PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Prevalence of prediabetes and undiagnosed diabetes in the Mollerussa prospective observational cohort study in a semi-rural area of Catalonia
AUTHORS	Falguera, Mireia; Vilanova, Maria; Alcubierre, Nuria; Granado-Casas, Minerva; Marsal, Josep Ramón; Miró, Neus; Cebrian, Cristina; Molló, Àngels; Franch-Nadal, Josep; Mata-Cases, Manel; Castelblanco, Esmeralda; Mauricio, Didac

VERSION 1 – REVIEW

REVIEWER	Dr. V. Mohan Madras Diabetes Research Foundation, Chennai
REVIEW RETURNED	26-Aug-2019

GENERAL COMMENTS	This article reports on the prevalence of undiagnosed diabetes and prediabetes in the general population in the Mollerussa cohort. The article is well written, but there are issues with the selection criteria. There are few issues that need to be addressed: • Abstract methods – sample size is not mentioned. • The selection of participants for the study is not clear. How were 2,226 subjects invited from the list of 24,666 eligible individuals were selected? How were 594 subjects recruited from the 2,226 invited subjects? • Follow-up details were obtained in only 166 out of 594 subjects recruited at baseline, which is 28% response rate – the low response rate is a major limitation and results cannot be generalized based on this. • Generalizing the results to the Mollerussa health care area, a Mediterranean semi-rural area in northeast Spain is questionable based on the small sample size. • The reported prevalence of prediabetes is high (39.3%) as the denominator (general population) does not include the self-reported diabetic subjects. • Abstract Conclusion is very vaguely written, especially the 2nd sentence, '…….few subjects with prediabetes progressed to diabetes'.

REVIEWER	Thaddäus Tönnies German Diabetes Center (DDZ), Düsseldorf, Germany
REVIEW RETURNED	24-Sep-2019

GENERAL COMMENTS	Mireira et al. report estimates of undiagnosed diabetes and prediabetes in a semi-rural area in Spain (Mollerussa). Furthermore, they identified variables associated with both conditions. In a longitudinal component, they also assessed

	predictors of returning from prediabetes to a normal glycaemic state.
	Overall, the paper is well written and the aims of the study are clear. However, there are some concerns that need consideration.
	
	Major points:
	1. The authors state that the study sample is representative for the Mollerussa general population. However, “cardiovascular disease (heart disease, heart failure, aortic stenosis), cancer, kidney disease, anaemia, hepatitis, gastrointestinal diseases, recent abdominal surgery, chronic pulmonary obstructive disease, chronic infectious diseases, use of systemic glucocorticoids or beta blockers or major psychiatric disorders with psychotic symptoms” were exclusion criteria. Hence, the study sample is probably much healthier than the general population, which is why I have doubts that the prevalence estimates can be transferred to the general population.
	2. With regard to the statistical analysis, I have a few concerns. As secondary objectives, the authors aimed to (i) identify variables associated with undiagnosed diabetes and prediabetes and (ii) describe changes in glycaemic status after follow-up of 1 year in participants with prediabetes.
	For (i), the authors report p-values in table 1 for variables potentially associated with prediabetes/undiagnosed diabetes. In my opinion, it would be preferable to report, in a separate table, some measure of association (e.g. prevalence ratio or prevalence difference) with corresponding confidence intervals. P-values alone bear little information about the strength of association. Furthermore, according to the STROBE reporting guideline, inferential statistics such as p-values should not be used to describe the study population (e.g. in table 1). In a next step, the authors performed a multivariable logistic regression using “the enter method with covariables that were clinically or statistically associated.” Please clarify what the “enter method” is and what the exact definition of “clinically and statistically associated” was. In general, I think it would be preferable to include variables in the regression model based on prior subject matter knowledge and/or a biological model/hypothesis rather than including variables based on associations observed in the data. For instance, it would appear reasonable to include known risk factors for diabetes in the model.
	
	For (ii), a “backward conditional logistic regression model was used to predict the normalization of the glycaemic state”. What is the reason for using a conditional logistic regression instead of a simple logistic regression? Again, I think prior subject matter knowledge and/or a biological model/hypothesis is better suited to select variables for the regression model. Furthermore, it might also be of interest to describe the probability of changing from the normoglycaemic state to prediabetes. One approach could be to estimate the transition probabilities of simple Markov model with the two states “prediabetes” and “normoglycaemic”. The transition probabilities could be estimated with logistic regression including all participants without undiagnosed diabetes. The outcome of this model would be “normoglycaemic at follow-up (yes/no)” and the predictor would be “normoglycaemic at baseline (yes/no)”. From this model one could predict the probability (with corresponding confidence intervals) of changing from normoglycaemic to prediabetes and vice versa. Perhaps this approach would provide more information than currently presented in table 3. In a next

	step, the authors assessed the predictive accuracy with the Hosmer-Lemeshow test and AUC-ROC. Usually these procedures are used to evaluate the performance of risk prediction model or a diagnostic test. Since neither was developed in the paper, I recommend the exclude these analyses. Minor points: - I recommend to round the confidence bounds of the prevalence estimates to the first decimal (in the abstract and the rest of the paper). - The first sentence of the introduction reads “Diabetes mellitus, a major problem …” – I think something like “Diabetes mellitus, a public health concern…” would be more appropriate. - “Additionally, multiple risk factors, such as family history, gestational diabetes, and certain ethnicities as well as combined risk factors such as metabolic syndrome, are known to predispose subjects to a higher risk for prediabetes and its progression to T2D.” – please provide a reference to support this statement. - Please describe briefly in the methods section, who is in the Primary Care Electronic Clinical Station. Is it everyone registered in health insurance? Only people that accessed health care facilities in the study region? - Please explain why people with certain diseases were excluded from the study. - “Sociodemographic variables were recorded, and a physical examination (weight, height, blood pressure and waist circumference) was carried out by researchers following a protocol for the inclusion of patients using a standardized baseline questionnaire for the clinical interview.” – From this sentence, I am not sure whether weight, height, blood pressure and waist circumference were measured by researchers or self-reported during a personal interview. Please specify. - Please describe continuous variables in table 1 either using mean and standard deviation or median and interquartile range. I cannot see, why for some variables the authors decided for the mean and for others chose the median. - “We observed a positive trend in age, BMI, waist circumference…” – What is meant by “positive trend”? Maybe use “association” instead. - In the results section, much details are given with regard to different kinds of prediabetes (Hba1c, FPG or both). To me the relevance of this distinction is not apparent. I suggest to either reduce complexity by only reporting results for prediabetes (as was done in table 1) or explain the relevance of this distinction more clearly. - “Third, although traditional factors such as hypertension, dyslipidaemia and obesity were included in the analysis models, the existence of unmeasured confounding variables cannot be entirely ruled out.” – To me, this statement at the end of the manuscript is confusing, because the authors do not state which estimated effect could be confounded by unmeasured variables. In addition, if confounding of a hypothesized effect is a concern of the authors, I would expect that it is explicitly stated in methods section (i) which effect is aimed to be estimated, (ii) what the potentially confounding variables are and (iii) which procedures were used to adjust for confounding. - In my opinion, the following conclusion drawn from this study is perhaps too strong: “…the identification of individuals with prediabetes provides an opportunity for intervention through lifestyle modification and pharmacological treatments not only to

	reduce the development of diabetes but also to prevent the development of chronic complications." While this statement is probably true, I cannot see how the results of the study contribute to this conclusion.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:Reviewer: 1Reviewer Name: Dr. V. MohanInstitution and Country: Madras Diabetes Research Foundation, ChennaiPlease state any competing interests or state 'None declared': None declaredPlease leave your comments for the authors below:This article reports on the prevalence of undiagnosed diabetes and prediabetes in the general population in the Mollerussa cohort. The article is well written, but there are issues with the selection criteria. There are few issues that need to be addressed:
• Abstract methods – sample size is not mentioned.

We thank the Reviewer for this comment. Now, we have included the overall number of participants in the Abstract section.

Abstract section
The study included 583 participants.

• The selection of participants for the study is not clear. How were 2,226 subjects invited from the list of 24,666 eligible individuals were selected? How were 594 subjects recruited from the 2,226 invited subjects?

We agree with the Reviewer and apologize for the incomplete information about recruitment. Now, we have completed the paragraph with all the required information in the Methods section.

Methods section – first paragraph
Then, from a total population of 24,666 potentially eligible individuals in the health-care area (subjects older than 25 years and attending any Primary Healthcare Centre in the same health area), 2,226 subjects were randomly selected using a randomiser programme (SPSS software V.16.0 for Windows; SPSS), following the principles of simple random sampling, and were then invited to participate by telephone contact. Based on their willingness to join the study, exclusion criteria, consent and baseline laboratory data, 594 subjects aged ≥ 25 years were finally included.
• Follow-up details were obtained in only 166 out of 594 subjects recruited at baseline, which is 28% response rate – the low response rate is a major limitation and results cannot be generalized based on this.

There has probably been a misunderstanding. Actually, the subjects that were invited to a follow-up visit were those with prediabetes at baseline, i.e. 229 participants. Among them, 166 patients (response rate of 72.5%) had a follow-up assessment. We are explaining this issue in the Methods section.

Methods section – first paragraph
Subjects with prediabetes at baseline (n=229) underwent a second visit 12 months after the baseline visit, and 166 (72.5%) of them had relevant information at follow up.
• Generalizing the results to the Mollerussa health care area, a Mediterranean semi-rural area in northeast Spain is questionable based on the small sample size.

Despite the statistical accuracy of the sample calculation, we agree with the Reviewer that we should

be more cautious regarding the representativeness of the study for the whole country. Now, we have added a piece of text in the paragraph on limitations in the Discussion section.

Discussion section – fifth paragraph
First, the number of participants in our study is smaller in comparison to other studies. In addition, the study may not be representative of urban areas in our region. Thus, the results may not be generalizable to other territories with different population characteristics in our country.
• The reported prevalence of prediabetes is high (39.3%) as the denominator (general population) does not include the self-reported diabetic subjects.

We appreciate this Reviewer's comment. We would like to point out that the methodology of the study precluded the inclusion of subjects with known diabetes, as we aimed at determining the prevalence of prediabetes in subjects without the disease, i.e. diabetes mellitus. Moreover, since the Mollerussa cohort project is part of a broader project that also included the detection of carotid atherosclerosis to identify asymptomatic cardiovascular disease, we carefully selected inclusion and exclusion criteria that have been previously published. However, to address the issue raised by the Reviewer, we have included a sentence in the Discussion section.

Discussion section – fifth paragraph

Second, our study sample is probably healthier than the general population, as we excluded subjects with already known diabetes and other comorbidities, a lower number of subjects were counted in the denominator, thus resulting in a higher prevalence of this condition.
• Abstract Conclusion is very vaguely written, especially the 2nd sentence, '……few subjects with prediabetes progressed to diabetes'.
We agree with the Reviewer's comment and have amended the final sentence, including the precise proportion of subjects with prediabetes that progressed to diabetes.

Abstract section
After a one-year follow-up, a small proportion of subjects (0.6%) with prediabetes progressed to diabetes, while a high proportion (41.6%) returned to normoglycaemia.
Reviewer: 2Reviewer Name: Thaddäus TönniesInstitution and Country: German Diabetes Center (DDZ), Düsseldorf, GermanyPlease state any competing interests or state 'None declared': nonePlease leave your comments for the authors belowMireira et al. report estimates of undiagnosed diabetes and prediabetes in a semi-rural area in Spain (Mollerussa). Furthermore, they identified variables associated with both conditions. In a longitudinal component, they also assessed predictors of returning from prediabetes to a normal glycaemic state. Overall, the paper is well written and the aims of the study are clear. However, there are some concerns that need consideration.Major points:1. The authors state that the study sample is representative for the Mollerussa general population. However, "cardiovascular disease (heart disease, heart failure, aortic stenosis), cancer, kidney disease, anaemia, hepatitis, gastrointestinal diseases, recent abdominal surgery, chronic pulmonary obstructive disease, chronic infectious diseases, use of systemic glucocorticoids or beta blockers or major psychiatric disorders with psychotic symptoms" were exclusion criteria. Hence, the study sample is probably much healthier than the general population, which is why I have doubts that the prevalence estimates can be transferred to the general population.

We appreciate this comment and fully agree with the Reviewer that the population of our study is healthier than the general population. Please, also see our response to a comment from the first reviewer. Thus, we have modified the sentences in the manuscript. We also added a comment on this as a limitation of the study in the Discussion section.

Discussion section – fifth paragraph
Second, our study sample is probably healthier than the general population, as we excluded subjects with already known diabetes and other comorbidities, a lower number of subjects were counted in the denominator, thus resulting in a higher prevalence of this condition.

2. With regard to the statistical analysis, I have a few concerns. As secondary objectives, the authors aimed to (i) identify variables associated with undiagnosed diabetes and prediabetes and (ii) describe changes in glycaemic status after follow-up of 1 year in participants with prediabetes.

For (i), the authors report p-values in table 1 for variables potentially associated with prediabetes/undiagnosed diabetes. In my opinion, it would be preferable to report, in a separate table, some measure of association (e.g. prevalence ratio or prevalence difference) with corresponding confidence intervals. P-values alone bear little information about the strength of association. Furthermore, according to the STROBE reporting guideline, inferential statistics such as p-values should not be used to describe the study population (e.g. in table 1).

We thank the Reviewer for his suggestion, and following his advice, we have replaced p values by the difference and the 95% CI of the clinical and sociodemographic characteristics of the prediabetic and diabetic groups with respect to the normoglycaemic group, to better show the strength of association. Thus, all the information, as requested by the Reviewer, is included in the new table 1.

In a next step, the authors performed a multivariable logistic regression using “the enter method with covariables that were clinically or statistically associated.” Please clarify what the “enter method” is and what the exact definition of “clinically and statistically associated” was. In general, I think it would be preferable to include variables in the regression model based on prior subject matter knowledge and/or a biological model/hypothesis rather than including variables based on associations observed in the data. For instance, it would appear reasonable to include known risk factors for diabetes in the model.

We greatly appreciate this comment. Actually, we realized that the expression "enter method" is not necessary at all as it refers to the multivariable logistic regression methodology; thus, we removed this term. In addition, we also agree with the Reviewer on the need to specify those variables that are known to be related to the risk of the disease. In the Methods section, we only stated that "clinically and statistically associated" were included; this statement was clearly insufficient. Now, we have added in the Methods section the list of variables included in the model to clarify this to the reader. This list also includes well known risk factors for diabetes. The variables included in the model were: age, sex, education level, physical activity, dyslipidaemia, hypertension, family history of diabetes, BMI, waist, glomerular filtration rate and fatty liver index.

Methods section – Statistical methods
In the prediabetes model, the variables used were age, sex, education level, physical activity, DLP, HT, family history of diabetes, BMI, waist, glomerular filtration rate and fatty liver index.

For (ii), a “backward conditional logistic regression model was used to predict the normalization of the glycaemic state”. What is the reason for using a conditional logistic regression instead of a simple logistic regression? Again, I think prior subject matter knowledge and/or a biological model/hypothesis is better suited to select variables for the regression model.

To clarify this, first we used simple logistic regression that included variables based on prior subject matter knowledge and/or a biological model/hypothesis (complete model). Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. Thus, we used a stepwise regression to drop variables that did not improve the prediction performance (for distinct reasons: collinearity, non-applicability or low association). Now, we have added the complete model as a supplementary table in the Results section.

Results section – Prediction of normalization
Backward conditional logistic regression, as described in the methods section, starting with the variables age, sex, waist circumference, BMI, hypertension, physical activity, family history of diabetes, education level, total cholesterol, HDL-cholesterol, FLI and HOMA2-IR, was performed to identify factors independently associated with the prediction of glycaemic status normalization (Supplementary table S4).

Furthermore, it might also be of interest to describe the probability of changing from the normoglycaemic state to prediabetes.
One approach could be to estimate the transition probabilities of simple Markov model with the two states “prediabetes” and “normoglycaemic”. The transition probabilities could be estimated with logistic regression including all participants without undiagnosed diabetes. The outcome of this model would be “normoglycaemic at follow-up (yes/no)” and the predictor would be “normoglycaemic at baseline (yes/no)”. From this model one could predict the probability (with corresponding confidence intervals) of changing from normoglycaemic to prediabetes and vice versa. Perhaps this approach would provide more information than currently presented in table 3.

We thank the Reviewer for this very interesting suggestion. Unfortunately, the study design did not include the follow-up of those subjects with baseline normal glucose tolerance as a secondary objective. This fact precluded the capture of this information; this is a clear limitation of the study. Therefore, we cannot perform the recommended analysis. We have added a comment on this issue as a limitation of the study in the Discussion section.

Discussion section – fifth paragraph
Fourth, we only followed up those participants with prediabetes. Thus, we could not analyse the probability of changing from normoglycaemia to prediabetes or diabetes in this study.

In a next step, the authors assessed the predictive accuracy with the Hosmer-Lemeshow test and AUC-ROC. Usually these procedures are used to evaluate the performance of risk prediction model or a diagnostic test. Since neither was developed in the paper, I recommend the exclude these analyses.

Following the reviewer’s request, we have removed this from the main text of the article. However, we have kept this as additional information in the online supplementary material (Figure S2).
Minor points:- I recommend to round the confidence bounds of the prevalence estimates to the first decimal (in the abstract and the rest of the paper).

Following the Reviewer’s advice, we rounded the confidence bounds of the prevalence estimates to the first decimal in all sections.
- The first sentence of the introduction reads “Diabetes mellitus, a major problem …” – I think something like “Diabetes mellitus, a public health concern…” would be more appropriate.

We thank the Reviewer for this comment. We have changed the sentence according to his suggestion.

Background section – first paragraph
Diabetes mellitus, a public health concern with an increasing incidence worldwide, is a great threat to general health and is leading to increased morbidity and mortality.
- “Additionally, multiple risk factors, such as family history, gestational diabetes, and certain ethnicities as well as combined risk factors such as metabolic syndrome, are known to predispose subjects to a higher risk for prediabetes and its progression to T2D.” – please provide a reference to

support this statement.

Following the Reviewer's comment, we have added the requested reference.

5. American Diabetes Association. 2. Classification and Diagnosis of Diabetes. Standards of Medical Care in Diabetes – 2019. Diabetes Care 2019;42:S13-S28. doi: 10.2337/dc19-S002

- Please describe briefly in the methods section, who is in the Primary Care Electronic Clinical Station. Is it everyone registered in health insurance? Only people that accessed health care facilities in the study region?

The Spanish health care system is based on the principles of universality, free access, equity and fairness of financing. Thus, all the population is passively included in the Primary Care Electronic Clinical record system and not only people attending the primary care centres. Now, we added text in the Methods section to clarify this.

Methods section – first paragraph
All the population is passively included in the Primary Care Electronic Clinical record according to the Spanish health system, which is based on the principles of universality, free access, equity and fairness of financing.[18]

18. Bernal-Delgado E, García-Armesto S, Oliva J, Sánchez Martínez FI, Repullo JR, Peña-Longobardo LM, Ridao-López M, Hernández-Quevedo C. Spain: Health system review. Health Syst Transit, 2018;20(2):1–179.

- Please explain why people with certain diseases were excluded from the study.

We appreciate the Reviewer's question on a controversial issue. The Mollerussa cohort project is part of a broader project that also included the detection of carotid atherosclerosis to identify asymptomatic cardiovascular disease, and therefore known cardiovascular disease was an exclusion criteria. In addition, we carefully selected inclusion and exclusion criteria that have been published previously. Thus, we have excluded high risk subjects or those with conditions accepted as criteria for routine screening for diabetes as a usual protocol in our primary health care system. Now, we have included a sentence in the Discussion section.

Discussion section – fifth paragraph
Second, our study sample is probably healthier than the general population, as we excluded subjects with already known diabetes and other comorbidities, a lower number of subjects were counted in the denominator, thus resulting in a higher prevalence of this condition.

- "Sociodemographic variables were recorded, and a physical examination (weight, height, blood pressure and waist circumference) was carried out by researchers following a protocol for the inclusion of patients using a standardized baseline questionnaire for the clinical interview." – From this sentence, I am not sure whether weight, height, blood pressure and waist circumference were measured by researchers or self-reported during a personal interview. Please specify.

Following the Reviewer's comment, we have clarified this information by adding the following text in the Methods section.

Methods section – third paragraph
Sociodemographic variables were recorded by researchers following a protocol for the inclusion of patients using a standardized baseline questionnaire during the clinical interview. In all cases a physical examination (including weight, height, blood pressure and waist circumference) was carried out by trained research staff.

- Please describe continuous variables in table 1 either using mean and standard deviation or median and interquartile range. I cannot see, why for some variables the authors decided for the mean and for others chose the median.

We agree with the Reviewer's comment, and following this request, we have added information to clarify this issue in the Methods section:

Methods section – Statistical methods
Descriptive statistics of the mean (standard deviation) or median [interquartile range] were estimated for quantitative variables with a normal or non-normal distribution, respectively. Qualitative variables were assessed using absolute and relative frequencies. Normally distributed data were analysed using the Shapiro-Wilk test.

- "We observed a positive trend in age, BMI, waist circumference…" – What is meant by "positive trend"? Maybe use "association" instead.

Following the Reviewer's advice, we changed the term "a positive trend" to "an association" in the Results section.

- In the results section, much details are given with regard to different kinds of prediabetes (Hba1c, FPG or both). To me the relevance of this distinction is not apparent. I suggest to either reduce complexity by only reporting results for prediabetes (as was done in table 1) or explain the relevance of this distinction more clearly.

We appreciate the Reviewer's point of view. However, please note that, in line with other studies similar to the current one, the results are analysed according to the different definitions, as it has consistently been shown that the characteristics of the groups differ according to which of the variables is used to classify the prediabetic state. Indeed, the differences found are consistent with the findings of other studies. Therefore, we kindly request the Reviewer to allow us to keep this information in the manuscript. Please, see in Results and Discussion sections.

Results section
Patients with both abnormal FPG and HbA1c were older, had larger waist circumference, had increased FLI and HOMA2-IR, were more likely to be overweight or obese and have hypertension, and had lower HOMA2-S.

Discussion section
The prevalence of prediabetes was three-fold higher based on HbA1c than that based on FPG. Subjects with prediabetes defined by both HbA1c and FPG criteria had unfavourable clinical and sociodemographic profiles related to increased cardiovascular risk.

- "Third, although traditional factors such as hypertension, dyslipidaemia and obesity were included in the analysis models, the existence of unmeasured confounding variables cannot be entirely ruled out." – To me, this statement at the end of the manuscript is confusing, because the authors do not state which estimated effect could be confounded by unmeasured variables. In addition, if confounding of a hypothesized effect is a concern of the authors, I would expect that it is explicitly stated in methods section (i) which effect is aimed to be estimated, (ii) what the potentially confounding variables are and (iii) which procedures were used to adjust for confounding.

We appreciate the Reviewer's suggestion. This is a general statement assuming that there could be possible biases and confounding due to the existence of unknown or unmeasured variables. This illustrates an important limitation of observational studies: they can only include measurable and available variables. As this is a general comment that can lead to misunderstandings, we decided to remove this sentence from the text.

- In my opinion, the following conclusion drawn from this study is perhaps too strong: "…the

identification of individuals with prediabetes provides an opportunity for intervention through lifestyle modification and pharmacological treatments not only to reduce the development of diabetes but also to prevent the development of chronic complications." While this statement is probably true, I cannot see how the results of the study contribute to this conclusion.

We agree with the Reviewer that not this cannot a conclusion of our study. This is a general statement related to the relevance of the intervention in this healthy and difficult to reach population. Therefore, we have rephrased this part of the conclusions.

Conclusions section
"…the identification of individuals with prediabetes provides an opportunity for intervention through lifestyle modification and pharmacological treatments to reduce the development of diabetes."

VERSION 2 – REVIEW

REVIEWER	Thaddäus Tönnies German Diabetes Center, Germany
REVIEW RETURNED	15-Nov-2019

GENERAL COMMENTS	I thank the authors for their thoughtful revision which improved the manuscript substantially. However, there are still issues that need consideration. 1. One point regarding table 1 needs clarification. The authors revised table 1 and added two columns with the difference between the glycaemic status groups. The differences do not correspond to the values in the first three columns. For instance, in the row describing the proportion of women in each group, the numbers are 57.8%, 59.0% and 65.0% for normoglycaemia, prediabetes and diabetes, respectively. The difference between prediabetes vs. normoglycaemia and diabetes vs. normoglycaemia is then reported with 0.6 and 6.6, respectively. However, these differences should be 1.2 and 7.2. 2. My previous comment: In a next step, the authors performed a multivariable logistic regression using "the enter method with covariables that were clinically or statistically associated." Please clarify what the "enter method" is and what the exact definition of "clinically and statistically associated" was. In general, I think it would be preferable to include variables in the regression model based on prior subject matter knowledge and/or a biological model/hypothesis rather than including variables based on associations observed in the data. For instance, it would appear reasonable to include known risk factors for diabetes in the model. Author's response: We greatly appreciate this comment. Actually, we realized that the expression "enter method" is not necessary at all as it refers to the multivariable logistic regression methodology; thus, we removed this term. In addition, we also agree with the Reviewer on the need to specify those variables that are known to be related to the risk of the disease. In the Methods section, we only stated that "clinically and statistically associated" were included; this statement was clearly insufficient. Now, we have added in the Methods section the list of variables included in the model to clarify this to the reader. This list also includes well known risk factors for diabetes. The variables included in the

	model were: age, sex, education level, physical activity, dyslipidaemia, hypertension, family history of diabetes, BMI, waist, glomerular filtration rate and fatty liver index. My repsonse: In the manuscript, the authors still refer to the enter method in the method section: “Multivariate logistic regression models were used to determine the association of variables with prediabetes, isolated FPG, isolated HbA1c and both FPG and HbA1c at baseline and were performed using the enter method with covariables that were clinically or statistically associated.” 3. My previous comment:: For (ii), a “backward conditional logistic regression model was used to predict the normalization of the glycaemic state”. What is the reason for using a conditional logistic regression instead of a simple logistic regression? Again, I think prior subject matter knowledge and/or a biological model/hypothesis is better suited to select variables for the regression model. Author's response: To clarify this, first we used simple logistic regression that included variables based on prior subject matter knowledge and/or a biological model/hypothesis (complete model). Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. Thus, we used a stepwise regression to drop variables that did not improve the prediction performance (for distinct reasons: collinearity, non-applicability or low association). Now, we have added the complete model as a supplementary table in the Results section. My response: In the revised version of the manuscript, the authors still state that they performed a conditional backward logistic regression in the methods and the results section. In the response to my previous comment, the authors state that they performed a simple stepwise regression. Please clearly state in the manuscript, how the independent variables were selected (backward elimination or stepwise selection?) and whether a simple or conditional logistic regression was performed.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:Reviewer: 2Reviewer Name: Thaddäus TönniesInstitution and Country: German Diabetes Center, GermanyPlease state any competing interests or state ‘None declared’: None declared Please leave your comments for the authors belowI thank the authors for their thoughtful revision which improved the manuscript substantially. However, there are still issues that need consideration.1. One point regarding table 1 needs clarification. The authors revised table 1 and added two columns with the difference between the glycaemic status groups. The differences do not correspond to the values in the first three columns. For instance, in the row describing the proportion of women in each group, the numbers are 57.8%, 59.0% and 65.0% for normoglycaemia, prediabetes and diabetes, respectively. The difference between prediabetes vs. normoglycaemia and diabetes vs. normoglycaemia is then reported with 0.6 and 6.6, respectively. However, these differences should be 1.2 and 7.2.

We appreciate the Reviewer's comment. We have modified Table 1 describing the mean and standard deviation instead of the median and the interquartile range as before. The differences in the

variables between the glycaemic states are the difference between the mean and 95%confidence intervals. This changes have clarified the content of Table 1.2. My previous comment: In a next step, the authors performed a multivariable logistic regression using “the enter method with covariables that were clinically or statistically associated.” Please clarify what the “enter method” is and what the exact definition of “clinically and statistically associated” was. In general, I think it would be preferable to include variables in the regression model based on prior subject matter knowledge and/or a biological model/hypothesis rather than including variables based on associations observed in the data. For instance, it would appear reasonable to include known risk factors for diabetes in the model.Author's response: We greatly appreciate this comment. Actually, we realized that the expression "enter method" is not necessary at all as it refers to the multivariable logistic regression methodology; thus, we removed this term. In addition, we also agree with the Reviewer on the need to specify those variables that are known to be related to the risk of the disease. In the Methods section, we only stated that "clinically and statistically associated" were included; this statement was clearly insufficient. Now, we have added in the Methods section the list of variables included in the model to clarify this to the reader. This list also includes well known risk factors for diabetes. The variables included in the model were: age, sex, education level, physical activity, dyslipidaemia, hypertension, family history of diabetes, BMI, waist, glomerular filtration rate and fatty liver index.My repsonse: In the manuscript, the authors still refer to the enter method in the method section: “Multivariate logistic regression models were used to determine the association of variables with prediabetes, isolated FPG, isolated HbA1c and both FPG and HbA1c at baseline and were performed using the enter method with covariables that were clinically or statistically associated.”

We thank also the Reviewer for this observation. We have now removed this term from the Methods section.
This sentence now reads as follows:

Multivariate logistic regression models were used to determine the association of variables with prediabetes, isolated FPG, isolated HbA1c and both FPG and HbA1c at baseline with covariables that were clinically or statistically associated.

3. My previous comment: For (ii), a “backward conditional logistic regression model was used to predict the normalization of the glycaemic state”. What is the reason for using a conditional logistic regression instead of a simple logistic regression? Again, I think prior subject matter knowledge and/or a biological model/hypothesis is better suited to select variables for the regression model.Author's response: To clarify this, first we used simple logistic regression that included variables based on prior subject matter knowledge and/or a biological model/hypothesis (complete model). Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. Thus, we used a stepwise regression to drop variables that did not improve the prediction performance (for distinct reasons: collinearity, non-applicability or low association). Now, we have added the complete model as a supplementary table in the Results section.My response: In the revised version of the manuscript, the authors still state that they performed a conditional backward logistic regression in the methods and the results section. In the response to my previous comment, the authors state that they performed a simple stepwise regression. Please clearly state in the manuscript, how the independent variables were selected (backward elimination or stepwise selection?) and whether a simple or conditional logistic regression was performed.

We thank the Reviewer for this comment. We have now modified the sentences in the Methods and Results sections for a better understanding.

In the Methods section
A stepwise method with selection of variables by backward elimination was used to build the final

logistic regression model to predict the normalization of the glycaemic state.

In the Results section
Logistic regression model, as described in the methods section, starting with the variables age, sex, waist circumference, BMI, hypertension, physical activity, family history of diabetes, education level, total cholesterol, HDL-cholesterol, FLI and HOMA2-IR, was performed to identify factors independently associated with the prediction of glycaemic status normalization (Supplementary file 2 Table 4).

VERSION 3 – REVIEW

REVIEWER	Thaddäus Tönnies German Diabetes Center (DDZ), Germany
REVIEW RETURNED	13-Dec-2019

GENERAL COMMENTS	No further comments