|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-F-006 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | N/A | Geneva, 30 May – 1 June 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Cardio Topic Driver | | |
| **Title:** | | TDD draft: TG-Cardio (Cardiovascular disease risk prediction) | | |
| **Purpose:** | | Approval | | |
| **Contact:** | | Benjamin R.H. Muthambi, DrPH, MPH  IEPH, Inc.  6224 Spring Knoll Drive  Harrisburg, PA 17111, USA | | Email: [brm5@caa.columbia.edu](mailto:brm5@caa.columbia.edu) |

|  |  |
| --- | --- |
| **Abstract:** | This document contains a skeleton of the Topic Driver Document for Cardiovascular disease risk prediction. This document was prepared by the secretariat on behalf of the TG driver and has not yet been reviewed. |

Table of Contents

[1 Introduction 3](#_Toc10740458)

[1.1 Topic Description 3](#_Toc10740459)

[1.1.1 Topic Objectives 3](#_Toc10740460)

[1.2 Ethical Considerations 4](#_Toc10740461)

[1.3 Existing AI Solutions 4](#_Toc10740462)

[1.4 Existing Work on Benchmarking 4](#_Toc10740463)

[2 AI4H Topic Group 4](#_Toc10740464)

[3 Method 5](#_Toc10740465)

[3.1 AI Input Data Structure 5](#_Toc10740466)

[3.1.1 Data availability 5](#_Toc10740467)

[3.2 AI Output Data Structure 5](#_Toc10740468)

[3.3 Test Data Labels 5](#_Toc10740469)

[3.4 Scores and Metrics 5](#_Toc10740470)

[3.5 Undisclosed Test Data Set Collection 5](#_Toc10740471)

# Introduction

## Topic Description

i) Topic Driver Team: Responsibilities of the topic driver may evolve over time but an interim list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document.
* Reviewing the input documents for the topic and moderating the integration in a dedicated session at each Focus Group meeting.
* Organizing regular phone calls to coordinate work on the topic description document between meetings.

ii) Topic Group Peer-Review & Advisory Forum: The preliminary objectives of the Topic Stakeholder Community are manifold:

* to provide a forum for open communication among various stakeholders,
* to agree upon the benchmarking tasks of this topic and scoring metrics,
* to facilitate the collection of high-quality labelled test data from different sources,
* to clarify the input and output format of the test data,
* to define and set-up the technical benchmarking infrastructure, and
* to coordinate the benchmarking process in collaboration with the Focus Group management and working groups.

The primary output of this topic group forum is one document that describes all aspects of how to perform the benchmarking for this topic. (The document will be developed in a cooperative way by suggesting changes as input documents for the next FG-AI4H meeting that will then be discussed and integrated into an official output document of this meeting. The process will continue over several meetings until the topic description document is ready for performing the first benchmarking.)

iii) Sub-topics & Technical Subgroups/Forums (Project Phase-Specific): The preliminary sub-topics & technical subgroups/forums subject to further calls for participants to be announced include:

* Conceptualization of Sub-Topics on CVD-related Predictive Analytics in Clinical & Public Health, & Applications thereof
* External Peer Review for Ethics (Pre-IRB) & Quality Assurance of Project Proceedings
* Methods: Epidemiology/Evaluation Study Design, Metrics & Statistical Methods
* Computer & Data Science: Machine Learning Algorithm Developer, Server Admins & Distributed Systems Engineering
* R Programming: Data Management, Predictive Analytics, Statistical Analyses, Shiny Web App
* Project Planning & Reproducible Reporting Tools & Writing

### Topic Objectives

This section introduces the subject matter of the topic including objectives and intended benchmarking task, relevance and data availability, followed by the next section on how to get involved.

#### Project Objectives/Problem to be addressed

Diabetics have higher CVD risk, hence improved CVD risk prediction is critical for better diabetes management and reducing mortality. The proposed project aims to:

a) To assess: i) CVD risk prediction accuracy of various machine learning (ML) methods benchmarked against CVD risk based on actually observed occurrence of first CVD event (truth) documented in diverse cohorts/populations data, and ii) replicability/reproducibility of ML prediction of CVD risk using 'external data' from diverse populations meeting prescribed criteria but 'not previously accessed' (undisclosed) data to the ML algorithms under evaluation;

b) Compare CVD risk prediction accuracy of ML algorithms referenced above in section 3.1. (a) to: i) several routine clinical-use CVD risk scoring tools/calculators, and ii) traditional multivariate statistical methods (in collaboration with other co-investigators who recently undertook similar risk prediction accuracy studies);

c) Determine which methods, if any, consistently show better predictive accuracy across diverse populations. Using the above-referenced methods, benchmarking, anticipated findings and peer-review thereof, the project expects to establish an evidence-based standards-setting blueprint.

#### Relevance/Background, significance and rationale

Cardiovascular disease (CVD) is the global leading cause of morbidity and mortality (WHO, 2014). CVD accounts for > 2/3 of mortality among type 2 diabetes patients (ADA, 2019). Widely used clinical CVD risk scoring tools/calculators incorporate several factors with well-established etiological associations with CVD such as age, sex, BMI, systolic blood pressure, smoking, A1C, lipid levels, age at diagnosis &/or onset of diabetes, diabetes duration, and antihypertensive and lipid-reducing drugs, but do not necessarily include a comparable set of predictors. In addition, these methods often fail to identify many people who would benefit from preventive treatment, while others receive unnecessary interventions. For example, approx. ~50% of myocardial infarctions (MIs) and strokes occur among persons predicted to be at risk of CVD by routinely-used risk calculators (Ridker et al, 2008). Highlighting the need for standardization, prior systematic CVD risk prediction accuracy studies often used: disparate study populations; either traditional multivariate statistical methods or disparate ML algorithms; incomparable sets of predictors often not considering the broader range of potential predictor data made possible by mining electronic health records, big data aggregation, or accounting for complex interactions that ML can handle more easily; and also used different measures of predictive accuracy. This study hypothesizes that ML algorithms can improve CVD predictive accuracy over CVD risk scoring tools/calculators used in the standard of practice across diverse populations. If demonstrated, ML-assisted DSS should be considered as the underlying approach for standard of practice in CVD risk prediction.

## Ethical Considerations

TBC

## Existing AI Solutions

TBC

## Existing Work on Benchmarking

TBC

# AI4H Topic Group

TBC

# Method

## AI Input Data Structure

### Data availability

#### Public Data

Anonymized data already acquired for preliminary work for this project are publicly-accessible academic training-use data extracted from the NIH-funded Framingham Heart Study which were in turn sourced from the BioLINCC data repository under NIH data sharing terms.

## AI Output Data Structure

TBC

## Test Data Labels

TBC

## Scores and Metrics

TBC

## Undisclosed Test Data Set Collection

'External data' meeting prescribed criteria but 'not previously accessed' (undisclosed data) to train the ML algorithms under evaluation, will be sourced for replication studies from various repositories identified as suitable potential data sources including multiple researcher-use data obtainable under well-established NIH-funded research data sharing terms applicable to the NHLBI/National Heart, Lung, and Blood Institute’s Open BioLINCC Biologic Specimen and Data Repository, i.e. NIH/NHLBI BioLINCC; UK NHS/CPRD data repository; and a diverse range of other data sources with suitable data identified in the literature, and other data sources still to be identified through a planned call for data-contributing project participants (per above-referenced AI for cardiovascular disease risk prediction topic group's Project Phase-Specific Technical Contributor Subgroups/Forums. The identified potential sources of 'not previously accessed/undisclosed data' contain clinical and other patient data used in routine clinical care by CVD risk scoring tools/calculators, and in research using traditional multivariate statistical methods or ML algorithms.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_