcare incident reports, moving beyond traditional clinical factors to encompass the broader systemic issues that impact patient safety.

Keywords

maternal care

*Corresponding Author:
Email Address: G.Cosma@lboro.ac.uk (Georgina Cosma)

Introduction

Background

A recent 2023 report by the UK Parliament's Women and Equalities Committee on Black maternal health highlights the stark disparities that exist in maternal mortality rates between ethnic groups in the UK [1]. Black women are nearly 4 times more likely to die during pregnancy or childbirth than White women. Asian women face almost double the risk of maternal mortality compared to White women. There are also significant differences depending on socioeconomic status, with women in the most deprived areas being 2.5 times more likely to die than those in the least deprived areas. In 2017, the UK government and NHS had set a goal to reduce stillbirths, newborn deaths, maternal deaths, and newborn brain injuries by 50% by 2025. However, there has been little progress on decreasing maternal mortality rates. When excluding deaths from COVID-19, the maternal mortality rate between 2010-2012 and 2018-2020 increased by 3% [1].

Patient safety

The Healthcare Safety Investigation Branch (HSIB) conducted independent investigations of patient safety incidents in NHS-funded care across England. Established in 2017 and funded by the Department of Health and Social Care, HSIB aimed to improve patient safety through its investigations. HSIB was hosted by NHS England and operated independently. In October 2023, HSIB transformed into two organisations: the Maternity and Newborn Safety Investigations (MNSI), which is hosted by the Care Quality Commission, and the Health Services Safety Investigations Body (HSSIB), which is an independent statutory body. HSIB carried out investigations into adverse incidents during pregnancy and birth. After each investigation, they produced a report discussing the investigation findings and recommendations. These reports were intended to provide insights and lessons learned for mothers and families affected. Maternity investigation work continues at MNSI.

Human factors

Addressing human factors in healthcare investigations is crucial for improving patient safety and outcomes. While clinical factors are important, a significant proportion of adverse events stem from systemic issues related to communication, teamwork, organisational culture, and other human and organisational factors. Current tools for analysing healthcare data predominantly focus on biomedical concepts, overlooking these critical human and system-level contributors to patient safety. This narrow focus limits our ability to identify and address the root causes of many preventable incidents.

Background into concept annotation tools

The increasing adoption of concept annotation tools in healthcare and other sectors stems from their ability to extract insights from unstructured text data. Manual annotation requires significant human effort and expertise. This demand has led to growing interest in automated annotation through machine learning and natural language processing. These techniques can accelerate the labelling of domain-specific concepts in text data, enabling more efficient intelligence extraction, search, and analysis. Several concept annotation tools have been specifically designed for the clinical field, including MetaMap [2], NCBO Annotator [3], cTAKES [4], Biomedical Named Entity Recognition (BINER) [5], UniversalNER [6], MedCat [7], and BioBERT [8]. These tools leverage medical ontologies to detect clinical concepts within unstructured text and have shown the potential to automate parts of the concept annotation process. These clinical concept annotation tools enable the extraction of biomedical insights but overlook human factors involved in healthcare delivery. Furthermore, significant challenges remain in developing fully automated annotation capabilities. Hybrid human-AI systems provide a promising approach for concept annotation by combining complementary strengths. The accuracy of human experts can be paired with the scalability of AI to efficiently generate high-quality labelled datasets. However, developing robust AI-powered annotation tools introduces several technical challenges. Specifically, large, annotated corpora, sufficient computational resources, and ongoing human validation are required to train and evaluate these machine learning models. With thoughtful co-design, machine learning automation and human intelligence can build upon each other to enable continued progress in extracting valuable insights from unstructured textual data.

Our contributions

We introduce the Intelligence Safety Incident Reporting and Annotation (I-SIRch) framework, a novel approach to human factors analysis in maternity incident investigations. I-SIRch employs advanced computational methods and machine learning algorithms to automatically annotate unstructured text from investigation reports using the Safety Intelligence Research (SIRch) taxonomy. The SIRch taxonomy provides a systematic methodology for extracting safety insights from healthcare investigations, focusing on work system designs' influence on patient safety outcomes. By analysing complex healthcare delivery systems, I-SIRch identifies opportunities to improve design and reduce safety incidents [9]. Unlike existing clinical annotation models that focus solely on biomedical concepts, I-SIRch addresses a critical gap by incorporating human factors insights. This holistic approach enables a more comprehensive understanding of patient safety incidents, facilitating more effective interventions and system improvements. To the best of our knowledge, our work demonstrates the first successful automatic annotation of incident reports with a human factors' taxonomy. By revealing the potential of human factors annotation using SIRch, we also open new research directions for understanding and improving socio-technical systems involved in healthcare delivery. The key contributions of this article are:

- Provides a computational framework, I-SIRch, that automatically annotates sentences (and text segments) with multiple applicable concepts. The framework was trained on sentences extracted from maternity incident investigation reports that were manually annotated by

patient safety experts. The experts assigned relevant human factor concept(s) from the SIRch taxonomy to each sentence. This mapping enabled the model to learn links between the incidents discussed and socio-technical factors implicated.

- The I-SIRch framework enables human-AI collaboration, allowing the machine learning model to continuously improve its knowledge by learning from new human expert annotations. This human-in-the-loop process helps develop a robust, adaptable model for extracting insights from reports by leveraging both human expertise and the model's growing knowledgebase.

- I-SIRch was tested on 818 synthetically generated sentences; and thereafter tested on an additional 1960 sentences extracted from 97 real reports, to evaluate its generalisability on unseen data. The synthetic sentences were semantically similar to those found in real reports. The real sentences were extracted from healthcare reports that contained sentences that were not previously seen by the model (i.e. not used for training).

- Statistical analysis was performed on the annotated text across demographic groups. The annotation results were analysed to identify insights and differences between ethnic groups regarding factors contributing to incidents.

I-SIRch Framework Highlights

Framework Overview

- I-SIRch: An innovative AI-powered framework designed to automatically identify human factors in maternity incident investigations, moving beyond traditional clinical focus.

- Utilises the SIRch taxonomy, combining established patient safety frameworks (SEIPS) with NHS England's Learn from Patient Safety Events (LFPSE) service categories. This integration allows for a comprehensive analysis that aligns with national reporting standards while incorporating systems-based approaches to patient safety.

- Trained and tested on real maternity investigation reports, ensuring direct relevance to clinical practice and patient safety improvement efforts.

- Employs a 'human-in-the-loop' approach, allowing for continuous refinement based on expert input and emerging safety insights.

- Demonstrates high accuracy in identifying human factors, with balanced accuracy of 90% on real-world data, indicating strong potential for practical application.

- Enables systematic, large-scale analysis of socio-technical factors in maternity care, potentially revealing systemic issues not easily identifiable through traditional review methods.

- Shows promise in identifying care disparities across different ethnic groups, supporting efforts to improve healthcare equity in maternity services.

- Offers a scalable solution to enhance learning from safety incidents, potentially leading to more targeted interventions and improved patient outcomes in maternity care.

Key Human Factors Identified
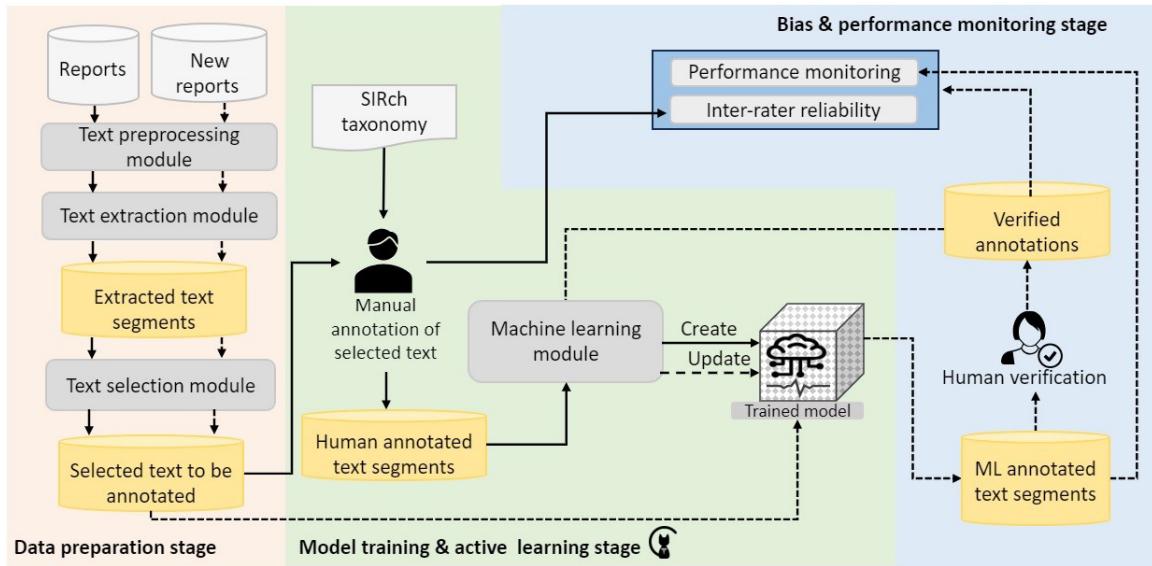
- Organisation-Teamworking in 155 reports, highlighting the critical role of effective collaboration in maternity care.

- Organisation-Communication factor in 159 reports, underscoring the importance of clear information exchange in healthcare settings.

- Assessment, investigation, testing, and screening issues 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 (99 reports) and decision errors (89 reports) were prominent, revealing human performance concerns.

- COVID-19 impacted 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.

Methods

Proposed I-SIRch framework

This section presents the machine learning based framework, I-SIRch, designed for the automatic annotation of maternity incident investigation reports using the human factors taxonomy, SIRch. Figure 1 illustrates the key components and stages of the I-SIRch framework, depicting its ability to develop a machine learning model capable of learning and adapting to additional inputs over time. A description of each component of the I-SIRch framework is provided below.

Figure 1: Proposed I-SIRch framework

Data preparation stage

Reports

The process starts with a dataset containing maternity investigation reports in PDF format. These reports are introduced into I-SIRch for initial training and subsequently processed through the various modules, as indicated by the solid line in Figure 1.

New reports

These reports are different from those used during the initial model training stage. They are introduced into I-SIRch post-training and subsequently processed through the various modules, as indicated by the dashed line in Figure 1.

SIRch taxonomy

SIRch codifies and combines the internationally recognised Systems Engineering Initiative for Patient Safety (SEIPS) method [10, 11, 12] with the incident categories used by NHS England and NHS Improvement's Learn from Patient Safety Events (LFPSE) service [13]. The SIRch taxonomy that was utilised for training the machine learning model is shown in Supplementary File 1.

Text preprocessing module

This module's purpose is to preprocess PDF files by eliminating unrecognisable fonts and symbols. It takes a collection of reports located in the 'Reports' repository as input. In the process, it extracts and decrypts text from PDF files using Unicode Transformation Format 8-bit (UTF-8) encoding, ensuring readability. It then eliminates unnecessary symbols such as inverted commas ('), dots (.), and unrecognised symbols. Subsequently, the extracted text undergoes cleaning to remove any extraneous sections or elements that do not contribute to the analysis. This accurate extraction of text from PDFs facilitates subsequent natural language processing

and makes the content accessible in a standardised machine-readable format. Following the preprocessing, the module stores the processed text in a file for subsequent processing by the 'Text extraction module'.

Text extraction module

The purpose of this module is to extract text from reports based on specific criteria, such as sections, pages, paragraphs, or the entire report. It takes a set of reports as input from the 'Text preprocessing module'. The user defines the target section of the report to be extracted. The 'text extraction' module then carries out the extraction based on the user's selection, whether it is a particular section, page, paragraph, or entire report. It is important to note that in structured reports, the user specifies a criterion for text selection (e.g. a section within an investigation report), which will be extracted from all reports, while in unstructured documents, the user has the option to extract text from the entire document. Following the extraction process, the module stores the extracted text in a file, preparing it for further processing by the 'Text selection module'.

Text selection module

The purpose of this module is to identify and select sentences with negative connotations, references to physical characteristics, and medicine names (dispensing medication). It takes the extracted text as input from the 'Text extraction module'. Initially, it uses a deep learning transformer model that was fine-tuned for identifying negated sentences. The module also maintains a list of negation keywords, including the terms 'not' and 'never', to catch any negated cases that might be missed by the model. Additionally, the module scans for the phrase 'in line with' in positive sentences and marks tokens following it as affirmed. Any sentence found to possess a negative meaning or contain references to physical characteristics or medication names is flagged for annotation. Subsequently, the module stores the selected sentences in a

file entitled 'Selected text to be annotated', preparing them for further training (i.e. retraining) by the 'Machine learning module'. Figure 2 depicts the process of how sentences are selected for automatic concept annotation.

Manual annotation of selected text

In this process, multiple human experts independently annotate the same set of selected sentences using the SIRch framework. This approach enables the evaluation of inter-rater reliability (IRR) and the consistency of annotations across various experts. These experts manually annotate (or code) the selected segments that highlight essential investigation findings. They accomplish this by assigning pertinent concepts from the specified taxonomy to the sentences or individual words. The annotation process can make use of the MedCATtrainer tool [7]. An illustrative example of manual annotation is provided in Supplementary Table 12. The resulting manually annotated sentences are systematically stored in a repository named 'Human annotated text segments'. These annotations serve as training data for the 'Machine learning module'.

Model training and active learning stage

Machine learning module

This module is responsible for training a natural language processing (NLP) model with the ability to automatically annotate concepts within selected text segments. It takes as input sentences manually annotated by human experts, stored in the 'Human annotated segments' repository. The 'machine learning module' encompasses several steps. Initially, it establishes a concept database (CDB) containing the SIRch concepts to be recognised. The 'Human annotated text segments' are then uploaded into the MedCATtrainer [7], where human experts use its graphical user interface to annotate the text segments. Once MedCATtrainer learns from a few annotated examples, it can make predictions, facilitating faster annotation through active learning. A concept needs to appear at least once in the training data to be considered for training. Human experts utilise MedCATtrainer to validate and correct automatic annotations. These validated examples contribute to fine-tuning the model, promoting an active learning process. The module's output is a 'Trained model' that is stored and ready for deployment in automatic annotation tasks, simplifying the recognition of concepts within textual data. When the trained model is employed to make predictions on previously unseen reports, the annotated sentences are stored in the 'ML annotated text segments' repository and subsequently utilised in the 'Bias and performance monitoring' stage.

Bias and performance monitoring stage

Human verification

The purpose of this stage is to assess the correctness of annotations generated by the trained model. It begins by taking the sentences annotated by the trained model as input for human evaluation. The process involves several steps. Initially, a new dataset is passed through the I-SIRch pipeline and processed through the 'Trained model,' which annotates the selected text. These annotated sentences are stored in the 'ML annotated text segments' repository. Human experts then access this repository to manually verify whether the model's predicted annotations for each sentence are correct or incorrect. These verified labels, indicating correctness or errors, are recorded as 'Verified annotations'. Any sentences with errors are isolated and used to retrain the model, thereby enhancing its performance. The updated model, after retraining with the verification data, becomes ready for deployment on new sentences. This verification-retraining loop allows for continuous performance monitoring and improvement of the model. Finally, the verified annotations (i.e. concepts) are passed on to the 'Bias and performance monitoring stage', where a more comprehensive evaluation of the model's biases, errors, and real-world effectiveness is conducted to ensure rigorous monitoring and ongoing enhancement of the model's performance and reliability.

Performance monitoring

The purpose of this module is to evaluate the performance of the trained machine learning model. This evaluation takes as input a dataset comprising verified human annotations, which forms the foundation for rigorous analysis. The process entails the calculation of various performance metrics, including balanced accuracy, precision, recall, and F-score, providing a comprehensive assessment of the model's concept annotation performance. Moreover, the model undergoes scrutiny for biased performance across diverse demographic groups or sentence types, ensuring fairness and impartiality. For instance, its performance is examined for systematic variations when annotating sentences from different ethnic groups.

Inter-rater reliability (IRR)

Inter-rater reliability (IRR) is calculated by comparing annotations from multiple human experts for the same sentences. High IRR indicates consistent human annotation, whereas low IRR may indicate ambiguity within the taxonomy. Ongoing monitoring of both model performance metrics and IRR provides valuable insights into potential refinements in the taxonomy, annotation process and model training approach. This iterative process, continuously adapting to new data, allows for ongoing improvements and updates to the model, while effectively addressing biases. The end goal is the development of a machine learning model that leverages both expert-annotated data and its own predictions to continually learn and enhance its text classification capabilities.

Experiment methodology

Four data batches were utilised for training and testing I-SIRch. Supplementary Table 1 holds information about the batches, and Supplementary Table 13 shows the frequencies of concepts found in the batches that were utilised for developing the model. Figure 3 shows the number of files (i.e. maternity investigation reports) in Batches 1, 2 and 3 containing each SIRch concept.

Figure 2: Text selection module
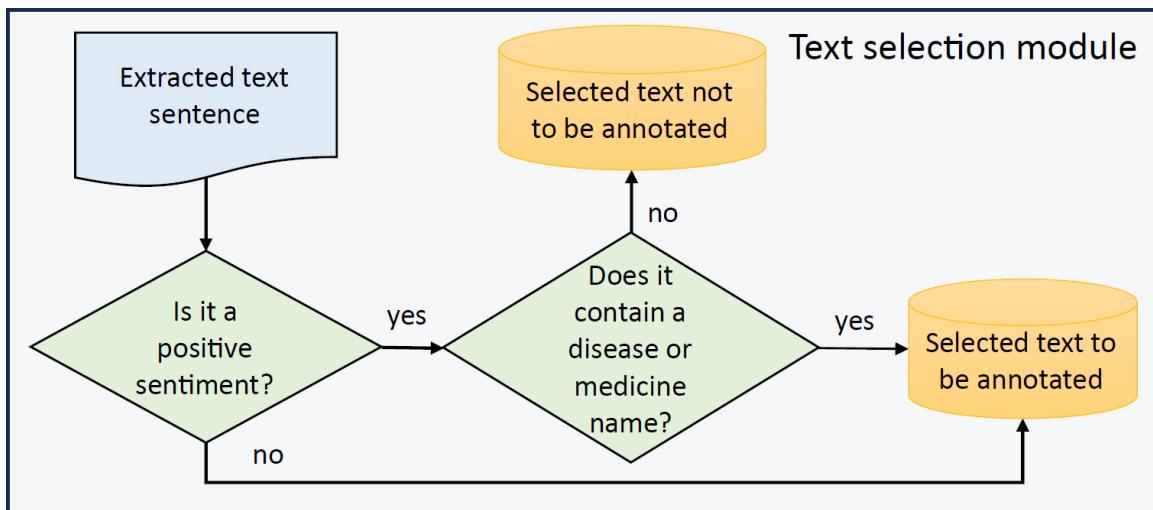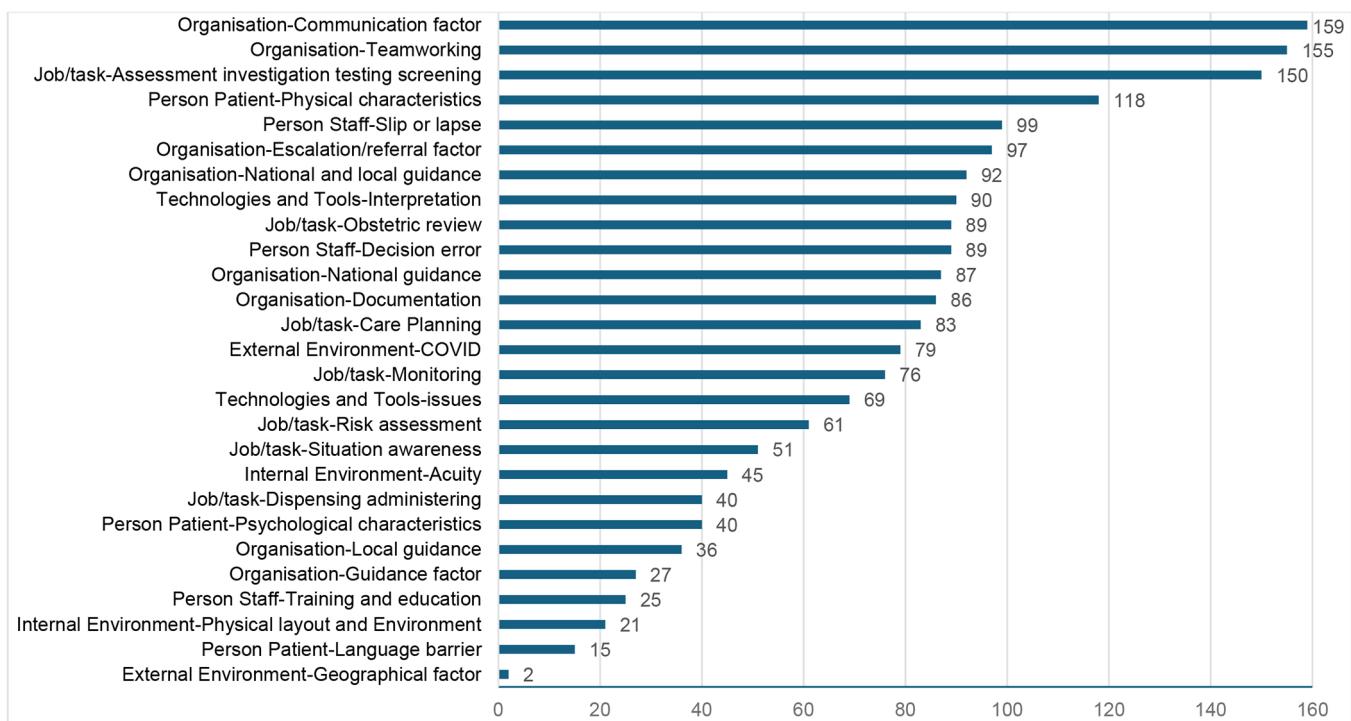
Figure 3: Number of reports with the concept

Data is shown in Supplementary Table 13.

Training

The maternity incident reports (Batch 1, n = 76) provided by HSIB were utilised for training the model, following the process described in Section *I-SIRch framework*. With regards to the 'Human annotation' stage, there were two experts from HSIB and a human factors expert from Loughborough University annotating the reports. Of those 76 maternity investigation reports, 20 reports (sentences = 184) were randomly selected for annotation by three human experts as discussed in the inter-rater reliability (IRR) step of the *Bias and performance monitoring stage*. The observed IRR among three annotators was 80.15% (see Supplementary Table 14). Ethnic groups were unknown and not considered during the human annotation stage to reduce any potential biases entering the system during human annotation. Performance was evaluated by comparing the predicted vs actual (i.e. relevant) concepts.

Testing (Test A)

The synthetic (test) dataset (Batch 4, n = 76) that was generated following the process described in Supplementary File 2, served as the set for testing the performance of the trained model on a previously unseen but semantically similar dataset. The relevant concepts for each sentence were known since they were mapped to the concepts of those sentences extracted from the real dataset. Ethnicity was considered when

analysing the performance of the model across the various ethnic groups.

Retraining 1

After testing with the synthetic dataset (Batch 4), a total of 15 real reports (Batch 2) were utilised to further train the model. These 15 reports were selected because they contained codes that were less frequent in the previous training set (Batch 1), to help improve model performance. The I-SIRch model was utilised to extract and automatically annotate 344 sentences from these 15 reports, and thereafter, a human annotator verified the correctness of the annotations.

Retesting 1 (Test B)

Thereafter a total of 97 real reports (Batch 3) were utilised for further testing. The I-SIRch model was utilised to automatically annotate 1960 sentences extracted from the reports, and thereafter a human annotator verified the correctness of the annotations. This enabled for evaluation of performance with regards to applying the trained model on a real unseen dataset. Ethnicity and healthcare outcomes were not provided with these reports. Retesting 1 is referred to as Test B.

Retraining 2

The 97 reports (Batch 3) were then used for retraining the model.

Retesting 2 (Test C)

After retraining the model with Batch 3, the model was tested using the synthetic dataset (Batch 4).

Metrics for evaluating the performance of machine learning based annotation

A set of evaluation metrics was employed to assess the trained model's performance in automatically annotating sentences from the test sets. In the context of a concept annotation task, the evaluation of system or annotator performance involves considering four key metrics: True Positives (TP), False Positives (FP), False Negatives (FN); and True Negatives (TN). These are described below, and provide the basis for calculating Recall, Precision, and the F-Score to assess the performance of the system.

True Positive (TP) is the count of concepts in a sentence that were correctly identified and annotated as a specific concept. In other words, these are instances where the system correctly recognised and marked the concept where it truly exists in the sentence.

False Positive (FP) is the count of concepts that were incorrectly annotated in a sentence, falsely identified as part of the concept when they are not. These are instances where the system made an error by including concepts that should not have been part of the annotated concept.

False Negative (FN) denotes the count of concepts that should have been annotated as part of the concept in a sentence but were missed or omitted during the annotation process. These are instances where the system failed to recognise and include concepts that should have been part of the concept.

True Negative (TN) is the count of concepts that were not annotated because they should not have been included in the concept. These are instances where the system correctly identified that certain concepts were not relevant to the annotated concept and left them out.

Recall measures the proportion of actual annotations that are correctly identified by the system. It assesses how comprehensive the system is in capturing all the relevant annotations. A higher recall indicates that the system can identify a greater number of actual annotations, although it may also include some false positives.

$$Recall = \frac{TP}{TP + FN}, \in (0, 1) \qquad (1)$$

High recall indicates that the system or annotator is good at capturing all instances of the concept in the text.

Precision measures the proportion of predicted annotations that are correct. It indicates how accurately the system can identify the relevant annotations without mislabelling irrelevant ones. A higher precision indicates that the system is more reliable in identifying the correct annotations.

$$Precision = \frac{TP}{TP + FP}, \in (0, 1) \qquad (2)$$

High precision indicates that when the system or annotator claims a code is part of the concept, it is highly likely to be correct.

F-score is a balanced measure that considers both precision and recall in a concept annotation task. It is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between ensuring all relevant concepts are included in the annotated concept (high recall) while maintaining high precision:

$$F\text{–}score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \in (0, 1) \qquad (3)$$

The *F*-Score is particularly useful when aiming to strike a balance between ensuring that all relevant concepts are included in the annotated concept (high recall) while maintaining high precision. It serves as an overall measure of the effectiveness of the concept annotation system or annotator.

Accuracy measures the overall correctness of the annotations made by a system. It is defined as the ratio of correctly annotated concepts (True Positives and True Negatives) to the total number of annotations (True Positives + True Negatives + False Positives + False Negatives), as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \in (0, 1) \qquad (4)$$

Accuracy provides an overall assessment of how well the system or annotator is performing in correctly identifying both positive and negative cases within the annotated concepts, thus quantifying the system's annotation correctness.

Balanced accuracy. Accuracy can be a misleading metric for imbalanced data sets. Given that some concepts are more frequent than others in the dataset, it is important to also report the balanced accuracy as a measure of accuracy. The balanced accuracy is the average between the sensitivity and the specificity, which measures the average accuracy obtained from both the minority and majority classes. The balanced accuracy measure was calculated as follows. First calculating the True Positive Rate (TPR) and True Negative Rate (TNR) as $TPR = TP/(TP + FN)$ and $TNR = TN/(TN + FP)$, respectively. Balanced Accuracy is thereafter calculated as follows.

$$Balanced\ Accuracy = \frac{TPR + TNR}{2}, \in (0, 1) \qquad (5)$$

Inter-rated reliability (IRR) is the degree of agreement among the human experts who annotated the initial reports before being used for initial model training. IRR results are shown in the Supplementary Table 14. IRR is calculated as follows.

$$IRR = \frac{Total\ number\ of\ concept\ agreements}{Total\ number\ of\ concepts \times Number\ of\ annotators} \qquad (6)$$

Results

This section describes the experiment results. Supplementary Table 1 shows the batches utilised for retraining and testing the model. The Experiment Methodology section described the retraining and testing process that was followed. I-SIRch was tested three times as follows: Test A: Training on Batch 1 (real-data) and testing on Batch 4 (synthetic data). Test B: Retraining on Batch 2 (real data), and testing on Batch 3 (real data). Test C: Retraining again on Batch 3 (real data) and testing on Batch 4 (synthetic data). The results of all tests are shown in Supplementary Tables 6, 7.

Dataset of maternity incident investigation reports

HSIB provided a random set of 188 investigation reports describing adverse maternity incidents (see Batches 1-3 in Supplementary Table 1). The reports were written between 2019 to 2022. The number of reports for each year is as follows: 4 reports in 2019, 115 reports in 2020, 42 reports in 2021, and 27 reports in 2022. Ethnicity was only provided for Batch 1 which comprised 76 reports. The discussion that follows focuses on Batch 1, since ethnicity was not available for the other batches.

Batch 1 analysis

The percentage of reports containing each concept/human factor across each ethnic group is shown in Supplementary Table 2. Figure 4 illustrates a description of the number of reports and concepts (human annotations using SIRch) found within the reports of Batch 1. The corresponding data can be found in Supplementary Table 3. As shown in Figure 5, 70.93% of the reports are of the White British ethnic group, compared to 81% nationally. The reports also

cover 8.97% Asian (vs. 9.6% nationally) and 8.35% Black (vs. 4.2% nationally) ethnic groups. The distribution of ethnic groups in the dataset closely mirrors the population of England based on the 2021 census [14]. This indicates that the reports provide a fairly representative sample across various ethnic groups in England, with a reasonable representation of the Black ethnic group compared to the national population [14] (Figure 5). The frequency distribution of concepts across reports for each ethnic group are provided in Supplementary Table 4.

Supplementary Table 5 provides a summary of the 76 reports describing adverse maternal and neonatal outcomes, categorised by ethnicity. These outcomes include instances of babies who received therapeutic hypothermia, early neonatal deaths, maternal deaths, and intrapartum stillbirths. Within the dataset, the most common adverse incident for reported cases for mothers of the White British ethnic group was babies receiving therapeutic hypothermia (30 out of 52 reports of their reports), whereas, for mothers of the Black ethnic group, it was maternal death (4 out of 7 of their reports). Please note that the dataset is a random set of reports extracted for testing the proposed concept annotation approach and demonstrating how the results can be analysed, and it is not a representation of the reports held by HSIB.

Model performance evaluation on the synthetic (test) set across ethnic groups

Test A

The I-SIRch model was initially trained on Batch 1 and tested on Batch 4. Table 1 shows the performance of the model when using the metrics described in the Experiment Methodology section. Supplementary Table 8 shows the performance of the model across concepts on the synthetic test dataset (Batch 4). I-SIRch was initially utilised for automatically coding a test set of 970 synthetic sentences (Batch 4). The process describing how the synthetic sentences were generated is found in Supplementary File 2. Metrics were applied to evaluate the performance of the model per sentence and overall model performance. The trained machine learning model demonstrated strong performance on the synthetic test set across all evaluation metrics. The precision of 0.87 indicates that the model has a low false positive rate in its annotations. The recall of 0.93 indicates effective identification of appropriate concepts with minimal false negatives. Finally, the F-score of 0.96 validates the overall effectiveness of the model by balancing both precision and recall. The high F-score indicates that the model was able to apply concepts where relevant while avoiding incorrect annotations. In summary, the high precision, recall, and F-score reflect the model's capability at multi-label annotation of sentences using the SIRch taxonomy. Further tests to explore the feasibility of the I-SIRch concept annotation framework on real data are described later in the paper (Tests B and C).

Performance across ethnic groups (Test A)

The synthetic dataset contains 818 sentences (970 concept annotations), with the distribution across ethnic groups matching those of Batch 1 (Figure 4). The distribution of

Figure 4: Total number of concepts and average number of concepts per report across ethnic groups
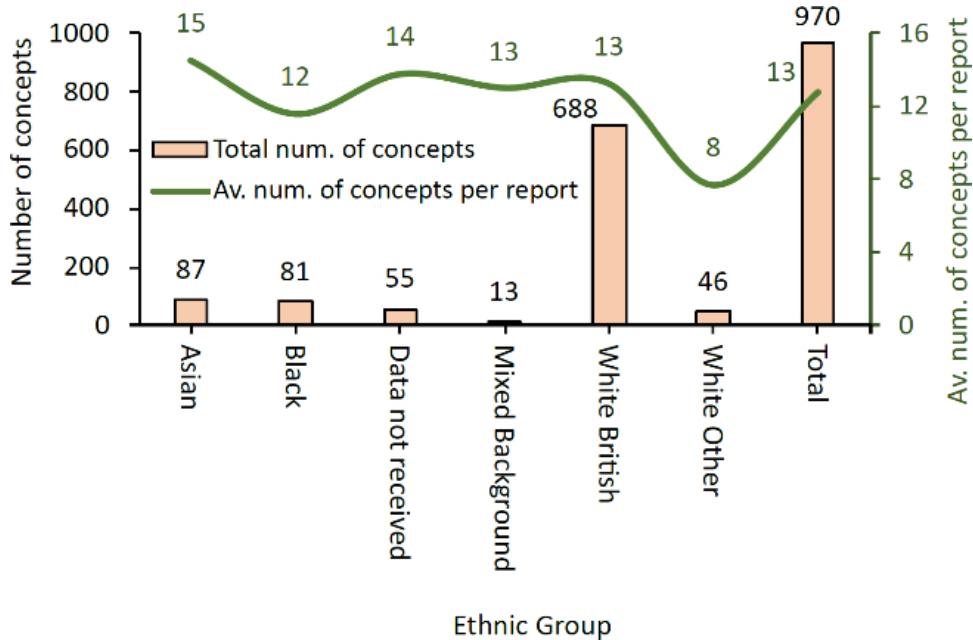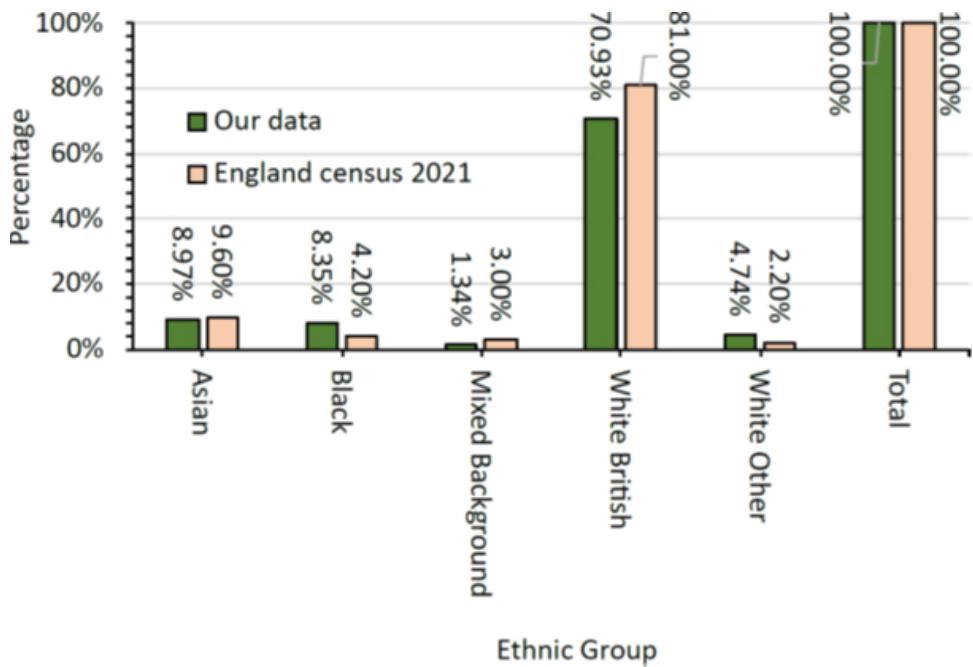
Figure 5: Percentage of mothers and families from each ethnic group in our dataset compared to England's 2021 census data

ethnic groups of the synthetic dataset maps to the real datasets because the synthetic sentences were generated from real sentences extracted from reports for which the ethnicity was known. Supplementary Table 9 (Test A) shows the results of the correct and incorrect annotations per ethnic group. The model correctly annotated 70.93% of sentences, with 686 correct annotations out of a total of 970 (see Supplementary Table 9). The model performed well across groups when ethnicity was considered (see Figure 6), showing no significant accuracy discrepancies (here accuracy refers to the percentage of correctness in identifying the TP annotations), as discussed in the following subsection. As shown in Supplementary Table 9, the sentences of the reports

of the White British ethnic group saw the highest accuracy of 72.82%, with 499 correct annotations out of 688. For minority ethnic groups, the model obtained accuracy values of 67.82% on Asian examples (59 correct out of 87), 66.67% on Black examples (54 correct out of 81), and 69.23% on the smaller Mixed Background set (9 out of 13). For the 'White Other' ethnic group, the model reached 69.57% accuracy (32 out of 46). Lower performance (i.e. 60% accuracy, with 33 correct out of 55) was observed in annotations where 'ethnicity data was not received', highlighting an area for improvement. However, the comparable accuracy across ethnic groups generally demonstrates the model's capability to learn meaningful patterns from diverse ethnic groups without
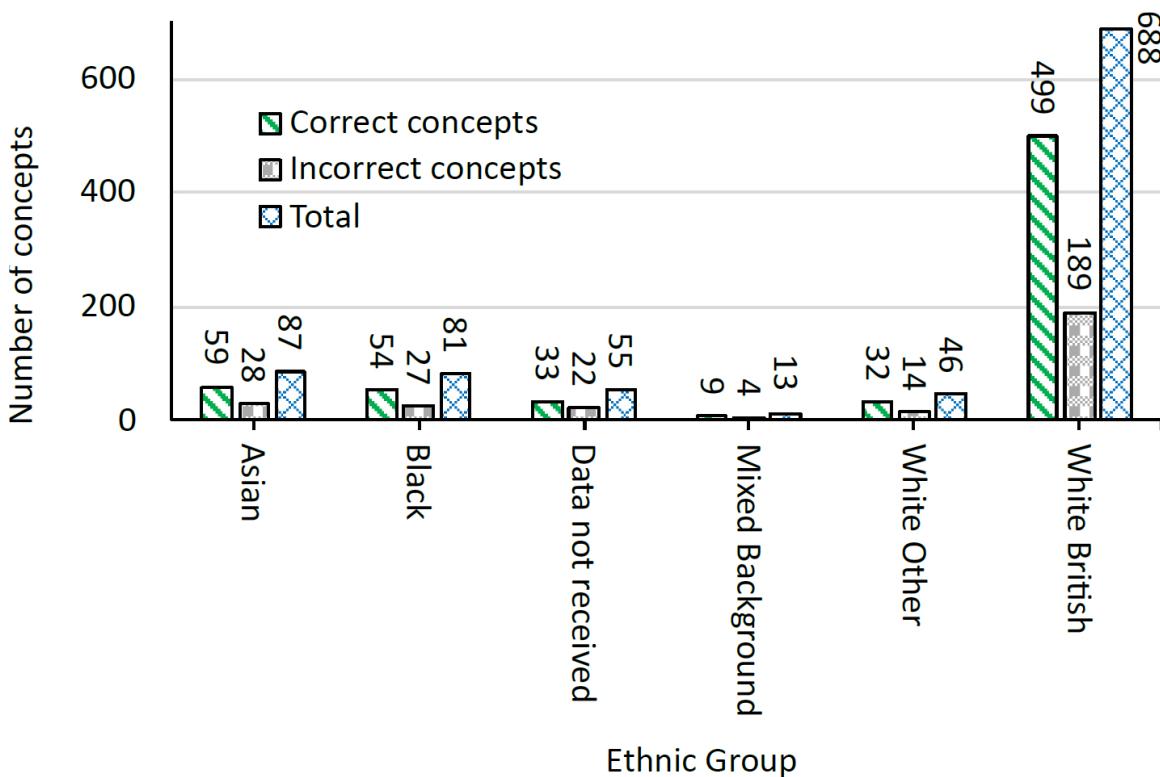
Table 1: Test A results: Performance of I-SIRch when trained on Batch 1 and tested on Batch 4

| Ethnic group	Precision	Recall	F-score	Misc.	Bal. Acc.
Asian	1.00 ± 0.00	0.60 ± 0.23	0.79 ± 0.09	0.40 ± 0.23	0.80 ± 0.11
Black	1.00 ± 0.00	0.59 ± 0.21	0.78 ± 0.08	0.41 ± 0.21	0.80 ± 0.10
Data not received	1.00 ± 0.00	0.59 ± 0.24	0.80 ± 0.08	0.41 ± 0.24	0.80 ± 0.12
Mixed Background	1.00 ± 0.00	0.65 ± 0.12	0.78 ± 0.09	0.35 ± 0.12	0.82 ± 0.06
Other White	1.00 ± 0.00	0.58 ± 0.25	0.79 ± 0.09	0.42 ± 0.25	0.79 ± 0.12
White British	1.00 ± 0.00	0.60 ± 0.23	0.79 ± 0.08	0.40 ± 0.23	0.80 ± 0.11
Average	1.00 ± 0.00	0.60 ± 0.21	0.78 ± 0.08	0.40 ± 0.21	0.80 ± 0.10

Table shows the mean and standard deviation values.
Bal. Acc.: Balanced Accuracy; Misc: Misclassification.

Figure 6: Average number of correct and incorrect concepts across various ethnic groups (Batch 1)

significant biases. Analysing by ethnic group provides valuable insights into model fairness, and the results indicate strong generalisability.

Are there any significant differences in the performance of I-SIRch on test data, across the Black and White ethnic groups?

A Wilcoxon signed-rank test was conducted to determine whether there were significant differences in the performance of I-SIRch's machine learning model between the Black and White British ethnic groups. The results showed no significant difference in model performance between Black (median $= 66.67\%$, SD $= 30.37\%$) and White British (median $= 70.14\%$, SD $= 7.5\%$) ethnic groups, $Z = -0.806$, $p = 0.42$. Based on these findings, it can be concluded that the

model demonstrated no significant variation in performance across the Black and White British ethnic groups analysed. The Wilcoxon signed-rank test indicates that the model performance was statistically comparable for the Black and White British ethnic groups represented in the dataset. However, the higher standard deviation for the Black ethnic group points to greater variability in model performance compared to the White British group. This suggests a more limited or inconsistent representation of text from the reports of the Black ethnic group. Expanding the diversity of the training data can enhance the model's ability to make fair and accurate predictions without unintended biases related to ethnicity and can reduce performance fluctuations across groups. Ongoing evaluation of the model will be performed and retrained over time to check for potential biases which is key to maintaining fairness as the model evolves.

Can performance improve when retraining the model with real data, and testing on real and synthetic data?

Test B

After the Retraining 1 phase (using Batch 2), the performance of the model was evaluated on a real dataset (Batch 3) comprising 1960 sentences extracted from 97 reports. The results are shown in Supplementary Table 6 (Test B). Note that ethnicity was not available for Batch 3 and hence a performance analysis of the model across the ethnic groups could not be conducted. The I-SIRch model reached the highest balanced accuracy of 0.90 $\pm$0.18 when tested on real data. The performance of I-SIRch on Test B was an improvement compared to its performance when tested on the synthetic dataset during tests A and C, with balance accuracy values of 0.80 ± 0.11 and 0.83 ± 0.08, respectively.

Test C

This test aims to evaluate the performance of the I-SIRch model when trained using all the available real data (Batches 1, 2 and 3) and tested on the synthetic data (Batch 4). Hence, during Retraining phase 2, Batch 3 was utilised to further retrain the model (i.e. model has already been trained with Batches 1 and 2), and testing was conducted using the synthetic data (Batch 4). Supplementary Table 9 shows the performance of I-SIRch when tested across each concept and ethnic group. The results of Test A and Test C can be directly compared because both tests were conducted on the synthetic dataset (Batch 4). Test B was conducted on Batch 3 and hence cannot be directly compared to the results of Test A and Test C. Comparing the results of Tests A and C in Supplementary Table 6 and Figure 7, there is an improvement in the performance of the model after retraining the model using Batch 2. The results of a Wilcoxon signed-rank test (see Supplementary Table 11) show that this improvement is not statistically significant. It is however worth noting that Supplementary Table 9 shows that there was an increase in the number of correctly annotated concepts across all ethnic groups.

Table 2 shows the average results across the ethnic groups when using various evaluation metrics and confidence intervals. For each metric, we computed the mean, standard deviation, and 95% confidence interval (CI). The 95% CIs were calculated using Student's t-distribution method, which is appropriate for smaller sample sizes and when the population standard deviation is unknown. The formula used was:

$$CI = \overline{x} \pm \left(t \times \frac{s}{\sqrt{n}} \right) \qquad (7)$$

where $\overline{x}$ is the sample mean, t is the critical value of the t-distribution for a 95% confidence level with $n - 1$ degrees of freedom, s is the sample standard deviation, and n is the sample size (number of concepts) for each ethnic group. We used specific t-values for each ethnic group based on their respective degrees of freedom: Asian ($n = 87$): $t_{(86)} \approx 1.988$; Black ($n = 81$): $t_{(80)} \approx 1.990$; Data not received ($n = 55$): $t_{(54)} \approx 2.005$; Mixed Background ($n = 13$): $t_{(12)} \approx 2.179$; Other White ($n = 46$): $t_{(45)} \approx 2.014$; White British ($n = 688$):

$t_{(687)} \approx 1.963$. In Table 2, the confidence intervals (CIs) provide an indication of the precision of our estimates. Narrow CIs suggest high confidence, indicating that the sample data provides a reliable estimate of the population parameter. Wide CIs indicate low confidence, reflecting greater uncertainty due to variability or smaller sample sizes. For precision, where all values were 1.00 with no variation, we report the CI as [1.00, 1.00]. This methodology accounts for the varying sample sizes across ethnic groups, providing more accurate confidence intervals, especially for groups with smaller sample sizes. It is worth noting that for the Mixed Background group, which has a notably smaller sample size ($n = 13$), the confidence intervals should be interpreted with caution due to the increased margin of error.

Discussion and conclusion

This paper introduces I-SIRch, a novel framework for automating the annotation of maternity incident investigation reports using human factors concepts. I-SIRch utilises the Safety Intelligence Research (SIRch) taxonomy developed by England's Healthcare Safety Investigation Branch to analyse how work system factors contribute to patient safety incidents. The key innovation is the annotation of reports based on human and system-level issues, providing insights into socio-technical dimensions of safety lapses in healthcare delivery. The framework uses computational methods to extract, prepare, and annotate unstructured text from investigation reports. A machine learning model annotates sentences with applicable SIRch concepts, implicating various human factors. This work establishes a foundation for extracting intelligence on work system contributors to patient harm, complementing clinical perspectives. As healthcare adopts AI automation, integrating human factors with biomedical understanding offers a comprehensive view of safety vulnerabilities. I-SIRch provides a step towards rigorous, scalable analysis of socio-technical dimensions in health services. The human-in-the-loop approach ensures continuous model evolution while incorporating human expertise. By extracting insights into human and system-level factors contributing to incidents, I-SIRch can inform targeted improvements to complex healthcare systems, impacting patient outcomes.

Potential impact on maternal healthcare practice and policy

I-SIRch could provide a more comprehensive view of adverse events by integrating seamlessly with existing clinical annotation tools and incident reporting systems. This holistic approach could lead to more targeted interventions, such as improved communication protocols or culturally sensitive care practices, particularly benefiting minority groups who face higher risks. Healthcare providers could use I-SIRch's outputs to identify systemic issues and inform staff training, while policymakers could leverage its insights for evidence-based decision making. To realise these benefits, future work should focus on validating I-SIRch in diverse healthcare settings and exploring its potential to reduce maternal health disparities through pilot studies and collaborations with healthcare

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

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]

Table shows results of I-SIRch performance across ethnic groups, showing number of annotations, mean values, standard deviations (SD), and 95% Confidence Intervals (CI).

institutions. Efforts should be made to transform I-SIRch into a user-friendly, integrated tool that healthcare professionals and researchers can easily implement in their workflows, including developing interfaces for seamless integration within incident reporting systems.

Limitations around patient and public involvement
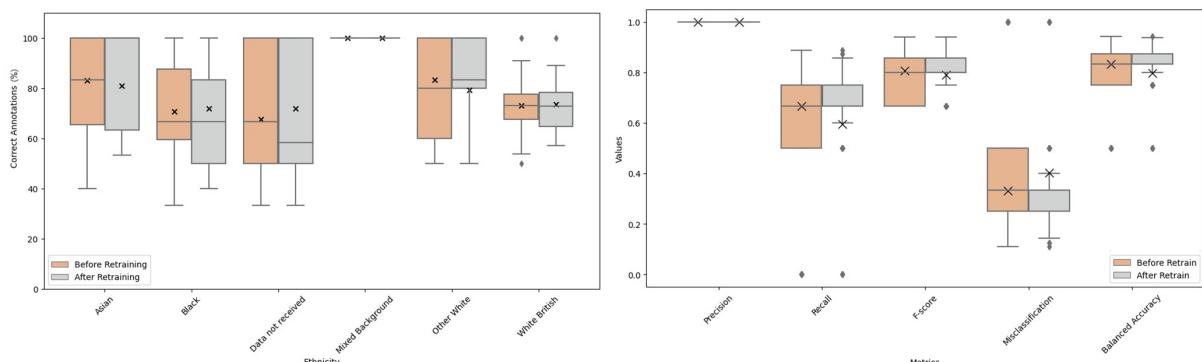
While the I-SIRch framework initially incorporated some Patient and Public Involvement (PPI), a significant limitation is the limited extent of PPI in the evaluation of its outputs. This constraint may impact the framework's ability to fully capture and address patient experiences and needs. Although the involvement of relevant stakeholders and subject matter experts has ensured pertinent analytical insights, increased PPI could further refine the identification and application of human factors concepts in real-world scenarios, potentially leading to more patient-centred outcomes. Further integration of PPI is

recognised as a critical component for future development, helping to mitigate risks such as bias amplification, potential misinterpretation of concepts, and the possible omission of crucial insights from the patient perspective. Enhanced PPI could also contribute to more targeted interventions for improving healthcare equity and outcomes, ensuring that the framework's findings align closely with patient needs and experiences. Expanding PPI in future iterations of the I-SIRch framework is crucial to enhance its relevance and effectiveness in addressing systemic challenges within maternity services.

Synthetic data generation usage and limitations

The paper describes the creation of a synthetic dataset comprising 818 sentences generated from 76 real maternity investigation reports. These synthetic sentences were designed to be semantically similar to the original text, with over 97.5% achieving a cosine similarity of more than 80% to their original counterparts. This synthetic dataset was used exclusively for testing purposes, not for training the

Figure 7: 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).

model. Whilst the synthetic data provided a valuable resource for testing, it is important to acknowledge its limitations. The synthetic sentences may not fully capture the nuanced complexity and variability inherent in real-world maternity incident investigation reports. There is a risk of introducing artificial patterns not present in actual reports or inadvertently replicating existing biases from the initial dataset. Performance on synthetic data may not accurately reflect the model's effectiveness on entirely new, unseen real-world data, potentially impacting its ability to generalise to diverse scenarios. Additionally, the synthetic dataset may not encompass the full spectrum of possible scenarios and nuances encountered in maternity care. To mitigate these limitations, the study included testing on real, unseen data alongside the synthetic data tests. This multi-stage approach helped validate the model's performance across different scenarios. However, continuous evaluation with diverse, real-world data remains crucial for ensuring the ongoing accuracy and generalisability of the I-SIRch framework.

Future work

Key future directions for I-SIRch include testing the framework on a larger and more diverse dataset, with particular focus on reports from Black and ethnic minority groups. This expanded testing is crucial to validate the tool's effectiveness and ensure its applicability across diverse populations, thereby addressing potential biases and improving its ability to capture unique challenges faced by these groups in maternity care. Efforts should be made to transform I-SIRch into a user-friendly, integrated tool that healthcare professionals and researchers can easily implement in their workflows. Expanding PPI throughout the process, especially from Black and ethnic minority communities, will be essential to enhance the I-SIRch framework's effectiveness in concept annotation, ultimately contributing to improved patient safety and care quality for all demographic groups.

Acknowledgements

Author contributions statement

G.C and M.K designed and conceived the experiments. M.K and G.C. conducted the experiments. G.C and M.K analysed the results and wrote the manuscript. P.W and T.J led and performed the human annotation of reports. J.B. provided expertise on safety investigations and the SIRch coding taxonomy. All authors reviewed the manuscript.

Ethics statement

HSIB would gain consent from families to investigate maternity incidents and this was governed by HSIB maternity investigations: directions 2018 - GOV.UK (www.gov.uk)

Data availability

Anonymised raw data were provided by HSIB. The Supplementary File provides the data that was generated from the incident investigation reports.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Women, Equalities Committee, H. o. C. W. & Commitee, E. Black maternal health. Third Rep. Sess. 2022-23 (18 April 2023). Available at:

https://committees.parliament.uk/publications/38989/documents/191706/default/

2. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36. https://doi.org/10.1136/jamia.2009.002733. PMID: 20442139; PMCID: PMC2995713.

3. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit Transl Bioinform. 2009 Mar 1;2009:56-60. PMID: 21347171; PMCID: PMC3041576.

4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13. https://doi.org/10.1136/jamia.2009.001560. PMID: 20819853; PMCID: PMC2995668.

5. Asgari M., Sierra-Sosa D, Elmaghraby AS. BINER: A low-cost biomedical named entity recognition. Inf. Sci. 602, 184–200 (2022). https://doi.org/10.1016/j.ins.2022.04.037

6. Zhou, W., Zhang, S., Gu, Y., Chen, M. & Poon, H. UniversalNER: Targeted distillation from large language models for open named entity recognition. arXiv preprint arXiv:2308.03279 (2023).

7. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, Mascio A, Zhu L, Folarin AA, Roberts A, Bendayan R, Richardson MP, Stewart R, Shah AD, Wong WK, Ibrahim Z, Teo JT, Dobson RJB. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021 Jul;117:102083. https://doi.org/10.1016/j.artmed.2021.102083. Epub 2021 May 1. PMID: 34127232.

8. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-1240. https://doi.org/10.1093/bioinformatics/btz682. PMID: 31501885; PMCID: PMC7703786.

9. HSIB. A thematic analysis of HSIB's first 22 national investigations. Accessed: 2023-22-09. Available at: https://www.hssib.org.uk/patient-safety-investigations/a-thematic-analysis-of-hsibs-first-22-national-investigations/.

10. Carayon P, Xie A, Kianfar S. Human factors and ergonomics as a patient safety practice. BMJ Qual Saf. 2014 Mar;23(3):196-205. https://doi.org/10.1136/bmjqs-2013-001812. Epub 2013 Jun 28. PMID: 23813211; PMCID: PMC3932984.

11. Holden RJ, Carayon P, Gurses AP, Hoonakker P, Hundt AS, Ozok AA, Rivera-Rodriguez AJ. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics. 2013;56(11):1669-86. https://doi.org/10.1080/00140139.2013.838643. Epub 2013 Oct 3. PMID: 24088063; PMCID: PMC3835697.

12. Carayon P, Schoofs Hundt A, Karsh BT, Gurses AP, Alvarado CJ, Smith M, Flatley Brennan P. Work system design for patient safety: the SEIPS model. Qual Saf Health Care. 2006 Dec;15 Suppl 1(Suppl 1):i50-8. https://doi.org/10.1136/qshc.2005.015842. PMID: 17142610; PMCID: PMC2464868.

13. England, N. & Improvement, N. Learn from patient safety events (LFPSE) service. (2021a). Available at: https://www.england.nhs.uk/patient-safety/patient-safety-insight/learning-from-patient-safety-events/learn-from-patient-safety-events-service/.

14. Office for National Statistics. Ethnic group, England and Wales: Census 2021 (2022). Available at: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest/.

Supplementary Information

Supplementary Table 1: Number of concepts and sentences per batch

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 2: Percentage of reports from Batch 1 containing each concept/human factor for each ethnic group calculated by using the formula shown at the bottom of this table

Concept	Asian	Black	DNR	MB	WO	WB	Average	SD
Organisation-Communication factor	83.33	100.00	75.00	100.00	66.67	75.00	83.33	13.94
External Environment-COVID	66.67	57.14	25.00	100.00	66.67	80.77	66.04	25.04
Organisation-Teamworking	83.33	71.43	50.00	100.00	33.33	57.69	65.96	23.99
Job/task-Assessment, investigation, testing, screening (e.g. holistic review)	66.67	42.86	50.00	100.00	50.00	63.46	62.16	20.60
Organisation-National guidance	16.67	71.43	50.00	100.00	16.67	71.15	54.32	33.22
Job/task-Care Planning	50.00	42.86	50.00	100.00	33.33	32.69	51.48	24.96
Person Patient-Physical characteristics	33.33	57.14	75.00	0.00	83.33	48.08	49.48	30.22
Technologies and Tools-issues	33.33	42.86	25.00	100.00	16.67	28.85	41.12	30.13
Job/task-Situation awareness (e.g. loss of helicopter view)	33.33	28.57	50.00	100.00	0.00	25.00	39.48	33.76
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	33.33	42.86	50.00	0.00	33.33	42.31	33.64	17.66
Organisation-Language barrier	50.00	0.00	0.00	100.00	16.67	0.00	27.78	40.37
Organisation-Documentation	50.00	14.29	25.00	0.00	16.67	53.85	26.63	21.21
Job/task-Monitoring	50.00	14.29	25.00	0.00	33.33	26.92	24.92	16.96
Organisation-Escalation/referral factor (including fresh eyes reviews)	16.67	28.57	50.00	0.00	16.67	36.54	24.74	17.52
Job/task-Obstetric review	33.33	14.29	25.00	0.00	33.33	34.62	23.43	13.83
Job/task-Dispensing, administering	0.00	0.00	100.00	0.00	0.00	13.46	18.91	40.09
Person Patient-Psychological characteristics (e.g. stress, mental health)	16.67	14.29	50.00	0.00	16.67	15.38	18.83	16.54
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	33.33	0.00	25.00	0.00	16.67	11.54	14.42	13.41
Person Staff-Training and education (e.g. attendance at antenatal classes)	16.67	0.00	50.00	0.00	0.00	15.38	13.68	19.45
Job/task-Risk assessment	33.33	28.57	0.00	0.00	0.00	13.46	12.56	15.24
Internal Environment-Physical layout and Environment	0.00	14.29	25.00	0.00	0.00	5.77	7.51	10.24
Person Staff-Decision error (errors in conscious judgments, decisions due to lack of knowledge and from misunderstanding of a situation)	0.00	0.00	0.00	0.00	16.67	23.08	6.62	10.46
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	0.00	14.29	0.00	0.00	0.00	21.15	5.91	9.40
External Environment-Geographical factor (e.g. Location of patient)	16.67	0.00	0.00	0.00	0.00	1.92	3.10	6.69

For example, the Organisation-Communication factor appeared in 7 out 7 reports of the Black ethnic group, hence 100% occurrence.
DNR: Data not received, WO: White Other, WB: White British, MB: Mixed background, SD: Standard Deviation.
Formula for computing the percentages: Number of reports in which a concept i appeared for ethnicity e/Total number of reports in ethnic group e.

Supplementary Table 3: Statistics about the investigation reports provided by HSIB (Batch 1)

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

Supplementary Table 4: File and concept frequencies

Concept	Asian	Black	DNR	MB	WO	WB	Sum
Organisation-Communication factor	5(20)	7(12)	3(4)	1(4)	4(6)	39(93)	59(139)
Job/task-assessment, investigation, testing, screening (e.g. holistic review)	4(7)	3(5)	2(3)	1(1)	3(4)	33(73)	46(93)
External Environment-COVID	4(5)	4(8)	1(1)	1(1)	4(6)	42(71)	56(92)
Organisation-Teamworking	5(9)	5(11)	2(8)	1(1)	2(3)	30(58)	45(90)
Organisation-National guidance	1(2)	5(7)	2(4)	1(2)	1(1)	37(73)	47(89)
Person Patient-Physical characteristics	2(2)	4(6)	3(4)	–(–)	5(6)	25(41)	39(59)
Organisation-Documentation	3(3)	1(1)	1(1)	–(–)	1(1)	28(43)	34(49)
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	2(3)	3(4)	2(3)	–(–)	2(4)	22(30)	31(44)
Job/task-Care Planning	3(5)	3(6)	2(6)	1(1)	2(2)	17(20)	28(40)
Organisation-Escalation/referral factor (including fresh eyes reviews)	1(3)	2(4)	2(2)	–(–)	1(1)	19(28)	25(38)
Job/task-Obstetric review	2(2)	1(1)	1(1)	–(–)	2(3)	18(24)	24(31)
Job/task-Monitoring	3(5)	1(1)	1(1)	–(–)	2(2)	14(21)	21(30)
Technologies and Tools-issues	2(2)	3(5)	1(2)	1(1)	1(2)	15(17)	23(29)
Job/task-Situation awareness (e.g. loss of helicopter view)	2(4)	2(2)	2(2)	1(1)	0(–)	13(18)	20(27)
Person Staff-Decision error (errors in conscious judgments, decisions due to lack of knowledge and from misunderstanding of a situation)	–(–)	–(–)	–(–)	–(–)	1(2)	12(17)	13(19)
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	–(–)	1(3)	–(–)	–(–)	–(–)	11(13)	12(16)
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	2(3)	–(–)	1(2)	–(–)	1(1)	6(9)	10(15)
Person Patient-Psychological characteristics (e.g. stress, mental health)	1(1)	1(1)	2(3)	–(–)	1(1)	8(9)	13(15)
Job/task-Dispensing, administering	–(–)	–(–)	4(4)	–(–)	–(–)	7(9)	11(13)
Person Staff-Training and education (e.g. attendance at antenatal classes)	1(1)	–(–)	2(3)	–(–)	–(–)	8(9)	11(13)
Job/task-Risk assessment	2(2)	2(3)	–(–)	–(–)	–(–)	7(7)	11(12)
Organisation-Language barrier	3(7)	–(–)	–(–)	1(1)	1(1)	–(–)	5(9)
Internal Environment-Physical layout and Environment	–(–)	1(1)	1(1)	–(–)	–(–)	3(4)	5(6)
External Environment-Geographical factor (e.g. Location of patient)	1(1)	–(–)	–(–)	–(–)	–(–)	1(1)	2(2)
Total	49(87)	49(81)	35(55)	9(13)	34(46)	415(688)	591(970)

Each cell of the table shows the number of reports (total number of concepts) across each ethnic group for Batch 1.
DNR: Data not received, MB: Mixed Background, WO: White Other, WB: White British – denotes that the concept did not appear in the particular ethnic group's reports.

Supplementary Table 5: Batch 1 reports and their associated outcomes across the ethnic groups

Ethnic group	TH	NND	MD	Stillbirth	Total
Asian	3	1	1	1	6
Black	3	–	4	–	7
Data not received	–	–	4	–	4
Mixed Background	1	–	–	–	1
White British	30	5	8	9	52
White Other	2	–	2	2	6
Total	39	6	19	12	76

This information was only provided with Batch 1.
Therapeutic Hypothermia: TH, Early Neonatal Death: NND, Mother's death: MD.
– denotes that the outcome did not appear in the particular ethnic group's reports.

Supplementary Table 6: Performance evaluation for each test

	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

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.
Avg: Average; SD: Standard deviation.

Supplementary Table 7: Percentage of correctly annotated concepts across all the tests A, B, and C

Concept	Test with real data	Tests with synthetic data	
	Test B	Test A	Test C
Organisation-Communication factor	82.70	45.59	62.32
Job/task-Assessment, investigation, testing, screening (e.g. holistic review)	93.39	60.97	61.26
External Environment-COVID	86.67	55.59	59.95
Organisation-Teamworking	90.53	74.40	60.58
Organisation-National guidance	94.02	78.22	73.25
Person Patient-Physical characteristics	90.42	66.81	71.98
Organisation-Documentation	86.73	79.05	79.59
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	91.27	80.00	87.78
Job/task-Care Planning	93.90	80.74	81.27
Organisation-Escalation/referral factor (including fresh eyes reviews)	92.73	61.86	57.10
Job/task-Obstetric review	95.10	48.36	52.36
Job/task-Monitoring	97.33	46.38	63.65
Technologies and Tools-issues	78.79	80.85	79.97
Job/task-Situation awareness (e.g. loss of helicopter view)	75.00	49.33	57.00
Person Staff-Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)	92.68	37.50	35.00
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	100.00	51.93	27.88
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	83.33	12.50	11.43
Person Patient-Psychological characteristics (e.g. stress, mental health)	72.73	52.73	52.73
Job/task-Dispensing, administering	92.86	36.37	36.36
Person Staff-Training and education (e.g. attendance at antenatal classes)	77.27	96.97	90.91
Job/task-Risk assessment	98.61	58.33	72.50
Organisation-Language barrier	–	0.00	22.22
Internal Environment-Physical layout and Environment	82.61	53.33	53.33
External Environment-Geographical factor (e.g. Location of patient)	–	50.00	50.00
Organisation-Guidance factor	100.00	–	–
Person Patient-Language barrier	95.24	–	–
Organisation-Local guidance	92.11	–	–
Organisation-National and local guidance	98.00	–	–
Average	89.77	56.57	58.35
Standard Deviation	7.85	22.02	20.50

– denotes that the concept did not appear in the particular test.

Supplementary Table 8: Percentage of correctly annotated concepts in the synthetic (test) sentences for each ethnic group

Code	Asian	Black	DNR	MB	OW	WB	Avg	SD
External Environment-COVID	64.29	83.33	50.00	0.00	66.67	69.23	55.59	29.24
External Environment-Geographical factor (e.g. Location of patient)	0.00	–	–	–	–	100.00	50.00	70.71
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	0.00	–	0.00	–	0.00	50.00	12.50	25.00
Internal Environment-Physical layout and Environment	–	0.00	100.00	–	–	60.00	53.33	50.33
Job/task-Assessment, investigation, testing, screening (e.g. holistic review)	57.14	75.00	80.00	0.00	80.00	73.68	60.97	31.03
Job/task-Care Planning	100.00	66.67	100.00	100.00	50.00	67.74	80.74	22.02
Job/task-Dispensing, administering	–	–	0.00	–	–	72.73	36.37	51.43
Job/task-Monitoring	66.67	0.00	0.00	–	100.00	65.22	46.38	44.56
Job/task-Obstetric review	0.00	50.00	50.00	-	60.00	81.82	48.36	30.00
Job/task-Risk assessment	0.00	100.00	–	–	–	75.00	58.33	52.04
Job/task-Situation awareness (e.g. loss of helicopter view)	100.00	33.33	33.33	0.00	-	80.00	49.33	40.17
Organisation-Communication factor	40.00	50.00	50.00	0.00	60.00	73.53	45.59	25.05
Organisation-Documentation	75.00	100.00	50.00	-	100.00	70.27	79.05	21.30
Organisation-Escalation/referral factor (including fresh eyes reviews)	60.00	66.67	100.00	–	0.00	82.61	61.86	37.88
Organisation-Language barrier	0.00	–	–	0.00	0.00	–	0.00	0.00
Organisation-National guidance	83.33	62.50	66.67	100.00	80.00	76.79	78.22	13.31
Organisation-Teamworking	66.67	66.67	53.85	100.00	83.33	75.86	74.40	15.99
Person Patient-Physical characteristics	100.00	66.67	0.00	–	100.00	67.39	66.81	40.83
Person Patient-Psychological characteristics (e.g. stress, mental health)	100.00	100.00	0.00	-	0.00	63.64	52.73	50.37
Person Staff-Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)	–	–	–	–	0.00	75.00	37.50	53.03
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	100.00	80.00	50.00	-	100.00	70.00	80.00	21.21
Person Staff-Training and education (e.g. attendance at antenatal classes)	100.00	–	100.00	–	–	90.91	96.97	5.25
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	–	50.00	–	–	–	53.85	51.93	2.72
Technologies and Tools-issues	100.00	100.00	66.67	100.00	50.00	68.42	80.85	21.94

Trained on real-data (Batch 1) and tested on synthetic data (Batch 4). These are the results from testing with Batch 4 (Test A). DNR: Data not received, WO: White Other, WB: White British, MB: Mixed background, SD: Standard Deviation, Avg: Average, – denotes that the concept did not appear in the particular ethnic group.

Supplementary Table 9: Test A: Training on Batch 1 and testing on Batch 4

Ethnicity	Test A			Test C		
	Correctly	Incorrectly	Correct (%)	Correctly	Incorrectly	Correct(%)
Asian	59	28	67.82	61	26	70.11
Black	54	27	66.67	54	27	66.67
Data not received	33	22	60.00	32	23	58.18
Mixed Background	9	4	69.23	9	4	69.23
Other White	32	14	69.57	34	12	73.91
White British	499	189	72.82	501	187	72.82
	Sum: 686	Sum: 284	Mean: 70.93($\pm$ 4.30)	Sum: 691	Sum: 279	Mean: 71.24 ($\pm$ 5.67)

Test C: Retraining on Batches 2 and 3 and testing on Batch 4.

Supplementary Table 10: Percentage of correctly annotated concepts in synthetic sentences for each ethnic group (after retraining with Batch 3)

Concepts	Asian	Black	DNR	MB	OW	WB	Avg	SD
External Environment-COVID	57.14	83.33	50.00	0.00	100.00	69.23	59.95	34.46
External Environment-Geographical factor (e.g. Location of patient)	0.00	–	–	–	–	100.00	50.00	70.71
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	0.00	0.00	0.00	–	0.00	57.14	11.43	25.56
Internal Environment-Physical layout and Environment	–	0.00	100.00	–	–	60.00	53.33	50.33
Job/task-Assessment, investigation, testing, screening (e.g. holistic review)	57.14	75.00	80.00	0.00	80.00	75.44	61.26	31.18
Job/task-Care Planning	100.00	66.67	100.00	100.00	50.00	70.97	81.27	21.68
Job/task-Dispensing, administering	–	–	0.00	–	–	72.73	36.36	51.43
Job/task-Monitoring	66.67	50.00	0.00	100.00	100.00	65.22	63.65	37.13
Job/task-Obstetric review	0.00	50.00	50.00	–	80.00	81.82	52.36	33.11
Job/task-Risk assessment	0.00	100.00	100.00	–	100.00	62.50	72.50	43.66
Job/task-Situation awareness (e.g. loss of helicopter view)	100.00	66.67	33.33	0.00	–	85.00	57.00	40.42
Organisation-Communication factor	53.33	40.00	50.00	100.00	60.00	70.59	62.32	21.09
Organisation-Documentation	100.00	75.00	50.00	–	100.00	72.97	79.59	21.06
Organisation-Escalation/referral factor (including fresh eyes reviews)	60.00	50.00	50.00	100.00	0.00	82.61	57.10	34.24
Organisation-Language barrier	0.00	–	-	0.00	0.00	88.89	22.22	44.44
Organisation-National guidance	83.33	62.50	33.33	100.00	80.00	80.36	73.25	22.90
Organisation-Teamworking	66.67	66.67	69.23	0.00	83.33	77.59	60.58	30.42
Person Patient-Physical characteristics	100.00	66.67	0.00	100.00	100.00	65.22	71.98	39.01
Person Patient-Psychological characteristics (e.g. stress, mental health)	100.00	100.00	0.00	–	0.00	63.64	52.73	50.37
Person Staff-Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)	0.00	–	100.00	0.00	0.00	75.00	35.00	48.73
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	100.00	100.00	50.00	100.00	100.00	76.67	87.78	20.73
Person Staff-Training and education (e.g. attendance at antenatal classes)	100.00	–	100.00	–	–	72.73	90.91	15.75
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	–	50.00	0.00	–	0.00	61.54	27.88	32.54
Technologies and Tools-issues	100.00	100.00	66.67	100.00	50.00	63.16	79.97	22.63

Trained with Batch 1, retrained with Batches 2 and 3, and tested on Batch 4. These are the results from testing with Batch 4 (Test C).
DNR: Data not received, WO: White Other, WB: White British, MB: Mixed background, SD: Standard Deviation, Avg: Average, - denotes that the concept did not appear in the particular ethnic group.

Supplementary Table 11: Wilcoxon signed-rank test results: Comparing Tests A and C

Ethnicity	W	p
Asian	1	0.29
Black	7	0.46
Data not received	1	0.29
Mixed Background	1.5	1.00
Other White	0	0.18
White British	44	0.59

Supplementary Table 12: Sample text segments annotations by Human Annotators using the SIRch framework

Report ID	Sentence	HA-1	HA-2	HA-3
ID-1	This meant that during pregnancy the Mother did not have an anaesthetic review, referral to a dietitian, or information shared regarding the risks associated with raised BMI during pregnancy and the need for high dose folic acid.	Assessment, investigation, testing, screening (e.g. holistic review)	Escalation/referral factor (including fresh eyes reviews)	Between staff and patient (verbal)
ID-2	The abnormal CTG was not recognised and did not prompt escalation to the obstetric team, in line with local guidance.	Interpretation (e.g. CTG)	Escalation/ referral factor (including fresh eyes reviews)	Local guidance
ID-3	Obstetric reviews of the Mother whilst in labour were not carried out in line with local guidance, these would have provided an earlier opportunity for a holistic review of the Mother's risk status and care pathway.	Obstetric review	Local guidance	Assessment, investigation, testing, screening (e.g. holistic review)

HA: Human Annotator.

Supplementary Table 13: Number of reports and concept frequencies found in the reports that were utilised for training the model

Concepts	Batch 1	Batch 2	Batch 3	Total
Organisation-Teamworking	45(90)	14(81)	96(359)	155(530)
Organisation-Communication factor	59(139)	15(48)	85(290)	159(477)
Job/task-Assessment, investigation, testing, screening (e.g. holistic review)	46(93)	14(46)	90(242)	150(381)
Person Patient-Physical characteristics	39(59)	10(22)	69(240)	118(321)
Technologies and Tools (e.g. CTG, use of screening tools)-Interpretation (e.g. CTG)	12(16)	13(35)	65(146)	90(197)
Person Staff-Slip or lapse (errors that tend to happen in routine tasks that people are doing without much conscious thought)	31(44)	7(18)	61(126)	99(188)
Person Staff-Decision error (errors in conscious judgements, decisions due to lack of knowledge and from misunderstanding of a situation)	13(19)	12(27)	64(123)	89(169)
Organisation-Documentation	34(49)	7(21)	45(98)	86(168)
Organisation-Escalation/referral factor (including fresh eyes reviews)	25(38)	8(10)	64(110)	97(158)
Organisation-National and local guidance	47(89)	8(11)	37(50)	92(150)
Job/task-Obstetric review	24(31)	7(14)	58(102)	89(147)
External Environment-COVID	56(92)	2(5)	21(45)	79(142)
Organisation-National guidance	0(0)	11(18)	76(117)	87(135)
Job/task-Care Planning	28(40)	6(10)	49(82)	83(132)
Job/task-Monitoring	21(30)	8(13)	47(75)	76(118)
Technologies and Tools-issues	23(29)	9(18)	37(66)	69(113)
Job/task-Risk assessment	11(12)	4(10)	46(72)	61(94)
Job/task-Situation awareness (e.g. loss of helicopter view)	20(27)	1(2)	30(48)	51(77)
Job/task-Dispensing, administering	11(13)	4(7)	25(42)	40(62)
Internal Environment-Acuity (e.g. capacity of the maternity unit as a whole)	10(15)	3(4)	32(36)	45(55)
Organisation-Local guidance	0(0)	9(17)	27(38)	36(55)
Person Patient-Psychological characteristics (e.g. stress, mental health)	13(15)	3(5)	24(33)	40(53)
Person Staff-Training and education (e.g. attendance at antenatal classes)	11(13)	2(5)	12(22)	25(40)
Organisation-Guidance factor	0(0)	1(1)	26(38)	27(39)
Internal Environment-Physical layout and Environment	5(6)	3(4)	13(23)	21(33)
Person Patient-Language barrier	5(9)	0(0)	10(21)	15(30)
External Environment-Geographical factor (e.g. Location of patient)	2(2)	0(0)	0(0)	2(2)
Total	591(970)	181(452)	1209(2644)	1981(4066)

Batch information is provided in Supplementary Table 1. The columns show the number of reports (concept frequency).

Supplementary Table 14: Inter-rater reliability among three human annotators (HA)

RN.	HA-1	HA-2	HA-3	Agreement
1	6	6	6	5
2	8	8	8	6
3	6	6	6	5
4	10	10	11	8
5	5	5	5	5
6	5	5	5	5
7	20	20	20	16
8	10	10	7	7
9	8	8	8	7
10	6	6	6	5
11	29	29	25	17
12	7	7	6	5
13	7	7	11	6
14	6	6	5	5
15	5	5	5	5
16	5	5	4	4
17	10	10	8	6
18	13	13	15	14
19	5	5	5	5
20	5	5	6	4

The columns show the number of sentences per report coded by each HA. The last column shows the number of sentences for which there was agreement on the annotations.
RN: Report number, HA: Human Annotator.
Total number of concept agreements: 420; Total number of concepts: 524; Number of annotators: 3. IRR: 80.15%.

Supplementary Table 15: Sample of original and synthetic (test) sentences

No.	Original sentences	Synthetic sentences
1	The Mother was appropriately risk assessed at booking for midwifery led care and the pregnancy progressed without complication until the induction of her labour at 41+5 weeks which was offered in line with national guidance.	The Mother's risk was assessed when she booked for midwifery led care and her pregnancy progressed without complication until she was induced at 41+5 weeks, which was offered in line with national guidance.
2	This would not have altered the outcome for the Baby as the Baby's birthweight was within the expected range.	Even if the outcome had been different, the baby's birthweight was still within the expected range.
3	This had an impact on the Mother's experience as she did not feel able to have in depth discussions over the telephone.	This impacted the Mother's experience as she felt unable to have in depth discussions over the telephone.
4	The Trust are encouraged to improve cross boundary working practices to ensure all mothers have equal access to local information about their care pathways.	The Trust should work on improving communication and collaboration between different departments to make sure all mothers have the same access to information about their care options.
5	The Mother was given the opportunity for a discussion about her choices and IOL processes on admission to the maternity unit.	The Mother was given the opportunity to discuss her choices and IOL processes with a staff member upon admission to the maternity unit.

In the dataset only Batch 4 contained synthetic sentences, and these were only used for testing purposes and not for training the model.

Supplementary File 1: 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 antenatal 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)

# Supplementary File 2. Synthetic dataset generation for testing purposes
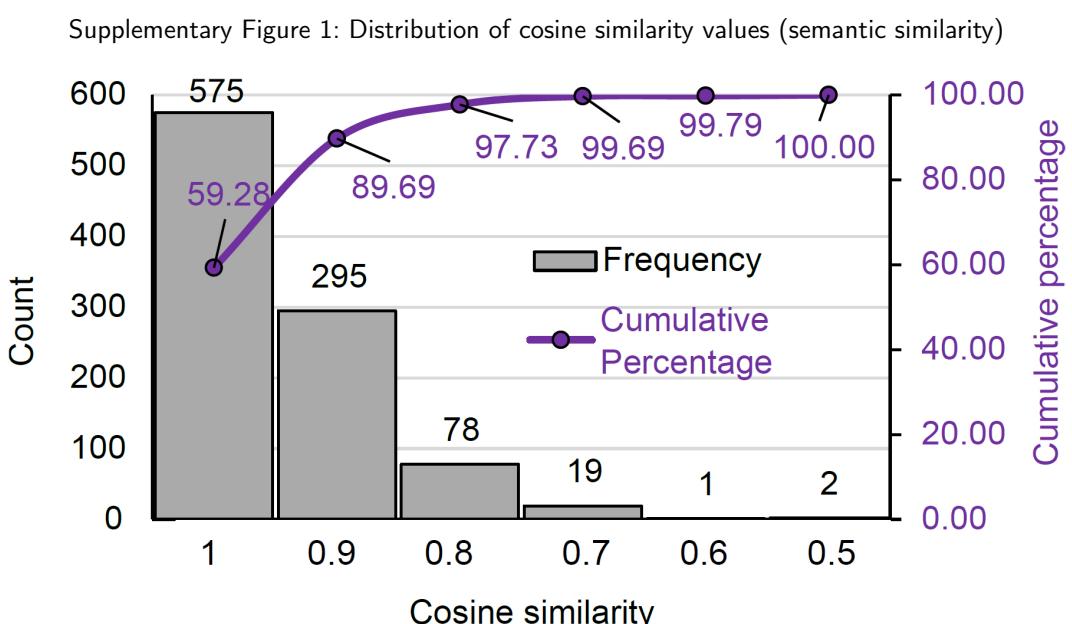
Generate synthetic sentences

To test the performance of the model on new sentences, a text generation tool was implemented that takes as input a set of original sentences from the 'Extracted text segments' and generates a new synthetic sentence per original sentence. The original and generated sentences were exported to a CSV file, with each row containing one original and its corresponding synthetic sentence. The synthetic sentences are then utilised for two purposes: 1) to further train the model by increasing its exposure to linguistic variability during training; and 2) to evaluate the performance of the trained model on new content that is semantically similar but uses different sentences. Supplementary Table 15 shows examples of original and synthetic sentences.

Sentence similarity

It is important to understand how semantically similar the original sentences are to their corresponding synthetic sentences to evaluate the quality of the sentences generated by the synthetic sentence generation tool. To compute the similarity of each pair (i.e. original and synthetic sentence) a Word2Vec model was utilised, sentences were represented as numerical vectors based on their constituent words, capturing contextual meanings. The average vector for each sentence was then calculated by combining its word vectors. A mathematical measure, cosine similarity, was employed to compare these average sentence vectors, quantifying their semantic similarity or difference. A high cosine similarity score indicates strong semantic resemblance, while a low score suggests dissimilarity. In essence, this process allowed for the assessment of the likeness or disparity in meaning between two sentences using their word representations and cosine similarity. The similarity values of each pair were written to a CSV file. A statistical overview of the similarity values is given in Supplementary Figure 1, which shows that more than 97.5% of the sentences have a cosine similarity of more than 80% indicating good preservation of semantic content by the synthetic data generation tool. Hence, the tool effectively generated unique sentences while preserving the core meaning of the original sentence.

Supplementary Figure 1: Distribution of cosine similarity values (semantic similarity)

The higher the similarity value the closer the semantic meaning of the original and synthetic sentence (Batch 1).