|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-C-017-R1 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | N/A | Lausanne, 22-25 January 2019 |
| **DOCUMENT** |
| **Source:** | WatIF Health & IEPH/Institutes of Epidemiology & Public Health |
| **Title:** | AI for Health Use-Case: Demonstration of Applicability of Cardiovascular Risk Prediction using Machine Learning & Routine Patient Data Across Diverse Populations: A Standards-Setting Approach |
| **Purpose:** | Discussion |
| **Contact:** | Nao Norman SipulaWatIF Health, 21 Miles Sharp RynfieldBenoni, South Africa 1501 | Tel: +27-74-537-9141Fax: Email: nao@WatIFHealth.com  |
| **Contact:** | Jason Paul, MSc, PMIWatIFHealth, #7/1 Saraswarthpur Kalaghatgi RoadDharwad Karnataka, India 580002 | Tel: +91-998-640-6691Fax: Email: jason@WatIFHealth.com |
| **Contact:** | Benjamin R.H. Muthambi, DrPH, MPHIEPH, Inc. 6224 Spring Knoll Drive Harrisburg, PA 17111, USA | Tel: +1-717-745-6343Fax: Email: brm5@caa.columbia.edu |

**TITLE: DEMONSTRATION OF APPLICABILITY OF CARDIOVASCULAR RISK PREDICTION USING MACHINE LEARNING & ROUTINE PATIENT DATA ACROSS DIVERSE POPULATIONS: A STANDARDS-SETTING APPROACH**

**PROJECT CONTEXT**: This document describes the WatIFHealth-IEPH partnership’s use-case for incorporation of Artificial Intelligence/*AI-assisted clinical decision support system (DSS) modules* in Electronic Health Record (EHR) systems implementable in low-middle income countries (LMICs) and resource-constrained settings towards improvement of primary health care. Specifically, the use of machine learning (ML) to improve Cardiovascular Disease (CVD) risk prediction as current standards of practice rely on less accurate CVD risk scoring tools/calculators used to inform selection of appropriate clinical management strategies for type 2 diabetes. This is a revision of the New York submission towards ITU-WHO development of a regulatory framework for AI in health.

**1. OVERVIEW/ABSTRACT** (p. 1 – 2)**:**

**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) Use routine patient data already collected from diverse populations through various retrospective cohort studies of type 2/adult onset diabetics (readily-obtainable under [US NIH](https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/nhlbi-policy-for-data-sharing-from-clinical-trials-and-epidemiological-studies) & [UK NHS/CPRD](https://www.cprd.com/) data sharing policies) – to assess CVD risk prediction accuracy of various methods (incl. ML) by comparing: i) *observed occurrence* of first CVD event in each cohort/population to ii) *‘blinded predictions’* (prediction methods blinded to the observed CVD status) using several clinical CVD risk scoring tools/calculators, traditional multivariate statistical methods, ML algorithms (in collaboration with several co-investigators who recently undertook similar studies);
b) 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. Specifically, ensure that performance of ML algorithms to be identified is demonstrated in ways which also address key *aspirational considerations* towards building confidence and consensus on methodological standards for adoption of AI in Health for the risk prediction domain, including explainability, interpretability, reproducibility of predictions using different ML methods and across diverse settings/populations, ability to deal with outliers, and amenability to human oversight.

**BACKGROUND, SIGNIFICANCE & RATIONALE:** Cardiovascular disease (CVD) is the global leading cause of morbidity and mortality ([WHO, 2014](https://www.who.int/nmh/publications/ncd-status-report-2014/en/)). CVD accounts for > 2/3 of mortality among type 2 diabetes patients ([ADA, 2019](https://doi.org/10.2337/dc19-S010) ). 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, ~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/)). 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 [ 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) ]. 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.
**METHODS**:
**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.
**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 time of 5-10 years.
**CVD risk prediction approaches to be compared & evaluated for predictive accuracy**: Within 3 domains of risk prediction approaches, methods to be compared for CVD risk prediction accuracy across diverse populations are:
a) 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/));
b) Multivariate statistical risk prediction methods incl. Cox Proportional Hazards and Multiple Logistic Regression;
c) 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);
**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 – 10 yr follow-up, entails comparing the following metrics:
a) Two simple metrics to be compared are: i) *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 ii) the *degree of agreement* between each risk prediction method vs. observed CVD events in each retrospective cohort used (Kappa statistic);
b) 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).

**2. RELEVANCE:**

**-**Globally, CVD is the leading cause of morbidity and mortality ([WHO, 2014](https://www.who.int/nmh/publications/ncd-status-report-2014/en/)). CVD accounts for more than 2/3 of mortality among type 2 diabetes patients ( [ADA Cardiovascular Disease and Risk Management: Standards of Medical Care in Diabetes, 2019](https://doi.org/10.2337/dc19-S010) ). Timely identification of type 2 diabetics with higher risk of CVD enables early implementation of preventative strategies for CVD and progressive morbidity and mortality ([WHO, 2005](http://www.who.int/chp/chronic_disease_report/full_report.pdf); [WHF, 2015](https://www.world-heart-federation.org/resources/risk-factors/)) To achieve this disease management objective, co-factors etiologically linked to CVD ([WHF, 2015](https://www.world-heart-federation.org/resources/risk-factors/)) are used in various CVD risk assessment and scoring tools/calculators ([Siontis et al, 2012](https://www.bmj.com/content/344/bmj.e3318)) of low predictive accuracy, specifically low sensitivity and specificity ([Van Staa et al, 2014](https://doi.org/10.1371/journal.pone.0106455); ) which are implemented in current standards of practice. Consequently, refinement and improvement of prediction strategies is critical.

**3. IMPACT (ANTICIPATED CLINICAL AND PUBLIC HEALTH IMPACT):**

**Better targeted and more effective programs through improvement of accuracy of CVD risk prediction**:

By combining a vast number of weak, strong, and complex predictors, ML algorithms are more likely to improve CVD predictive accuracy. Improved predictive accuracy over the current status quo of less accurate CVD risk scoring tools/calculators which rely on routine demographic and clinical indicators will enable primary health care providers predict and offer prevention to more ‘true positive’ patients at risk for CVD, while reducing the number of ‘false positives’ currently treated unnecessarily. Better targeted interventions improve the timeliness and effectiveness of diabetes prevention and care programs.

**Devolution/cascading of chronic disease management to lower-skilled health care workers**:

To overcome the shortage of a health workforce in LMICs and resource-constrained settings, the use of ML in disease management is a potential health workforce multiplier by facilitating incorporation of community healthcare workers into mainstream clinical work, tasked with the responsibility of identifying and offering timely prevention interventions to a large numbers of patients at risk for CVD, thus freeing up-skilled nurses to oversee the management of a larger cohort of diabetes patients, and only refer those at high risk of complications to a physician. Broader application of this approach may well lead to evidence based, consistent early identification of people at risk of complications, with little or no variability of care, and the added benefit of early referral to a specialist physician with shorter referral delays.

**Improved efficiency &** **cost- savings in Type 2 Diabetes Management which ultimately reduces mortality:**

AI-assisted DSS has great potential to help reduce physician information processing needs, while delivering more accurate patient-specific DSS for clinical management of disease and contributing to the reduction of microvascular complications commonly associated with diseases such as diabetes. This could potentially allow increase clinical efficiency that saves thousands of people from such complications, with resultant cost saving due to a reduction in the burden of disease progression, expensive hospitalizations, and catastrophic health expenditure for people in LMICs.

**4. STATE OF THE ART & EXISTING/ONGOING WORK IN THE AREA OF THE PROJECT:**

**State of the Art & Current Standard of Practice in the Area of the Problem to be Addressed:**

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 particular, 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/)). This project hypothesizes that certain ML algorithms can improve CVD predictive accuracy over CVD risk scoring tools/calculators used in the standard of practice across diverse populations, and should thus be adopted as the standard of practice.

**Existing/Ongoing Work in the Area of the Problem to be Addressed:**

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 including assessments of 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. Notwithstanding, CVD risk prediction studies of accuracy of ML suggest substantial improvements of predictive accuracy are possible compared to the standard of practice [ [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)], and such improved predictive accuracy has also been demonstrated for other biomedical applications of ML ([Rahimian et al, 2018](https://dx.doi.org/10.1371/journal.pmed.1002695) ).

**5. FEASIBILITY: CONSIDERATION OF PROJECT IMPLEMENTATION & FEASIBILITY THEREOF:**

**Demographically Diverse Study Populations**:

The project will be implemented across 3 - 4 different retrospective cohorts of type 2 (adult onset) diabetic patients studied in recent CVD risk prediction studies which contain routine clinical data, used traditional multivariate statistical methods or ML algorithms. Researcher-use data on these populations are readily-obtainable upon request 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) . Similarly, UK NHS data is also readily-obtainable following [UK NHS/CPRD](https://www.cprd.com/) data sharing policies.

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 observed endpoint of first CVD event over a 10-yr follow-up period.

**Study Design Requirements**:

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 until occurrence of the endpoint of first CVD or censoring of follow-up time after 10 years.

**Key Resources Needed/Available:**

*Data & Computing Resources:* Datasets (researcher-use data readily-accessible); R Programming Software & use of publicly-accessible ML algorithms (No cost anticipated); Computing equipment (No cost anticipated for exploratory analyses; possible need for upgraded computing equipment);

*Human Resources*: Data Scientists, specifically Research Assistant, Epidemilogist/Evaluation Analyses Programmer, Computer Scientist for ML algorithm adjustment/optimization if needed (Funding potentially needed for Data Scientists & Research Assistant Time); EHR application developer for incorporation of computation code and selected ML algorithm into EHR (EHR application developer time can be funded by WatifHealth, possible funding need if special coding of interphase is required);

**Tentative Potential Project Implementation Milestones:**

*Study Approvals, Data Acquisition & Preparation:* IRB Application & Approvals; Acquisition of Data & Documentation; Optimization & Selection of Predictors & End-Point Data; Data Management including Recoding and Labeling (Months 1 – 4);

*Analyses:* R programming for exploratory/Preliminary Analyses using data in hand (Months 3 - 5); Final Analyses (Months 5 – 8);

*Reporting:* Summarizing of Result/Findings; Literature Review for Discussion of Findings; Report/Manuscript Writing; Internal Peer Review; External Peer Review & Publication of Findings (Months 8 – 10).

**6. 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.

**7. DATA QUALITY:**

Abstracts, dictionaries & quality information on data needed 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/)), etc.

**8. DATA ANNOTATION/LABEL QUALITY:**

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. Descriptions of each identified study including abstracts, data dictionaries, and quality information are provided through web pages hyperlinked under the “Data Quality” section above.

**9. DATA PROVENANCE:**

Anonymized US data sources for this project are all qualifying US NIH-funded studies which are required to submit data to the NIH/NHLBI BioLINCC data repository. Before data collection, all these studies are initially subject to ethics review and have been approved by an Institutional Review Board (IRB) which meets ‘[federal-wide assurance](https://www.hhs.gov/ohrp/register-irbs-and-obtain-fwas/fwas/fwa-protection-of-human-subjecct/index.html)’ requirements for human research protections. The same is true for UK data to be sourced from the NHS/CPRD data repository. *All readily-obtainable data identified for this project have been successfully used in research requiring routine clinical and other patient data used by CVD risk scoring tools\calculators in routine clinical care*. To ensure applicability in this project’s requirement for data from diverse populations in the evaluation of CVD predictive risk accuracy of ML algorithms and other methods to be compared, studies/data sources selected for this project were undertaken in demographically diverse populations.

Anonymized data already acquired for preliminary work to train ML algorithms 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). The project will follow [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) & [UK NHS/CPRD data-sharing policies](https://www.cprd.com/) to acquire additional selected data required for completion of this project. The provenance of NIH data from the time of collection by original study investigators to the point of deposition with the relevant NIH data repository is summarized by [Giffen et al, 2015](https://dx.doi.org/10.1089/bio.2014.0050) . In turn, this project will use R Markdown to document provenance, enable reproducibility and external peer review of data acquired for this project 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 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.

**10. BENCHMARKING:**

**CVD risk prediction approaches to be compared & evaluated for predictive accuracy**: Within 3 domains of risk prediction approaches, methods to be compared for CVD risk prediction accuracy across diverse populations include:
a) 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/));
b) Multivariate statistical risk prediction methods such as Cox Proportional Hazards and Multiple Logistic Regression;
c) 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);

**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:
a) Two simple metrics to be compared are: i) *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 ii) the *degree of agreement* between each risk prediction method vs. observed CVD events in each retrospective cohort used (Kappa statistic);
b) 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](https://dx.doi.org/10.1155/2014/137896); [Luo et al, 2016](https://dx.doi.org/10.2196/jmir.5870)).

**11. ORGANIZERS OF PROJECT:**

**Motivation, Ongoing Work, & Future Plans**: The WatIFHealth portal has been a self-funded, designed and developed EHR over a period of ten uninterrupted years by the same team of ICT engineers and Sub-Saharan African doctors experienced in the management of NCD, HIV/AIDS and TB in resource constrained primary health care settings.

The WatifHealth-IEPH partnership’s interest and passion for this area of health care is informed by the many years of practical experience and first-hand exposure to the devastating impact of NCDs, HIV/AIDS and TB on communities we serve; and observation of many shortcomings faced by government lead programs, whose success depends on rapid adoption of enabling technology and ability and capacity to innovate for emerging global pandemics.

Building on the team’s motivation and experience with the complexities of Sub-Saharan health care challenges, the following outlines prior experience and planned work on development and implementation of the WatifHealth EHR portal demonstration project, and future work on AI-assisted DSS for improvement of disease risk prediction:

*Related Projects, Milestones Completed*:

1. Successful development of the WatifHealth EHR portal;
2. Development and incorporation into the EHR portal of add-on DSS modules with built-in health analytics dashboards for clinician use including CVD and other disease risk scoring algorithms;
3. Completion of a proof-of-concept implementation of the WatifHealth EHR portal in the Eastern Cape Province of South Africa;

*Work In-Progress;*

1. Successful scale-up of deployment and ongoing implementation of the WatifHealth EHR portal in over 60 community health clinics in Sub-Saharan Africa;
2. Development of the project proposal (in this document) for the purpose of improvement of currently used CVD risk prediction accuracy

*Future Related Work Planned;*

1. Demonstration of peer-reviewed improvement in CVD and other disease risk prediction accuracy using ML;
2. Real world demonstration of identified ML algorithms into a model EHR such as the WatifHealth portal which is LMIC-optimized and incorporates evidence-based AI-assisted DSS for CVD risk prediction to mitigate these challenges in LMICs, and
3. Demonstration of achievement of the above-referenced impact at scale, through:
4. Better targeted and more effective programs through improvement of accuracy of CVD risk prediction;
5. Devolution/cascading of chronic disease management to lower-skilled health care workers; and
6. Improved efficiency & cost- savings in Type 2 Diabetes Management which ultimately reduces morbidity and mortality;

**Project Team & Organizational Profile**: The WatifHealth & IEPH partnership on this project brings together no less than twelve masters level ICT engineers, doctoral-level Epidemiologists and Computer Scientists with interests in health analytics and artificial intelligence, primary health care physicians qualified in NCD, HIV/AIDS and TB management with many years of field work experience in this area. For this project, an advisory and co-investigator team is also in formation, consisting of faculty and research investigators who have recently undertaken and published studies on related topics, including disease risk prediction using ML.

**WatifHealth ICT International Recognition:** The WatIF Health EHR Portal has received [several international recognition awards](http://watifhealth.com/) such as the Frost and Sullivan best South African technology award for NCD, The Global Mobile Award nomination for top five best contribution to SDG among others.

**Corresponding Co-Principal Investigator & Roles:**

* Dr. N.N Sipula (MBChB) is a Co-Principal Investigator, physician and Medical Director responsible for conceptualization of the WatifHealth EHR portal, implementation of AI-assisted DSS, and clinical concept of the proposed use case. He is the Founder, CEO, and medical and clinical Systems Architect for incorporation of AI-assisted clinical DSS development at WatIF Health. He is the corresponding co-investigator on aspects of the proposed project related to his responsibilities in the projects.
* Mr. J. Paul is the Co-Principal investigator who is an experienced masters level ICT engineer and the project manager and leader of the ICT team responsible for software programming and development of the WatifHealth EHR, and software programming to incorporate therein, the AI-assisted clinical DSS to be developed by the project. He is the corresponding co-investigator on aspects of the proposed project related to his responsibilities in the projects.
* Dr. B. Muthambi (DrPH) is the Co-Principal Investigator & doctoral-level epidemiologist responsible for the conceptualization, design and conduct of the proposed use case study to evaluate use of ML to improve CVD risk prediction accuracy, establishing a team of study advisors and co-investigators (in formation), leading analyses & health analytics dashboard development displaying CVD risk prediction, and reporting of findings of the proposed project and submission thereof for peer-review. He is a Snr Scholar-in-Residence at the Institutes of Epidemiology & Public Health/IEPH, with more than 20 years’ experience as an applied public health epidemiologist. He is the corresponding co-investigator on aspects of the proposed project related to his responsibilities in the projects.

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