

PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://bmjopen.bmj.com/site/about/resources/checklist.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: The US Veterans Administration Diabetes Risk (VADR) National Cohort
AUTHORS	Avramovic, Sanja; Alemi, Farrokh; Kanchi, Rania; Lopez, Priscilla; Hayes, Richard; Thorpe, Lorna; Schwartz, Mark

VERSION 1 – REVIEW

REVIEWER	Alyson Littman VA Puget Sound Health Care System and University of Washington, Seattle, WA United States
REVIEW RETURNED	30-Jun-2020

GENERAL COMMENTS	<p>This article provides a nice overview of a cohort study that will likely be of interest to many researchers and the larger scientific community. The details provided in the Appendix will be especially appreciated by those who plan to use the cohort or those who want to model a study off of this one. I think the paper could be strengthened by providing more details on a few planned analyses. The authors talk in very broad strokes about how the data may be used, but presumably, there are numerous studies underway. It would be helpful to provide some details about a few of those studies. My other comments are either minor or suggest providing more details.</p> <p>Abstract</p> <ul style="list-style-type: none">- Minor: Write out VADR in the purpose- Please more clearly state how follow-up worked. This may be a function of the headings used, that specify “Participants” and not “Methods” but it was not clear to me when follow-up began.- Minor: In the strengths and limitations section, revise “is” to “are” since data are plural <p>Introduction</p> <ul style="list-style-type: none">- The authors include reference #18 for evidence on the collaboration, but the paper referenced does not seem related to diabetes. Please clarify.- Consider revising the introduction to state what the cohort aims to do (e.g., in the future). It read a bit strangely to state what it is doing without providing support/references for those statements. <p>Cohort description</p> <ul style="list-style-type: none">- The authors state that patients were enrolled through December 31, 2016 and then later states that subject entry is ongoing. Please clarify. For this paper, was the last entry date in 2016 but that is not the case more broadly?
-------------------------	--

	<p>- Page 3, lines 13-15 – It is not clear what aspect of the definition the authors are referring to. Does this sentence refer to the exclusion of metformin or acarbose alone?</p> <p>- Page 3, line 17: The authors begin the sentence “for etiologic analyses” which makes it seem like it is contrast to other kinds of analyses. It would be helpful if the authors specified those analyses and how the eligibility criteria differed for them.</p> <p>Predictor variables and covariates</p> <p>- Do the authors consider predictor variables measured as far back as 1999? Could two measures be used that were say 15 years apart?</p> <p>- The authors note that they use the two most recent measures, or a single measure if only one is available. Please clarify the reason for using the average of the two most recent measures, rather than the most recent measure only.</p> <p>- The authors state that the “first” address on file per patient were used. Please clarify whether VA has a historic record of addresses. What was the time difference of these addresses compared to the date of cohort entry?</p> <p>Profile in a nutshell</p> <p>- The number of medical centers and community-based outpatient clinics listed here (151 and 800) is different than the number listed in the body of the paper (170 and more than 1,000). Please clarify which is correct.</p> <p>Table 1</p> <p>- Revise the third column header from “Veterans without incidence diabetes” to “Veterans without incident diabetes”</p> <p>- There is inconsistent precision of percentages – sometimes to the whole number and sometimes to the tenth. Please revise for consistency</p> <p>- For age, the authors have a separate row for the continuous variable, but this is not the approach for HbA1c or BMI. Consider revising to present as was done with age.</p> <p>- It is not clear when variables were measured. Please state this explicitly. This is especially important for a variable like HbA1c. My understanding was that this was measured prior to their diabetes diagnosis/onset of diabetes.</p> <p>- It would be helpful to explicitly state what percentage of persons are missing for each variable.</p> <p>Appendix</p> <p>- The detail provided in the appendix will be very appreciated for those who plan to use the data or evaluate studies derived from this cohort.</p> <p>- Some additional detail would be helpful. For example, the authors note that “implausible values removed” in several places but do not specify what the plausible range is or what publications served as a guide. Please add that information.</p>
--	---

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Alyson Littman

Institution and Country:

VA Puget Sound Health Care System and University of Washington, Seattle, WA
United States

Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below

This article provides a nice overview of a cohort study that will likely be of interest to many researchers and the larger scientific community. The details provided in the Appendix will be especially appreciated by those who plan to use the cohort or those who want to model a study off of this one. I think the paper could be strengthened by providing more details on a few planned analyses. The authors talk in very broad strokes about how the data may be used, but presumably, there are numerous studies underway. It would be helpful to provide some details about a few of those studies. My other comments are either minor or suggest providing more details.

We appreciate the reviewer's acknowledgement. We have added descriptions of analyses underway and planned at the end of the Introduction section.

Abstract

- Minor: Write out VADR in the purpose

We wrote out The Veterans Administration Diabetes Risk (VADR) cohort.

- Please more clearly state how follow-up worked. This may be a function of the headings used, that specify "Participants" and not "Methods" but it was not clear to me when follow-up began.

We clarified the follow-up period for the cohort subjects in the abstract in the Participants section.

- Minor: In the strengths and limitations section, revise "is" to "are" since data are plural

We revised "is" to "are."

Introduction

- The authors include reference #18 for evidence on the collaboration, but the paper referenced does not seem related to diabetes. Please clarify.

To show the collaboration, we referenced the Diabetes LEAD (Location, Environmental Attributes, and Disparities) Network, which is a CDC-funded collaboration between Drexel University, Geisinger-Johns Hopkins University, New York University School of Medicine, and University of Alabama at Birmingham. The coauthors of this submitted manuscript are collaborators on the CDC-funded study, and we are using the VADR Cohort for our analytic contributions to the Network. The primary goal of the Network is to further understanding of the role of community-level factors and geographic differences in diabetes incidence across the US and across demographic groups.

- Consider revising the introduction to state what the cohort aims to do (e.g., in the future). It read a bit strangely to state what it is doing without providing support/references for those statements.

As noted in our response above, this cohort is being used for the CDC-funded investigation of individual and community-level factors (e.g. local food, housing, and socioeconomic environment) and geographic differences in diabetes incidence in the US. We are also now funded to use the VADR cohort to study the interaction between the COVID-19 pandemic and outcomes among the more than 900,000 veterans with incident diabetes in the cohort.

Cohort description

- The authors state that patients were enrolled through December 31, 2016 and then later states that subject entry is ongoing. Please clarify. For this paper, was the last entry date in 2016 but that is not the case more broadly?

Thank you for pointing out this apparent discrepancy. Indeed, for this cohort, enrollment ended December 31, 2016 and we have chosen December 31, 2018 as the endpoint for follow-up in our analyses to ensure that records are complete and are less likely to change due to updates. However, since the VA Corporate Data Warehouse is updated daily, for future studies we are able to extend both the enrollment and follow-up periods indefinitely.

We have clarified these dates in the Cohort Description section of the manuscript.

- Page 3, lines 13-15 – It is not clear what aspect of the definition the authors are referring to. Does this sentence refer to the exclusion of metformin or acarbose alone?

The sentence refers to the overall definition, not just the exclusion of subjects prescribed metformin or acarbose alone. We clarified this in the manuscript.

- Page 3, line 17: The authors begin the sentence “for etiologic analyses” which makes it seem like it is contrast to other kinds of analyses. It would be helpful if the authors specified those analyses and how the eligibility criteria differed for them.

Thank you for your comment, we removed “for etiologic analyses” and instead of that the sentence starts with: “For the analytic cohort”.

Predictor variables and covariates

- Do the authors consider predictor variables measured as far back as 1999? Could two measures be used that were say 15 years apart?

For eligibility we went to 1999. For predictor variables, we took any measure prior to cohort entry date. If only one measure was available, we used that one value. If more than one was available, we took the average of the last two measures taken closest to the cohort entry date. The only variable that constrained to going back only 2 years, was blood pressure. We expanded the definition to averaging multiple weights as a sensitivity analysis and got similar values across the cohort for weight.

- The authors note that they use the two most recent measures, or a single measure if only one is available. Please clarify the reason for using the average of the two most recent measures, rather than the most recent measure only.

We used the average of the two most recent measures, because we wanted to minimize the influence of potentially aberrant extreme values. However, we also wanted to include patients with only 1 measure because we wanted to preserve the sample size and reduce missing data. This is a common decision in many studies, including NHANES studies on blood pressure, etc. We also did sensitivity analysis to confirm that including veterans with only 1 measure did not results in meaningful differences in average values for all comorbidities.

- The authors state that the “first” address on file per patient were used. Please clarify whether VA has a historic record of addresses. What was the time difference of these addresses compared to the date of cohort entry?

We have an historic record of addresses for most veterans in the cohort (99%). Of these, 83% (n=5,049,074) of veterans have a geocodable address; 41.13% (n=2,076,549) of which had an address before or on cohort entry date, and 58.8% (n=2,972,525) had an address after cohort entry date. A total of 56% had an address within 2 years of cohort entry date. For the purposes of characterizing the cohort, we used all veterans in the cohort and their first address on record. Decisions regarding use of dates and assigning exposure on a particular address in the historical record are made by researchers based on their specific research questions. For example, assigning neighborhood-based exposures can be assigned using the address closest to cohort entry date and steps can be taken to establish temporality.

Profile in a nutshell

- The number of medical centers and community-based outpatient clinics listed here (151 and 800) is different than the number listed in the body of the paper (170 and more than 1,000). Please clarify which is correct.

Thank you for highlighting this discrepancy. The numbers in the body of the paper are correct.

Table 1

- Revise the third column header from "Veterans without incidence diabetes" to "Veterans without incident diabetes"

Thank you for your comment, we revised the column header.

- There is inconsistent precision of percentages – sometimes to the whole number and sometimes to the tenth. Please revise for consistency

Thank you for your comment, we revised the table so the precision is the same throughout.

- For age, the authors have a separate row for the continuous variable, but this is not the approach for HbA1c or BMI. Consider revising to present as was done with age.

Thank you for your comment, we revised the table as suggested.

- It is not clear when variables were measured. Please state this explicitly. This is especially important for a variable like HbA1c. My understanding was that this was measured prior to their diabetes diagnosis/onset of diabetes.

All variables in the table were measured when the patient entered the cohort. We changed the title of the table to clarify this as, "Cohort demographics and clinical characteristics at cohort entry by incident diabetes status."

- It would be helpful to explicitly state what percentage of persons are missing for each variable.

Thank you for your comment. We added the number and percentage "Not measured or Missing" for each variable in the Table 1.

Appendix

- The detail provided in the appendix will be very appreciated for those who plan to use the data or evaluate studies derived from this cohort.

Thank you, we hope so as well.

- Some additional detail would be helpful. For example, the authors note that “implausible values removed” in several places but do not specify what the plausible range is or what publications served as a guide. Please add that information.

Thank you for your comment. For lipids data cleaning (dealing with multiple measure in the same time) we followed: <https://www.ahajournals.org/doi/full/10.1161/JAHA.118.011869>"

For weight and height implausible measures and data cleaning (multiple measures at the same day) we followed: <https://academic.oup.com/milmed/article/179/10/1119/4159656>"

To determine implausible measures of lipids, A1C, BP, ALT and AST, we followed NHANES protocols (<https://wwwn.cdc.gov/Nchs/Nhanes/>)