PULSE

PARTICIPATORY URBAN LIVING FOR SUSTAINABLE ENVIRONMENTS

D5.1 List of state-of-the-art models and variables used by them

PULSE project

H2020 - 727816

UNIPD

October 2017

DOCUMENT INFO

0.1 AUTHORS

Author	Organization	e-mail
Martina Vettoretti	University of Padova	martina.vettoretti@dei.unipd.it
Andrea Facchinetti	University of Padova	andrea.facchinetti@dei.unipd.it
Yan Li	New York Academy of Medicine	yli@nyam.org
Jose Pagan	New York Academy of Medicine	jpagan@nyam.org

0.2 DOCUMENT KEYDATA

Key words	H2020 – 727816 – PULSE Deliverable 5.1	
Editor info	Name	Martina Vettoretti
	Organization	UNIPD
	e-mail	martina.vettoretti@dei.unipd.it

0.3 DOCUMENT HISTORY

Date	Version	Contributor	Change	Status
28/07/2017	1.0	UNIPD	First template	Draft
20/09/2017	1.1	UNIPD	Draft with partial contents	Draft
29/09/2017	1.2	UNIPD, NYAM	Draft with partial contents	Draft
19/10/2017	1.3	UNIPD	First complete draft	Draft
31/10/2017	1.4	UNIPD	Final revision	Final

0.4 DISTRIBUTION LIST

Date	Issue	Distribution list
28/07/2017	Circulate first template of the deliverable and assign sections to partners	Cecilia Vera Muñoz, Riccardo Bellazzi, NYAM, ASPB
20/09/2017	Circulate updated draft	NYAM
19/10/2017	Circulate first complete draft for internal revision	Cecilia Vera Muñoz, Riccardo Bellazzi, NYAM
31/10/2017	Final version	All Consortium and the European Commission

TABLE OF CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of this deliverable, entitled “List of state-of-the-art models and variables used by them”, is to report the activities of task 5.1 in the identification of state-of-the-art risk models for the prediction of type 2 diabetes and asthma onset and the selection of models to implement and/or recalibrate, as part of activities of next task 5.2, based on the datasets available in the Consortium.

In particular, the deliverable is structured in four sections. Section 1 is dedicated to an introduction about type 2 diabetes and asthma and the use of prediction models in the PULSE project to quantify the risk of developing such diseases. In section 2, prediction models for type 2 diabetes and asthma onset that were proposed in the literature are reviewed focusing on the methodologies used for model development, the variables included in the models as risk factors and model validation techniques and results. Then, in section 3, the datasets available in the Consortium for models implementation and development are described and the state-of-the-art models that can be implemented and/or recalibrated in these datasets are selected. Finally, in section 4, some considerations about dataset pre-processing for model implementation, including variable homogenization and training, test and validation datasets definition are reported. Dataset pre-processing and model implementation will be a part of task 5.2 activities and related deliverable D5.2.

1. INTRODUCTION

Type 2 diabetes (T2D) is a chronic metabolic disease characterized by persistently high blood glucose concentration due to reduced insulin production and/or insulin resistance. Diagnosis of T2D is performed by blood tests such as fasting plasma glucose, oral glucose tolerance test or glycated haemoglobin (HbA1C). Diabetes affects 387 million people worldwide of which about 90% suffers from T2D. As demonstrated by several studies, the development of T2D is related to a combination of genetic factors and lifestyle factors, such as being overweight, lack of physical activity, poor diet and stress. Starting from the evidence of these studies, prediction models were proposed in the literature in order to predict the risk of developing T2D based on identified risk factors [1]-[3]. In particular, some prediction models were used to define T2D risk scores, like the Finnish Diabetes Risk Score [13], which allows a simple and practical quantification of risk of developing T2D. Such prediction models and risk scores can be very useful for clinical decision support, disease surveillance and population health management.

Another common chronic disease is asthma, a long-term inflammatory disease of lung airways characterized by reversible airflow obstruction, bronchospasm and recurring symptoms, like wheezing, coughing, chest tightness and shortness of breath. Diagnosis of asthma is performed based on symptoms, response to therapy over time and spirometry. Currently, asthma affects about 358 million people worldwide. In particular, 9.4% of children and 8.2% of adults in the European Union are affected by asthma, with a slightly lower percentage of individual affected by the disease in the United States (8.4% of children and 7.6% of adults). According to the onset of the disease, two different types of asthma can be identified: childhood asthma, in which the disease appears during the childhood; and adult-onset asthma, in which the disease appears for the first time in adulthood. The two kinds of asthma are characterized by different pathogenesis. In particular, childhood asthma is mainly driven by genetic risk factors, while adult-onset asthma is more influenced by lifestyle and environmental risk factors. An early diagnosis of asthma is important to prevent under-treatment and improve asthma control and progression. For this purpose, many prediction models of the risk of childhood asthma were proposed in the literature [4], some of which were used to define risk scores like the Asthma Predictive Index [5]. Conversely, the prediction of adult-asthma onset is a topic that was less investigated in the literature, and in particular, less is known about the risk factors for adult-onset asthma. Risk factors for asthma exacerbations were also studied in the literature and prediction models to identify asthmatic subjects at risk of exacerbations were proposed [6]-[12].

One of the objective of PULSE is the development of advanced models for the detection of risks associated with the onset of T2D and asthma. Such models will be implemented in predictive analytics and decision support systems that will take advantage of integration of multiple data sources, including personal data and data provided by mobility sensors, air quality sensors, and satellites, to provide individuals and public health agencies (PHA) with information on T2D and asthma risk, particularly focusing on environmental and behavioural risk factors. In particular, four use cases of PULSE analytics and decision support systems were defined, i.e., asthma patient (1), pre-diabetic patient (2), citizen (3) and PHA (4), which were described in detail in deliverable D1.1. Both asthma and T2D onset prediction models can find useful application in the PULSE system for PHA, which will provide PHA officers (use case 4) with information on T2D and asthma risk distribution in the population, allowing them to make better informed decisions about initiatives to prevent these diseases. T2D risk models can also be applied in the

PULSE app to provide pre-diabetic subjects (use case 2) with a measure of their risk of developing T2D and encouraging them to make health lifestyle choices in order to minimize such risk. The PULSE app can also incorporate risk models of asthma exacerbations in order to provide patients with asthma (use case 1) with advices about good practice for asthma management and minimize such risk.

The development, test and validation of prediction models to assess risk of T2D onset and asthma adult-onset is object of PULSE WP5. In particular, in this WP, current state-of-the-art prediction models will be implemented (task 5.2) and, then, enhanced by addition of new variables suspected to influence the risk of T2D/asthma onset (task 5.3). The present deliverable concerns the identification of the state-of-the-art risk models for the prediction of T2D onset and asthma adult-onset that will be implemented within task 5.2.

2. STATE-OF-THE-ART MODELS FOR THE PREDICTION OF TYPE 2 DIABETES AND ASTHMA ONSET

In this section, we will present a description of the risk models for the prediction of T2D onset (subsection 2.1) and asthma adult-onset (subsection 2.2) that represent current state of the art. In particular, for each model we will discuss methodologies used for model development, variables included in the model for risk assessment and model validation.

2.1. RISK MODELS FOR THE PREDICTION OF TYPE 2 DIABETES ONSET

A wide literature about T2D onset risk factors and prediction models is present. A review of T2D onset prediction models can be found in Buijsse et al. [1], Noble et al. [2] and Cichosz et al. [3]. In work by Abbasi et al. [14] and Kengne et al. [15], the performances of existing T2D prediction models were compared in external validation cohorts. Most of existing T2D prediction models were developed by using logistic regression or the Cox proportional hazard model. Some of them are based only on non-clinical risk variables, e.g., age, gender, family history of diabetes, physical activity etc., while others include also clinical variables requiring laboratory tests, e.g., fasting plasma glucose, cholesterol level etc. Many prediction models were translated into simple risk scores of easy calculation in which each risk factor is assigned a certain number of points and the final risk score is obtained as the sum of the risk factor points. In the following subsections, we will present the main T2D prediction models and risk scores that were proposed in the literature in chronologic order, particularly focusing on the variables used by them, the method for model development and their validation.

2.1.1. THE MODELS BY STERN ET AL.

A first study to determine if multivariable models, accounting for multiple diabetes risk factors, are effective for identifying people at risk of developing T2D was performed by Stern et al. [16], which derived two T2D prediction models based on the multiple logistic regression. Such models were developed using the data of 1791 Mexican Americans and 1112 non-Hispanic whites gathered in the San Antonio Heart Study, who resulted non-diabetics at baseline. Seven to eight years after the baseline exam, these subjects underwent a follow-up examination at which incidence of T2D was assessed. Such

data were used to fit a multiple logistic regression model with incident T2D at follow-up as dependent variable and diabetes risk factors at baseline as independent variables. In particular, Stern et al. derived two models, i.e. a full model and a simpler clinical model, which were both tested with and without considering the plasma glucose measured 2-h post oral glucose tolerance test (OGTT) as independent variable. The independent variables included by the two models are listed in Table 1. In work by Stern et al. [16], model parameters estimates are shown only for the clinical model without 2h-OGTT, which are here reported in Table 2.

Table 1. Risk factors included as independent variables in the T2D prediction models by Stern et al.

Variable	Full model with 2h-OGTT	Full model without 2h-OGTT	Simple model with 2h-OGTT	Simple model without 2h-OGTT
Age	X	X	X	X
Sex	X	X	X	X
Systolic blood pressure	X	X	X	X
Diastolic blood pressure	X	X	-	-
Total cholesterol	X	X	-	-
LDL cholesterol	X	X	-	-
HDL cholesterol	X	X	X	X
Triglyceride level	X	X	-	-
BMI	X	X	X	X
Parental or sibling history of diabetes	X	X	X	X
Fasting glucose	X	X	X	X
2h post OGTT glucose	X	-	X	-

The models were internally validated and their discriminatory ability in predicting the 7.5-year incidence of T2D was compared by assessing the area under the receiver-operating characteristic curve (AUC). Including 2h post OGTT glucose level, the full and the clinical model performed similarly with AUC equal to 0.859 and 0.857, respectively. Removing the 2h post OGTT glucose level from the models, the models' performance only slightly deteriorated with AUC equal to 0.845 and 0.843 for the full and clinical models, respectively. These results demonstrated, for the first time, that a simple clinical model with readily

available clinical measurements is effective in predicting T2D onset without the necessity of time-consuming and expensive tests like the OGTT.

The clinical model by Stern et al. was implemented in several subsequent literature studies. In particular, in work by McNeely et al. [17], the model was externally validated as in its original formulation and after recalibration. Furthermore, in work by Stern et al [18], the model was extended to include the metabolic syndrome as an independent variable.

Table 2. Coefficients of the logistic regression clinical model by Stern et al. without 2h post OGTT plasma glucose.

Variable	Model coefficient
Age [years]	0.028
Sex [Female/Male]	0.661
Ethnicity [Mexican American/non-Hispanic white]	0.412
Fasting glucose [mg/dl]	0.079
Systolic blood pressure [mmHg]	0.018
HDL cholesterol [mg/dl]	-0.039
BMI [kg/m^2]	0.070
Family history of diabetes [Boolean]	0.481
Intercept	-13.415

2.1.2. FINDRISC

The Finnish Diabetes Risk Score (FINDRISC) is a risk assessment tool for T2D onset developed in Finland, which is based on easily available individual information that can be collected by questionnaire on medical history and health behaviour and a simple clinical examination without any laboratory tests.

The FINDRISC was developed in work by Lindström and Tuomilehto [13] by using the data of 4746 Finnish subjects, not on antidiabetic drug therapy, who underwent a baseline survey in 1987. Incidence of drug-treated diabetes in this sample was assessed over a 10-year follow-up period based on data collected in the Social Insurance Institution drug register. In particular, drug-treated diabetes was observed in 196 subjects during the follow-up period. These data were used to fit a logistic regression model with drug-treated diabetes during follow-up as the dependent variable and 7 known risk factors for diabetes onset, categorized as in Table 3 (second column), as independent variables. Based on the estimated β coefficients of the logistic regression, a risk score was assigned to each risk factors with the criterion reported in Table 4. The resulting score for each variable category is reported in Table 3 (third column). Finally, the FINDRISC was defined as the sum of the risk of each variable and, as such, varies from 0 to 20.

Table 3. FINDRISC: variables and related points

Variables	Values	Points
Age [years]	45-54	2
	55-64	3
BMI [kg/m^2]	25 to <30	1
	≥30	3
Waist circumference [cm]	Men: 94 to <102 Women: 80 to <88	3
	Men: ≥102 Women: ≥88	4
Use of blood pressure medication [Boolean]	Yes	2
History of high blood glucose [Boolean]	Yes	5
Physical activity [hours/week]	<4	2
No daily consumption of vegetables, fruits or berries [Boolean]	Yes	1

Table 4. Point assignment criterion in the FINDRISC

β coefficient	Points
0.01 – 0.2	1
0.21 – 0.8	2
0.81 – 1.2	3
1.21 – 2.2	4
>2.2	5

In work by Lindström and Tuomilehto [13], both internal and external validation of the FINDRISC were performed. In particular, the external validation involved 4615 not drug-treated subjects that underwent a baseline survey in 1992 and were observed over a follow-up of 5 years for incidence of drug-treated diabetes. Drug-treated diabetes was developed by 67 subjects during follow-up. The model presented good discriminatory ability with AUC equal to 0.87. The FINDRISC value of 9 was chosen as cut-off value to

classify subjects that will develop T2D from those who will not develop the disease. With this cut-off, the prediction model presented sensitivity of 0.81 and specificity of 0.76 in the external validation dataset.

Besides the full model reported in Table 3, Lindström and Tuomilehto [13] also proposed a concise model which contains all the independent variables of the full model except physical activity and fruit and vegetables consumption that did not result statistically significant association with T2D onset. In the internal validation, the discriminatory ability of the concise model (AUC=0.857) was only slightly lower than that of the full model (AUC=0.860).

The FINDRISC full model is widely used in the literature as a tool to determine risk of T2D onset [19]-[23]. In addition, the FINDRISC was used as a reference algorithm by many subsequent studies aiming at modelling risk of T2D onset (see for example Noto et al. [24]).

2.1.3. ARIC DIABETES RISK MODELS

The ARIC diabetes risk models are multivariable predictive models of T2D onset derived in work by Schmidt et al. [25] based on data of collected in the Atherosclerosis Risk in Communities (ARIC) study. In particular, the analysis was performed on 7915 subjects (of age 45-64 years) participating in the study that were free of diabetes at baseline and completed the follow-up examinations.

Based on these data, three multivariable logistic regression models predicting the 9-year incidence of T2D were derived: a simple risk model including variables that do not require blood test and two additional models including also fasting glucose and fasting glucose + lipids concentration. Model parameters were fitted in a training test, approximately equal to half of the entire dataset. In Table 5, Table 6 and Table 7, we report the variables included in the three models and related model coefficients.

Table 5. Logistic regression coefficients for variables in the simple ARIC diabetes risk model.

Variables	Model coefficients
Age [years]	0.0271
Black race [Boolean]	0.2295
Parental history of diabetes [Boolean]	0.5463
Systolic blood pressure [mmHg]	0.0161
Waist circumference [cm]	0.0412
Height [cm]	-0.0115
Intercept	-7.3359

Table 6. Logistic regression coefficients for variables in the ARIC diabetes risk model with fasting glucose.

Variables	Model coefficients
Age [years]	0.0168
Black race [Boolean]	0.4433
Parental history of diabetes [Boolean]	0.5088
Fasting glucose [mmol/l]	1.6445
Systolic blood pressure [mmHg]	0.0120
Waist circumference [cm]	0.0328
Height [cm]	-0.0261
Intercept	-12.2555

Table 7. Logistic regression coefficients for variables in the ARIC diabetes risk model with fasting glucose and lipids concentration.

Variables	Model coefficients
Age [years]	0.0173
Black race [Boolean]	0.4433
Parental history of diabetes [Boolean]	0.4981
Fasting glucose [mmol/l]	1.5849
Systolic blood pressure [mmHg]	0.0111
Waist circumference [cm]	0.0273
Height [cm]	-0.0326
HDL cholesterol [mmol/l]	-0.4718
Triglyceride [mmol/l]	0.2420
Intercept	-9.9808

The three models were validated in a test set obtained excluding training set data from the original dataset. The validation pointed out that the simple ARIC diabetes risk model presented discriminatory ability significantly lower than the models including fasting glucose and fasting glucose + lipids concentration, as measured by the AUC that was equal to 0.71, 0.78 and 0.80 for the three models, respectively.

External validation of the ARIC diabetes risk model including fasting glucose and fasting glucose + lipids concentration was performed with good results in work by Sun et al. [26] in the Taiwan population.

2.1.4. FRAMINGHAM DIABETES RISK SCORE

The Framingham diabetes risk score is a tool to predict the development of T2D in middle-aged adults based on personal information, which can be collected by a questionnaire, and simple clinical measurements requiring blood test. The Framingham diabetes risk score was developed in work by Wilson et al. [27] using the data collected in the mid-1990s in the Framingham Offspring study, i.e. a longitudinal study directed by the National Heart, Lung and Blood Institute committed to identify the common factors contributing to cardiovascular disease [28].

The baseline examination included self-reported information on medications and parental history of diabetes, a physical examination including measurements of blood pressure, height, weight and waist circumference, a fasting blood sample and 2-hour oral glucose tolerance test (OGTT). People presenting T2D at baseline (defined as use of oral hypoglycemic medications or insulin, fasting plasma glucose >126 mg/dl or post-OGTT plasma glucose >200 mg/dl) were excluded from the study resulting in a sample of 3140 subjects (53.9% female) with a mean age of 54 years. Participants were observed for an average follow-up of 7 years during which they were characterized as developing T2D if they started receiving oral hypoglycemic medication or insulin or they had a fasting plasma glucose level >126 mg/dl.

These data were used to fit a logistic regression model with onset of T2D diabetes during follow-up as dependent variable and diabetes risk factors as independent variables. Specific risk factors included in the model and their categories are reported in Table 8 (first and second column). Then, a point score system based on the logistic regression β-coefficients was defined to assess the 8-year risk of developing T2D. The point scores assigned to each risk factor as reported in the third column of Table 8. The Framingham diabetes risk score is obtained as the sum of the points assigned to the single risk factors and, as such, varies from 0 to 30. Wilson et al. [27] also provided guideline on how to convert the Framingham diabetes risk score into the percentage risk of developing T2D in 8 years (see Table 9).

Table 8. Framingham diabetes risk score: variables and related points

Model variables	Values	Points
BMI [kg/m^2]	25 to <30	2
	≥30	5
Fasting glucose level [mg/dl]	100-126	10
HDL-C level [mg/dl]	Men: <40 Women: <50	5
Parental history of diabetes mellitus [Boolean]	Yes	3
Triglyceride level [mg/dl]	>150	3
Hypertension [mmHg] or taking antihypertensive treatment [Boolean]	>130/85 or yes	2

Table 9. Conversion of the Framingham diabetes risk score to the percentage 8-year risk of T2D.

Points	8-year risk of T2D [%]
≤10	≤3
11	4
12	4
13	5
14	6
15	7
16	9
17	11
18	13
19	15
20	18
21	21
22	25
23	29
24	33
≥25	>35

In Wilson et al. [27], the Framingham diabetes risk score was tested with the same data used for model development, showing good discriminatory ability (AUC=0.85). In work by Nichols and Brown [32], the Framingham diabetes risk score was assessed in an independent dataset, including 20,644 adults aged 26-82 years of the Kaiser Permanente Northwest (KPNW), a health maintenance organization located in Portland, Oregon. In particular, Nichols and Brown showed that, when applied to the KPNW population, the Framingham diabetes risk score correctly estimates the relative risks, but significantly underestimates the incidence of T2D in this population. The problem was not present when Nichols and Brown applied the Framingham diabetes risk score to a subsample of the KPNW population presenting T2D incidence similar to the Framingham Offspring Study Cohort. This indicates that the Framingham diabetes risk score need to be recalibrated before using it in populations with different T2D incidence.

The need for model recalibration was also pointed out in work by Xu et al. [33], who applied the Framingham diabetes risk score to predict 4-year incident diabetes in older Chinese. In particular, Xu et al. showed that by model recalibration they were able to improve the discriminatory ability of the Framingham risk score in the Chinese population tested (AUC=0.740 for the score, AUC=0.779 for the recalibrated model).

A possible reason for which the Framingham diabetes risk score often require recalibration for its use in different populations is that the Framingham Offspring Study Cohort from which it was derived included 99% white people. Furthermore, the Framingham Offspring Study Cohort was composed of volunteers, who may have been healthier than nonvolunteers.

Despite the problem of recalibration, the Framingham diabetes risk score, as the FINDRISC model, is one of the most popular tools to predict risk of T2D onset and plays the role of reference method in several literature studies aiming at development of T2D prediction models (see for example Kahn et al. [34] and Noto et al. [24]).

Besides the main score, Wilson et al. [27] also proposed a personal model, including only age, sex, parental history of diabetes and BMI, and three complex clinical models, including the variables of the Framingham diabetes risk score and additional complex clinical variables like 2h post OGTT glucose level and HOMA indices of insulin resistance and β-cell function. Internal validation of these models showed that the personal model has lower discriminatory ability (AUC=0.724), while the complex clinical models have discriminatory ability comparable to the Framingham diabetes risk score (AUC=0.850-0.854).

2.1.5. GERMAN DIABETES RISK SCORE

The German diabetes risk score (GDRS) is a risk score developed by Schultze et al. [29] to predict the development of T2D based on noninvasive measurements like anthropometric, dietary and lifestyle risk factors. The risk score was derived based on the data collected in the European Prospective Cancer and Nutrition (EPIC)-Potsdam study, a longitudinal study including 27,548 men and women aged 40-65 years at baseline. In particular, the analysis was performed on a database including 9,729 men and 15,438 women who were free of diabetes at baseline and underwent an average of 7 years of follow-up. On these data, a Cox regression model of time to diabetes onset was fitted using diabetes risk factors as covariates. The estimated β-coefficients were used to assign a score value to each variable. The GDRS was then obtained as the sum of these scores by the following formula:

$$GDRS = 7.4 \cdot W - 2.4 \cdot H + 4.3 \cdot A + 46 \cdot HT + 49 \cdot RM - 9 \cdot WGB - 4 \cdot C - 20 \cdot MA - 2 \cdot PA + 24 \cdot FS + 64 \cdot CHS \,. \quad (1)$$

The description of variables in eq. (1) is provided in Table 10. The range of the GDRS goes from 118 to 983 points, with average score of 446 points.

Table 10. Description of variables in the GDRS of eq. (1).

Variable	Description
W	Waist circumference [cm]
H	Height [cm]
A	Age [years]
H	Hypertension [1=yes, 0=no]
RM	Red meat consumption [150 g/day]

Variable	Description
WGB	Consumption of whole-grain bread [50 g/day]
C	Coffee consumption [150 g/day]
MD	Moderate alcohol consumption [1=10-40 g/day, 0=otherwise]
PA	Physical activity [h/week]
FS	Former smoker (≥20 cigarettes/day) [1=yes, 0=no]
CHS	Current heavy smoker (≥20 cigarettes/day) [1=yes, 0=no]

In work by Schultze et al. [29], external validation of the GDRS was performed with data of the EPIC-Heidelberg study. In this dataset, the GDRS presented AUC=0.82. With cut-off of 550 points, the GDRS had sensitivity of 79.7% and specificity of 79.3% in predicting incident T2D.

Interestingly, the GDRS includes several lifestyle factors not used by previous risk scores. However, the study by Schultze et al. [29] did not provide evidence of how much such lifestyle factors can improve the performance of previous risk scores. This analysis was performed by Schwartz et al. [30] for the FINDRISC. Schwartz et al. [30] demonstrated that by adding lifestyle variables (smoking, alcohol consumption and physical activity) to the FINDRISC concise score, the AUC increased nonsignificantly from 0.795 to 0.805.

The GDRS was subsequently updated in work by Mühlenbruch et al. [31] by including family history of diabetes among predictive variables. The extended GDRS was validated on data of the Multinational MONItoring of trends and determinants in Cardiovascular diseases (MONICA)/ Cooperative Health Research in the Region of Augsburg (KORA) study. The equation for calculating the extended GDRS (eGDRS) is:

$$eGDRS = 7.6 \cdot W - 2.4 \cdot H + 5 \cdot A + 46 \cdot HT + 58 \cdot RM - 9 \cdot WGB - 4 \cdot C - 18 \cdot MA - 2 \cdot PA + 36 \cdot FS + 66 \cdot CHS + 56 \cdot OPD + 106 \cdot BPD + 48 \cdot SD \quad (2)$$

where OPD, BPD and SD are Boolean variables equal to 1 if one parent, both parents and at least one sibling have diabetes, respectively. Mühlenbruch et al. [31] demonstrated that including parent and sibling history of diabetes improves the discriminatory ability of the GDRS from AUC=0.848 to AUC=0.856.

2.1.6. RISK SCORING SYSTEMS BY KAHN ET AL.

Two risk scoring systems were developed by Kahn et al. [34] to identify adults at high risk of T2D by using longitudinal data from the ARIC Study: a basic risk score including anthropometric characteristics, sex, parental history of diabetes and other clinical variables that do not require a blood specimen; an enhanced risk score that additionally includes glucose concentration and other analytics commonly assessed in a fasting blood sample.

The ARIC Study included 15,792 white and black adults aged 45 to 64 at baseline (1987-1989) which were followed up for 14.9 years. Patients with prevalent diabetes at baseline were excluded from the analysis.

The remaining database was divided into two parts: 75% of the subjects were used for model development, while the remaining 25% was used for model validation. The effect of diabetes risk factors on incident T2D was modelled by Weibull proportional hazard regression models, in which continuous variables were categorized into quintiles to which simplified point scores were assigned. Model variables and related values and point scores are reported in Table 11 and Table 12 for the basic and enhanced model respectively. In particular, the basic risk score ranges from 0 to 100, while the enhanced risk score ranges from 0 to 99.

In the validation dataset, the ability of the two prediction models to estimate the 10-year incidence of T2D was assessed. The basic scoring system had AUC=0.71, with maximum sensitivity+specificity at a basic score of 38 (sensitivity=69%, specificity=64%). Better performance were achieved by the enhanced scoring system which presented AUC=0.79, with maximum sensitivity+specificity at enhanced score of 38 (sensitivity=74%, specificity=71%). In study by Kahn et al. [34], the enhanced scoring system outperformed the Framingham diabetes risk score that was tested in the validation dataset in its original formulation driving to AUC=0.76. This result was confirmed in work by Abbasi et al. [14], in which the methods were compared in the Dutch cohort of the European Prospective Investigation into Cancer and Nutrition cohort study. In this study the enhanced model presented AUC equal to 0.88, significantly higher than the Framingham diabetes risk score, for which AUC was equal to 0.81. Similarly, in work by Schmid et al. [35] the enhanced risk score by Kahn et al. achieved an AUC value higher than the Framingham diabetes risk score (0.899 vs 0.830).

However, it is important to note that since the two scoring systems by Kahn et al. [34] were derived using data collected in patient aged 45-64 either black or white, the performance of the scoring systems could not be satisfactory when applied to subjects from other age groups or racial groups.

Table 11. Kahn basic diabetes prediction model: variables and related points

Model variables	Values	Points
Mother with diabetes	Yes	13
Father with diabetes	Yes	8
Hypertension	Yes	11
Black race	Yes	6
Ever smoker	Yes	4
Waist circumference [cm]	Men: 90 to <95 Women: 81 to <81	10
Waist circumference [cm]	Men: 95 to <100 Women: 88 to <96	20
	Men: 100 to <106 Women: 96 to <105	26

Model variables	Values	Points
	Men: ≥106 Women: ≥105	35
Height [cm]	Men: <171 Women: <157	8
	Men: 171 to <175 Women: 157 to <161	6
	Men: 175 to <178 Women: 161 to <164	3
Resting pulse [beats/min]	Men: ≥68 Women: ≥70	5
Weight [kg]	Men: ≥86.4 Women: ≥72.7	5

Table 12. Kahn enhanced diabetes prediction model: variables and related points

Model variables	Values	Points
Mother with diabetes	Yes	8
Father with diabetes	Yes	6
Hypertension	Yes	3
Black race	Yes	6
Age	55-64	2
Never drinker or former drinker	Yes	2
Waist circumference [cm]	Men: 90 to <95 Women: 81 to <81	4
	Men: 95 to <100 Women: 88 to <96	10
	Men: 100 to <106 Women: 96 to <105	14
Waist circumference [cm]	Men: ≥106 Women: ≥105	20
Height [cm]	Men: <171	4

Model variables	Values	Points
	Women: <157	
	Men: 171 to <178 Women: 157 to <164	2
Resting pulse [beats/min]	Men: ≥68 Women: ≥70	2
Fasting plasma glucose [mg/dl]	95 to <100	7
	100 to <106	13
	≥106	30
Triglyceride level [mg/dl]	Men: 130 to <179 Women: 112 to <151	3
	Men: ≥179 Women: ≥151	7
HDL cholesterol [mg/dl]	Men: <40 Women: <53	5
Uric acid [mg/dl]	Men: ≥7.8 Women: ≥6.4	3

2.1.7. QDSCORE

The QDScore is a diabetes risk score derived to predict the 10-year risk of developing T2D in England and Wales. Interestingly, the score was developed in a large ethnically and socioeconomically diverse population and takes into account such differences by including ethnicity and a measure of social deprivation in the score. The QDScore does not include any laboratory of clinical measurements and, thus, it can be cost effectively implemented.

The QDScore was originally developed in work by Hippisley-Cox et al. [36] using the data collected from 355 general practices in England and Wales including 2,540,753.00 patients aged 25-79. The effect of diabetes risk factors on T2D onset was estimated by Cox proportional hazards model separately for men and women. Variables included in the model and their estimated hazard ratios are reported in Table 13. Note that fractional polynomial were included for age and BMI to model the non-linear risk relations with these variables. In addition, significant interactions between age and BMI, age and family history of diabetes, and age and smoking status were found and thus included in the model. The Townsend score in Table 13 is a measure of social deprivation accounting for unemployment, non-car ownership, non-home ownership and household overcrowding [37].

The QDScore was validated with data collected in 176 separate practices, including 1,232,832 patients. The AUC was equal to 0.853 in women, 0.834 in men. Hippisley-Cox et al. also compared the performance of the QDScore with those of analogous models obtained removing ethnicity, Townsend score, or both these variables from the score by using the Bayes information criterion. The analysis showed that the QDScore was more superior to the other models, demonstrating that both ethnicity and social deprivation are important factors to consider for risk prediction of the T2D onset.

The QDScore was implemented and externally validated also in work by Collins and Altman [38] and Kengne et al. [15].

Table 13. Variables in the QDScore and estimated hazard ratios of Cox model in women and men.

Variables	Hazard ratios in women	Hazard ratios in men
White	1	1
Indian	1.710	1.929
Pakistani	2.152	2.538
Bangladeshi	4.071	4.532
Other Asian	1.264	1.894
Black Caribbean	0.798	0.955
Black African	0.805	1.695
Chinese	1.961	1.414
Other	0.889	1.199
Women: $(Age/10)^{1/2}$ Men: log(age/10)	84.059	105.666
$(Age/10)^3$	0.995	0.996
Women: (BMI/10) Men: $(BMI/10)^2$	37.293	3.168
$(BMI/10)^3$	0.934	0.832
Townsend score [per 1 SD increase]	1.201	1.140
Family history of diabetes in first degree relative [Boolean]	2.358	2.725
Current smoker [Boolean]	1.268	1.249
Treated hypertension [Boolean]	1.787	1.711
Diagnosis of cardiovascular	1.458	1.500

Variables	Hazard ratios in women	Hazard ratios in men
disease [Boolean]		
Treatment with corticosteroids [Boolean]	1.412	1.259

2.1.8. DPORT

The Diabetes Population Risk Tool (DPoRT) is a population-based risk prediction tool developed by Rosella et al. [39] to predict T2D onset using national survey data. The DPoRT was derived using the data of the Ontario participants of the 1996/7 National Population Health Survey conducted by Statistics Canada. Such data included the records of 9177 male and 10618 female subjects free of diabetes at baseline who could be individually linked to a registry of physician-diagnosed diabetes. The data were used to fit a Weibull survival model separately for women and men. The variables included in the model and their estimated model coefficients are reported in Table 14 for women, Table 15 for men. Note that the DPoRT includes some social factors, like immigrant status and education level, as predictive variables of T2D onset.

In work by Rosella et al. [39], the ability of the DPoRT to predict the 9-year onset of T2D was validated in two validation cohorts: the Manitoba 1996/7 National Population Health Survey and the 2000/1 Canadian Community Health Survey. In these two external cohorts the DPoRT score showed good calibration and discriminatory ability, with AUC equal to 0.80 and 0.76 for women, 0.79 and 0.77 for men.

The DPoRT was externally validated also in work by Kengne et al. [15].

Table 14. DPoRT: variables and model coefficients for women.

Variable	Value	Model coefficient
Hypertension [Boolean]	Yes	-0.2865
Non-white ethnicity [Boolean]	Yes	-0.4309
Immigrant status [Boolean]	Yes	-0.2930
Education	Post-secondary or higher	0.2042
BMI [kg/m^2] & age [years]	BMI=23-24, age<45	-0.5432
	BMI=25-29, age<45	-0.8453
	BMI=30-34, age<45	-1.4104
	BMI≥35, age<45	-2.0483
	BMI missing, age<45	-1.1328
	BMI=23-24, age=45-64	0.0711

Variable	Value	Model coefficient
	BMI=25-29, age=45-64	-0.7011
	BMI=30-34, age=45-64	-1.4167
	BMI≥35, age=45-64	-2.2150
	BMI missing, age=45-64	-2.2695
BMI [kg/m^2] & age [years]	BMI<23, age≥65	-1.0823
	BMI=23-24, age≥65	-1.1419
	BMI=25-29, age≥65	-1.5999
	BMI=30-34, age≥65	-1.9254
	BMI≥35, age≥65	-2.1959
	BMI missing, age≥65	-1.8284
Intercept	-	10.5474

Table 15. DPoRT: variables and model coefficients for men.

Variable	Value	Model coefficient
Hypertension [Boolean]	Yes	-0.2624
Non-white ethnicity [Boolean]	Yes	-0.6316
Heart disease [Boolean]	Yes	-0.5355
Current smoker [Boolean]	Yes	-0.1765
Education	Post-secondary or higher	0.2344
BMI [kg/m^2] & age [years]	BMI=23-24, age<45	-1.2378
	BMI=25-29, age<45	-1.5490
	BMI=30-34, age<45	-2.5437
	BMI≥35, age<45	-3.4717
	BMI<23, age≥45	-1.9749
	BMI=23-24, age≥45	-2.4426
	BMI=25-29, age≥45	-2.8588
	BMI=30-34, age≥45	-3.3179
	BMI≥35, age≥45	-3.5857
Intercept	-	10.5971

2.1.9. AUSDRISK

AUSRISK is a diabetes risk score developed by Chen et al. [40] to predict the 5-year incidence of T2D in the Australian population. The AUSRISK was derived using the data collected in the Australian Diabetes, Obesity and Lifestyle study in 6060 subjects aged 25 years or older, free of diabetes at baseline. Similar to the work by Lindstrom et al. [13], a logistic regression model with incident T2D as dependent variable and diabetes risk factors as independent variables were fitted on these data. Then, the logistic regression parameters were converted to a simple risk score by assigning some points to each variable category, as reported in Table 16. The final AUSRISK is obtained as the sum of each variable points and can vary between 0 and 38.

Internal validation of the AUSRISK score showed good discriminatory ability of the score that presented AUC equal to 0.78. With cut-off of 12 points, the score was able to predict 5-year T2D onset with sensitivity equal to 74% and specificity equal to 67.7%. In work by Chen et al. [40], the AUSRISK score was also validated in two independent Australian cohorts: the Blue Mountains Eye Study and the North West Adelaide Health Study. In these cohorts, AUC was equal to 0.66 and 0.79, respectively.

Table 16. AUSRISK score: variables and related points.

Variable	Values	Points
Age [years]	35-44	2
	45-54	4
	55-64	6
	≥65	8
Gender	Male	3
Ethnicity	Aboriginal, Torres Strait Islander, Pacific Islander or Maori descendent	2
Country of birth	Asia, Middle East, North Africa, Southern Europe	2
Either parents or any brother or sister diagnosed with diabetes [Boolean]	Yes	3
History of high blood glucose [Boolean]	Yes	6
Antihypertensive medication [Boolean]	Yes	2
Currently smoking [Boolean]	Yes	2
Eat vegetables every day [Boolean]	No	1

Variable	Values	Points
Physical activity [hours/week]	<2.5	2
Waist circumference [cm]	Asian, Aboriginal or Torres Strait Islander: Men: 90-100; Women: 80-90. Others: Men: 102-110; Women: 88-100	4
	Asian, Aboriginal or Torres Strait Islander: Men: >100; Women: >90 Others: Men: >110; Women: >100	7

2.1.10. NYAM-DIABETES MODEL

The New York Academy of Medicine Diabetes Simulation (NYAM-DS) Model is an individual-based stochastic simulation model of diabetes disease progression [41][42]. The model consists of a series of health states representing the development and consequences of diabetes and related complications (e.g., neuropathy, nephropathy, retinopathy, cardiovascular disease). It also incorporates evidence-based equations that guide dynamic changes of individual-level biomarkers (e.g., body mass index, HbA1c) and health factors (e.g., smoking status, hypertension), as well as the impact of these factors on health outcomes. The model logic and embedded equations come from NYAM research as well as critical analysis of existing diabetes simulation models and risk calculators, such as the CDC-RTI Diabetes Cost-effectiveness Model, the Michigan Model for Diabetes, and the United Kingdom Prospective Diabetes Study (UKPDS) Risk Engine. The model is implemented by using advanced simulation software—AnyLogic 7.

The NYAM-DS Model provides a graphical user interface that can help policymakers 1) define a population of interest, 2) define system parameters, 3) visualize dynamic changes of health factors and outcomes, 4) predict health and cost outcomes for a user-defined intervention, 5) perform cost-effectiveness analysis, and 6) report simulation results in a graphical way. These capabilities enable policymakers to easily evaluate and compare different prevention or treatment strategies to inform complex policy questions. The NYAM-DS Model provides an innovative, cost-effective tool to predict diabetes risk and help design population specific interventions. It will help promote evidence-based decision making in health and health care organizations to improve population health. However, at difference of previous models that can be used on single individuals, the NYAM-DS model cannot be used to predict the risk of having future T2D in single individuals, but it can only be used to predict the proportion of subjects expected to develop T2D in a certain population.

The input variables for the model are the characteristics of the population including:

- Population Size
- Age Distribution (Mean, Standard Deviation, Minimum, Maximum)
- Female Fraction (%)
- Current Smoker Fraction (%)
- BMI (kg/m2)
- HbA1c (%)
- History of Hypertension (%)
- History of High Cholesterol (%)
- History of Type 2 Diabetes (%)
- History of Myocardial Infarction (%)
- History of Stroke (%)

The model predicts the cumulative numbers of diabetes onset and related complications:

- Type 2 Diabetes
- Blindness
- End-Stage Renal Disease
- Foot Amputation
- Myocardial Infarction
- Stroke
- Death due to Diabetes or Diabetic Complications

The model also predicts the following outcomes to conduct economic analysis related to diabetes interventions:

- Life Expectancy
- QALYs
- Discounted QALYs
- Medical Costs (also available for major cost categories if needed)
- Discounted Medical Costs (also available for major cost categories if needed)

2.2. RISK MODELS FOR THE PREDICTION OF ASTHMA ADULT-ONSET

Most of literature studies on asthma onset are focused on childhood asthma. Indeed, several prediction models of childhood asthma onset were proposed in the literature [43]-[46]. Conversely, as reported in recent reviews **Error! Reference source not found.**[48], the causes of asthma onset in adults have not been extensively investigated. A special case is that of occupational asthma, i.e., asthma caused by specific agents that are found only in the workplace of certain occupations, which accounts for 10-25% of overall adult-onset asthma.

Concerning nonwork-related adult-onset asthma, no prediction model or risk score to predict the onset of such condition is available in the literature. The risk factors for nonwork-related adult-onset asthma were

investigated in few studies. A review of evidences collected can be found in work by Jeebhay et al. [48] and Ilmarinen et al. [49]. Some studies, in particular, investigated the relationship between air pollution and asthma onset, despite this seems to be less important in adult-onset asthma than childhood-onset asthma. A review of these studies can be found in Jacquemin et al. **Error! Reference source not found.** and Le Moual et al. **Error! Reference source not found.**.

Despite a real prediction tool for the onset of asthma in adults was never proposed, some multivariable logistic regression models that assess the influence of individuals' risk factors on adult-onset asthma were derived, like the models by Thomsen et al. [51], Jamrozik et al. [52] and Antó et al. [53] described in subsections 2.2.1, 2.2.2 and 2.2.4, respectively.

Other studies aimed at assessing the influence of specific risk factors on adult-onset asthma taking into account other covariates in multivariable logistic regression models. Here, we report the models derived in two of these studies, in which also the covariates' contribution to the dependent variable was reported in the papers [54][55] (subsections 2.2.5 and 2.2.6).

Finally, in subsection 2.2.3 the asthma score is presented, which have been used in work by Sunyer et al. [56] to predict adult-onset asthma, despite this is not truly a score to predict asthma onset by rather a score to quantify the presence of asthma-like symptoms and thus to identify patients with probable undiagnosed asthma.

2.2.1. THE MODEL BY THOMSEN ET AL.

In work by Thomsen et al. [51], a study to establish the risk factors for the development of asthma in young adults was performed using the longitudinal data collected in The Danish Twin Registry for birth cohorts over the period 1953-1982. In particular, the data of 19,349 subjects with no history of asthma in 1994 who answered to the follow-up questionnaire in 2002 were selected for the analysis. The age at baseline of selected subjects was 12-41 years.

A logistic regression model was applied to investigate the association of possible risk factors at baseline with asthma onset at follow-up. The analysis was performed separately for subjects of age 12-19 years and 20-41 years. For subjects of age 12-19 years a significant age/sex interaction was found. The variables statistically significantly associated with asthma onset in this age group and the respective logistic regression odd ratios are reported in Table 17. For subjects aged 20-41, no significant age-sex interaction was found, but a statistically significant association of BMI with incident asthma was found. Variables and odd ratios of the logistic regression model for subjects of age 20-41 are reported in Table 18.

Thomsen et al. [51] also shown that subjects with a history of hay fever and/or eczema at baseline had up to 8-fold higher likelihood of developing asthma at follow-up compared to subjects who never had such diseases.

Note that the aim of work by Thomsen et al. [51] was to identify risk factors for asthma onset in young adults, not to develop a tool to predict asthma onset. Indeed, Thomsen et al. did not test the ability of the proposed logistic regression model to correctly predict future onset of asthma.

Table 17. Variables included in the logistic regression model by Thomsen et al. for subjects aged 12-19 years and corresponding odd ratios.

Variables	Values	Odd ratios
Age [years]	Male subjects	0.85
	Female subjects	0.99
Smoking habits	Current daily smoker	1.32
	Occasional smoker	0.96
	Former smoker	1.58
	Never smoker	1
Physical activity	Light	0.84
	Moderate	1
	Heavy	0.98

Table 18. Variables included in the logistic regression model by Thomsen et al. for subjects aged 20-41 years and corresponding odd ratios.

Variables	Values	Odd ratios
Sex	Male	1
	Female	1.49
BMI [kg/m^2]	Per unit	1.05
Smoking habits	Current daily smoker	1.15
	Occasional smoker	1.31
	Former smoker	1.10
	Never smoker	1
Physical activity	Light	1.09
	Moderate	1
	Heavy	1.14

2.2.2. THE MODEL BY JAMROZIK ET AL.

Another study to investigate risk factors related to new onset of asthma in adults was performed by Jamrozik et al. [52]. Based on the data of 1554 adults (average age at baseline of 45 years) who

participated in the Busselton Health Study in 1981 (baseline) and 1994-1995 (follow-up) and never reported asthma before 1981, Jamrozik et al. used the logistic regression analysis to identify baseline variables significantly associated with new asthma onset.

Among the baseline variables considered as possible risk factors there were age, sex, history of respiratory diseases, tobacco smoking, alcohol drinking, weight, height, spirometric measures of lung function and atopy assessed by skin-prick test. Jamrozik et al. also assessed if changes in baseline variables were associated with new asthma onset. The final multivariate logistic regression model including only components that were found statistically significantly associated with asthma incidence in a stepwise selection procedure is reported in Table 19. The respiratory count in Table 19 is defined as the number of respiratory problems ever experienced by the subjects at baseline including wheeze, bronchitis, pneumonia, pleurisy, sinusitis, rhinitis, shortness of breath and cough. FEV1 in Table 19 stands for Forced expiratory volume in the 1st second and is a measure of lung function in spirometry.

As for the study of Thomsen et al. [51], the aim of work by Jamrozik et al. [52] was to identify risk factors associated with asthma onset, not to develop a tool to predict asthma onset. Thus, Jamrozik et al. did not test the ability of the proposed logistic regression model to correctly predict future onset of asthma.

Table 19. Logistic regression model by Jamrozik et al.: variables and odd ratios.

Variables	Values	Odd ratios
Respiratory history count	0	1.0
	1	2.0
	2	6.7
	≥3	9.8
FEV1 [%]	<80	3.0
New hay fever/allergic rhinitis	Yes	2.4
New wheeze/chest tightness	Yes	3.5
New habitual snore	Yes	2.4
FEV1 [%] decline	Yes	1.02

2.2.3. THE ASTHMA SCORE

The asthma score consists in the simple sum of the positive answers to 5 respiratory symptoms: (i) breathless while wheezing in the last 12 months, (ii) woke up with a feeling of chest tightness in the last 12 months, (iii) attack of shortness of breath at rest in the last 12 months, (iv) attack of shortness of breath after exercise in the last 12 months, (v) woken by attack of shortness of breath in the last 12 months. In work by Sunyer et al. [56], the prediction ability of the asthma score was assessed by using the longitudinal data collected in the European Community Respiratory Health Study (ECRHS). Specifically, the analysis included 8956 subjects who were followed up in the period 1998-2001.

The prediction ability of the score was assessed measuring the association of the score at baseline in relation to markers of asthma at follow-up using logistic regression. The analysis showed that increasing score values at baseline are associated with increasing risk of asthma onset at follow up (see the logistic regression odds ratios reported in Table 20).

In addition, Sunyer et al. [56] assessed the association of known asthma risk factors with changes in the asthma score. The assessed risk factors were those previously reported in one of the ECRHS baseline study [57]. Risk factors for which a statistically significant association was found are: female gender, IgE to cat, passive smoking in men, rhinitis, education level, bronchial reactivity, body mass index, smoking during the follow-up and increase of BMI during the follow-up.

Table 20. Association of asthma score at baseline and incidence of asthma at follow-up (logistic regression odds ratios).

Score value	Odds ratio
0	1
1	2.18
2	2.97
3	4.20
4	7.74
5	22.46

2.2.4. THE MODEL BY ANTÓ ET AL.

In the work by Antó et al. [53], the risk factors for new onset of asthma in adults were studied by using the data collected in the two surveys of the ECRHS. The baseline survey, ECRHS I, was conducted in 1990-1995 and included subjects aged 20-44 randomly selected from 25 centres in 13 countries (12 were European). A follow-up survey, ECRHS II, was performed 9 years later in the period 1998-2003. In the study by Antó et al. [53], the 4588 subjects who were free of asthma and never had asthma and respiratory symptoms in ECRHS I were selected for the analysis. Of this sample, 179 subjects reported new onset of asthma in ECRHS II.

Such data were used to fit a multivariate logistic regression model with asthma onset at follow-up as dependent variable and asthma risk factors as independent variables. Risk factors included in the multivariate logistic regression model are those that resulted statistically significantly associated with new asthma onset in at least one of the three univariate logistic regression analysis performed in all subjects and atopics and nonatopics separately. Such risk factors and their respective multivariate logistic regression odds ratio are reported in Table 21. Note that bronchial hyperresponsiveness was defined as a fall of at least 20% in FEV1 associated with a methacholine dose of 1 mg or less. Predicted FEV1 was calculated according to ECSC prediction equations. Atopy was defined as a specific IgE level higher than

0.35 kU/l to at least one of four allergens (cat, house dust mite, timothy grass and *Cladosporum herbarum*) or any positive skin prick test to seven allergens. Regarding occupation, subjects who worked exclusively in professional, clerical or administrative jobs were classified as having low-risk occupations. High-risk occupations were defined according to Kogevinas et al. [58] and included nurses, cleaners, spray-painters, bakers and farmers.

As for previous studies, the aim of work by Antó et al. [53] was to identify risk factors associated with asthma onset, not to develop a tool to predict asthma onset. Therefore, the ability of the multivariable logistic regression model of Table 21 to correctly predict future onset of asthma was not tested by Antó et al. [53].

Table 21. Multivariate logistic regression model by Antó et al.: variables and odd ratios.

Variables	Values	Odds ratio
Gender	Female	1.97
Bronchial hyperresponsiveness	Yes	3.33
Atopy	Yes	1.59
Predicted FEV1	<100%	1.88
BMI	<20	1.01
	25-30	1.20
	≥30	1.69
Nasal allergies	Yes	1.96
Maternal asthma	Yes	1.92
Smoking	Ex-smoker	1.32
	Current smoker	0.93
Respiratory infections before age of 5	Yes	1.41
High-risk occupation	Yes	1.25

2.2.5. THE MODEL BY MCDONNELL ET AL.

The relationship between adult asthma incidence and ambient ozone exposure was investigated in a prospective study conducted by McDonnell et al. [54]. The study involved 3091 nonsmokers, aged 27-87 years who were followed up for a 15-year period as part of the Adventis Health Air Pollution Study in California. In this study, monthly indices of ambient air pollution measured at monitoring stations throughout California were interpolated to zip codes centroids according to residence and work location

histories as described in Abbey et al. [59]. As primary exposure variable, the 8-h average ozone concentration between 9 am and 5 pm (working hours) was calculated. Association between 8-h average ozone concentration and new onset of asthma was estimated by a multivariate logistic regression model accounting for other possible confounding factors, which was fitted separately for men and women. Variables and estimated odds ratios for the logistic regression model are reported in Table 22 for men, Table 23 for women. As shown by the model coefficients in Table 22 and Table 23, the ambient ozone concentration was found positively associated with asthma onset in men but not in women.

The model by McDonnel et al. [54] is interesting because, at difference of models previously presented, it includes a marker of environment pollution as risk factors. However, the model was derived in a population of non-smoker subjects and may require revision when applied to a general population.

Table 22. Variables and coefficients of the multivariate logistic regression model by McDonnel et al. for men.

Variables	Values	Model coefficient
Age [years]	16-year IQR increase	0.0024
Education [years]	4-year IQR increase	-0.0290
Ozone 8-h average [ppb]	27-ppb IQR increase	0.0277
Pneumonia/bronchitis before age of 16 [Boolean]	Yes	-5.9809
History of smoking	Yes	0.8975
Intercept	-	-4.7052

Table 23. Variables and coefficients of the multivariate logistic regression model by McDonnel et al. for women.

Variables	Values	Model coefficient
Age [years]	16-year IQR increase	-0.0331
Education [years]	4-year IQR increase	-0.0473
Ozone 8-h average [ppb]	27-ppb IQR increase	-0.0058
Pneumonia/bronchitis before age of 16 [Boolean]	Yes	1.1669
Years worked with smoker	7-year IQR increase	0.0277
Intercept	-	-0.7748

2.2.6. THE MODEL BY VERLATO ET AL.

Verlato et al. [55] investigated the association of smoking with new onset of asthma in adults, taking into account also other variables as confounding factors. For this purpose, data collected in 3 population cohorts extracted from the Italian Study on Asthma in Young Adults and the Italian Study on the Incidence of Asthma were used. In particular, 5241 subjects without history of asthma at baseline were selected. Subjects participated in a follow-up survey on average 9 years after the baseline survey.

On these data, a multivariate logistic regression model was fitted using asthma onset at follow-up as dependent variable, smoking habits and other confounders as dependent variables. Independent variables and their odds ratios are reported in Table 24. In particular, asthma-like symptoms included wheezing, tightness in the chest and shortness of breath in the last 12 months. Current smokers were defined as subjects smoking at least 1 cigarette/day or 1 cigar/week for 1 year and also in the last month, while ex-smokers were defined as subjects smoking at least 1 cigarette/day or 1 cigar/week for 1 year but not in the last month.

As for previous studies, the aim of work by Verlato et al. [55] was to study risk factors associated with asthma onset, not to develop a tool to predict asthma onset. Therefore, the prediction ability of the multivariable logistic regression model of Table 24 was not tested by Verlato et al. [55].

Table 24. Multivariate logistic regression model by Verlato et al.: variables and odd ratios.

Variables	Values	Odds ratios
Sex	Female	1.18
Age class	30-39	1.13
	40-54	0.85
Occupation	Clerk	1
	Housewife	0.88
	Businessman	0.95
	Unemployed	1.02
	Worker	0.84
Occupation	Student	1.15
	Other	1.1
Asthma-like symptoms	Yes	1.81
Chronic bronchitis	Yes	0.93
Allergic rhinitis	Yes	4.00
Smoking	Ex-smoker	1.28
	Current smoker	1.01

2.3. DISCUSSION

A review of the literature was performed to identify state-of-the-art models of T2D and asthma onset of interest for PULSE. Several prediction models of T2D onset were proposed in the literature. Many of them were translated into risk scores of simple calculation. General patient information, anthropometrics, clinical variables and lifestyle indicators are common risk factors included in these models.

As discussed in section 1, PULSE will implement T2D risk models both in systems for PHA, as tools to identify the high-risk population and plan prevention initiative (e.g., to target the most prevalent risk factors), and in the PULSE app, as tools to provide pre-diabetic users with a measure of their risk of developing the disease and stimulating them to improve their lifestyle. Since all PULSE products aim at promoting healthy behaviours, modifiable risk factors, i.e., risk factors related to lifestyle and behaviour, are most of interest for PULSE. Among these, one of the most important prevention measure of T2D is physical activity, which will also be tracked in PULSE by use of mobility sensors and their integration with the PULSE app. In addition to modifiable risk factors, personal information and simple clinical variables that can be collected by the PULSE app are also of interest for PULSE. Conversely, complex clinical variables, e.g., OGTT results, are not of interest for PULSE, since such information, which require complex and expensive laboratory tests for their collection, will likely not be available to PULSE use cases.

Taking into account the previous considerations, the T2D prediction models that include physical activity as independent variable, like the FINDRISC and the GDRS, are particularly of interest for PULSE. Of interest are also the models including socio-economic indicators, like the DPoRT and the QDScore. Models including personal information and simple clinical variables, like the ARIC model, the Framingham simple clinical model and the risk scores by Kahn et al., are also interesting to investigate in PULSE, as the benefits of adding behavioural variables, like physical activity, to these models can be assessed in task 5.3. An interesting model for use in PHA only is the NYAM-DS model, which can be used to predict the proportion of subjects expected to develop T2D in a certain population, but cannot be used to predict the risk of having future T2D in single individuals. Of scarce interest for PULSE is the AUSRISK model, since the variables of this model were all included in previous literature models, despite here adapted to account for the ethnic composition of the Australian population. Finally, complex clinical models, like the model by Stern et al. with OGTT and the Framingham complex clinical model, are not of interest for PULSE.

Regarding asthma, only few studies on multivariable models to predict adult-onset asthma were found in the literature. As anticipated in section 2.2, most of these studies aimed at identifying risk factors for adult-onset asthma rather than developing tools to predict the incidence of the disease, thus none of the models was tested in the prediction of future incident asthma. The identified state-of-the-art models include general subject information, anthropometrics, lifestyle indicators, clinical variables and exposure to environmental pollutants as risk factors.

As discussed in section 1, the PULSE use case that will take advantage of adult-onset asthma risk models is PHA (use case 4). Prediction models of adult-onset asthma will be implemented in the PULSE system for PHA as tools to stratify the population based on risk assessment and plan prevention initiative for the subgroups at higher risk (e.g., to target the most prevalent risk factors). The PULSE system will also integrate information on air quality. For this reason, the risk factors of adult-onset asthma that are most of interest for PULSE are modifiable risk factors related to lifestyle and those related to environment pollution.

Taking into account the previous considerations, the prediction models of adult-onset asthma that take into account lifestyle factors and environmental factors are most of interest for PULSE. These includes the models proposed by Thomsen et al., Antó et al. McDonnell et al. and Verlato et al. Of scarce interest for PULSE are the model by Jamrozik et al., which only includes information on asthma-like symptoms and spirometry test results, and the asthma score, which is based only on asthma-like symptoms.

3. SELECTED MODELS BASED ON AVAILABLE DATASETS

In this section, we will first present the clinical datasets available in the Consortium for prediction models implementation and development (subsection 3.1). Then, based on the characteristics of available dataset, particularly in terms of variables collected, we will define the state-of-the-art models that can be implemented within task 5.2 (subsection 3.2).

3.1. AVAILABLE DATASETS

3.1.1. HEALTH AND RETIREMENT STUDY

The Health and Retirement Study (HRS) is a longitudinal study of health, retirement and aging conducted in the United States. The HRS is supported by the National Institute of Aging (NIA U01AG009740) and the Social Security Administration. The HRS dataset includes nationally representative public survey data collected every 2 years since 1992 (most recent data were collected in 2014) in males and females of age 51 or older. In addition, biomarkers were collected in 2006 and 2010 for a subgroup of participants, in 2008 and 2012 for another subgroup of participants.

Public survey data includes the following 29 sections: preload; coverscreen; demographics; physical health; cognition; family structure; parents, siblings and transfers; functional limitations and helpers; housing; physical measures; employment; last job; job history; disability; health services and insurance; expectations; assets and income; asset change; widowhood and divorce; wills and life insurance; asset reconciliation; modules; event history, internet use and social security; thumbnails; repeat cognition; time calculations; interviewer observations; leave-behind questionnaires. A codebook with details about variables in each section is available for each study wave at http://hrsonline.isr.umich.edu/index.php?p=avail&_ga=2.160380989.1410054135.1500378358-963515926.1499241364.

Biomarker data includes total cholesterol [mg/dl], HDL cholesterol [mg/dl], HbA1c [%], C-reactive protein [ug/ml], cystatin C [mg/L].

The study involves a representative sample of approximately 20,000 people. New onset of T2D can be identified, at each interview, by the answer to questions “Has a doctor ever told you that you have diabetes or high blood sugar?” and “In what year was your diabetes first diagnosed?”. In a subsample of 8,738 who were interviewed in 2006/2008 and underwent biomarker collection both in 2006/2008 and 2010/2012, presence of diabetes can be quantitatively assessed since HbA1c was measured. This will

allow us to identify also subjects with undiagnosed diabetes. In this subsample, 852 subjects who were free of diabetes in 2006/2008 developed diabetes during the 4-year follow-up period.

Regarding asthma, the HRS questionnaire includes only questions about childhood asthma like “Before you were 16 years old, did you have asthma?”, “At what age were you first diagnosed with asthma?”, “Until what age did you have it?”. No question about asthma onset in adulthood is present in the HRS survey, thus is not possible to use the HRS data for implementation of adult-onset asthma prediction models.

Public survey data collected in HRS are public and can be downloaded from the HRS website by any registered user. Biomarker data are not public because considered sensitive health data and can be accessed only upon signature of a sensitive data access use agreement and a sensitive data order form. Such data are currently available to UNIPD and NYAM, i.e. the partners of the PULSE Consortium who signed the abovementioned agreement and form.

3.1.2. MULTI-ETHNIC STUDY OF ATHEROSCLEROSIS

The Multi-Ethnic Study of Atherosclerosis (MESA) study is a large longitudinal study funded by the National Heart, Lung, and Blood Institute (NHLBI) starting in July 2000. The MESA study investigates subclinical cardiovascular disease (CVD) in a sample (n=6,814) of population consisting of African-Americans (27.8%), Hispanics (21.9%), Chinese (11.8%), and Whites (38.5%). Participants enrolled were 45-84 years old and 53% were female. Data were collected from 6 U.S. communities: Baltimore City and Baltimore County, Maryland; Chicago, Illinois; Forsyth County, North Carolina; Los Angeles County, California; Northern Manhattan and the Bronx, New York; and St. Paul, Minnesota. Details regarding MESA’s design and procedures can be found in a previous study [60].

The first visit (Exam 1) was conducted in 2000–2002. Follow-up visits (Exams 2, 3, 4, and 5) were done in 2002–2004, 2004–2005, 2005–2007, and 2010–2012, respectively. The retention rate was 92% at Exam 2, 89% at Exam 3, 87% at Exam 4, and 76% at Exam 5. At each exam, subjects were interviewed about their health and lifestyle and underwent some clinical assessments. In Table 25, we summarize the questionnaires and procedures/assessments that were performed in each exam of the study.

Table 25. Questionnaires and procedures/assessments performed at each MESA exam.

	Exam 1	**Exam 2**	**Exam 3**	**Exam 4**	**Exam 5**
Questionnaires					
Personal history	X	X	X	X	X
Medical history	X	X	X	X	X
Medications	X	X	X	X	X
Family history		X			
Sleep history		X		X	X

Residential/neighbourhood		X			
Psycho-social	X	X	X	X	X
Occupation/employment	X	X	X		
Physical activity	X	X	X		X
Food frequency (diet)	X				X
	Exam 1	**Exam 2**	**Exam 3**	**Exam 4**	**Exam 5**
Procedures/Assessments					
Anthropometry	X	X	X	X	X
Phlebotomy collection	X	X	X	X	X
Urine collection	X	X	X		X
Genotyping				X	
Cognitive function					X
Blood pressure	X	X	X	X	X
Ankle brachial index	X		X		X
Electrocardiogram	X				X
Arterial wave form	X				X
Retinal photography		X			X
Vision/Refraction		X			X
US endothelial function	X				
US carotid IMT	X	X	X	X	X
US carotid distensibility	X				
MRI cardiac	X	X		X	X
MRI carotid		X			
CT coronary (chest)	X	X	X	X	X
CT aortic (abdomen)		X	X		
Spirometry			X	X	X

New onset of T2D can be assessed at each exam by answer to question “Has a doctor ever told you that you have diabetes?” or by fasting plasma glucose assessment that was included in all the MESA exams and will allow to identify also undiagnosed diabetes. According to two previous studies [61][62], 5931 subjects were free of diabetes at Exam 1 and some of them developed T2D over the follow-up years from

2000-2012. In particular, 414 participants were diagnosed with incident diabetes in Exam 4 (6.6 years of follow-up) and 695 participants were diagnosed with incident diabetes in Exam 5 (11.4 years of follow-up).

As far as asthma is concerned, new onset of asthma can be identified at each exam by the answer to question “Has a doctor ever told you that you had asthma?”. In Exams 3-5, a spirometry test and a spirometry questionnaire including questions on asthma-like symptoms were introduced in a subsample of 3600 subjects, but it is not clear from the documentation available to us in this moment if these procedures would allow to perform new diagnosis of asthma. According to a previous study [63], 6125 subjects were free of asthma at Exam 1. No evidence is available in the literature about the number of subjects who developed new asthma during the follow-up years in MESA.

Data collected in MESA are available to investigators upon submission of a research proposal to the MESA Coordinating Center and its approval. NYAM and UNIPD authored a joint research proposal that was approved by the MESA Coordinating Center. At the time of writing (end of October 2017), UNIPD is waiting for the Ethics Board review of the research proposal, which is required to finalize the MESA Data Distribution Agreement and finally obtain access to the data.

3.2. SELECTED MODELS

According to the variables available in each of the dataset described in section 3.1, we identified the state-of-the-art models of interest for PULSE that can be implemented/re-calibrated using the available data within task 5.2 of the project. Such models are listed in Table 26 and briefly discussed below.

On HRS data, only T2D prediction models/risk scores can be implemented, since no information on adult-onset asthma was gathered in the study. Unfortunately, two risk factors commonly included in state-of-the-art T2D prediction models/risk scores, i.e., family history of diabetes and fasting plasma glucose, are not available in HRS. Consequently, only few T2D prediction models can be implemented using HRS data. In particular, these include the FINDRISC concise model, the FINDRISC concise model with addition of physical activity and the DPoRT model. Modified versions of other literature models can be implemented if family history of diabetes and/or fasting plasma glucose are not considered, e.g., the Framingham personal model, the simple ARIC model, the Kahn basic models and the clinical model by Stern et al.

On MESA data, both T2D and asthma prediction models/risk scores can be implemented, since information on both diabetes onset and asthma onset was collected during the study. For what concerns diabetes, most of the variables included in the literature models/risk scores were collected in MESA, thus most of the models identified in section 2.1 can be implemented/re-calibrated. Specifically, these models includes the clinical model by Stern et al., the FINDRISC, the ARIC risk models, the Framingham personal and clinical models, the GDR score, the simple risk score by Kahn et al. and the DPoRT model. Note that, according to the available documentation about MESA exams, family history of diabetes was recorded starting from Exam 2. In addition, the immigration status, which is required to implement the DPoRT model, was not directly collected in MESA. However, such information can be derived based on respondent’s birth place that was collected in the MESA questionnaire.

Regarding asthma prediction models, only the models by Thomsen et al. [51] and Verlato et al. [55] can be implemented in MESA. In particular, the model by Thomsen et al. [51] requires information on the subject's age, anthropometry, smoking and activity habits, which is available in MESA since Exam 1. The model by Verlato et al. [55] includes also information on asthma related symptoms and respiratory problems or infections as risk factors. In MESA, a questionnaire on respiratory health was performed since Exam 3, thus the model by Verlato et al. [55] can be implemented only using the data of Exams 3-5. The other models, i.e., the models by Jamrozik et al. [52], Antó et al. [53] and McDonnell et al. [54], cannot be implemented since some of the variables required by them were not collected in MESA. Specifically, snoring is missing for the model by Jamrozik et al. [52], bronchial hyperresponsiveness, atopy and respiratory infections before age of 5 are missing for the model by Antó et al. [53] and ozone exposure is missing for the model by McDonnell et al. [54]. Finally, the asthma score [56], which accounts only for asthma-like symptoms, cannot be implemented because not all the symptoms of the score were collected in MESA.

Table 26. Literature prediction models of T2D and asthma onset that can be implemented in the HRS and MESA datasets available in PULSE.

HRS	**MESA**	
T2D prediction models	**T2D prediction models**	**Asthma prediction models**
• FINDRISC concise model • FINDRISC concise model + physical activity • DPoRT model	• Clinical model by Stern et al. • FINDRISC • ARIC models • Framingham personal and clinical models • GDR score • Simple risk score by Kahn et al. • DPoRT	• Model by Thomsen et al. • Model by Verlato et al.

4. DATASET PREPROCESSING

Of the two datasets identified for risk models implementation, recalibration and development, which are described in section 3.1, only the HRS dataset is actually available at the time of writing. As anticipated, the MESA dataset is not available in the Consortium at the time of writing (end of October 2017), but a research proposal on its use for risk models development has been approved by the MESA Coordinating Center and submitted to the Ethics Board for review. The Ethics Board review is expected to be completed by end of October 2017, so the MESA dataset should become available in the Consortium in November 2017.

As soon as both the datasets are available, suitable data preprocessing will be performed. Specifically, the variables required for model implementation will be selected in each dataset and appropriately homogenised, i.e. their definition and unit of measures will be aligned. In each dataset, the subjects with T2D/asthma at baseline and those without follow-up information on T2D/asthma development will be discarded. The remaining data will be divided into training set (approximately 90% of the entire dataset) and validation set (approximately 10% of the entire dataset). In the training test, missing values will be imputed by the k-Nearest Neighbour algorithm. After imputation, continuous variables will be quantized has required by the models to implement. More details on dataset preprocessing will be provided in deliverable D5.2 related to activities of task 5.2.

5. REFERENCES

[1] Buijsse B., Simmons R.K., Griffin S.J., and Schulze M.B. Risk Assessment Tools for Identifying Individuals at Risk of Developing Type 2 Diabetes. Epidemiol Rev. 2011 Jul; 33(1): 46–62.

[2] Noble D., Mathur R., Dent T., Meads C., Greenhalgh T. Risk Models and Scores for Type 2 Diabetes: Systematic Review. BMJ 2011; 343:d7163.

[3] Cichosz S.L., Johansen M.D., Hejlesen O. Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. J Diabetes Sci Technol. 2015 Oct 14; 10(1):27-34.

[4] Smit H.A., Pinart M., Antó J.M., Keil T., Bousquet J., Carlsen K.H., Moons K.G., Hooft L., Carlsen K.C. Childhood asthma prediction models: a systematic review. Lancet Respir Med. 2015 Dec; 3(12):973-84.

[5] Castro-Rodríguez J.A., Holberg C.J., Wright A.L., Martinez F.D. A clinical index to define risk of asthma in young children with recurrent wheezing. Am J Respir Crit Care Med. 2000 Oct; 162(4 Pt 1):1403-6.

[6] Juniper E.F., O'Byrne P.M., Guyatt G.H., Ferrie P.J., King D.R. Development and validation of a questionnaire to measure asthma control. Eur Respir J. 1999 Oct; 14: 902–7.

[7] Greenberg S. Asthma exacerbations: predisposing factors and prediction rules. Curr Opin Allergy Clin Immunol. 2013 Jun; 13:225–36.

[8] Osborne M.L., Pedula K.L., O'Hollaren M., Ettinger K.M., Stibolt T., Buist A.S., Vollmer W.M. Assessing future need for acute care in adult asthmatics: the Profile of Asthma Risk Study: a prospective health maintenance organization-based study. Chest. 2007 Oct; 132:1151–61.

[9] Eisner M.D., Yegin A., Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest. 2012 Jan; 141:58–65.

[10] Miller M.K., Lee J.H., Blanc P.D., Pasta D.J., Gujrathi S., Barron H., Wenzel S.E., Weiss S.T.; TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J. 2006 Dec; 28:1145–55.

[11] Bateman E.D., Buhl R., O'Byrne P.M., Humbert M., Reddel H.K., Sears M.R., Jenkins C., Harrison T.W., Quirce S., Peterson S., Eriksson G. Development and validation of a novel risk score for asthma exacerbations: The risk score for exacerbations. J Allergy Clin Immunol. 2015 Jun; 135:1457–64.

[12] Loymans R.J., Honkoop P.J., Termeer E.H., Snoeck-Stroband J.B., Assendelft W.J., Schermer T.R., Chung K.F., Sousa A.R., Sterk P.J., Reddel H.K., Sont J.K., Ter Riet G. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax. 2016 Sep; 71(9): 838-46.

[13] Lindström J. and Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care. 2003 Mar; 26(3):725-31.

[14] Abbasi A., Peelen L. M., Corpeleijn E., van der Schouw Y.T., Stolk R.P., Spijkerman A.M.W., van der A D.L., Moons K.G., Navis G., Bakker S.J., Beulens J.W. Prediction Models for Risk of Developing Type 2 Diabetes: Systematic Literature Search and Independent External Validation Study. BMJ. 2012 Sep; 345: e5900.

[15] Kengne A.P., Beulens J.W., Peelen L.M., Moons K.G., van der Schouw Y.T., Schulze M.B., Spijkerman A.M., Griffin S.J., Grobbee D.E., Palla L., Tormo M.J., Arriola L., Barengo N.C., Barricarte A., Boeing H., Bonet C., Clavel-Chapelon F., Dartois L., Fagherazzi G., Franks P.W., Huerta J.M., Kaaks R., Key T.J., Khaw K.T., Li K., Mühlenbruch K., Nilsson P.M., Overvad K., Overvad T.F., Palli D., Panico S., Quirós J.R., Rolandsson O., Roswall N., Sacerdote C., Sánchez M.J., Slimani N., Tagliabue G., Tjønneland A.,

Tumino R., van der A D.L., Forouhi N.G., Sharp S.J., Langenberg C., Riboli E., Wareham N.J. Non-Invasive Risk Scores for Prediction of Type 2 Diabetes (EPIC-InterAct): A Validation of Existing Models. Lancet Diabetes Endocrinol. 2014 Jan; 2(1):19-29.

[16] Stern M. P., Williams K., Haffner S.M. Identification of Persons at High Risk for Type 2 Diabetes Mellitus: Do We Need the Oral Glucose Tolerance Test? Ann Intern Med. 2002 Apr; 136(8):575-581.

[17] McNeely M.J., Boyko E.J., Leonetti D.L., Kahn S.E., Fujimoto W.Y. Comparison of a clinical model, the oral glucose tolerance test, and fasting glucose for prediction of type 2 diabetes risk in Japanese Americans. Diabetes Care. 2003 Mar; 26(3):758–763.

[18] Stern M.P., Williams K., González-Villalpando C., Hunt K.J., Haffner S.M. Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease? Diabetes Care. 2004 Nov; 27(11):2676–2681.

[19] Schmiedel K., Mayr A., Fießler C., Schlager H., Friedland K. Effects of the lifestyle intervention program GLICEMIA in people at risk for type 2 diabetes: a cluster-randomized controlled trial. Diabetes Care. 2015 May; 38(5):937-9.

[20] Fizelova M., Jauhiainen R., Stančáková A., Kuusisto J., Laakso M. Finnish Diabetes Risk Score is associated with impaired insulin secretion and insulin sensitivity, drug-treated hypertension and cardiovascular disease: a follow-up study of the METSIM cohort. PLoS One. 2016 Nov 16; 11(11):e0166584.

[21] Väätäinen S., Cederberg H., Roine R., Keinänen-Kiukaanniemi S., Saramies J., Uusitalo H., Tuomilehto J., Martikainen J. Does future diabetes risk impair current quality of life? A cross-sectional study of health-related quality of life in relation to the Finnish Diabetes Risk Score (FINDRISC). PLoS One. 2016; 11(2):e0147898.

[22] Breeze P.R., Thomas C., Squires H., Brennan A., Greaves C., Diggle P.J., Brunner E., Tabak A., Preston L., Chilcott J. The impact of Type 2 diabetes prevention programmes based on risk-identification and lifestyle intervention intensity strategies: a cost-effectiveness analysis. Diabet Med. 2017 May; 34(5):632-640.

[23] Tankova T., Chakarova N., Atanassova I., Dakovska L. Evaluation of the Finnish Diabetes Risk Score as a screening tool for impaired fasting glucose, impaired glucose tolerance and undetected diabetes. Diabetes Res Clin Pract. 2011 Apr; 92(1):46-52.

[24] Noto D., Cefalù A.B., Barbagallo C.M., Falletta A., Ganci A., Sapienza M., Cavera G., Nardi I., Pagano M., Notarbartolo A., Averna M.R. Prediction of incident type 2 diabetes mellitus based on a twenty-year follow-up of the Ventimiglia heart study. Acta Diabetol. 2012 Apr; 49(2):145-51.

[25] Schmidt M.I., Duncan B.B., Bang H., Pankow J.S., Ballantyne C.M., Golden S.H., Folsom A.R., Chambless L.E., Atherosclerosis Risk in Communities Investigators. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. Diabetes Care. 2005 Aug; 28(8):2013-8.

[26] Sun F., Tao Q., Zhan S. An accurate risk score for estimation 5-year risk of type 2 diabetes based on a health screening population in Taiwan. Diabetes Res Clin Pract. 2009 Aug; 85(2):228-34.

[27] Wilson P. W., Meigs J.B., Sullivan L., Fox C.S., Nathan D.M., D'Agostino R.B. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007 May; 167(10):1068-74.

[28] Mahmood, S.S., Levy, D., Vasan, R.S., Wang, T.J. The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. Lancet. 2014 Mar; 383(9921), 999–1008.

[29]Schulze M.B., Hoffmann K., Boeing H., Linseisen J., Rohrmann S., Möhlig M., Pfeiffer A.F., Spranger J., Thamer C., Häring H.U., Fritsche A., Joost H.G. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007 Mar; 30(3):510-5.

[30]Schwarz P.E., Li J., Wegner H., Bornstein S.R., Lindström J., Tuomilehto J. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes: response to Schulze et al. Diabetes Care. 2007 Aug; 30(8):e87.

[31]Mühlenbruch K., Ludwig T., Jeppesen C., Joost H.G., Rathmann W., Meisinger C., Peters A., Boeing H., Thorand B., Schulze M.B. Update of the German Diabetes Risk Score and external validation in the German MONICA/KORA study. Diabetes Res Clin Pract. 2014 Jun; 104(3):459-66.

[32]Nichols G.A., Brown J.B. Validating the Framingham Offspring Study equations for predicting incident diabetes mellitus. Am J Manag Care. 2008 Sep; 14(9):574-80.

[33]Xu L., Jiang C.Q., Schooling C.M., Zhang W.S., Cheng K.K., Lam T.H. Prediction of 4-year incident diabetes in older Chinese: Recalibration of the Framingham diabetes score on Guangzhou Biobank Cohort Study. Prev Med. 2014 Dec; 69:63-8.

[34]Kahn H.S., Cheng Y.J., Thompson T.J., Imperatore G., Gregg E.W. Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults aged 45 to 64 years. Ann Intern Med. 2009 Jun; 150(11):741-51.

[35]Schmid R., Vollenweider P., Bastardot F., Waeber G., Marques-Vidal P. Validation of 7 type 2 diabetes mellitus risk scores in a population-based cohort: CoLaus study. Arch Intern Med. 2012 Jan; 172(12):188-9.

[36]Hippsley-Cox J., Coupland C., Robson J., Sheikh A., Brindle P. Predicting Risk of Type 2 Diabetes in England and Wales: Prospective Derivation and Validation of QDScore. BMJ. 2009 Jan; 338:b880.

[37]Townsend, P., Phillimore, P. and Beattie, A. Health and Deprivation: Inequality and the North. Routledge, London, 1988.

[38]Collins G.S., Altman D.G. External validation of QDSCORE(®) for predicting the 10-year risk of developing Type 2 diabetes. Diabet Med. 2011 May; 28(5):599-607.

[39]Rosella L.C., Manuel D.G., Burchill C., Stukel T.A., for the PHIAT-DM team. A Population-Based Risk Algorithm for the Development of Diabetes: Development and Validation of the Diabetes Population Risk Tool (DPoRT). J Epidemiol Community Health. 2011 Jul; 65(7):613-620.

[40]Chen L., Magliano D.J., Balkau B., Colagiuri S., Zimmet P.Z., Tonkin A.M., Mitchell P., Phillips P.J., Shaw J.E. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010 Feb; 192(4):197-202.

[41]Li Y., Padron N., Mangla A.T., Russo P.G., Schlenker T., and Pagán J.A. Using Systems Science to Improve Population Health Strategies in Local Health Departments: A Case Study in San Antonio, Texas. Public Health Reports. 2017 Aug; 132(5):549-555.

[42]Li Y., Kong N., Lawley M., Weiss L., and Pagán J.A. Advancing the Use of Evidence-Based Decision Making in Local Health Departments with Systems Science Methodologies. Am J Public Health 2015 Apr; 105(s2):s217-s222.

[43]Smit H.A., Pinart M., Antó J.M., Keil T., Bousquet J., Carlsen K.H., Moons K.G., Hooft L., Carlsen K.C. Childhood asthma prediction models: a systematic review. Lancet Respir Med. 2015 Dec; 3(12):973-84.

[44]Hafkamp-de Groen E., Lingsma H.F., Caudri D., Levie D., Wijga A., Koppelman G.H., Duijts L., Jaddoe V.W., Smit H.A., Kerkhof M., Moll H.A., Hofman A., Steyerberg E.W., de Jongste J.C., Raat H. Predicting asthma in preschool children with asthma-like symptoms: Validating and updating the PIAMA risk score. J Allergy Clin Immunol. 2013 Dec; 132(6):1303-10.

[45]Pescatore A.M., Dogaru C.M., Duembgen L., Silverman M., Gaillard E.A., Spycher B.D., Kuehni C.E. A simple asthma prediction tool for preschool children with wheeze or cough. J Allergy Clin Immunol. 2014 Jan; 133(1):111-8.e1-13.

[46]Grabenhenrich L.B., Reich A., Fischer F., Zepp F., Forster J., Schuster A., Bauer C., Bergmann R.L., Bergmann K.E., Wahn U., Keil T., Lau S. The Novel 10-Item Asthma Prediction Tool: External Validation in the German MAS Birth Cohort. PLoS One. 2014 Dec; 9(12): e115852.

[47]Jacquemin B., Schikowski T., Carsin A.E., Hansell A., Krämer U., Sunyer J., Probst-Hensch N., Kauffmann F., Künzli N. The role of air pollution in adult-onset asthma: a review of the current evidence. Semin Respir Crit Care Med. 2012 Dec; 33(6):606-19.

[48]Jeebhay M.F., Ngajilo D., le Moual N. Risk factors for nonwork-related adult-onset asthma and occupational asthma: a comparative review. Curr Opin Allergy Clin Immunol. 2014 Apr; 14(2):84-94.

[49]Ilmarinen P., Tuomisto L.E., Kankaanranta H. Phenotypes, Risk Factors, and Mechanisms of Adult-Onset Asthma. Mediators Inflamm. 2015 Oct; 2015:514868.

[50]Le Moual N., Jacquemin B., Varraso R., Dumas O., Kauffmann F., Nadif R. Environment and asthma in adults. Presse Med. 2013 Sep; 42(9 Pt 2):e317-33.

[51]Thomsen S.F., Ulrik C.S., Kyvik K.O.,Larsen K., Skadhauge L.R., Steffensen I., Backer V. The Incidence of Asthma in Young Adults. Chest. 2005 Jun; 127(6):1928-34.

[52]Jamrozik E., Knuiman M.W., James A., Divitini M., Musk A. Risk Factors for Adult-Onset Asthma: A 14-Year Longitudinal Study. Respirology. 2009 Aug; 14(6):814-821.

[53]Antó J.M., Sunyer J., Basagaña X. , Garcia-Esteban R., Cerveri I., de Marco R., Heinrich J., Janson C., Jarvis D., Kogevinas M., Kuenzli N., Leynaert B., Svanes C., Wjst M., Gislason T., Burney P. Risk Factors of New-Onset Asthma in Adults: A Population-Based International Cohort Study. Allergy. 2010 Aug; 65(8):1021-1030.

[54]McDonnell W.F., Abbey D.E., Nishino N., Lebowitz M.D. Long-Term Ambient Ozone Concentration and the Incidence of Asthma in Nonsmoking Adults: The Ahsmog Study. Environ Res 1999; 80: 110-121.

[55]Verlato G., Nguyen G., Marchetti P., Accordini S., Marcon A., Marconcini R., Bono R., Fois A., Pirina P., de Marco R. Smoking and New-Onset Asthma in a Prospective Study on Italian Adults. Int Arch Allergy Immunol. 2016 Aug; 170(3):149-57.

[56]Sunyer J., Pekkanen J., Garcia-Esteban R., Svanes C., Künzli N., Janson C., de Marco R., Antó J.M., Burney P. Asthma Score: Predictive Ability and Risk Factors. Allergy. 2007 Feb;62(2):142-148.

[57]Basagaña X., Sunyer J., Kogevinas M., Zock J.P., Duran-Tauleria E., Jarvis D., Burney P., Anto J.M.; European Community Respiratory Health Survey. Socioeconomic Status and Asthma Prevalence in Young Adults: the European Community Respiratory Health Survey. Am J Epidemiol. 2004 Jul; 160(2):178-88.

[58]Kogevinas M., Zock J.P., Jarvis D., Kromhout H., Lillienberg L., Plana E., Radon K., Torén K., Alliksoo A., Benke G., Blanc P.D., Dahlman-Hoglund A., D'Errico A., Héry M., Kennedy S., Kunzli N., Leynaert B., Mirabelli M.C., Muniozguren N., Norbäck D., Olivieri M., Payo F., Villani S., van Sprundel M., Urrutia I., Wieslander G., Sunyer J., Antó J.M. Exposure to Substances in the Workplace and New-Onset Asthma: An International Prospective Population-Based Study (ECRHS-II). Lancet. 2007 Jul; 370(9584):336-41.

[59]Abbey D.E., Moore J., Petersen F., Beeson W.L. Estimating Cumulative Ambient Concentrations of Air Pollutants: Description and Precision of Methods Used for an Epidemiological Study. Arch Environ Health. 1991 Sep-Oct; 46(5):281-287.

[60]Bild D.E., Bluemke D.A., Burke G.L., Detrano R., Diez Roux A.V., Folsom A.R., Greenland P., Jacob D.R. Jr, Kronmal R., Liu K., Nelson J.C., O'Leary D., Saad M.F., Shea S., Szklo M., Tracy R.P. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. Am J Epidemiol. 2002 Nov; 156(9):871–881.

[61]Abraham S., Shah N.G., Diez Roux A., Hill-Briggs F., Seeman T., Szklo M., Schreiner P.J., Golden S.H. Trait Anger but not Anxiety Predicts Incident Type 2 Diabetes: The Multi-Ethnic Study of Atherosclerosis (MESA). Psychoneuroendocrinology. 2015 Oct; 60:105–113.

[62]Abiemo E.E., Alonso A., Nettleton J.A., Steffen L.M., Bertoni A.G., Jain A., Lutsey P.L. Relationships of the Mediterranean dietary pattern with insulin resistance and diabetes incidence in the Multi-Ethnic Study of Atherosclerosis (MESA). Br J Nutr. 2013 Apr; 109(08):1490–1497.

[63]Tattersall M.C., Guo M., Korcarz C.E., Gepner A.D., Kaufman J.D., Liu K.J., Barr R.G., Donohue K.M., McClelland R.L., Delaney J.A., Stein J.H. Asthma Predicts Cardiovascular Disease Events: The Multi-Ethnic Study of Atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology. 2015 Apr; 35(6):1520-152