|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-G-006 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | New Delhi, 13-15 November 2019 |
| **DOCUMENT** |
| **Source:** | TG-Cardio topic driver |
| **Title:** | TDD update: TG-Cardio (Cardiovascular disease risk prediction) |
| **Purpose:** | Discussion |
| **Contact:** | Benjamin R.H. Muthambi, DrPH, MPH IEPH, Inc. Topic Driver | Email: brm5@caa.columbia.edu  |

|  |  |
| --- | --- |
| **Abstract:** | This document provides the Topic Driver Document update for Cardiovascular disease risk prediction. NOTE – harmonization / possible consolidation with the content proposed for the TG subtopic in [G-021](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-G-021.docx) [*Sub-topic Coronary CT*, ShinKun Technology, MIIT (China)] is still need. |

Table of Contents

[1 Introduction 3](#_Toc24451692)

[1.1 Topic description 3](#_Toc24451693)

[1.1.1 Topic objectives 3](#_Toc24451694)

[1.2 Ethical considerations 4](#_Toc24451695)

[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: 4](#_Toc24451696)

[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) 5](#_Toc24451697)

[1.2.3 TG Cardio Reserves the Rights to Terminate Association with Any Projects for Research Ethics Violations 6](#_Toc24451698)

[1.3 Existing AI solutions 6](#_Toc24451699)

[1.4 Existing work on benchmarking 7](#_Toc24451700)

[1.4.1 CVD risk prediction approaches to be compared & evaluated for predictive accuracy: 7](#_Toc24451701)

[2 AI4H Topic Group Cardio – TG Cardio 7](#_Toc24451702)

[2.1 Mandate of Topic Group: Primary topic group output & mandate of the topic group teams, fora & technical subgroups 7](#_Toc24451703)

[2.1.1 Primary Topic Group Output (Topic Description Document) and Document Development Process 7](#_Toc24451704)

[2.1.2 Mandate of the Topic Group Teams, Fora & Technical Subgroups 7](#_Toc24451705)

[2.2 Subtopics 9](#_Toc24451706)

[2.3 Topic Group Participation and "Call for Participation" 9](#_Toc24451707)

[2.4 Status of this TDD & Topic Group: 10](#_Toc24451708)

[2.5 Focus Group & Topic Group Meetings & Collaboration 10](#_Toc24451709)

[3 Methods 10](#_Toc24451710)

[3.1 AI input data structure 10](#_Toc24451711)

[3.1.1 Data availability 11](#_Toc24451712)

[3.2 AI Output data structure 11](#_Toc24451713)

[3.3 Test data labelling 12](#_Toc24451714)

[3.3.1 Protection of confidentiality in data labelling: 12](#_Toc24451715)

[3.3.2 Quality of data annotation/labelling: 12](#_Toc24451716)

[3.3.3 Recoding, creation and labelling of standardized datasets for algorithm training and reproducible/replicable analyses: 13](#_Toc24451717)

[3.4 Scores and metrics 13](#_Toc24451718)

[3.5 Undisclosed test data set collection 13](#_Toc24451719)

# Introduction

## Topic description

### 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:

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 ML algorithms referenced above (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 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

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 methods/procedures

The following are examples and not an exhaustive list of methods/procedures which may be employed in "research" subject to IRB review and approval:

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

#### IRB submission shall also be required when the following is applicable 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 Reserves the 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

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

### 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);

# AI4H Topic Group Cardio – TG Cardio

## Mandate of Topic Group: Primary topic group output & mandate of the topic group teams, fora & technical subgroups

### Primary Topic Group Output (Topic Description Document) and Document Development Process

* **Purpose**: The primary topic 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 topic.
* **TDD Development Process**: Further development of this TDD beyond this 1st draft will be done in a cooperative way whereby any topic group participant may submit suggested TDD changes to the Topic Driver Team Chair (Topic Group Convenor & Secretariat) at any time through a designated online collaboration portal (to be announced) providing for asynchronous submission of comments/suggested changes. The Topic Driver Team Chair (Topic 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 consisting of the Topic Driver Team Chair (Forum Chair) along with 1 leader/subtopic (1 vote per subtopic, but permitting the Topic Driver Team Chair to cast a tie-breaker vote if needed). Suggested TDD changes shall be submitted during a 30-day period following each FG-AI4H meeting to allow for 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 the Topic Group Teams, Fora & Technical Subgroups

i) Topic Driver Team (Topic Group Convenor & Secretariat):

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, 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 Topic 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 Topic Driver Team (Topic 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*, Watif Health

ii) Subtopic Drivers' Forum

The Subtopic Drivers' Forum consists of one leader/subtopic along with the Topic Driver Team Chair (Topic Group Convenor, also serving as Chair of this Forum). At minimum, this forum shall consist of 2 persons if there is at least 1 Subtopic, along with the Topic Driver Team Chair, as was the case at the time of establishment of TG-Cardio. 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) by Members of the Subtopic Drivers' Forum consisting of the Topic Driver Team Chair (Forum Chair) along with 1 leader/subtopic (1 vote per subtopic, but permitting the Topic Driver Team Chair to cast a tie-breaker vote if needed).

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 #1 Co-Investigator(s) representative: Dr. Nao Sipula
* Subtopic #2 Co-Investigator(s) representative: Dr. Nathan Guo

iii) TG-Cardio Stakeholder Community Forum (Topic Group & Benchmarking Peer-Review & Technical Advisory Forum:

This forum is open to all participants and stakeholders in TG-Cardio proceedings. The preliminary objectives of the is Topic Group 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 topic Driver Team (Topic Group Convenor & Secretariat), Focus Group management and working groups
* To establish technical advisory subgroups for:
* Conceptualization of Subtopics on CVD-related Predictive Analytics in Clinical & Public Health, & Applications thereof
* External Peer Review for Ethics (Pre-IRB) & Quality Assurance of Each Project's Proceedings
* Methods: 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
* Project Management Planning & Infrastructure, Reproducible Reporting Tools & Writing

TG-Cardio Stakeholder Community Forum is accepting expressions of interest (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.

## Subtopics

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 subtopics/ benchmarking use cases which share similar methods especially in terms of metrics for evaluating accuracy of AI methods applied in comparable contexts. However, in some cases, it is also conceivable that a Topic Group may include Subtopics which pursue different functional focus areas within a common broader health topic. For example, under TG-Cardio, the higher-level unifying subject matter is currently applications AI in various aspects of cardiovascular disease management, hence subtopics which have been grouped under TG Cardio may appear to be somewhat disparate usage of AI but may potentially share similar metrics for accuracy of AI algorithms used in different contexts, i.e.:

* **Subtopic #1**: The 1st subtopic focusses on accuracy of AI algorithms using data inputs from routinely measured clinical indicators to predict risk of adverse cardiovascular disease sequelae such as stroke [Subtopic Co-Investigator(s): Drs. Nao Sipula & Benjamin Muthambi]
* **Subtopic #2**: The 2nd subtopic focusses on accuracy of AI algorithms for image recognition using inputs of images from computerized tomography (CT) scans of the heart to detect coronary calcium deposits, an indicator of atherosclerosis, which is in turn predictive of risk of adverse cardiovascular disease sequelae such as stroke. [Subtopic Co-Investigator(s): Dr. Nathan Guo]

## Topic Group Participation and "Call for Participation"

The participation in both the focus and Topic 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 the each Topic Group. For this topic group (TG-Cardio), 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. Notices of invitation of comments/input according to a time-phased schedule of online asynchronous/synchronous events/meetings, and information on how to access a collaboration forum portal will be sent to TG-Cardio participants via email after the FG-AI4H meeting in New Delhi on 12-13 November 2019.

## Status of this TDD & Topic Group:

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

**Topic Group Status**: With the publication of this TDD incorporating a "call for participation" the current Topic Group members, Ada Health and Your.MD, started to share it within their networks of field experts. Some already declared general interest and are expected to join official via input documents at meeting D or E. Before the initial submission of the first draft of this TDD it was jointly edited by the current Topic Group members. Some of the approached experts started working on own contributions that will soon be added to the document. For the missing parts of the TDD where input is needed the Topic Group will reach out to field expects at the upcoming meetings and the in between.

## 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).
* **Topic Group Meetings & Collaboration (Online):**
* Topic Group e-Meetings will be conducted through asynchronous/synchronous online communications per announcements to be posted on the Topic Group TG-Cardio [web page](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/cardio.aspx).
* Online collaboration forum portal access information will be sent to topic group participants via email.

# Methods

## 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

'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.

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