International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.7

FG-AI4H Topic Description Document for the Topic Group on maternal and child health (TG-MCH)

Summary

This topic description document (TDD) aims to specify a standardized benchmarking for AI-based maternal and child health. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence; benchmarking; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; maternal health; child health

Change Log

This document contains Version 1 of the Deliverable DEL10.7 on "*FG-AI4H Topic Description Document for the Topic Group on maternal and child health (TG-MCH)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editors: Alexandre Chiavegatto Filho
TG-MCH
University of Sao Paulo, Brazil
Email: alexdiasporto@usp.br

Gabriel Silva
TG-MCH
University of São Paulo, Brazil
Email: gabriel8.silva@usp.br

Raghu Dharmaraju
TG-MCH
Wadhwani AI, India
Email: raghu@wadhwaniai.org

CONTENTS

List of Tables

List of Figures

ITU-T FG-AI4H Deliverable 10.7

FG-AI4H Topic Description Document for the Topic Group on maternal and child health (TG-MCH)

1 Introduction

Improving the health and well-being of mothers, infants, and children is one of the most important public health goals worldwide. Every day, an estimated 810 women die from causes related to pregnancy or childbirth and over 15,000 children die from preventable diseases. Despite notable recent improvements for most countries, the Millenium Development Goal (MDG) target for 2015 of reducing child mortality globally by two thirds was not achieved.

Artificial intelligence methods have the potential of improving maternal and child health decisions, especially in low-resource settings. As advances are made in the collection and availability of data on maternal and child health, the possibility of using this data to improve health decisions increases, especially when access to specialized professionals is scarce. Several challenges in the maternal and child area can be overcome with adequate preventive methods, which in turn depend on the establishment of risk scores for the development of targeted public policies, especially in low income countries where the available resources for these policies are lower.

The aim of this document therefore is to develop a standardised benchmarking approach for AI for maternal and child health, with a focus on developing countries. This topic description document specifies the standardized benchmarking for maternal and child health (MCH) systems. It serves as the final deliverable ITU/WHO Focus Group on AI for Health (FG-AI4H).

2 About the FG-AI4H topic group on Maternal and Child Health

The introduction highlights the potential of a standardized benchmarking of AI systems for maternal and child health to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TGMCH at the meeting Brasília, Brazil.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. Alexandre Chiavegatto Filho from the Department of Epidemiology of the University of São Paulo, Brazil, was nominated as topic driver for the TG-MCH.

2.1 Documentation

This document is the TDD for the TG-MCH. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for maternal and child health (MCH). It describes the existing approaches for assessing the quality of maternal and child health systems and provides the details that are likely relevant for setting up a new standardized benchmarking. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The final version of this TDD will be released as deliverable "DEL 10.07 Maternal and Child Health (TG-MCH)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic Group output documents

Number	Title
FGAI4H-L-07-A01	Latest update of the Topic Description Document of the TG-MCH
FGAI4H-L-07-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-L-0y-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-MCH

2.2 Status of this topic group

The following subsections describe the final update of the collaboration within the TG-MCH for the official focus group meetings.

2.2.1 Status update for meeting

The topic group had two online meetings during the 2021-2022 cycle. Due to the serious consequences brought about by the COVID-19 pandemic in the countries of the topic leaders (India and Brazil), there was a difficulty in holding more frequent meetings during this period of worsening of the crisis. However, online meetings served to guide forward the topic group and the establishment of priorities for the next documents. From the meeting held in 2020, professor Alexandre Chiavegatto Filho from the University of São Paulo, Brazil, assumed the responsibility of writing this document according to the current standardized format.

2.3 Topic Group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding 'Call for TG participation' (CfTGP) can be found here:

– https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-MCH.pdf

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

– https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-MCH.aspx

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG 'zoom' link:

– https://itu.zoom.us/my/fgai4h

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

– https://itu.int/go/fgai4h/join

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

– https://itu.int/go/fgai4h

3 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in maternal and child health and how this can help to solve a relevant 'real-world' problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise.

3.1 Maternal and Child Health

3.1.1 Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to DEL3 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

Improving the health and well-being of mothers, infants, and children is one of the most important public health goals worldwide, and progress has been slower than expected. From 2018 to 2030, there will be an estimated 27.8 million worldwide deaths in the first month of life in case every country maintains their current rate of reduction. Malnourished children, particularly those with severe acute malnutrition, have a higher risk of death from common childhood illness such as diarrhoea, pneumonia, and malaria. Nutrition-related factors contribute to about 45% of deaths in children under-5 years of age.

Almost half of under-five deaths will be of newborns whose deaths could be effectively prevented by providing high quality antenatal care, skilled care at birth, postnatal care for mothers and their babies, and care of small and sick newborns. Every year, around 94% of maternal deaths and over 80% of under-5 deaths occur in low and lower middle-income countries.

AI as a tool in maternal and child health care will benefit individuals and communities across the world, especially in low-resource settings. Some examples of potential AI-based applications in this area include:

i. Predictions during pregnancy:

   a. Foetal growth prediction.

   b. Final gestational age prediction.

   c. Incidence of relevant comorbidities during pregnancy.

   d. Mortality risk:

      d1. Foetal mortality.

      d2. Maternal mortality.

      d3. Neonatal mortality.

      d4. Infant mortality.

ii. Hospital warning systems for the necessity of:

   a. Labour rooms.
   b. Neonatal intensive care units.
   c. Healthcare emergency specialists.

iii. Patient-centric health screening tools: IOT tools to screen for the risk of common diseases such as pneumonia, jaundice, anaemia, etc.

iv. Post-natal predictions:

  a. Risk of neonatal and post-neonatal mortality given the newborn's characteristics.

  b. Risk of rehospitalization after childbirth discharge.

In developing countries, the burden of delivery of health services pertaining to maternal and child is the responsibility of frontline health workers, who in some cases have limited skills and training and are often overworked and underpaid. In this scenario, AI can help close the expertise gap and lead to better monitoring and accountability by enabling easy, automatic, accurate and tamper-proof screening.

While there are several research and commercial groups working on AI applications in this area, the lack of consistent standardization makes it difficult for organisations like the WHO, governments, and other key players to adopt symptom assessment systems as part of their solutions to address global health challenges. The implementation of a standardised benchmarking for these classes of applications as part of the WHO/ITU's AI for Health Focus Group will therefore be an important step towards addressing this issue.

3.1.2 Current gold standard

While there are already several research and commercial groups working on AI applications in maternal and child health, the lack of consistent standardization makes it difficult for organisations to adopt assessment systems as part of their solutions to address global health challenges. The implementation of a standardised benchmarking for these classes of applications as part of the WHO/ITU's AI for Health Focus Group will therefore be an important step towards addressing this issue.

In recent years there has been a strong growth in AI solutions for maternal and child health issues. However, it is important to consider some growing and important challenges for the area. First, rigorous validation of these techniques is required to ensure that issues commonly present in machine learning projects are not present, such as data leakage. Second, it is important to ensure that all the data is available for AI solutions to be applied where they are most needed, i.e. in developing countries that have less data collection and a still incipient use of electronic medical records.

3.1.3 Relevance and impact of an AI solution

Avoiding the occurrence of negative child outcomes often involves low-cost interventions, but require an alert early enough to prevent its occurrence. Predictive machine learning algorithms can assist in this task by analyzing the patient's maternal and child characteristics and providing an early risk score for the occurrence of an adverse effect such as neonatal mortality.

3.1.4 Existing AI solutions

A few scientific studies have been performed in recent years that apply machine learning algorithms to predict maternal and child health events. A 2019 study used machine learning to predict postpartum hospital admission in the first 12 weeks after delivery found a high predictive performance for hospitalization from hypertensive disorders (AUC = 0.879) (Betts et al., 2019). Another analysis from 2020 found that machine learning was able to predict height-for-age z-scores in children from a rural area of Pakistan (Harrison et al., 2020). A more recent study from 2021 fond that machine learning algorithms were able to predict with reasonable accuracy the risk of readmission for complications of hypertensive disorders of pregnancy (Hoffman et al., 2021).

Real-time fetal electrocardiogram recordings can be used by patients and clinicians to monitor fetal status. AI-based sensors can be used to monitor blood glucose and blood pressure, which are especially useful in low-resource settings. This has become even more important since the COVID-19 pandemic, where healthcare professionals are overwhelmed by the huge wave of patients because of which patients with pre-existing conditions are not being able to seek help and timely follow-up (Oprescu et al., 2020). This is also related to one of the most promising areas of AI applications, namely mobile health (mHealth). mHealth is extremely useful for prenatal care, especially in low-resource settings, where community health workers can facilitate care, monitor health, and enable patient self-management. Even in high-resource settings, mHealth allows for personalized monitoring to support pregnant women (Davidson et al., 2021).

4 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable DEL1 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL1 and reflects the ethical considerations of the TG-MCH.

Topics to consider:

- Technical robustness of the algorithms according to:
    - Racial differences.
    - Dataset shifts.
    - Patient income.
- Overall predictive performance of the algorithms.
- Differences in predictive performance according to vulnerable subgroups.
- Availability of data for training the algorithm in low and middle-income countries in terms of:
    - Data-collection quality.
    - Enough variables to perform an accurate prediction.
- Data governance (storage, access and security) and privacy.
- Bias and fairness of training datasets.
- Generalization ability:
    - From large urban areas to remote rural areas.
    - From high-income to low-income areas.
    - According to patient's characteristics.
    - According to differences in local clinical protocols.
- Explainability.
- Accountability.

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and maternal and child health for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

5.1 Maternal and child health

5.1.1 Publications on benchmarking systems

While a representative comparable benchmarking for maternal and child health does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL7 *"AI for health evaluation considerations,"* DEL7.1 *"AI4H evaluation process description,"* DEL7.2 *"AI technical test specification,"* DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL7.4 *"Clinical Evaluation of AI for health"*.

5.1.2 Benchmarking by AI developers

All developers of AI solutions for maternal and child health implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

5.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL7.5 *"FG-AI4H assessment platform"* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the maternal and child health AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL6 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 Subtopic

The benchmarking of maternal and child health was developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines the benchmarking that has been implemented and the rationale behind it. It serves as an introduction to the subsequent sections.

6.1.1 Benchmarking version 1.

This section includes all technological and operational details of the benchmarking process for the benchmarking version 1.

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version 1.

Neonatal mortality in low- and middle-income countries remains a significant global public health challenge. These countries face multiple obstacles, including limited access to quality prenatal care, inadequate healthcare infrastructure, and a shortage of specialized healthcare professionals. Preventing neonatal mortality is crucial for improving collective health indicators and has a profound impact on families and communities. In this context, the use of machine learning algorithms can play an important role in predicting and preventing these outcomes. These algorithms can analyse large volumes of data, identify patterns and risk factors, and provide valuable insights for healthcare professionals, enabling proactive and targeted interventions. Implementing effective machine learning models has the potential of reducing neonatal mortality and improving public health in low-and middle-income countries.

To address these challenges, we used data from the Global Network's Maternal Newborn Health Registry (MNHR), a population-based observational study designed to quantify and understand trends in pregnancy outcomes in specific low-resource geographic areas over time. The study collected data from approximately 500,000 pregnancies between 2010 and 2019, in three rounds of data collection. The data were collected in eight different countries: Argentina, Zambia, Guatemala, Kenya, Pakistan, India, the Democratic Republic of the Congo (DRC), and Bangladesh. The variables collected included gestational information, delivery details, and a 42-day follow-up after delivery, encompassing the five minimum indicators suggested by the World Health Organization.

Using the MNHR data, we evaluated the performance of machine learning algorithms in predicting the risk of neonatal mortality and identified potential training strategies using multicentric data from low- and middle-income countries. We employed three initial training frameworks: 1) a general algorithm for all countries, 2) country-specific algorithms and 3) biggest train sample size country. These algorithms were trained on data from 2010 to 2016 and tested on data from 2017 to 2019, utilizing five different algorithms. The initial predictors tested were the five basic indicators suggested by the WHO: maternal age, place of delivery, type of delivery, birth weight, and gestational age. The predicted outcome was neonatal mortality from the first day of birth to the 42nd day after delivery. Preliminary results indicate that training the algorithm in a general manner, i.e. using data from all countries together, is preferable possibly due to the larger sample size and diversity of examples.

6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version 1. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

6.1.1.2.1 Benchmarking system architecture

The benchmarking system architecture for the neonatal mortality prediction study was developed by the topic group for scientific purposes and to identify different training strategies to predict neonatal death. The system utilized Python 3.9.12 as the programming language for training the algorithms, on its interface with Jupyter Notebook. Several libraries were employed, including pandas and numpy for data manipulation, matplotlib for data visualization, and scikit-learn, catboost, xgboost, lightgbm, and boruta for machine learning algorithms. Furthermore, the interpretation of the algorithm was enhanced through the utilization of the shap package.

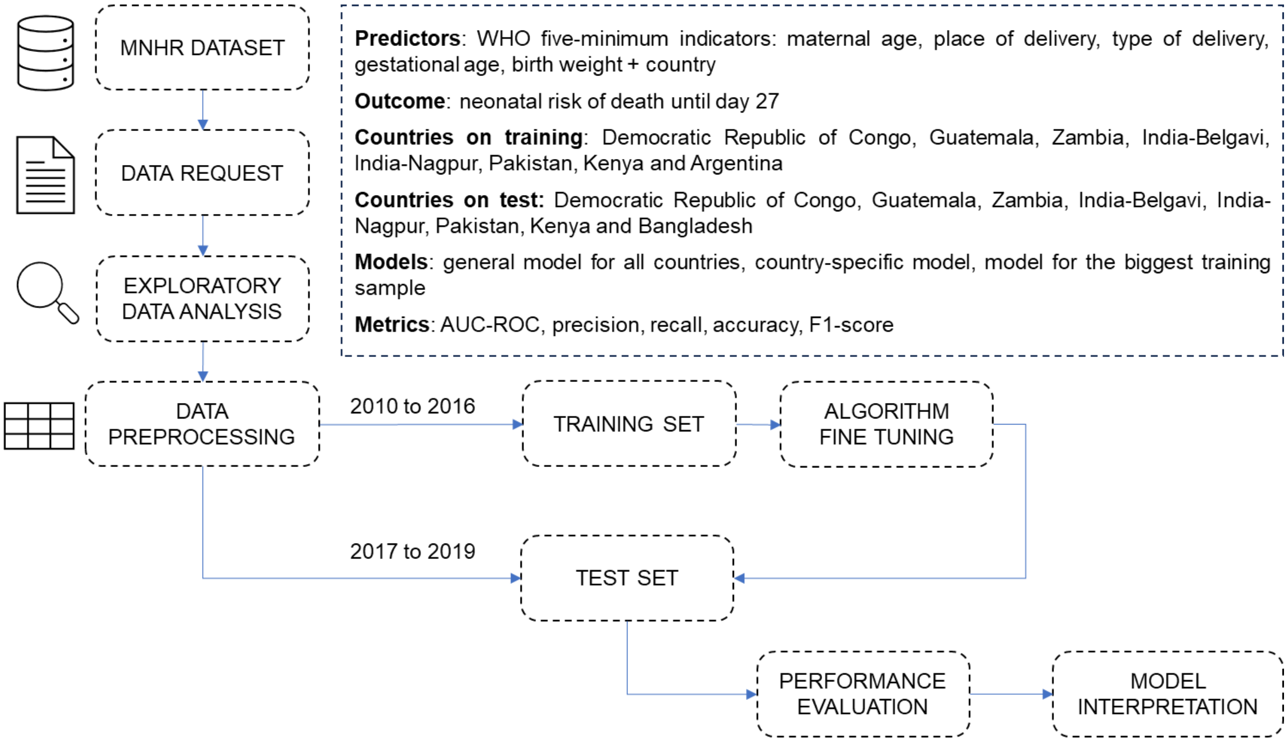
6.1.1.2.2 Benchmarking system dataflow

The workflow for this study involved several key steps to develop and evaluate predictive models for neonatal risk of death using the MNHR dataset. The initial phase included a data request to access the MNHR dataset, which served as the primary source of information. The dataset underwent exploratory data analysis to gain insights into its structure, distributions, and potential patterns. Following this, data preprocessing techniques were applied to clean and transform the dataset, ensuring its suitability for model training.

The next step involved creating a training set that consisted of data from eight countries: Democratic Republic of Congo, Guatemala, Zambia, India-Belgavi, India-Nagpur, Pakistan, Kenya, and Argentina. The training set was used to train different models based on the WHO five-minimum indicators, which included maternal age, place of delivery, type of delivery, gestational age, birth weight, and the country variable. Three types of models were developed: a general model for all countries, a country-specific model, and a model using the largest training sample available.

The models underwent algorithm fine-tuning to optimize their performance and enhance their predictive capabilities. The fine-tuning process involved adjusting various parameters and hyperparameters to achieve the best possible results. Once the models were trained and fine-tuned, they were evaluated using a separate test set consisting of data from the same countries as the training set, with the addition of Bangladesh. Performance evaluation was conducted using several metrics, including AUC-ROC, precision, recall, accuracy, and F1-score. These metrics provided a comprehensive assessment of the models' predictive accuracy, ability to identify positive cases correctly, overall model performance, and balance between precision and recall. Model performance was compared and analyzed to identify the most effective approach for predicting neonatal risk of death.

Finally, model interpretation was carried out to gain insights into the factors influencing model predictions. The Shap library was used to facilitate the interpretation process, providing valuable insights into feature importance and the contribution of each predictor variable. This allowed researchers to better understand the underlying relationships and mechanisms driving model prediction. The workflow involved accessing the MNHR dataset, performing exploratory data analysis, preprocessing the data, creating a training set, fine-tuning the algorithms, evaluating the models' performance, and interpreting the results. By following this comprehensive workflow, the study aimed to develop accurate predictive models for neonatal risk of death, leveraging the WHO five-minimum indicators and country-specific information from a diverse set of countries.



Predictors: WHO five-minimum indicators: maternal age, place of delivery, type of delivery, gestational age, birth weight + country

Outcome: neonatal risk of death until day 27

Countries on training: Democratic Republic of Congo, Guatemala, Zambia, India-Belgavi, India-Nagpur, Pakistan, Kenya and Argentina

Countries on test: Democratic Republic of Congo, Guatemala, Zambia, India-Belgavi, India-Nagpur, Pakistan, Kenya and Bangladesh

Models: general model for all countries, country-specific model, model for the biggest training sample

Metrics: AUC-ROC, precision, recall, accuracy, F1-score

Figure 1: Workflow development for neonatal risk of death prediction using machine learning algorithms and MNHR data.

6.1.1.2.3 Safe and secure system operation and hosting

In this section, we address security considerations regarding the storage and hosting of data, as well as safety precautions to protect against data manipulation, leakage, or loss in the context of the benchmarking study. Since the data for the study is locally hosted after data request, it is necessary to implement appropriate security measures to safeguard the integrity and confidentiality of the data.

To prevent data loss, regular backups of the benchmarking results and reports were performed. This ensures that in the event of a system failure or other unforeseen circumstances, data can be restored, and no critical information is permanently lost. Access controls have been implemented to restrict unauthorized access to the data. Only authorized personnel, such as researchers and healthcare professionals directly involved in the study, are granted access to manipulate or view the data. This helped to maintain data privacy and prevent any unauthorized modifications or disclosures.

Furthermore, in line with the security requirements of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Data and Specimen Hub (DASH), necessary precautions have been taken to ensure the benchmark's security. These requirements serve as guidelines for protecting sensitive data and maintaining a secure environment for data storage and sharing.

Overall, the implementation of these security measures is essential to mitigate risks associated with data manipulation, leakage, or loss. By adopting these precautions, the benchmarking study can maintain the confidentiality, integrity, and availability of the data, fostering a secure environment for the analysis and interpretation of results.

6.1.1.2.4 Benchmarking process

The participants included in the dataset were independently approached as part of a previous study conducted by NICHD. The data consists of individual collections from pregnant participants, spanning from prenatal care to 42 days after delivery. The authors of the benchmark declare no conflicts of interest, as the objective of the project is purely scientific and aimed at promoting public health. The benchmarking process ensures transparency and impartiality in its execution, analysis, and interpretation of results.

Conflicts that may arise during the benchmarking process are addressed through a systematic resolution approach. These conflicts are carefully examined and resolved through rigorous analysis and discussions among the benchmarking team, ensuring fairness and adherence to scientific standards. Upon completion of the benchmarking process, the final results will be published in scientific materials and disseminated to the community. This includes the publication of research papers that undergo peer review for scientific validation. Additionally, the results will be shared through community outreach initiatives, such as articles, reports, and other forms of public dissemination.

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of maternal and child health. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic.

This section provides a complete and precise description of the input data structure for the benchmarking of maternal and child health, excluding the encoding of labels for expected outcomes. It outlines the data format, coding, handling of missing values, standardization, algorithms used, training strategy, hyperparameter tuning, and cross-validation approach.

In the benchmarking of maternal and child health, the AI input data structure includes the following variables:

- Maternal age in years (numeric variable)
- Place of delivery (coded variable: 1=Hospital, 2=Clinic/health center, 3=Home/Other)
- Mode of delivery (coded variable: 1=Vaginal, 2=Vaginal, assisted, 3=C-section, 4=Miscarriage, 5=Medical termination of pregnancy (MTP))
- Birth weight in grams (numeric variable)
- Gestational age in weeks (numeric variable)
- Country (coded variable: 1 = 01 Argentina, 2 = 02 DRC, 3 = 03 Zambia, 6 = 06 Guatemala, 7 = 07 Bangladesh, 8 = 08 Belagavi, 9 = 09 Pakistan, 11 = 11 Nagpur, 12 = 12 Kenya)

To prepare the data for benchmarking, the categorical features were encoded using one-hot encoding, which converts them into binary vectors to enable the AI algorithms to process them effectively. Missing values in continuous variables were imputed using mean imputation, where the missing values were replaced with the mean value of the respective feature. For maternal age, birth weight, and gestational age, a Z-score standardization was applied. This standardization technique transforms the values of these variables into a standard normal distribution by subtracting the mean and dividing by the standard deviation.

The benchmarking process applied a range of AI algorithms for structured data, namely Adaboost, XGBoost, CatBoost, LightGBM, and Random Forest. These algorithms have been carefully selected to ensure comprehensive evaluation and comparison of performance. The training strategy followed a time-period hold-out approach, where the training set comprised data collected from 2010 to 2016, and the test set consisted of data collected from 2017 to 2019. This approach enables the assessment of algorithm performance on different time periods, providing insights into their effectiveness in predicting neonatal risk of death across multiple years.

Hyperparameters for the algorithms were tuned using random search, which involves sampling random combinations of hyperparameters to find the optimal configuration. Model performance was assessed using a 10-fold stratified cross-validation approach, which divides the data into 10 subsets while ensuring proportional representation of different classes. This process was repeated for 50 iterations.

Three general approaches were tested for algorithm construction: general algorithms for all countries, country-specific algorithms, and an algorithm trained on the country with the largest sample in the training set. These approaches were designed to evaluate the best training strategies for the specific dataset and its transferring capabilities across different countries.

6.1.1.4 AI output data structure

The target variable of interest was neonatal status between 1 and 42 days after birth, which is categorized as either "alive" or "death." As a binary feature, it consists of two distinct categories. During the training of the AI models, the focus was on predicting the risk of the neonate experiencing the specified outcome in the future. This prediction is represented as a probabilistic measure. By establishing an initial threshold, typically set at 0.5, the expected outcome for the neonate was determined based on the interaction with this measure.

6.1.1.5 Test data label/annotation structure

While the algorithms can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called 'labels') for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

6.1.1.6 Scores and metrics

Different metrics were used to evaluate the performance, robustness, and general characteristics of the AI algorithms under examination, such as AUC-ROC (Area Under the Receiver Operating Characteristic curve), precision, recall, accuracy, and F1-score. The AUC-ROC metric is particularly significant for comparing different approaches. It provides an overall assessment of the ability of the model to distinguish between the positive and negative classes, making it a suitable criterion for decision-making. Additionally, precision measures the proportion of true positive predictions out of all positive predictions, recall evaluates the proportion of true positive predictions out of the actual positive instances, accuracy determines the overall correctness of the predictions, and the F1-score combines precision and recall assessing the model's overall performance. By employing these metrics, a comprehensive evaluation of the algorithms can be achieved, allowing comparison between distinct countries and training strategies.

6.1.1.7 Test dataset acquisition

Data was obtained from the Global Network Maternal Newborn Health Registry, which is a community-based registry of pregnancy outcomes across multiple countries. The dataset can be accessed through the National Institutes of Health (NIH) data platform, specifically the NICHD Data and Specimen Hub (DASH). DASH serves as a centralized resource where researchers can share and access de-identified data from studies funded by the NICHD. While the data is publicly available and anonymized, a formal data request process is required.

The data request necessitates providing various information, including requester details such as email address, name, job position, phone number, institution name, and address. Additionally, information about the study, including project title, request description, and design and analysis plan, is required. Funding information, such as the funding source, funding type, and identifying number, must also be provided, as well as principal investigator information, including name, email address, institution, and institution address. Furthermore, the names and emails of authorized representatives/institutional business officials and affiliates, if applicable, should be included. To complete the data request, a NICHD DASH Data Use Agreement and Institutional Review Board (IRB) approval is required.

The data acquisition process undergoes an independent audit by the NIH to assess the accuracy of the applicant data and the dataset provided. Once the data request is submitted, it is internally reviewed, and if deemed appropriate, the data is made available to the requester. However, before accessing the data, the requester must sign a non-disclosure agreement to ensure the confidentiality of the data. The quality of the dataset is evaluated through exploratory data analysis, which involves examining metrics such as missing values, outliers, frequency analysis, and basic statistics.

6.1.1.8 Data sharing policies

Data sharing follows the guidelines of the NICHD DASH Data Use Agreement. Thus, no data can be shared through this benchmark. The study is public, and access to the data is free upon request through a data request on the NICHD website.

6.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

6.1.1.10 Reporting methodology

The dissemination of the benchmarking results is an important aspect discussed in this section. Initially, the results will be shared through research papers published in scientific journals. This allows for rigorous scrutiny and validation by the scientific community. Once the findings have undergone scientific dissemination and received validation, they will be communicated to the general public through various channels. These may include newspaper articles, podcasts, magazine reviews, and interviews. By utilizing these mediums, the main outcomes of the benchmarking study can reach a wider audience and contribute to public awareness. It is worth noting that the participants involved in the study will have indirect access to the results, as the dataset undergoes appropriate anonymization measures to ensure their confidentiality.

6.1.1.11 Result

The performance of different machine learning algorithms was evaluated for each country, considering three different model approaches: general, country-specific, and using the largest training size. The models were trained using lightgbm (LGBM), xgboost (XGB), adaboost, catboost and random forest, with hyperparameters optimized through tuning.

In the case of Kenya (Table 1), the LGBM Tuned algorithm achieved an AUC-ROC of 0.794 [0.762, 0.823] when using the General model approach. Comparatively, the Country-specific approach with the same LGBM Tuned algorithm yielded a slightly lower AUC-ROC of 0.788 [0.757, 0.819]. These results suggest that the general algorithm is the more effective approach for achieving better predictive results in Kenya. Similar trends can be observed for other countries. In the Democratic Republic of the Congo (DRC), both the General and Country-specific models utilized the LGBM Tuned algorithm, with AUC-ROCs of 0.789 [0.763, 0.817] and 0.783 [0.758, 0.810], respectively. Although the difference is marginal, the General model approach tends to outperform the Country-specific approach.

For Guatemala, the LGBM Tuned algorithm achieved AUC-ROCs of 0.786 [0.763, 0.807] and 0.779 [0.756, 0.801] for the General and Country-specific models, respectively. Once again, the General model approach exhibits slightly better performance. In the case of Zambia, the General model approach with the LGBM Tuned algorithm achieved an AUC-ROC of 0.789 [0.752, 0.829], while the Country-specific approach with the XGB Tuned algorithm outperformed with an AUC-ROC of 0.795 [0.758, 0.830]. These results indicate that the choice of algorithm may have a greater impact on predictive performance than the model approach itself.

In India-Belgavi, both the General and Country-specific models, using the LGBM Tuned and XGB Tuned algorithms, yielded comparable AUC-ROCs of 0.771 [0.738, 0.801] and 0.770 [0.737, 0.798], respectively. These findings suggest that the choice between algorithms may not significantly affect the predictive performance in this case.

Overall, the results highlight that the choice of algorithm and model approach can influence the predictive performance for each country. While the General model approach generally yields favourable results, there are instances where the Country-specific approach or a different algorithm may be more effective. It is important to carefully consider the specific context and data characteristics when selecting the most suitable algorithm and model approach for each country.

Table 1: Benchmark results for predictive models of neonatal risk of death

Country	Model	Algorithm	Support	Positive Outcome	CI AUC-ROC	CI Recall	Accuracy	Precision	Specificity	F1-score
General[1]	General	LGBM Tuned	134,241	3,219	0.811 [0.803, 0.820]	0.212 [0.198, 0.226]	0.978	0.637	0.997	0.318
General[1]	Biggest train size	XGB Tuned	134,241	3,219	0.773 [0.764, 0.784]	0.180 [0.167, 0.195]	0.978	0.674	0.998	0.284
DRC	General	LGBM Tuned	18,581	458	0.789 [0.763, 0.817]	0.264 [0.226, 0.304]	0.979	0.720	0.997	0.387
DRC	Country-specific	XGB Tuned	18,581	458	0.783 [0.758, 0.810]	0.244 [0.207, 0.284]	0.979	0.723	0.998	0.365
DRC	Biggest train size	XGB Tuned	18,581	458	0.781 [0.757, 0.809]	0.242 [0.202, 0.285]	0.978	0.649	0.997	0.353
Guatemala	General	LGBM Tuned	25,252	624	0.786 [0.763, 0.807]	0.205 [0.174, 0.237]	0.979	0.766	0.998	0.324
Guatemala	Country-specific	XGB Tuned	25,252	624	0.779 [0.756, 0.801]	0.193 [0.164, 0.226]	0.978	0.716	0.998	0.305
Guatemala	Biggest train size	XGB Tuned	25,252	624	0.732 [0.708, 0.757]	0.141 [0.115, 0.168]	0.978	0.838	0.999	0.241
Zambia	General	LGBM Tuned	18,684	222	0.789 [0.752, 0.829]	0.248 [0.197, 0.305]	0.990	0.663	0.998	0.361
Zambia	Country-specific	XGB Tuned	18,684	222	0.795 [0.758, 0.830]	0.225 [0.177, 0.280]	0.989	0.568	0.998	0.323
Zambia	Biggest train size	XGB Tuned	18,684	222	0.751 [0.712, 0.794]	0.238 [0.190, 0.300]	0.989	0.582	0.998	0.339
India-Belgavi	General	LGBM Tuned	16,085	328	0.771 [0.738, 0.801]	0.265 [0.215, 0.313]	0.983	0.744	0.998	0.391
India-Belgavi	Country-specific	XGB Tuned	16,085	328	0.770 [0.737, 0.798]	0.262 [0.212, 0.309]	0.983	0.735	0.998	0.387
India-Belgavi	Biggest train size	XGB Tuned	16,085	328	0.770 [0.737, 0.798]	0.262 [0.212, 0.309]	0.983	0.735	0.998	0.387
India-Nagpur	General	LGBM Tuned	18,294	339	0.809 [0.780, 0.839]	0.221 [0.179, 0.264]	0.982	0.568	0.997	0.318
India-Nagpur	Country-specific	LGBM Tuned	18,294	339	0.809 [0.782, 0.839]	0.194 [0.156, 0.237]	0.983	0.595	0.997	0.293
India-Nagpur	Biggest train size	XGB Tuned	18,294	339	0.809 [0.781, 0.837]	0.203 [0.165, 0.246]	0.982	0.552	0.997	0.297
Pakistan	General	LGBM Tuned	16,599	916	0.770 [0.747, 0.782]	0.179 [0.154, 0.202]	0.946	0.522	0.990	0.267
Pakistan	Country-specific	LGBM Tuned	16,599	916	0.765 [0.747, 0.783]	0.183 [0.159, 0.208]	0.946	0.525	0.990	0.272
Pakistan	Biggest train size	XGB Tuned	16,599	916	0.753 [0.734, 0.771]	0.145 [0.122, 0.166]	0.950	0.727	0.997	0.242
Kenya	General	LGBM Tuned	19,657	300	0.794 [0.762, 0.823]	0.157 [0.117, 0.201]	0.986	0.595	0.998	0.248
Kenya	Country-specific	LGBM Tuned	19,657	300	0.788 [0.757, 0.819]	0.153 [0.114, 0.195]	0.986	0.676	0.999	0.250
Kenya	Biggest train size	XGB Tuned	19,657	300	0.765 [0.731, 0.800]	0.120 [0.085, 0.159]	0.985	0.562	0.999	0.198
Bangladesh[2]	General	LGBM Tuned	1,089	32	0.860 [0.787, 0.926]	0.156 [0.038, 0.306]	0.971	0.500	0.995	0.238
Bangladesh[2]	Biggest train size	XGB Tuned	1,089	32	0.853 [0.776, 0.928]	0.125 [0.027, 0.250]	0.973	0.800	0.999	0.216

Regarding the analysis of the most important predictors, we found a similar behaviour to what has been previously observed in the literature (Figure 2). The variables Birth Weight and Gestational Age were the most relevant for predicting the risk of neonatal death, when considering the general model, where the LGBM algorithm performed the best. The inclusion of variables related to the countries of origin revealed the differences in the contribution of each country to the composition of the final probability of outcomes. In other words, including the country of residence of the pregnant women in the algorithm presented important predictive information.

These findings highlight the importance of these variables in assessing the risk of neonatal mortality and emphasize the importance of considering country-specific factors in the prediction models. It is crucial to take into account the variations in healthcare systems, socio-economic conditions, and maternal health indicators across different countries, as they can greatly influence the outcomes. The results underscore the potential value of using machine learning algorithms to predict neonatal mortality and improve the identification of high-risk pregnancies. By incorporating both general predictors and country-specific variables, more accurate risk assessments can be made, enabling targeted interventions and resource allocation to areas with the highest vulnerability.

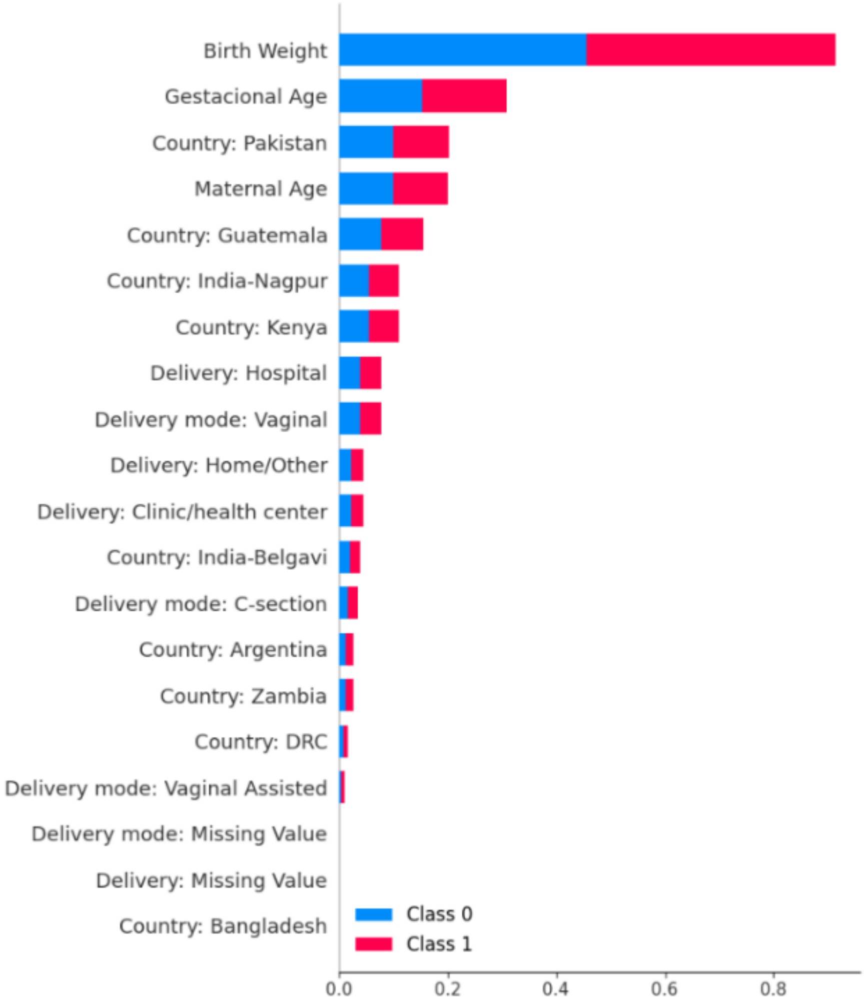


Figure 2: Mean absolute Shapley-values bar plot for predictors of neonatal risk of death using LGBM Tuned algorithm in the general approach.

Considering the algorithm trained with data from India-Belgavi (XGB Tuned) (Figure 3), which had the largest sample size in the training set, the variables Birth Weight, Gestational Age, and Maternal Age were observed to be the strongest predictors, followed by place of delivery and type of delivery. This result reinforces the importance of collecting the five minimum indicators recommended by the WHO for assessing the neonatal situation in countries. With these indicators, it is possible to build predictive tools that can contribute to clinical decision-making in the neonatal population.

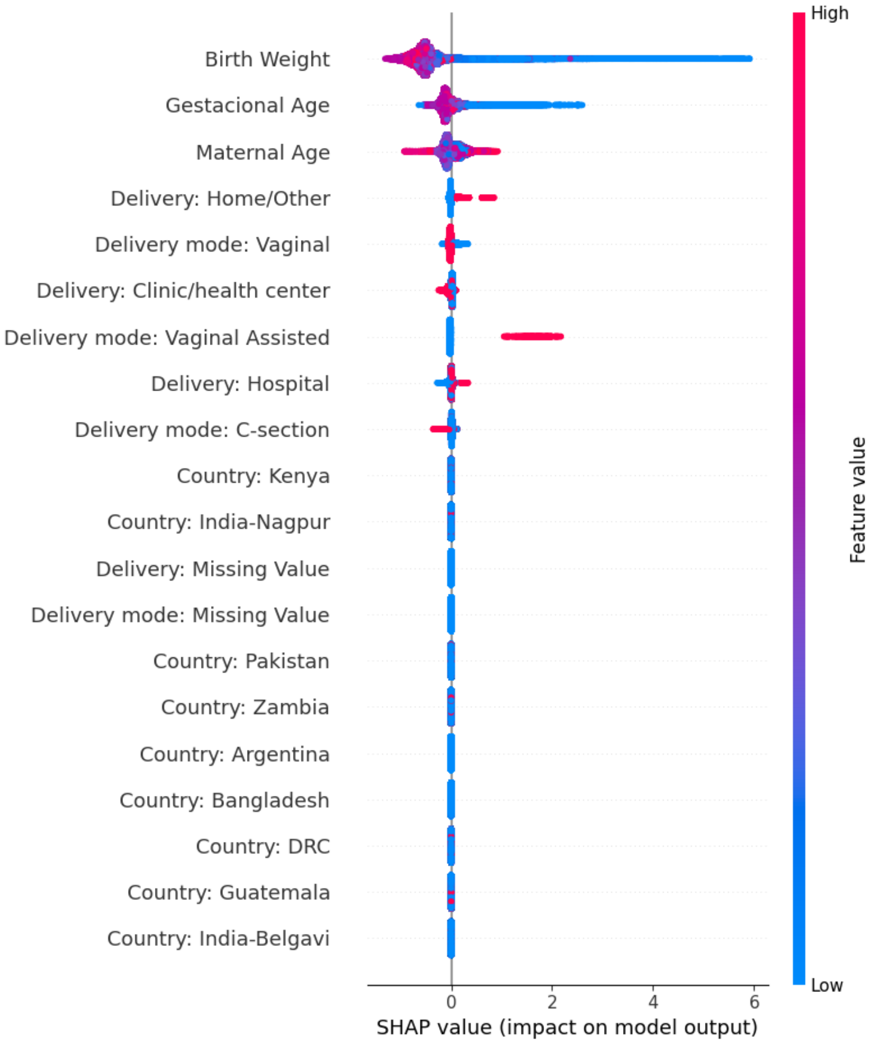


Figure 3: Shapley-values plot for predictors of neonatal risk of death using XGB Tuned algorithm in the biggest training sample size approach (India-Belgavi).

6.1.1.12 Discussion of the benchmarking

The benchmarking process involved evaluating different algorithms and approaches to assess their performance in predicting the outcome of neonatal risk of death. The outcomes of these benchmarking iterations were analysed to gain valuable insights and draw meaningful conclusions. One of the key findings from the benchmarking iterations was the identification of important predictors for neonatal risk of death. Variables such as Birth Weight, Gestational Age, and Maternal Age emerged as strong predictors, indicating their significance in predicting the outcome. Additionally, factors such as the place of delivery and type of delivery were also found to be influential predictors. Moreover, the benchmarking process highlighted the importance of collecting the minimum five indicators recommended by the World Health Organization (WHO) for assessing the neonatal situation in different countries. These indicators provide valuable information for constructing predictive tools that can aid in clinical decision-making for neonatal populations.

During the benchmarking process, a comprehensive evaluation of algorithms, including LGBM and XGB was conducted using different approaches, such as the general and the country-specific. The performance of these algorithms was carefully assessed using key metrics, such as AUC-ROC, recall, accuracy, precision, specificity, and F1-score. The benchmark results clearly indicate that the choice of algorithm and approach plays a crucial role in determining the predictive performance across different countries. Notably, the general approach, employing the LGBM algorithm, consistently demonstrated superior performance, surpassing other approaches in terms of the AUC-ROC. Furthermore, it is noteworthy that the XGB algorithm, utilized in the biggest training sample size approach (India-Belgavi), showed impressive predictive capabilities, indicating the viability for transferring an algorithm trained in one country to others.

Overall, the benchmarking process provided valuable insights into the predictive modelling of neonatal risk of death. The findings emphasize the importance of specific predictors, the inclusion of country-specific variables, and the selection of appropriate algorithms and approaches. These insights can contribute to the development of more accurate and effective predictive models for supporting clinical decision-making in neonatal populations.

6.1.1.13 Retirement

Once the benchmarking activity is concluded, the handling of the AI system and data requires careful consideration. In this context, the database used for the benchmarking process can serve various purposes, including traceability and future research endeavours. Retaining the data allows for deeper studies and analysis, enabling researchers to gain further insights and potentially uncover additional patterns or trends. In this regard, the data will be stored under the responsibility of the principal investigator and the data custodian, subject to a formal data request submitted to the NIH.

7 Overall discussion of the benchmarking

The efforts of this topic group highlight the considerable challenges of establishing a machine learning benchmark within an area as complex as maternal and child health, considering its importance and need for local strategic planning and methodological precision. A first challenge arises in identifying and delineating the most pressing health issues regarding the well-being of both the mother and the child, each with its own particular needs. The complexity of this task is amplified by the diverse socio-economic contexts of countries with different economic capacities.

There is a need for further examination of the availability of epidemiological data, health indices, socio-cultural determinants, and prevailing healthcare systems across diverse geopolitical contexts. Beyond health concerns, there is not only a need for rigorous quantitative analysis, but equally an awareness of qualitative factors such as prevailing cultural norms, data quality for the machine learning models and societal perceptions.

Further efforts need to focus on the development of benchmarks that are not only theoretically sound but also relevant to the dynamics of each locale. The essence of these machine learning benchmarks depends on the robustness of the data that is provided to them, necessitating a collaboration with local institutions, healthcare providers, and relevant authorities to ensure higher accuracy, comprehensiveness, and reliability.

In conclusion, the complex nature of this task necessitates closer relationship between theoretical constructs, empirical insights, and contextual details. There is a need for machine learning benchmarks that serve as generalizable tools for the evaluation and improvement of maternal and child health standards across diverse socioeconomic realities.

8 Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *"Regulatory considerations on AI for health" (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL2.1 *"Mapping of IMDRF essential principles to AI for health software"*, and DEL2.2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL4 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG-MCH.

8.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for maternal and child health.

8.2 Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

8.3 Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

8.4 Regulatory approach for the topic group

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL2 *"AI4H regulatory considerations."*

References

Betts KS, Kisely S, Alati R. Predicting common maternal postpartum complications: leveraging health administrative data and machine learning. BJOG 2019;126(6):702-9.

Harrison E, Syed S, Ehsan L, Iqbal NT, Sadiq K, Umrani F, Ahmed S, Rahman N, Jakhro S, Ma JZ, Hughes M, Ali SA. Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth - a four-year prospective study. BMC Pediatr 2020;20(1):498.

Hoffman MK, Ma N, Roberts A. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. Am J Obstet Gynecol 2021;3:100250.

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
DEL	Deliverable	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
PII	Personal identifiable information	
SaMD	Software as a medical device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group MCH
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B:
Declaration of conflict of interests

None declared:

Prof. Dr. Alexandre Chiavegatto Filho
School of Public Health
University of Sao Paulo