|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FGAI4H-J-006-A01 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plen | E-meeting, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | TG-Cardio Topic Driver |
| **Title:** | Att.1 – TDD update (TG-Cardio) [same as Meeting H] |
| **Purpose:** | Discussion |
| **Contact:** | Benjamin MuthambiWatIF Health, South Africa | Email: brm5@caa.columbia.edu |

|  |  |
| --- | --- |
| **Abstract:** | This document provides the **Topic Driver Document (TDD) update for TG-Cardio | Subtopic: Cardiovascular disease (CVD) risk prediction using AI**. **NOTE**: A 2nd subtopic was added to the topic group TG-Cardio on use of AI in Cardiovascular disease (CVD) management, i.e. *Subtopic on Coronary CT image processing using AI [Subtopic Driver(s) Guo et al].* With the advent of the addition of a subtopic not primarily addressing the originally envisaged limited scope of TG-Cardio which focused on CVD risk prediction, the scope of topic group TG-Cardio has thus been effectively expanded into a broader cardiac organ system-based cluster of potential subtopics to accommodate Coronary CT image processing using AI, and potentially other applications of AI in Cardiovascular disease (CVD) management.**SUBTOPIC CATEGORIES UNDER TG-CARDIO**: Inclusive of the subtopic on CVD risk prediction using AI, the broadening of the scope of TG-Cardio means that other subtopics could potentially still be proposed addressing a range of other applications of AI in CVD management. Potential subtopics are broadly classified by [Yan et al (2019](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748906/pdf/jgc-16-08-585.pdf)) into the 4 subtopic categories listed below:[X] **CLINICAL PREDICTIONS -** ***Cardiovascular disease (CVD) Risk Prediction***. (Subtopic for TDD in this document – Subtopic Driver(s) Muthambi et al.)[ ] **CARDIAC IMAGE ANALYSES – *Coronary CT Image Processing/Image Recognition for Coronary CT angiography (CCTA) in coronary artery disease (CAD) diagnosis***. (Subtopic TDD in separate document pending - Subtopic Driver(s) Guo et al.)[ ] **INTELLIGENT ROBOTS** – ***Surgical Robot Technologies incl. AI-assisted Minimally Invasive Cardiac Surgery*** (Subtopic not yet proposed)[ ] **PRECISION MEDICINE** – ***AI-assisted Individualized Medicine and healthcare customized for each patient***. (Subtopic not yet proposed)This version of the TDD is the same as seen in Meeting H (FGAI4H-H-006-A01), reproduced for easier reference as a Meeting J document.  |

Table of Contents

TDD for TG-Cardio Subtopic on Cardiovascular Risk Prediction

[\*For efficiency, Section 1.2. (marked with asterisk) may be similar across TG-cardio subtopic TDDs]

[1 Introduction 4](#_Toc30167047)

[1.1 Topic/Subtopic Description – Cardiovascular diseases (CVD) Risk Prediction 4](#_Toc30167048)

[1.1.1 Subtopic objectives 4](#_Toc30167049)

[1.2 \*Ethical Considerations for TG-Cardio 5](#_Toc30167050)

[1.2.1 "Research" requiring review and approval by an Institutional Review Board (IRB) including 'equivalent research-review & approval authorities' (ERR&AA), for purposes of human research protections 5](#_Toc30167051)

[1.2.2 Requirement for IRB review and approval in advance of commencement of "research" activities in any of the above-referenced types of "research" involving human subjects (incl. other projects requiring IRB approval as referenced above) 6](#_Toc30167052)

[1.2.3 TG Cardio Rights to Terminate Association with Any Projects for Research Ethics Violations 7](#_Toc30167053)

[1.3 Existing AI solutions (TG-Cardio | Subtopic: CVD Risk Prediction) 7](#_Toc30167054)

[1.4 Existing work on benchmarking (TG-Cardio | Subtopic: CVD Risk Prediction) 8](#_Toc30167055)

[1.4.1 CVD risk prediction approaches to be compared & evaluated for predictive accuracy: 8](#_Toc30167056)

[2 Subtopic Group for CVD Risk Prediction using AI (TG-Cardio | Subtopic: CVD Risk Prediction) 8](#_Toc30167057)

[2.1 Mandate of CVD Risk Prediction Subtopic Group: Primary subtopic group output & mandate of the subtopic group teams, fora & technical subgroups 8](#_Toc30167058)

[2.1.1 Primary CVD Risk Prediction Subtopic Group Output (Topic Description Document) and Document Development Process for TG Cardio | Subtopic CVD Risk Prediction 8](#_Toc30167059)

[2.1.2 Mandate of CVD Risk Prediction Subtopic Group Teams, Fora & Technical Subgroups 9](#_Toc30167060)

[2.2 Common Projects on Subtopic on CVD Risk Prediction using AI incl. Replication Studies 10](#_Toc30167061)

[2.3 Participation in Subtopic Group on CVD Risk Prediction using AI and "Call for Participation" 11](#_Toc30167062)

[2.4 Status of the TDD & Subtopic Group on CVD Risk Prediction using AI: 11](#_Toc30167063)

[2.5 Focus Group & Topic Group Meetings & Collaboration 11](#_Toc30167064)

[3 Methods for CVD Risk Prediction 12](#_Toc30167065)

[3.1 AI *input data* structure 12](#_Toc30167066)

[3.1.1 Data availability 13](#_Toc30167067)

[3.2 AI *Output data* structure 13](#_Toc30167068)

[3.3 Test data labelling 14](#_Toc30167069)

[3.3.1 Protection of confidentiality in data labelling: 14](#_Toc30167070)

[3.3.2 Quality of data annotation/labelling: 14](#_Toc30167071)

[3.3.3 Recoding, creation and labelling of standardized datasets for algorithm training and reproducible/replicable analyses: 14](#_Toc30167072)

[3.4 Scores and metrics 15](#_Toc30167073)

[3.5 Undisclosed test data set collection for replication studies across diverse populations 15](#_Toc30167074)

# Introduction

## Topic/Subtopic Description – Cardiovascular diseases (CVD) Risk Prediction

### Subtopic 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

Cardiovascular diseases (CVD) is a major threat to human health and the leading cause of death worldwide (WHO, 2014; ADA, 2019). Certain subgroups including diabetics have higher CVD risk (ADA, 2019), hence improved CVD risk prediction is also critical for better diabetes management and reducing mortality. The proposed project aims to:

1. To assess:
2. 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
3. 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;
4. Compare CVD risk prediction accuracy of several ML algorithms referenced above under (a) to:
5. Several routine clinical-use CVD risk scoring tools/calculators, and
6. Traditional multivariate statistical methods (in collaboration with other co-investigators who recently undertook similar risk prediction accuracy studies);
7. 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 metrics/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 for TG-Cardio

This section provides a framework for protection of the rights, welfare, and wellbeing of human subjects involved in research conducted within the TG Cardio framework.

### "Research" requiring review and approval by an Institutional Review Board (IRB) including 'equivalent research-review & approval authorities' (ERR&AA), for purposes of human research protections

Projects to be undertaken within the terms of reference of TG Cardio shall be deemed to be "research" requiring review and approval by an IRB (incl. an ERR&AA), henceforth collectively referred to simply as "IRB", IF they meet the following definitions &/or conditions.

#### Definition of "Research" requiring review and approval by an IRB

For purposes of determining what will be deemed to be "Research" vs. "Non-Research", the term "Research" shall be understood to mean *a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge*.

#### Definition of "Non-Research" investigational activities/projects NOT requiring review and approval by an IRB

Projects NOT deemed to be research which fall under the "non-research" category may include the following:

* Scholarly and journalistic activities (e.g., oral history, journalism, biography, literary criticism, legal research, and historical scholarship), including the collection and use of information, that focus directly on the specific individuals about whom the information is collected;
* Public health surveillance, other investigations or interventional activities duly authorized/required by statutes or regulations applicable to the jurisdiction in which pertinent data collection is to be undertaken;
* Other *routine* public health or clinical data collection including routine documentation of public health services or clinical care rendered, &/or *routine* public health, health care or prevention procedure and service activities undertaken as part of *routine* health services consistent with prevailing standards of care/prevention or *routine* public health services, and thus specifically undertaken for *routine* purposes and NOT for "research" purposes;
* Certain activities authorized by law or court order solely for criminal justice or criminal investigative purposes;
* Certain activities in support of intelligence, homeland security, defense, or other national security missions;

#### Human subjects to whom the above-referenced human research protections apply

In this context, usage of the term **Human subject,** shall be understood to mean a living *or deceased* individual about whom an investigator (whether professional or student) conducting research may:

* Obtain information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes, or generates identifiable private information; or
* By intent or inadvertently, obtain/access, use, study, analyze or generate identifiable private information or identifiable biospecimens;

#### Research procedure examples which often raise IRB review requirements

The following are examples and not an exhaustive list of procedures which may be undertaken for “research” purposes that is subject to IRB review and approval in contexts not meant for administrative or public health investigations (authorized by governmental statutes/regulations) or for routine healthcare or recording/monitoring thereof, i.e. whereby such activities would not be undertaken except to fulfil research objectives:

* Interviews
* Questionnaires/Surveys
* Focus Groups
* Observations
* Records Reviews/Abstractions (patient medical records, school student records, etc)
* Tests/Tasks
* Medical procedures
* Blood draws, genetic tests, saliva samples
* Secondary Data Analysis

#### Other examples of activities requiring IRB review within the TG Cardio projects framework

* Student activities that are not research but present > minimal risk to participants, with "minimal" vs. "> minimal risk" determined by the IRB at the student's institution or other institutional guidelines;
* Genetic Testing (NYS 79-l definition) using anonymous human biological samples
* Research involving deidentified data from a repository and/or dataset that requires IRB approval (e.g. Framingham heart study data from dbGap)
* Research involving data for which the provider requires compliance with strict security requirements (e.g. FISMA requirements or CMS data)

### Requirement for IRB review and approval in advance of commencement of "research" activities in any of the above-referenced types of "research" involving human subjects (incl. other projects requiring IRB approval as referenced above)

Prior to commencement of any "research" activities studies/projects within the TG Cardio Framework (incl. data collection/acquisition, interventions, or execution of any of the above methods/procedures for "research" purposes), all studies utilizing data derived from "human subjects" to be conducted within the Topic Group (TG Cardio) framework, including AI algorithm training projects or replication studies, shall be subject to the above-referenced ethics review and approval by a bona-fide Institutional Review Board (IRB) deemed competent and recognized by a health science/medical research regulatory national authority for purposes of reviewing research compliance with human research protections and granting project approvals in/for the jurisdiction that the data are to be sourced &/or jurisdiction in which the study is to be conducted. Notwithstanding IRB approvals granted to projects to be undertaken, the lead investigator/principal investigator (PI) of each project shall be responsible for ensuring that IRB approvals correspondence is copied (cc'd) directly from the approving IRB to the TG Cardio secretariat (to be established) in addition to a copy of the project protocol approved by the IRB, and a TG Cardio form letter (To be developed) signed by the PI of the project acknowledging that TG Cardio (and its coordinators), ITU, WHO and their associated agencies are not a party to any of the studies approved and shall be fully indemnified from any and all liability which may be incurred by the approved study, which shall also be reflected in any agreements with sources of existing data procured and/or consent forms signed by patients where applicable. Subsequent to obtaining IRB approval, the responsible PI of any project within the TG Cardio framework shall ensure that TG Cardio is kept informed of any subsequent adverse changes in the IRB approval status of the project such as adverse restrictions/modifications of terms or revocation of approval, including termination of approval due to completion of the project.

### TG Cardio Rights to Terminate Association with Any Projects for Research Ethics Violations

TG Cardio shall at all times reserve the right to terminate its association with any project for wilful or repeated negligent violation of the above-referenced research ethics requirements.

## Existing AI solutions (TG-Cardio | Subtopic: CVD Risk Prediction)

The number of CVD risk prediction studies demonstrating potential AI/machine learning solutions is growing rapidly, and a number of health organizations are developing, piloting and implementing their own proprietary AI/ML-based clinical decision support sub-systems embedded in real world health system settings. These predictive algorithms are thus used to help identify patients at particular risk levels for adverse outcomes (or those with complex health needs), including use as adjuncts to existing standard-of-care CVD risk scoring tools/prediction calculators.

Several challenges have plagued the plethora of emerging AI/ML-based clinical decision support sub-systems. Highlighting the need for standardization of evaluation of predictive fidelity of these algorithms is, most notably, racial bias recently revealed in an evaluation of an algorithm which is live and deployed at scale in the management of the health of large populations across the United States [Obermeyer et al., 2019; Paul, 2019]. This algorithm is reportedly one of the most widely used among typical examples of a class of commercial risk-prediction tools that is, by industry estimates, said to be applied, each year, to nearly 200 million people across the United States. Unaware of racial biases recently uncovered in this algorithm, which could be adversely impacting millions of African Americans as reported in the above-referenced evaluation study, large health systems and health insurance claims payers depend on this algorithm to target care management programs designed to patients predicted to be at "high-risk" of various adverse outcomes (or those with complex health needs).

Further, highlighting the need for standards-setting and robust evaluation prior to adoption of AI in health, 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 [ relevant studies: Narain et al, 2016 (FHS-USA); Fox et al, 2016 (JHS-USA); Ambale-Venkatesh et al, 2017 (MESA-USA); Weng et al, 2017 (NHS-CPRD-UK); Unnikrishnan et al, 2016 (BMES-AUS); & related methodology: Rahimian et al, 2018 ; Luo et al, 2016; Bal et al, 2014 ].

More specifically, public domain ML algorithm-based solutions previously studied for disease risk prediction accuracy can be loosely categorized as:

* Simple linear (Linear Discriminant Analysis/LDA),
* Nonlinear (Classification and Regression Trees/CART; K-Nearest Neighbors/kNN; & gradient boosting classifier/GBC), &
* Complex nonlinear methods (Support Vector Machines/SVM; Random Forest/RF; & Artificial Neural Networks/ANNs).

Of note, none of the CVD risk prediction algorithms studied are known to have yet been approved by FDA or other countries' regulatory authorities for use as the standard-of-care for clinical decision support in patient care/individualized healthcare.

## Existing work on benchmarking (TG-Cardio | Subtopic: CVD Risk Prediction)

### CVD risk prediction approaches to be compared & evaluated for predictive accuracy:

1. Within 3 domains of risk prediction approaches, methods to be compared for CVD risk prediction accuracy across diverse populations include:
Clinical CVD risk scoring tools/calculators such as ACC/AHA ([Goff et al, 2014](https://www.ncbi.nlm.nih.gov/pubmed/24222018/)), QRISK2 ([Hippisley-Cox et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18573856/)), Framingham ([D'Agostino et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18212285/)), SCORE ([Conroy et al, 2003](https://doi.org/10.1016/S0195-668X%2803%2900114-3)), DECODE ([Balkau et al, 2004](https://doi.org/10.1007/s00125-004-1574-5)), Reynolds Risk Score ([Ridker et al, 2007](https://www.ncbi.nlm.nih.gov/pubmed/17299196/)); UKPDS ([Simmons et al, 2009](http://care.diabetesjournals.org/content/32/4/708) ; [UKPDS risk engine](https://www.dtu.ox.ac.uk/riskengine/) ), Swedish NDR 5-yr risk equation ([Cederholm et al, 2008](http://care.diabetesjournals.org/content/31/10/2038#ref-6); [Jackson R, 2008](http://dx.doi.org/10.1136/hrt.2007.138040)), & WatifHealth algorithms ([Sipula N, 2018](http://diabetescare.africa/));
2. Multivariate statistical risk prediction methods such as Cox Proportional Hazards and Multiple Logistic Regression;
3. Public domain ML algorithms previously used for disease risk prediction loosely categorized as simple linear (Linear Discriminant Analysis/LDA), nonlinear (Classification and Regression Trees/CART; K-Nearest Neighbors/kNN; & gradient boosting classifier/GBC) & complex nonlinear methods (Support Vector Machines/SVM; Random Forest/RF; & Artificial Neural Networks/ANNs);

# Subtopic Group for CVD Risk Prediction using AI (TG-Cardio | Subtopic: CVD Risk Prediction)

## Mandate of CVD Risk Prediction Subtopic Group: Primary subtopic group output & mandate of the subtopic group teams, fora & technical subgroups

### Primary CVD Risk Prediction Subtopic Group Output (Topic Description Document) and Document Development Process for TG Cardio | Subtopic CVD Risk Prediction

* **Purpose**: The primary subtopic group's output is ***this document*** which is an ITU-mandated topic group output named a *topic description document (TDD).* This TDD is supposed to describe all aspects of how to perform the benchmarking for this subtopic.
* **TDD Development Process**: Further development of this TDD beyond this draft will be done in a cooperative way whereby *active participants* in technical advisory/consultative subgroups of this subtopic group will develop consensus (or vote) on proposed changes to the section of this subtopic’s TDD that’s assigned to the subgroup for peer-review and refinement. Proposed changes generated by the subgroup’s peer-review and refinement activities will be uploaded to the relevant online collaboration portal discussion (for asynchronous discussion) and if possible, subsequently presented to the Subtopic Driver Team at scheduled live webinar events for the given subgroup to provide inputs. The Subtopic Driver Team (Subtopic Group Convenor & Secretariat) shall moderate comments/suggested changes and seek a determination of adoption/non-adoption by way of consultation and consensus (or vote) by Members of the Subtopic Drivers' Forum. Suggested TDD changes submitted during the 30-day period following each FG-AI4H meeting will be included in updated versions getting produced and adopted in time for the next FG-AI4H meeting at which the latest version of the TDD shall be submitted. The process will continue over several meetings until the topic description document is ready for performing the first benchmarking.

### Mandate of CVD Risk Prediction Subtopic Group Teams, Fora & Technical Subgroups

#### Subtopic Driver Team (Subtopic Group Convenor & Secretariat):

Responsibilities of the subtopic driver may evolve over time but an interim list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document (TDD) for this subtopic.
* Reviewing the input documents for the subtopic, including suggested TDD changes/updates, and moderating the integration of suggested changes as described above under TDD Development Process in sub-section 2.1.1.
* Organizing online asynchronous/synchronous consultative mechanisms to facilitate and coordinate the TDD development process between Subtopic Group adoption and submission of official TDD updates for each FG-AI4H meeting.
* Facilitation of the benchmarking process in collaboration with the Focus Group management and working groups.

The Subtopic Driver Team (Subtopic Group Convenor & Secretariat) consists of the following members:

* Dr. Benjamin Muthambi, DrPH, MPH(Epidemiology); *Managing Director & Snr. Fellow-in-Residence*, IEPH/Institutes of Epidemiology & Public Health
* Dr. Nao Sipula, MBChB, PDH(Global Health); *CEO*, WatifHealth
* Mr. Jason Paul, M.Sc, PMI; *ICT Engineer*, WatifHealth

#### Subtopic Drivers' Forum

At minimum, the *Subtopic Drivers' Forum* shall consist of at least 2 persons, consisting of the above-referenced 2 members of the Subtopic Driver Team. The preliminary objectives of the Subtopic Drivers' Forum are as follows:

* Determination of adoption/non-adoption of input to TG-Cardio from topic group participants, fora and stakeholder subgroups. Such determinations will be done by way of consultation and consensus (or vote) of the steering committee of the Subtopic Drivers' Forum consisting of the 3 members listed below (1 vote/member of steering committee; with the Topic Driver Team Chair casting a tie-breaker vote if consensus not achieved).

The *Subtopic Drivers' Forum* is only open to eligible TG-Cardio participants as referenced above, and membership of this forum is thus constituted as follows:

* *Subtopic Drivers' Forum* Chair: Dr. Benjamin Muthambi
* Subtopic Co-Investigator(s) representative: Dr. Nao Sipula
* TG-Cardio | CVD Risk Prediction Subtopic Stakeholder Community/Consultative Forum representative (each given month’s Rotating Chair of the Stakeholder Community/Consultative Forum): To be Determined

#### TG-Cardio | CVD Risk Prediction using AI - Subtopic Stakeholder Community/Consultative Forum (Benchmarking, Peer-Review & Technical Advisory/Consultative Participants):

This forum is open to all participants and stakeholders in TG-Cardio with an interest in the proceedings of the subtopic CVD risk prediction. The preliminary objectives of the Subtopic’s Stakeholder Community are manifold:

* To provide a forum for open communication among various stakeholders
* To provide peer-review and advisory support to the benchmarking process set out in the TG-Cardio TDD in collaboration with the Subtopic Driver Team (Topic Group Convenor & Secretariat), Focus Group management and working groups
* To establish technical advisory/consultative subgroups responsible for peer-review and refinement of:
* The objectives section of the CVD risk prediction subtopic TDD and IRB protocol for this subtopic’s project incl. replication studies;
* The literature review, background and rationale section of the TDD and IRB protocol for this subtopic’s project incl. replication studies;
* Sub-sections of the Methods section of the TDD and IRB protocol for this subtopic’s project: Epidemiology/Evaluation Study Design, Undisclosed Data Management incl. Procurement and Preparation Procedures; Statistical Methods, Metrics, & Programming Tools; and Benchmarking Procedures and Infrastructure:
* Agree on benchmarking tasks of this topic and scoring metrics
* Facilitate the collection of high-quality labelled test data from different sources,
* Clarify the input and output format of the test data
* Define and set-up the technical benchmarking infrastructure
* R Programming: Data Management, Predictive Analytics, Statistical Analyses, Shiny Web App
* Ethics section of this subtopic’s TDD and IRB protocol for the project (Pre-IRB) for this subtopic & quality assurance of each replication project's proceedings
* Project management planning & infrastructure, reproducible reporting tools & writing templates in R Markdown
* Requirements for undisclosed data contributions for use in replication studies for ITU/WHO external benchmarking
* Final version of TDD document

TG-Cardio Stakeholder Community Forum is accepting expressions of interest in each of the above-referenced specific roles/tasks (see below under sub-section 2.3. Topic Group Participation). Membership of this forum will be posted [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/cardio.aspx) in future.

## Common Projects on Subtopic on CVD Risk Prediction using AI incl. Replication Studies

To facilitate efficiency of focus group proceedings and help limit the number of similar use-case specific inputs at FG-AI4H meetings, TG-Cardio attempts to bring together benchmarking use cases which share similar objectives and methods including metrics for evaluating accuracy of AI methods applied in comparable contexts. Those use cases that share objectives and methods will be subject to the same benchmarking as described in this TDD as replication studies. However, in some cases, there may be CVD risk prediction of different end-points along the continuum of CVD management which may be best handled as sub-objectives under the same subtopic of CVD risk prediction also subject to this TDD’s methodological approach in general.

## Participation in Subtopic Group on CVD Risk Prediction using AI and "Call for Participation"

Participation in both the Focus and Topic/Subtopic Group is generally open and free of charge. Anyone who is from a member country of the ITU may participate. On 14 March 2019, ITU published an official "call for participation" document outlining the process for joining the Focus Group and each Topic/Subtopic Group. For this subtopic group (TG-Cardio | CVD risk prediction), the call for topic group participation is [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/cardio.aspx) .

A number of potential participants have already submitted expressions of interest in participation in TG-Cardio, including experts in Cardiovascular Disease risk prediction and related AI/Machine Learning, etc. Updated information on joining, accessing the collaboration site, and participating in tasks of the Subtopic Group has now been clarified in clearer steps in the call for participation issued in advance of the FG-AI4H meeting in Brasilia, Brazil planned for January 2020.

## Status of the TDD & Subtopic Group on CVD Risk Prediction using AI:

**TDD Status**: The status of successive drafts of the TDD will be outlined at each FG-AI4H meeting.

**Subtopic Group Status**: With the initial publication of the TDD and a "call for participation", the current Subtopic group members, WatifHealth, Inc. and IEPH started to share it within their respective networks of field experts and potential participants. Some already declared interest in general or technical aspects and are expected to identify specific roles/tasks towards a) peer review and refinement of sections of the TDD which match their interest and technical expertise, b) data-contributions for replication studies, or c) demonstration projects. Before the initial submission of the first and current draft of this TDD it has been jointly edited by the convening Subtopic Driver Team. Some of the approached experts started working on own contributions that will soon be added to the document. Further TDD refinement work is expected to be undertaken asynchronously through the online collaboration site which can be accessed as clarified in the latest version of the “call for participants” in the CVD Risk Prediction Subtopic Stakeholder Community/Consultant Forum.

## Focus Group & Topic Group Meetings & Collaboration

* **Focus Group Meetings (Online and On-Site)**: The Focus Groups meets about every two months online and on site (at changing locations). An up-to-date list can be found at the official ITU FG AI4H [website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).
* **CVD Risk Prediction Subtopic Group Meetings & Collaboration (Online):**
* Subtopic Group e-Meetings will be conducted through asynchronous/synchronous online communications per announcements to be posted on the [web page](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/cardio.aspx) of the TG-Cardio Subtopic on CVD Risk Prediction and subtopic list;
* Information on how to access the online collaboration forum site has been clarified in the version (updated January 2020) of the “call for participants” posted on the above-referenced web page of the CVD Risk Prediction Subtopic Stakeholder Community/Consultant Forum.

# Methods for CVD Risk Prediction

While varying by country or region of the world, prevailing standards-of-care include use of a variety of clinical CVD risk scoring tools/calculators (WHO, 2019) which 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, ~50% of myocardial infarctions (MIs) and strokes occur among persons predicted to be at risk of CVD ([Ridker et al, 2008](https://www.ncbi.nlm.nih.gov/pubmed/18997196/%22%20%5Ct%20%22_blank)). Highlighting the need for standardization, prior systematic CVD risk prediction accuracy studies often used: disparate study populations; either traditional multivariate statistical methods or disparate AI algorithms, specifically 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 [ relevant studies: [Narain et al, 2016](https://doi.org/10.2147/PPA.S108203) (FHS-USA); [Fox et al, 2016](https://doi.org/10.1001/jamacardio.2015.0300) (JHS-USA); [Ambale-Venkatesh et al, 2017](https://doi.org/10.1161/CIRCRESAHA.117.311312) (MESA-USA); [Weng et al, 2017](https://dx.doi.org/10.1371/journal.pone.0174944) (NHS-CPRD-UK); [Unnikrishnan et al, 2016](https://dx.doi.org/10.1155/2016/3016245) (BMES-AUS); & related methodology: [Rahimian et al, 2018](https://dx.doi.org/10.1371/journal.pmed.1002695) ; [Luo et al, 2016](https://dx.doi.org/10.2196/jmir.5870); [Bal et al, 2014](https://dx.doi.org/10.1155/2014/137896) ]. It is hypothesized that certain ML algorithms can improve CVD predictive accuracy over CVD risk scoring tools/calculators currently used in the standard of practice, and also do so across diverse populations. If demonstrated, ML-assisted clinical Decision Support Systems (DSS) should be considered for use as the standard of practice in CVD risk prediction. The methods section of the Topic Description Document on this subtopic seeks to provide a standards-setting blueprint for evaluation/benchmarking of accuracy performance in CVD risk prediction using AI/ML-assisted clinical DSS (vs. various CVD risk scoring tools/calculators used in prevailing standards-of-care).

## AI *input data* structure

1. **Data sources/study populations**: Anonymized US & UK data sources for this project include 3 - 4 different retrospective cohorts of type 2 (adult onset) diabetic patients in recent CVD risk prediction studies with sufficient sample size which contain routine clinical data, used traditional multivariate statistical methods or ML algorithms, and researcher-use data are readily-obtainable under well-established NIH-funded research [data sharing terms](https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/nhlbi-policy-for-data-sharing-from-clinical-trials-and-epidemiological-studies) applicable to the [NIH/NHLBI's Open BioLINCC Biologic Specimen and Data Repository](https://biolincc.nhlbi.nih.gov/media/guidelines/handbook.pdf?link_time=2019-01-13_01:34:37.368596#1.2%20BioLINCC) & [UK NHS/CPRD](https://www.cprd.com/) data sharing policies. NIH/NHLBI's BioLINCC provenance information on these data was summarized by [Giffen et al, 2015](https://dx.doi.org/10.1089/bio.2014.0050) . Abstracts, dictionaries & quality information on data needed are publicly-accessible for the Framingham Heart Study ( [FHS Original](https://biolincc.nhlbi.nih.gov/studies/framcohort/); [FHS Offspring](https://biolincc.nhlbi.nih.gov/studies/framoffspring/) cohorts; & [FHS-Academic Use Dataset](http://courses.washington.edu/b513/datasets/datasets.php?class=513) ); [Jackson Heart Study (JHS)](https://biolincc.nhlbi.nih.gov/studies/jhs); [Multi-Ethnic Study of Atherosclerosis (MESA)](https://biolincc.nhlbi.nih.gov/studies/mesa/); [Action to Control Cardiovascular Risk in Diabetes (ACCORD)](https://biolincc.nhlbi.nih.gov/studies/accord/) , [Cardiovascular Health Study (CHS)](https://biolincc.nhlbi.nih.gov/studies/chs/), & the [UK-NHS Study](https://doi.org/10.1371/journal.pone.0174944) ([CPRD data repository](https://www.cprd.com/)), etc. Across selected cohorts, CVD risk predictors identified by each method will be pooled and considered for models to be compared for accuracy. Similarly, standard criteria will be set for the endpoint of first CVD event over a 5-10-yr follow-up period.
2. **Study design**: The envisaged epidemiologic study design will assemble retrospective cohorts in each data source study with 10-year follow-back to identify pre-CVD type-2 diabetes patients, allowing sufficient follow-up time for occurrence of diagnoses of CVD or censored follow-up at the 5- and 10-year time points.
3. **Predictors & main outcome conditions**: Systematic reviews show the 7 core risk factors taken into account among categories of predictors mainly used in clinical CVD risk scoring tools/prediction calculators, namely demographics (such as sex, age, race), physical examination (incl. BMI), systolic blood pressure, lipid levels & other blood variables, comorbidities (incl. history of diabetes), lifestyle (incl. smoking status), antihypertensive treatment, family history, and genetics [Dahagam et al, 2016; Alaa et al, 2019]. Lipid lowering agents, such as statins have not been historically included among predictors in these CVD risk predictions. Beyond the 7 core risk factors widely used, risk predictions using ML algorithms entail computational complexity arising from exponentially increasing the number of predictor variables to more than 400 [Alaa et al, 2019]. To distinguish the CVD risk prediction accuracy gain derived from using ML risk prediction algorithms from that derived from just using more variables, the more complex ML risk prediction using more variables can be compared to a simpler ML risk prediction using the same 7 core predictors typically used by CVD risk scoring tools/prediction calculators.
The main outcome conditions of interest are first fatal or non-fatal CVD events, defined by any of these ICD-10 diagnoses codes: I20-I25 (coronary/ischaemic heart disease), I50 (heart failure events, including acute and chronic systolic heart failure), I60-I69 (cerebrovascular disease), and F01 (vascular dementia), or any of these ICD-9 codes: 410-414 (ischemic heart disease), 436-438, and 430-434 (cerebrovascular disease).

### Data availability

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](http://courses.washington.edu/b513/datasets/datasets.php?class=513) which were in turn sourced from the BioLINCC data repository [under NIH data sharing terms](https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/nhlbi-policy-for-data-sharing-from-clinical-trials-and-epidemiological-studies). Additional datasets to be sourced under the same terms from the NIH/NHLBI BioLINCC & UK NHS/CPRD data repositories for completion of this project include 3 - 4 different retrospective cohorts of type 2 (adult onset) diabetic patients in recent CVD risk prediction studies. These datasets 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. The afore-mentioned identified researcher-use data are readily-obtainable under well-established NIH-funded research [data sharing terms](https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/nhlbi-policy-for-data-sharing-from-clinical-trials-and-epidemiological-studies) applicable to the [NHLBI/National Heart, Lung, and Blood Institute's Open BioLINCC Biologic Specimen and Data Repository](https://biolincc.nhlbi.nih.gov/media/guidelines/handbook.pdf?link_time=2019-01-13_01:34:37.368596#1.2%20BioLINCC). Specific datasets obtainable with sufficient sample size include the Framingham Heart Study ( [FHS Original](https://biolincc.nhlbi.nih.gov/studies/framcohort/); [FHS Offspring](https://biolincc.nhlbi.nih.gov/studies/framoffspring/) cohorts; & [FHS-Academic Training Use Dataset](http://courses.washington.edu/b513/datasets/datasets.php?class=513) ); [Jackson Heart Study (JHS)](https://biolincc.nhlbi.nih.gov/studies/jhs); [Multi-Ethnic Study of Atherosclerosis (MESA)](https://biolincc.nhlbi.nih.gov/studies/mesa/); [Action to Control Cardiovascular Risk in Diabetes (ACCORD)](https://biolincc.nhlbi.nih.gov/studies/accord/) , [Cardiovascular Health Study (CHS)](https://biolincc.nhlbi.nih.gov/studies/chs/), & [UK-NHS Study](https://doi.org/10.1371/journal.pone.0174944) ([UK NHS data sharing terms & NHS/CPRD data repository](https://www.cprd.com/)), etc.

#### 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

CVD risk prediction models including AI models proposed are generally designed to generate results assigning predicted levels of risk stratifications/categories to each individual whose risk of an adverse CVD outcome is being predicted (viz. risk of deaths due to myocardial infarctions, MI). Risk strata/categories are primarily represented as gradations of progressive levels of severity of predicted risk of a given adverse CVD outcome, such as:

* Low risk
* Normal risk
* Intermediate risk
* Very high risk

In a validation study demonstrating accuracy of a CVD risk prediction AI model (Narain et al, 2016), the proposed model was applied to a retrospective sub-cohort of actually observed CVD deaths due to myocardial infarction and coronary heart disease in the Framingham Study, and an output of summary results for each group presents percentages of recorded death assigned to predicted risk levels classified into the above risk strata/categories. As an example, the table below from Narain et al (2016) presents the results output structure summarized into the above-referenced AI-predicted risk strata/categories:



Source: [Narain et al, 2016](https://doi.org/10.2147/PPA.S108203) .

## Test data labelling

Benchmarking of AI/ML-based CVD risk prediction algorithms invariably requires use of labelled test data.

### Protection of confidentiality in data labelling:

In all instances, high standards of human research protections will be upheld in data labelling to ensure privacy of personal health data derived from human subjects in accordance with standard IRB requirements. More specifically, as data to be used will be primarily derived from retrospective patient medical records and previously collected for research or routine clinical care purposes, standard protocols for anonymizing/deidentifying all the data used must be followed to effectively strip all personal identifier labels from these data.

### Quality of data annotation/labelling:

Data dictionaries provided with datasets accessible through the relevant NIH data repository include data annotation/labels submitted by contributing NIH-funded studies which meet NIH data submission quality standards. Similar standards are in place for data in the NHS/CPRD data repository. Using the submitted annotations/labels, specific data identified for this project have been successfully and widely reused in the past. Abstracts, dictionaries, labeling & quality information of earmarked data sources are provided in the study descriptions for the Framingham Heart Study ([FHS Original](https://biolincc.nhlbi.nih.gov/studies/framcohort/); [FHS Offspring](https://biolincc.nhlbi.nih.gov/studies/framoffspring/) cohorts; & [FHS-Academic Use Dataset](http://courses.washington.edu/b513/datasets/datasets.php?class=513) ); [Jackson Heart Study (JHS)](https://biolincc.nhlbi.nih.gov/studies/jhs); [Multi-Ethnic Study of Atherosclerosis (MESA)](https://biolincc.nhlbi.nih.gov/studies/mesa/); [Action to Control Cardiovascular Risk in Diabetes (ACCORD)](https://biolincc.nhlbi.nih.gov/studies/accord/) , [Cardiovascular Health Study (CHS)](https://biolincc.nhlbi.nih.gov/studies/chs/), & [UK-NHS Study](https://doi.org/10.1371/journal.pone.0174944) ([CPRD data repository](https://www.cprd.com/)), and other data to be acquired from diverse populations for replication studies.

### Recoding, creation and labelling of standardized datasets for algorithm training and reproducible/replicable analyses:

The initiative will use R Markdown to document provenance, recode, create and label standardized datasets sourced from diverse sources for reproducible/replicable analyses of various undisclosed data acquired for this project, external peer review, and validation following key published guidelines for repurposed data (such as those summarized by [Bache et al, 2013](https://doi.org/10.2218/ijdc.v8i2.262) ). This approach will also enable this project to also contribute R programs and data from this project's data repository to ITU-WHO AI-for-Health initiative to facilitate broader access and further re-use to address other health topics.

## Scores and metrics

Metrics of predictive accuracy: Observed first CVD will be used as a reference (benchmark) event across each retrospective cohort study population, for comparison of CVD risk prediction methods within each of the above-referenced three domains of risk prediction approaches. Identification of methods with better CVD risk prediction accuracy for first CVD event after 5-yr or 10-yr follow-up, entails comparing the following metrics:

1. Two simple metrics to be compared are:
2. accuracy of each risk prediction method (defined as the number of correctly predicted CVD cases divided by the total number of actually observed CVD diagnoses in each retrospective cohort used; multiplication by 100 gives a percentage, e.g. 95% accurate), and
3. the degree of agreement between each risk prediction method vs. observed CVD events in each retrospective cohort used (Kappa statistic);
4. Advanced metrics to be compared include area under the curve/AUC (area under the receiver operating characteristic, AUROC); sensitivity; specificity; positive predictive value (PPV); negative predictive value (NPV).

Further consideration will also be given to multidisciplinary guidelines for developing and reporting of ML predictive models in biomedical research (Bal et al, 2014; Luo et al, 2016).

## Undisclosed test data set collection for replication studies across diverse populations

'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. A diverse range of other data sources will also be used 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.

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