unite for diabetes

Predict Onset of Diabetes
Big Data System Design - Final Report

John Carneiro - johnmcarneiro@gmail.com
Jan Zaloudek - honzicekz@gmail.com
Janakiram Sundaraneedi - janakiram_sundaraneedi@student.uml.edu

Outline

- Introduction
- Motivations
- Related Work
- Proposed Approach
- Evaluation
- Timeline
- Experimental Results and Discussions
- Conclusions and Future Work
- References

Introduction

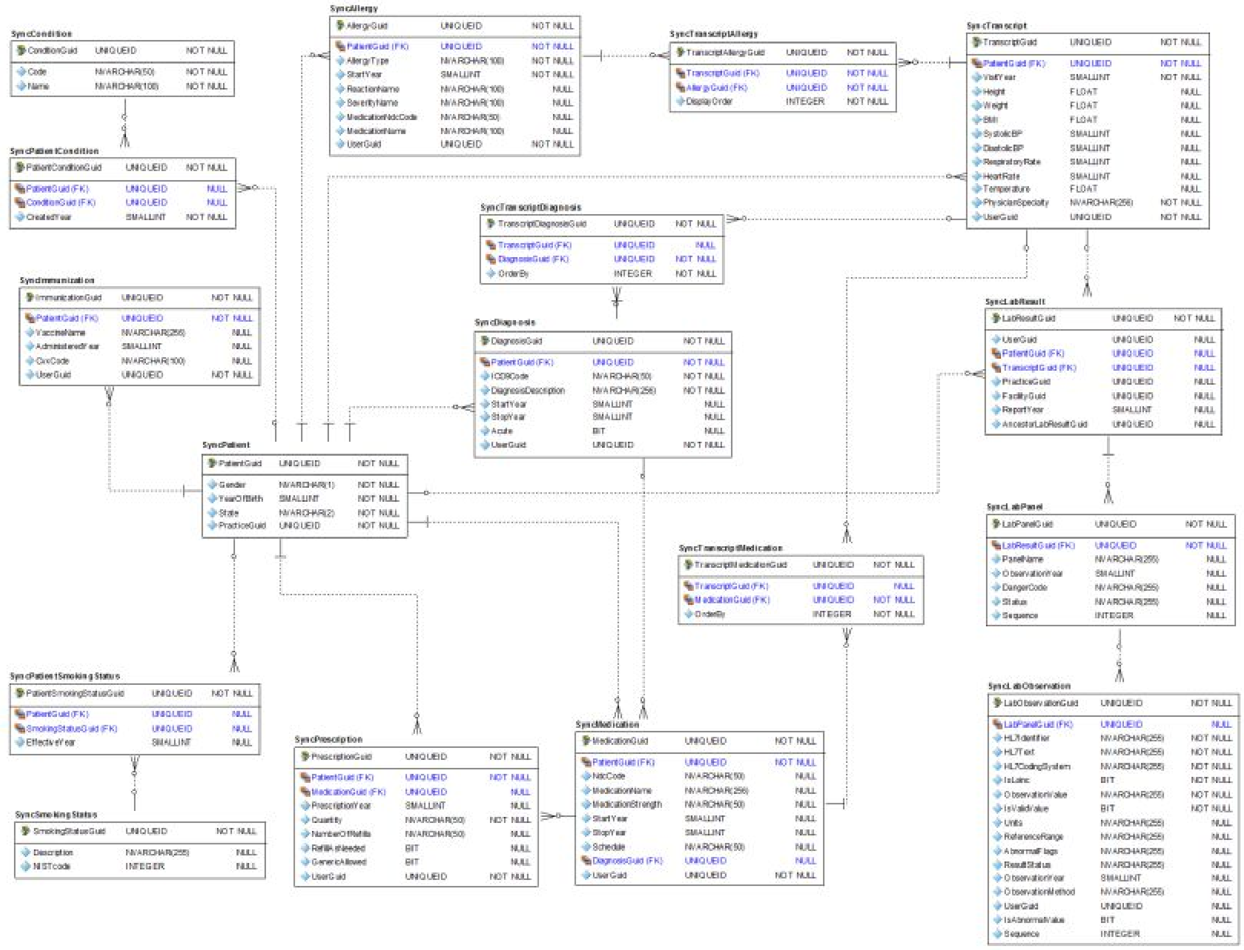
The research/development question is to predict based on diagnostic measurements whether a patient has type 2 diabetes using the Kaggle based Practice Fusion Diabetes Classification dataset.

The CSV text dataset contains records of 10,000 de-identified medical records. Data includes pregnancy count, glucose concentration, blood pressure, tricep skinfold thickness, insulin resistance, body mass index, diabetes family history factor, age, and diabetes positive.

The ultimate goal of your proposed system is to perform this prediction in an automated way using big data system design.

Hadoop big data ecosystem was used (MongoDB, Machine Learning, Processing techniques such as MapReduce, Keras, TensorFlow).

Practice Fusion De-Identified Data Set



SyncCondition

ConditionGuid	UNIQUEID	NOT NULL
Code	NVARCHAR(50)	NOT NULL
Name	NVARCHAR(100)	NOT NULL

SyncPatientCondition

PatientConditionGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NULL
ConditionGuid (FK)	UNIQUEID	NULL
CreatedYear	SMALLINT	NOT NULL

SyncImmunization

ImmunizationGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
VaccineName	NVARCHAR(255)	NULL
AdministeredYear	SMALLINT	NULL
CvxCode	NVARCHAR(100)	NULL
UserGuid	UNIQUEID	NOT NULL

SyncAllergy

AllergyGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
AllergyType	NVARCHAR(100)	NOT NULL
StartYear	SMALLINT	NOT NULL
ReactionName	NVARCHAR(100)	NULL
SeverityName	NVARCHAR(100)	NULL
MedicationNdcCode	NVARCHAR(50)	NULL
MedicationName	NVARCHAR(100)	NULL
UserGuid	UNIQUEID	NOT NULL

SyncTranscriptAllergy

TranscriptAllergyGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NOT NULL
AllergyGuid (FK)	UNIQUEID	NOT NULL
DisplayOrder	INTEGER	NOT NULL

SyncTranscript

TranscriptGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
VisitYear	SMALLINT	NOT NULL
Height	FLOAT	NULL
Weight	FLOAT	NULL
BMI	FLOAT	NULL
SystolicBP	SMALLINT	NULL
DiastolicBP	SMALLINT	NULL
RespiratoryRate	SMALLINT	NULL
HeartRate	SMALLINT	NULL
Temperature	FLOAT	NULL
PhysicianSpecialty	NVARCHAR(255)	NOT NULL
UserGuid	UNIQUEID	NOT NULL

SyncTranscriptDiagnosis

TranscriptDiagnosisGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NULL
DiagnosisGuid (FK)	UNIQUEID	NOT NULL
OrderBy	INTEGER	NOT NULL

SyncDiagnosis

DiagnosisGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
ICD9Code	NVARCHAR(50)	NOT NULL
DiagnosisDescription	NVARCHAR(255)	NOT NULL
StartYear	SMALLINT	NULL
StopYear	SMALLINT	NULL
Acute	BIT	NULL
UserGuid	UNIQUEID	NOT NULL

SyncLabResult

LabResultGuid	UNIQUEID	NOT NULL
UserGuid	UNIQUEID	NULL
PatientGuid (FK)	UNIQUEID	NULL
TranscriptGuid (FK)	UNIQUEID	NULL
PracticeGuid	UNIQUEID	NULL
FacilityGuid	UNIQUEID	NULL
ReportYear	SMALLINT	NULL
AncestorLabResultGuid	UNIQUEID	NULL

SyncPatient

PatientGuid	UNIQUEID	NOT NULL
Gender	NVARCHAR(1)	NOT NULL
YearOfBirth	SMALLINT	NOT NULL
State	NVARCHAR(2)	NOT NULL
PracticeGuid	UNIQUEID	NOT NULL

SyncLabPanel

LabPanelGuid	UNIQUEID	NOT NULL
LabResultGuid (FK)	UNIQUEID	NOT NULL
PanelName	NVARCHAR(255)	NULL
ObservationYear	SMALLINT	NULL
DangerCode	NVARCHAR(255)	NULL
Status	NVARCHAR(255)	NULL
Sequence	INTEGER	NULL

SyncTranscriptMedication

TranscriptMedicationGuid	UNIQUEID	NOT NULL
TranscriptGuid (FK)	UNIQUEID	NULL
MedicationGuid (FK)	UNIQUEID	NOT NULL
OrderBy	INTEGER	NOT NULL

SyncPatientSmokingStatus

PatientSmokingStatusGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NULL
SmokingStatusGuid (FK)	UNIQUEID	NULL
EffectiveYear	SMALLINT	NULL

SyncSmokingStatus

SmokingStatusGuid	UNIQUEID	NOT NULL
Description	NVARCHAR(255)	NULL
NISTcode	INTEGER	NULL

SyncPrescription

PrescriptionGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
MedicationGuid (FK)	UNIQUEID	NULL
PrescriptionYear	SMALLINT	NULL
Quantity	NVARCHAR(50)	NOT NULL
NumberOfRefills	NVARCHAR(50)	NULL
RefillAsNeeded	BIT	NULL
GenericAllowed	BIT	NULL
UserGuid	UNIQUEID	NOT NULL

SyncMedication

MedicationGuid	UNIQUEID	NOT NULL
PatientGuid (FK)	UNIQUEID	NOT NULL
NdcCode	NVARCHAR(50)	NULL
MedicationName	NVARCHAR(255)	NULL
MedicationStrength	NVARCHAR(50)	NULL
StartYear	SMALLINT	NULL
StopYear	SMALLINT	NULL
Schedule	NVARCHAR(50)	NULL
DiagnosisGuid (FK)	UNIQUEID	NULL
UserGuid	UNIQUEID	NOT NULL

SyncLabObservation

LabObservationGuid	UNIQUEID	NOT NULL
LabPanelGuid (FK)	UNIQUEID	NULL
HL7Identifier	NVARCHAR(255)	NOT NULL
HL7Text	NVARCHAR(255)	NOT NULL
HL7CodingSystem	NVARCHAR(255)	NOT NULL
IsLoinc	BIT	NOT NULL
ObservationValue	NVARCHAR(255)	NOT NULL
IsValidValue	BIT	NOT NULL
Units	NVARCHAR(255)	NULL
ReferenceRange	NVARCHAR(255)	NULL
AbnormalFlags	NVARCHAR(255)	NULL
ResultStatus	NVARCHAR(255)	NULL
ObservationYear	SMALLINT	NULL
ObservationMethod	NVARCHAR(255)	NULL
UserGuid	UNIQUEID	NULL
IsAbnormalValue	BIT	NULL
Sequence	INTEGER	NULL

Motivations

- This is interesting question for us due to a lot of our family members have diabetes
- It is important to develop a system to be able to predict the occurrence of diabetes if the patient’s lifestyle and eating habits are kept the same and possibly prevent the occurrence of diabetes if preventative steps are taken by patient
- 415 million people have diabetes worldwide
- 8.3% of the world adult population (equal parts men/women) have diabetes and is rising
- Diabetes doubles a person's risk of early death.
- 5 million deaths occur worldwide each year because of diabetes.
- The global economic cost of diabetes is estimated to be US $612 billion.

Related Work

The following are sources of existing research/development work has tried to answer the same or a similar question:

- https://www.kaggle.com/c/pf2012-diabetes (data set)
- Using the ADAP learning algorithm to forecast the onset of diabetes mellitus
- Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting

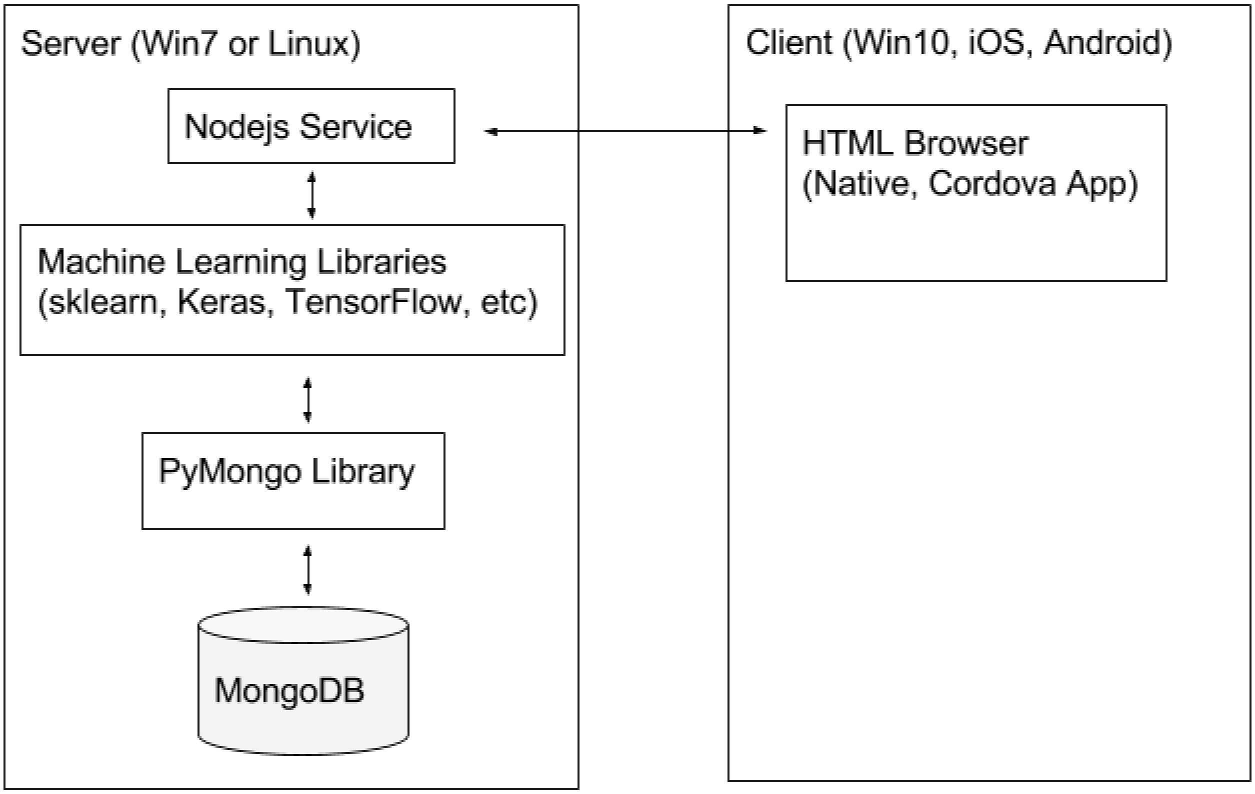
A highly reliable and accurate automated prediction algorithm is still not available.
Also, it would be self-evident that an accurate automated prediction algorithm should work regardless of gender and ethnic background.

Proposed Approach

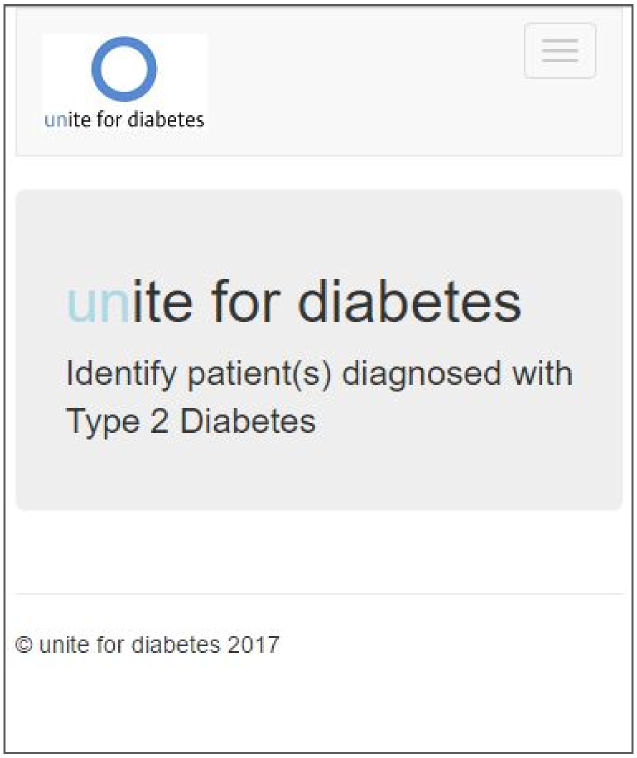
- Plan for working out the solutions to the question:
- Research dataset
  - Survey knowledge experts (UML doctor, nursing staff, and people with the diabetes)
  - Improve via algorithms accuracy and precision level on missing values
  - Research ML algorithms, current solutions on kaggle.com
  - Use rapidminer.com, and Python ML libraries
  - Performance test ML algorithms
- Main features in your proposed system:
  - Hadoop based backend (HDFS, MongoDB (i.e. NoSQL DB), Python, Apache Web Server)
  - HTML5, JS, Cordova mobile development front end (desktop, iOS, Android)
  - Data entry and/or upload data set
  - Matplotlib, plotly, Google Chart, D3, etc.
  - Based on input data, one or more users can be told whether or not he/she has or is susceptible to get type 2 diabetes (data can be optionally added to learning dataset to improve algorithm accuracy)
- Implementation approach:
  - MongoDB
  - Python, RapidMiner
  - HTML5 - JS
  - NN (Spark, Keras, TensorFlow), Scikit Learn, MapReduce
  - Machine Learning Algorithm(s) (Classification, Regression, Clustering, Dimensionality reduction, Model selection, Preprocessing) - those found exhibiting best performance
    - Keras based neural network
    - Gaussian Naive Bayes
    - K-Nearest Neighbor Classifier
    - Decision Tree Classifier

- Random Forest Classifier
- Ada Boost Classifier
- Gradient Boosting Classifier

Block Diagram



Mobile Application

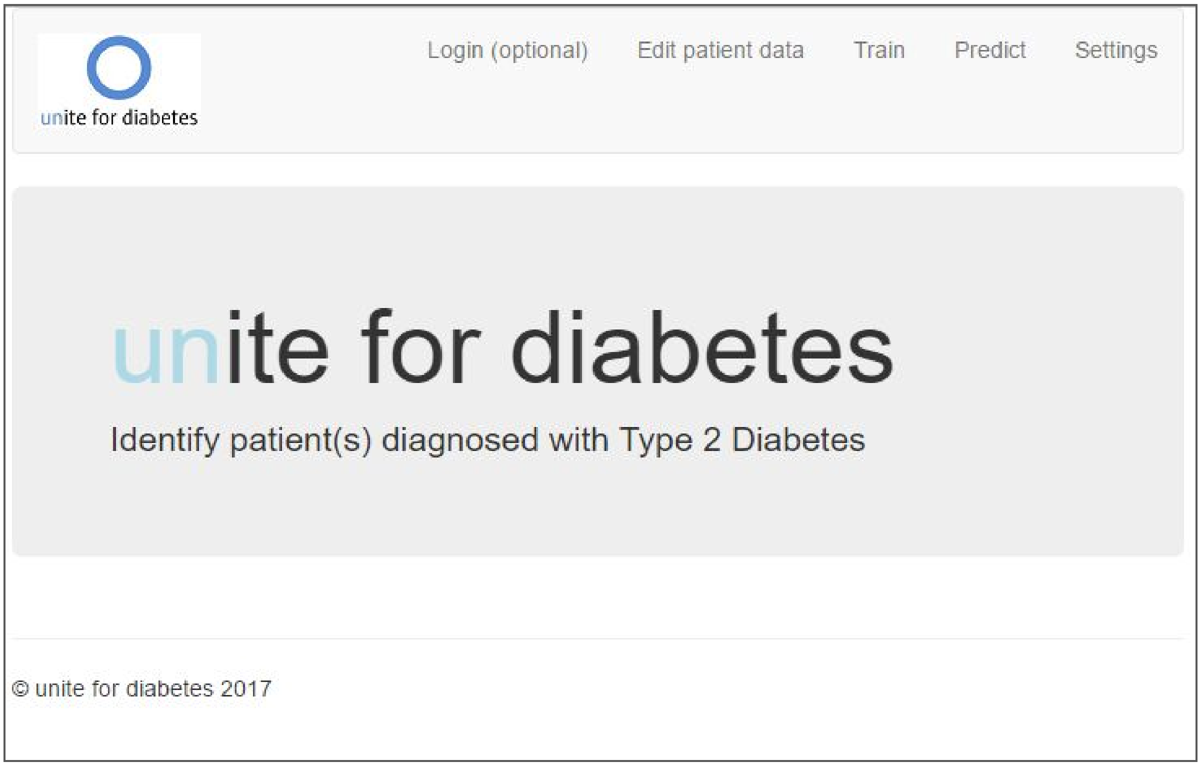


unite for diabetes

Identify patient(s) diagnosed with Type 2 Diabetes

© unite for diabetes 2017

Tablet/Desktop Application



unite for diabetes

Identify patient(s) diagnosed with Type 2 Diabetes

© unite for diabetes 2017

Evaluation

We evaluated our solution as follows (to demonstrate that our solution/answer is good/reasonable):

- Test ML algorithms locally with the stock dataset
- Preprocess dataset
- Find the algorithm that performs the best in a reasonable amount of processing time at a small scale first
- Setup a MongoDB environment to handle the increase in size
- Integrate MongoDB component
- Test the scaled up components at the backend server side
- Test the scaled up components at the client side
- Compare all results with those available online in previous studies
- Publish results on sites like github and kaggle.com to verify and respond to feedback

Timeline

- Week 1 (ending 2017-02-02) - Proposal presentation
- Week 2 (ending 2017-02-09) - Pre data set analysis and Github setup
- Week 3 (ending 2017-02-16) - Subject matter expert surveys
- Week 4 (ending 2017-02-23) - ML algorithm analysis
- Week 5 (ending 2017-03-02) - ML algorithm testing
- Week 6 (ending 2017-03-09) - ML algorithm testing
- Week 7 (ending 2017-03-16) - Scale up dataset
- Week 8 (ending 2017-03-23) - Hadoop ecosystem integration
- Week 9 (ending 2017-03-30) - Data plot development
- Week 10 (ending 2017-04-06) - Front end development
- Week 11 (ending 2017-04-13) - System performance testing
- Week 12 (ending 2017-04-20) - System performance testing
- Week 13 (ending 2017-04-27) - Final presentation

Experimental Results and Discussions

Discuss the research/development results and quantify the performance study

Conclusions and Future Work

Summarize your work, give your conclusions if possible, and point out the possible future work (how the work can be further improved/extended)

References

[1] Practice Fusion Diabetes Classification Challenge Background and Dataset
https://www.kaggle.com/c/pf2012-diabetes

[2] Deep Learning with Python by Jason Brownlee
https://machinelearningmastery.com/deep-learning-with-python

[3] Keras: Deep Learning library for Theano and TensorFlow
https://keras.io

[4] scikit-learn - Machine Learning in Python library
http://scikit-learn.org

[5] MongoDB - NoSQL database
https://www.mongodb.com

[6] PyMongo - MongoDB API
https://api.mongodb.com/python/current

[7] Cordova - HTML5 based Cross Platform Mobile App platform
https://cordova.apache.org

[8] American Diabetes Association. Economic costs of diabetes in the U.S. in 2012. Diabetes Care April 2013. 36(4):1033-1046
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3609540/

[9] Centers for Disease Control and Prevention: Diabetes successes and opportunities for population-based prevention and control. U.S. Department of Health and Human Services; 2011.
http://www.cdc.gov/chronicdisease/resources/publications/aag/ddt.htm
https://www.cdc.gov/diabetes/pdfs/newsroom/diabetes-materials-for-study-or-distribution-by-health-care-professionals.pdf

[10] Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, Marks JS. Prevalence of obesity, diabetes, and obersity-related health risk factors. JAMA. 2003; 289(1):76-79.
https://www.ncbi.nlm.nih.gov/pubmed/12503980

[11] National Diabetes Information Clearinghouse. Am I at risk for type 2 diabetes? Taking Steps to Lower Your Risk of Getting Diabetes. National Institute of Diabetes and Digestive and

Kidney Diseases. NIH Publication No 12-4805. June 2012.
https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-type-2-diabetes

[12] Haffner S, Lehto S, Ronnemaa T, Pyorala K. Mortality from Coronary Heart Disease in Subjects with Type 2 Diabetes and in Nondiabetes subjects with and without prior Myocardial Infarction. N Engl J Med. 1998; 339:229-234
http://www.nejm.org/doi/full/10.1056/NEJM199807233390404#t=article