

# Harnessing registry data to identify socio-demographic and socio-economic gaps in HIV care in the Netherlands

Corresponding Author: Dr Vita Jongen

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper demonstrates the utility of linkage of population health and socio-demographic data to the ATHENA HIV cohort. While data linkages of HIV cohorts have previously been performed, including in the assessment of continuum of care (<https://doi.org/10.7448/IAS.18.1.20634>), this analysis provides a focus on association of broader socio-economic and other health related determinants. This method and findings are also comparable to recently published work regarding loss to follow up in HCV care (DOI: 10.1111/liv.15729). While there is a substantial body of work highlighting the importance of socio-economic determinants in HIV outcomes (DOI: 10.1016/S2468-2667(16)30002-0, 10.1057/s41599-024-04121-y, 10.23889/ijpds.v10i1.2496), this publication substantiates those findings in a large well characterised population.

The Statistics Netherlands data provides rich whole of population data. The data extracted for analysis is the “most” recent. The rationale of the paper to identify characteristics that support continued engagement would be better served by analysing data as close to, or prior to, HIV diagnosis. This would obviate possibly reverse causation arising from decline in socio-economic status, and increased stigma and discrimination that may follow HIV diagnosis. Additionally, assessment of impact of change in these factors over-time would be valuable.

The authors note that younger age had poorer linkage quality, as has been documented in other data linkage papers (DOI: 10.1186/s12913-018-3495-x). Considering this is primary finding of the paper sensitivity analysis of the impact of poor linkage should be considered.

The linkage methodology is only briefly described. Additional information should be provided including who performed the linkage. Does consent to ATHENA include consent for data linkage?

The analysis is stratified according to MSM, cis heterosexual men and women. While this is common in HIV research, it would be good to provide additional context to the rationale for these groupings recognising the conflation of gender and sexual behaviour.

Specific comments

- was there sufficient power to assess socio-economic determinants of disengagement from care among women?
- consider a sensitivity analysis of  $\leq 1000$  copies HIV-RNA (WHO definition of suppressed VLD) or discuss how many people fell between 200 and 1000 copies and possible impact on findings.
- consider a sensitivity analysis of including deaths as failures.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

This is an interesting paper that analyses the cascade of care (focusing on viral load suppression and engagement in care) among HIV positive patients from the Athena cohort in the Netherlands. Data from the cohort are linked probabilistically with registry data from Statistics Netherlands to get more detailed information on education, socio-economic information, migration background and household composition.

The paper is well written but has a number of limitations as outlined below.

Not much is presented on the probabilistic linkage:

Both files should be described in more detail; a short description of the Statistics Office registry should be provided (how were data collected when; are there any updates done over time?). At what time point were the data from the statistics office collected; since some of the variables may change over time this should be described in detail and also any potential limitations in terms of the timing should be discussed.

Linkage procedure and data management should be described in detail: information on cleaning, blocking weights and threshold should be provided. How does the choice of threshold affect the results? How were missing values handled?

Quantify linkage quality with validation and flow diagrams

Software used for linkage

It is not clear why transgender individuals and those with HIV2 were excluded; please explain

Provide a brief rationale for using Bayesian logistic regression (versus a normal logistic regression approach). How were missing variables handled at the analysis stage?

The variables presented in table 1 and the variables eventually included in the analysis seem not to be the same. Were there any additional variable selections done and how was this done? This should be described in detail (including also any transformation and categorization of variables). While collinearity prevented the inclusion of all variables in the same model initially, I was still wondering why at the end not also a combined model was developed.

The split into one model that includes only socio-demographic and economic variables and the other one including only health-related variables seems a bit arbitrary. Also, the second model does not even include some basic information such as age (that is usually included almost in any regression analysis as a consequence).

Why was age included linearly? Did you also consider the use of splines of age categories?

I am not sure it is novel enough to be published in Nature Communication but would see it more appropriate for a more specialized journal.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

All requests for revision have been thoroughly addressed and incorporated in to the manuscript appropriately. Our acknowledgement to the authors for their rigour.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

The authors have satisfactorily addressed all comments. I would therefore recommend the article for publication.

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

**Reviewer #1:**

1. This paper demonstrates the utility of linkage of population health and socio-demographic data to the ATHENA HIV cohort. While data linkages of HIV cohorts have previously been performed, including in the assessment of continuum of care (<https://doi.org/10.7448/IAS.18.1.20634>), this analysis provides a focus on association of broader socio-economic and other health related determinants. This method and findings are also comparable to recently published work regarding loss to follow up in HCV care (DOI: 10.1111/liv.15729). While there is a substantial body of work highlighting the importance of socio-economic determinants in HIV outcomes (DOI: 10.1016/S2468-2667(16)30002-0, 10.1057/s41599-024-04121-y, 10.23889/ijpds.v10i1.2496), this publication substantiates those findings in a large well characterised population.

**Re. We would like to thank the reviewer for their helpful assessment of our manuscript. We would also like to thank the reviewer for pointing out these important references.**

**In the meta-analysis by Medland NA, et al (<https://doi.org/10.7448/IAS.18.1.20634>), the linkage to data was mostly on HIV-related or clinical outcomes and not socioeconomic factors. We agree that the results from the ASTRA study have been very insightful with the respect to certain indicators of poverty (reference 10.1016/S2468-2667(16)30002-0); however, as this study relies on questionnaire data, there is inherent response bias and the risk of not including those with severely disadvantaged backgrounds is likely higher than in our study. The study by Dinçer M, et al (reference 10.1057/s41599-024-04121-y) does provide inference on mostly income and human capital related to HIV related deaths, but these results are from an ecological study and risks certain biases without using individual level data. Lastly, the study from the COAST cohort (10.23889/ijpds.v10i1.2496) of individuals with and without HIV has a very similar format as to our study, yet this was a cohort profile and to our knowledge, the main results have not yet been published. This cohort profile does demonstrate the increasing importance of more comprehensively evaluating the socio-economic determinants of HIV care.**

**We agree that our study builds on these previous findings, but at the same time, our study has a more comprehensive set of socioeconomic and health indicators, has a wider representation of individuals with HIV at country level, and is the most timely to report on these outcomes.**

2. The Statistics Netherlands data provides rich whole of population data. The data extracted for analysis is the “most” recent. The rationale of the paper to identify characteristics that support continued engagement would be better served by analysing data as close to, or prior to, HIV diagnosis. This would obviate possibly reverse causation arising from decline in socio-economic status, and increased stigma and discrimination that may follow HIV diagnosis. Additionally, assessment of impact of change in these factors over-time would be valuable.

Re. We thank the reviewer for this suggestion. Our logic was that by including the most recent visit, we would be able to understand the current state of individuals who are not in care, yet still alive in 2023. We fully agree that this type of analysis does not consider the changes over time after HIV diagnosis and could represent reverse causation more than anything else.

To address this concern, we conducted a sensitivity analysis in which we modeled the time until disengagement into care with time-updated variables after HIV diagnosis. We also included time since HIV diagnosis into the model. Since the dataset from Statistics Netherlands only included data from 2012 onward, we could only include a subset of individuals diagnosed after 2012. These analyses did identify a broader set of statistically significant determinants of disengagement from care, likely because of time-varying risks for certain exposures and increased statistical power. Nevertheless, the associations from the main analysis (please refer to the comments below) had similar directions compared to this sensitivity analyses.

We have included this sensitivity analysis in the Methods, Results, Supplementary Table 4, and limitation section of the Discussion.

Original text	Revised text, Methods (p 13)
-	In the second sensitivity analysis, we modelled the time from HIV diagnosis to disengagement from care with time-updated covariates. The hazard ratio (HR) comparing the hazards across levels of determinants were obtained along with its 95% CI using a piecewise exponential survival model. Due to data availability, only individuals diagnosed with HIV after 2012 could be included in this analysis.
Original text	Revised text, Results (p 6)
-	In the sensitivity analysis assessing the time to disengagement from care among individuals diagnosed with HIV after 2012, the piecewise exponential survival model identified a broader set of significant exposures associated with disengagement from care (Supplementary Table 4). For all three populations, increased years since HIV diagnosis, younger age, and a low-middle income or income below the poverty line

	<p>were associated with a higher rate of disengagement from care. For MSM and cisgender heterosexual men, only having primary education, a first or second generation migration background, or being a single parent or living institutionalized or in another type of household were associated with a higher rate of disengagement from care. For MSM, receiving social welfare and using antidepressants were associated with a lower rate of disengagement from care.</p>
<b>Original text</b>	<b>Revised text, Discussion (p 9)</b>
-	<p>Fourth, our main analysis used cross-sectional data from 2023. If, for example, there were declines in income or changes in employment status after HIV diagnosis, which then contributed to an increased risk of disengagement from care, the temporality of exposure on outcome would be unclear in a cross-sectional analysis. In a sensitivity analysis using longitudinal data on time from HIV diagnosis to disengagement from care and time-updated variables, a broader set of determinants were identified, but the direction of these associations (particularly the key associations age and income) was consistent among MSM and cisgender heterosexual men. Notably, income was not associated with disengagement from care among women in the cross-sectional analysis, but it was in the longitudinal analysis.</p>

3. The authors note that younger age had poorer linkage quality, as has been documented in other data linkage papers (DOI: 10.1186/s12913-018-3495-x). Considering this is primary finding of the paper sensitivity analysis of the impact of poor linkage should be considered.

Re. We thank the reviewer for pointing out this pertinent article. Similar to that analysis , younger individuals were indeed less likely to be linked to data from Statistics Netherlands. Sensitivity analysis for this type of issue is exceedingly difficult as it relies on completely non-missing data (or in the case of the cited article, nearly perfectly linked data). Unfortunately, these data are not at our disposition. To correct for the potential bias introduced by differential linkage, we redid our analyses using a Heckman probit model instead of Bayesian logistic regression. The Heckman probit model jointly estimates the probability of being successfully linked to registry data from Statistics Netherlands (in a "selection" equation) and outcomes (i.e., not being virally suppressed and disengagement from care).

In the selection equation we included age, being born in the Netherlands, and level of urbanization of residence. In our initial analysis, we had to separate determinants by socio-demographic and socio-economic factors in one model and health related factors in another model due to computational issues of the model. With the Heckman probit model, this issue is no longer present and we are able to include all variables in a single model. While the results from these new models differed slightly from the original analysis, the main results remained comparable. We have adapted our Methods and Results to reflect these changes:

Original text	Revised text, Methods (p 12-13)
We modeled and assessed socio-demographic, -economic, and health related determinants of not achieving two major milestones of the HIV care continuum, i.e., suppressed viral load (HIV-1 RNA <200 copies/mL) and engagement in care (at least one HIV care visit in 2023), using multivariable Bayesian logistic regression. For each determinant, we specified a prior distribution of odds ratios (ORs) as uniform (i.e., noninformative). Using this prior together with the data, we estimated the a posterior distribution of ORs with Markov Chain Monte Carlo (MCMC) methods from the "bayes" prefix commands in Stata. The median of this distribution defined the parameter estimate (termed "posterior OR") and the 2.5% and 97.5 % quantiles defined the 95% credible intervals (CrI), quantifying the uncertainty in the parameter estimate. As gender and sexual preference largely affected outcomes, analyses were stratified by key populations	We assessed socio-demographic, -economic, and health related determinants of not achieving two major milestones of the HIV care continuum, i.e., suppressed viral load (HIV-1 RNA <200 copies/mL) and engagement in care (at least one HIV care visit in 2023), <b>using multivariable Heckman probit regression. This model jointly estimates the probability of being successfully linked to the registry data from Statistics Netherlands (in a "selection" equation) and the outcomes (in an "outcome" equation), and hence reduces selection bias from individuals who were not linked to data from Statistics Netherlands. Age, being born in the Netherlands, and level of urbanization of residence were included in the selection equation. We added individual covariates to the outcome equation to obtain univariable coefficients and 95% confidence intervals (CI) comparing the probability of having the outcome across</b>

<p>[i.e., men who have sex with men (MSM), women, and heterosexual cis-gender men]. When including all variables in the multivariable model, high autocorrelation between MCMC runs was observed for most parameter estimates. Therefore, we constructed two multivariable models, which reduced this autocorrelation, including (i) only socio-demographic and -economic variables, and (ii) only health related variables.</p>	<p>levels of covariates. Missing values were included in the models as a separate category. Determinants with <math>p &lt; 0.2</math> in univariable analysis were included in a full multivariable model. Variables that did not significantly improve the model fit based on the likelihood ratio test (<math>p &gt; 0.05</math>) were removed sequentially from the model in backwards, stepwise fashion. Age was included <i>a priori</i> in both the selection and outcome equations in the multivariable model.</p> <p>We stratified analyses by key population based on sex assigned at birth and probable mode of HIV acquisition, as the epidemiology, prevention, and care of HIV has been known to differ between these key populations.<sup>5</sup> Specifically, we categorized individuals as men who have sex with men (MSM, assigned male at birth, not transgender, and likely acquired HIV through sex with another man), cisgender women (assigned woman at birth, not transgender), or heterosexual cisgender men (assigned male at birth, not transgender, acquired HIV through sex with a woman).</p>
--	---

4. The linkage methodology is only briefly described. Additional information should be provided including who performed the linkage. Does consent to ATHENA include consent for data linkage?

**Re. Our apologies for the lack of details in the previous version of the manuscript. Statistics Netherlands possesses continuously updated data collected through municipal registries, tax systems and healthcare reimbursement records. These data are updated several times a year and for this study, we used the most recently available data (2023). Statistics Netherlands releases data in the form of so-called “versions”, which guarantees reproducibility of research results. We have listed the versions of the files used in our manuscript in the Supplementary Materials.**



Probabilistic linkage between datasets from the ATHENA and Statistics Netherlands registries was based on date of birth, sex at birth, and four digits of the postal code. Statistics Netherlands performed exact matching. Any linkage error was the result of measurement error (e.g., mis-registered data in one of the data registries) or the inability to derive an exact match (e.g., two people with the exact same birth and sex registered at a single postal code). Any data with linkage error were discarded. Data linkage was performed by Statistics Netherlands within their secure environment. We have added this information to the Methods.

Participants of the ATHENA cohort provided consent to use their data for research purposes and for data linkage to other sources (<https://www.hiv-monitoring.nl/en/what-we-do/information-people-living-hiv/patient-information-sheet>). Information for participants about the currently active data linkages within SHM can also be found on our website (<https://www.hiv-monitoring.nl/en/research-using-our-data/datakoppelingen>). We have now added this to the Methods section of the manuscript.

Original text	Revised text (p 10-11)
Data from the ATHENA cohort were linked to data from Statistical Netherlands using a probabilistic approach based on individual date of birth, postal code of last known residence and sex at birth.	Data from the ATHENA cohort were uploaded to the secure Remote Access environment hosted by Statistics Netherlands. Data linkage between data from ATHENA and microdata from Statistics Netherlands was facilitated by Statistics Netherlands using a probabilistic approach based on individual date of birth, first four digits of the postal code of last known residence and sex at birth. Statistics Netherlands performed exact matching. Any linkage error would be the result of measurement error (e.g., mis-registered data in one of the data registries) or the inability to perform exact matching (e.g., two people with the exact same date of birth and sex registered at a single postal code). Any data with linkage error were discarded. Participants from the ATHENA cohort provided consent for use of their data for data linkage purposes and information about active data linkages is available on the SHM website. The Remote Access environment is only available for researchers authorized by SHM and Statistics Netherlands. All output

	from the Remote Access environment is independently verified by Statistics Netherlands to ensure data cannot be traced back to individuals.
--	---

5. The analysis is stratified according to MSM, cis heterosexual men and women. While this is common in HIV research, it would be good to provide additional context to the rationale for these groupings recognising the conflation of gender and sexual behaviour.

**Re. Agreed. To be clear, we constructed key populations from variables collected within the ATHENA Cohort (i.e., sex assigned at birth and probable mode of HIV acquisition). This categorization is used in our government-commissioned report on HIV in the Netherlands as well as commonly used in Europe, mainly because these key populations are known to differ in HIV epidemiology, HIV prevention and care for this context. We have included further rationale for this categorization to the Methods.**

**We should also mention that obtaining information on sexual identify and behaviors is uncommon in large, clinical-based cohorts and is a limitation shared across most large-scale studies. This lack of data makes it difficult to provide variables that fully encompass the complexity of gender and sexual identity. We have acknowledged this issue in the limitations of the Discussion.**

**The following changes have been made:**

Original text	Revised text (p 13)
As gender and sexual preference largely affected outcomes, analyses were stratified by key populations [i.e., men who have sex with men (MSM), women, and heterosexual cis-gender men].	We stratified analyses by key population based on sex assigned at birth and probable mode of HIV acquisition, as the epidemiology, prevention and care of HIV has been known to differ between these key populations. <sup>5</sup> Specifically, we categorized individuals as men who have sex with men (MSM, assigned male at birth, not transgender, and likely acquired HIV through sex with another man), cisgender women (assigned woman at birth, not transgender), or heterosexual cisgender men (assigned

	male at birth, not transgender, acquired HIV through sex with a woman).
Original text	Revised text (p 8-9)
Third, we cannot regroup individuals to specific key subpopulations (e.g., MSM who inject drugs), hence these results only represent the overall key populations and may not be generalizable to specific subpopulations.	Third, we cannot regroup individuals to specific key subpopulations (e.g., MSM who inject drugs), hence these results only represent the overall key populations and may not be generalizable to specific subpopulations. Furthermore, these groupings could conflate gender identity, sexual orientation and sexual behavior. Future research could incorporate more inclusive and nuanced measures of gender identity and sexuality to more specifically reflect the experiences of all individuals.

#### Specific comments

- was there sufficient power to assess socio-economic determinants of disengagement from care among women?

**Re. Indeed, there could be a risk of insufficient power for this specific group. We prefer not to perform *post hoc* power calculations, namely because the parameter estimate is itself not a fixed entity (i.e., noisy) and any power calculation from this entity will have high uncertainty (see Hoenig JM, Heisey DM, The American Statistician, 2001; <https://statmodeling.stat.columbia.edu/2018/09/24/dont-calculate-post-hoc-power-using-observed-estimate-effect-size/>). We have added this as a limitation:**

Original text	Revised text (p 9)
-	Fifth, there could have been insufficient statistical power to identify determinants in some subgroups, namely women.

-consider a sensitivity analysis of  $\leq 1000$  copies HIV-RNA (WHO definition of suppressed VLD) or discuss how many people fell between 200 and 1000 copies and possible impact on findings.

**Re: Agreed and thanks for this suggestion. We have added a sensitivity analysis in which we defined viral suppression at an HIV RNA  $< 1000$  copies/mL. Using this definition resulted in 833**

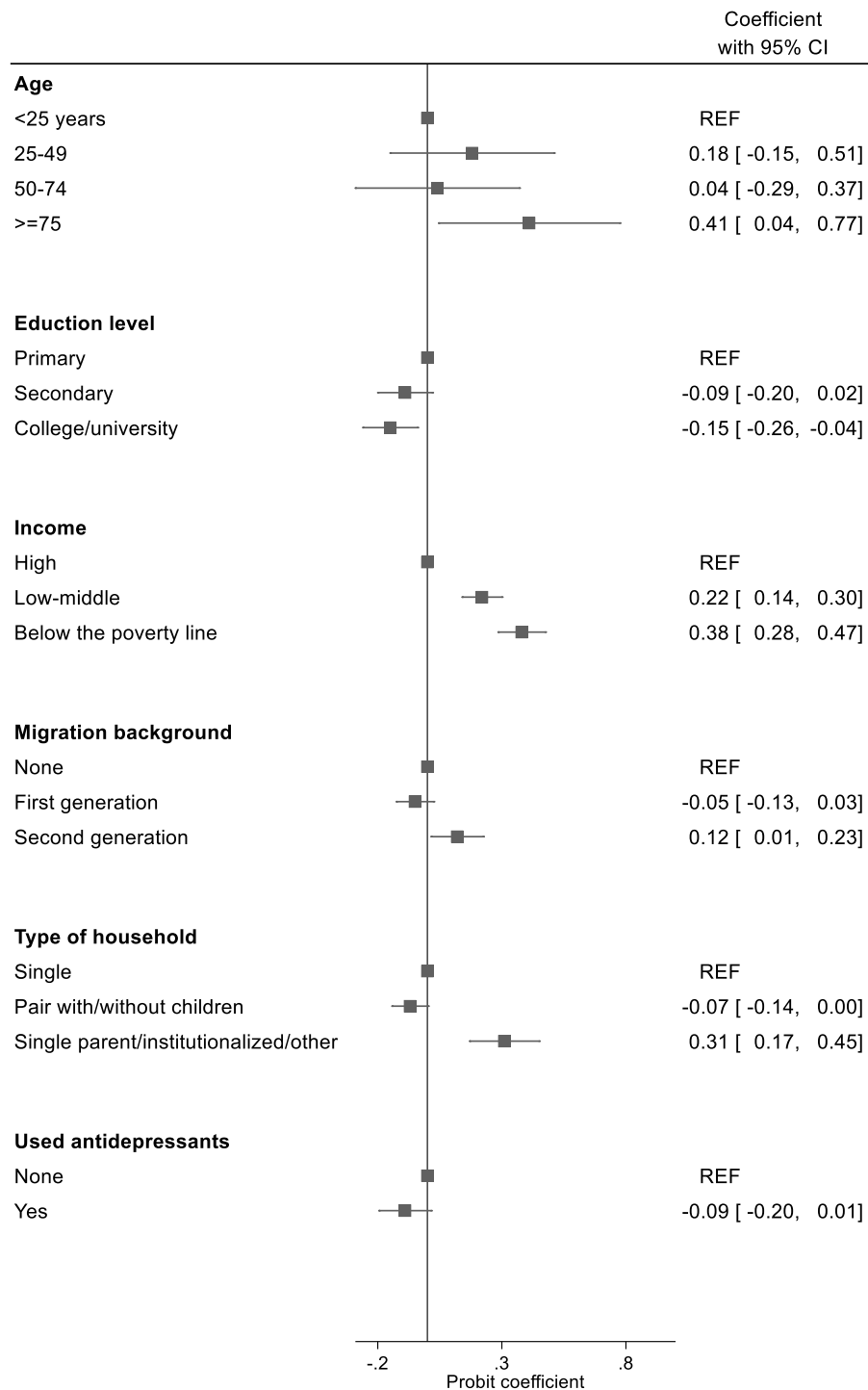
MSM, 398 women, and 413 cis-gender heterosexual men who were not virally suppressed. Changing this threshold, however, did not result in substantial changes compared to the main analysis (Supplementary Table 3). We have added this analysis to the Methods and Results.

Original text	Revised text, Methods (p 13)
-	In the second sensitivity analysis, we increased the threshold of viral suppression to HIV-1 RNA <1000 copies/mL. <sup>1</sup>
Original text	Revised text, Results (p 6)
-	When increasing the threshold for viral suppression to HIV-1 RNA<1000 copies/mL, 833 MSM, 413 cisgender heterosexual men, and 398 women had a detectable viral load. Increasing the threshold slightly changed the results from the main analysis for MSM and cis-gender heterosexual men. For MSM, age and use of antidepressants were no longer associated with an increased probability of a detectable viral load in MSM (Supplementary Table 3). For cis-gender heterosexual men, having only primary education and no migration background were associated with increased probability of a detectable viral load.

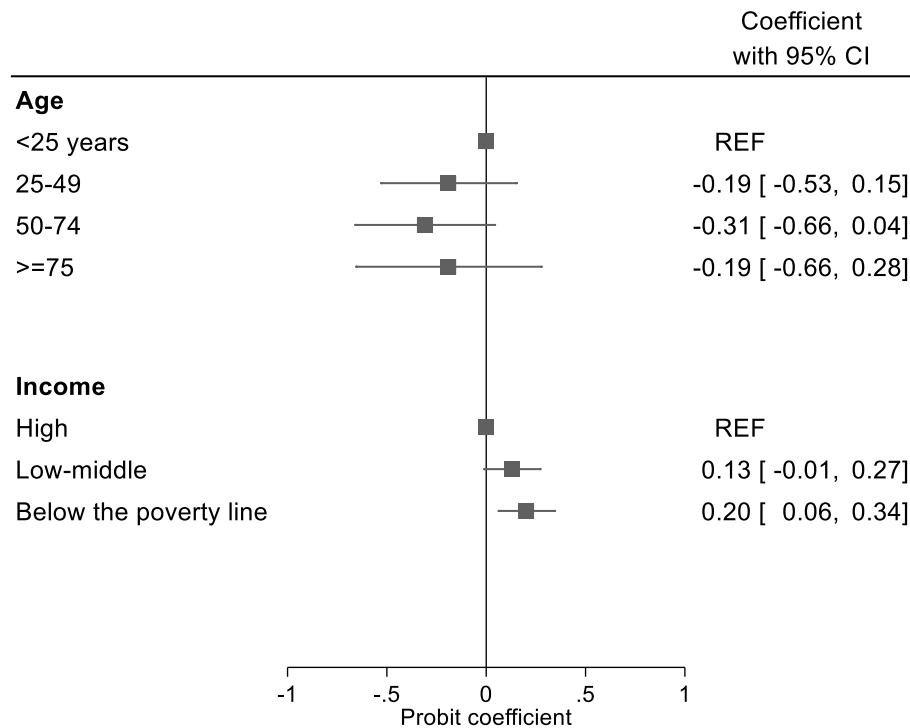
-consider a sensitivity analysis of including deaths as failures.

Re. Thank you for this suggestion. We conducted an analysis including deaths as part of a composite outcome with not being virally suppressed:

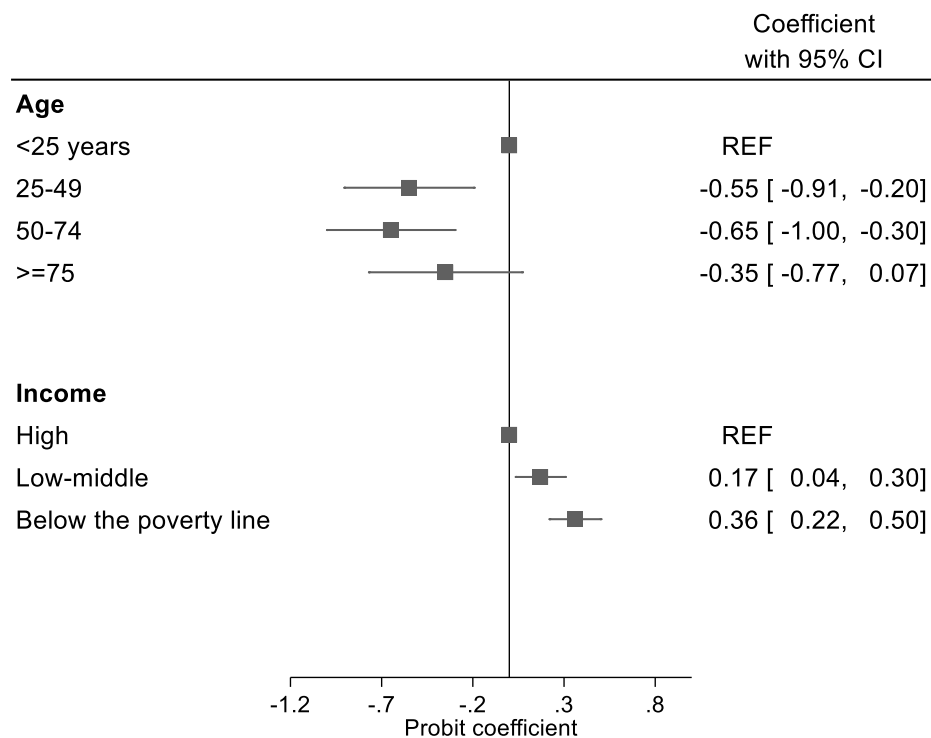
## Men who have sex with men



## Women



## Cis-gender heterosexual men



**Rebuttal Figure 1. Socio-demographic, economic, and health related determinants of having a detectable viral load for MSM (A), women (B), and cisgender heterosexual men (C). A detectable viral load was defined as having a HIV RNA >200 copies/mL or having died.**

The change in outcome resulted in an additional 1,016 failures for MSM, 470 for women and 512 for heterosexual cis-gender men. The results from the multivariable models using this composite endpoint were largely comparable to the main analysis, with no noticeable changes in effect estimates.

We should point out that the majority of the deaths are unlikely to be the result of virological failure and more related to chronic comorbidities common to ageing populations).<sup>2</sup> We used no virological suppression to highlight challenges associated with HIV-related care and including deaths might mask this relationship. For this reason, we decided not to include this analysis in the manuscript.

**Reviewer #2:**

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

**Re: Thank you for reviewing our manuscript and helping early career researchers to become more perceptive peer reviewers.**



**Reviewer #3:**

1. This is an interesting paper that analyses the cascade of care (focusing on viral load suppression and engagement in care) among HIV positive patients from the Athena cohort in the Netherlands. Data from the cohort are linked probabilistically with registry data from Statistics Netherlands to get more detailed information on education, socio-economic information, migration background and household composition.

The paper is well written but has a number of limitations as outlined below.

**Re. Thank you for taking the time to review our manuscript and the constructive feedback, particularly on the methods of this analysis.**

2. Not much is presented on the probabilistic linkage:

Both files should be described in more detail; a short description of the Statistics Office registry should be provided (how were data collected when; are there any updates done over time?). At what time point were the data from the statistics office collected; since some of the variables may change over time this should be described in detail and also any potential limitations in terms of the timing should be discussed.

Linkage procedure and data management should be described in detail: information on cleaning, blocking weights and threshold should be provided. How does the choice of threshold affect the results? How were missing values handled?

Quantify linkage quality with validation and flow diagrams

Software used for linkage

**Re. Linkage was performed between the ATHENA cohort and individual-level data from Statistics Netherlands. Statistics Netherlands possesses continuously updated data collected through municipal registries, tax systems and healthcare reimbursement records. These data are updated several times a year and for this study, we used the most recently available data (2023). Statistics Netherlands releases data in the form of so-called “versions”, which guarantees reproducibility of research results. We have listed the versions of the files used in our manuscript in the Supplementary Materials (Appendix A).**

**Probabilistic linkage between datasets from the ATHENA and Statistics Netherlands registries was based on date of birth, sex at birth, and four digits of the postal code. CBS performed exact matching. Any linkage error was the result of measurement error (e.g., mis-registered data in one of the data registries) or the inability to derive an exact match (e.g., two people with the exact same birth and sex registered at a single postal code). Any data with linkage error were discarded. Data linkage was performed by Statistics Netherlands within their secure environment. We have added this information to the Methods.**

Please also see our response to Reviewer 1, comment 4 with respect to linkage error.

The quality of linkage was assessed by comparing characteristics (collected in the ATHENA cohort) of individuals who could and who could not be linked. As shown in Supplementary Table 1, these individuals differ on some points. Most importantly, individuals not matched were younger and less often had a suppressed viral load in 2023. To correct for this, we employed a Heckman probit model, which is further explained in our response to Reviewer 3, comment 3.

3. It is not clear why transgender individuals and those with HIV2 were excluded; please explain

Re. Our apologies. One of the stipulations for reporting results from Statistics Netherlands is that the risk of identification must be kept at a minimum. As a result, they have requirements that we have to respect to ensure minimum risk of identification (e.g., no fewer than 10 individuals in a denominator of a summary statistic). There were only 331 transgender individuals with HIV in care as of 2023,<sup>2</sup> which made it nearly impossible to respect these reporting guidelines. We believed it best not to include these individuals.

Similarly, the number of individuals with HIV-2 is small (59 people with HIV-2 were still in care by the end of 2023)<sup>2</sup>. Hence we encountered the same problem with identification. Moreover, the clinical progression and treatment recommendations are very different from HIV-1, which makes some of the pillars of the cascade-of-care less evident.

4. Provide a brief rationale for using Bayesian logistic regression (versus a normal logistic regression approach). How were missing variables handled at the analysis stage?

Re. The choice for a Bayesian versus frequentist approach was more a preference than anything else. We believed that many of the frequentist hypotheses would be overpowered for this analysis, making hypothesis testing a clinically meaningless exercise. Since (1) the analysis has greatly changed since the last version (i.e., included a Heckman probit model instead of logistic regression, please refer to comment 3 from Reviewer 1) and (2) future analyses from this collaboration will be using frequentist methods and we prefer keeping this consistent, we are now using a frequentist approach.

Observations with missing data were included in the regression model as a separate category. We have included this in the Methods section:

Original text	Revised text, Methods (p 12)
---------------	------------------------------

-	Missing values were included in the models as a separate category.
---	--

5. The variables presented in table 1 and the variables eventually included in the analysis seem not to be the same. Were there any additional variable selections done and how was this done? This should be described in detail (including also any transformation and categorization of variables). While collinearity prevented the inclusion of all variables in the same model initially, I was still wondering why at the end not also a combined model was developed.

**Re. Our apologies for the confusion between tables. The variables included in the final model did indeed differ from Table 1.**

**From our revised analysis approach, we have reported all variables included in the univariable and multivariable regression models in the Tables (Table 2 and Supplementary Tables 2-4). We have also included footnotes under the Figure 1 and the Tables showing the variables included in the initial multivariable model, before backwards selection was employed.**

**We included the following footnotes to Table 2 and Figure 1:**

Original text	Revised text, footnote Table 2
-	<ol style="list-style-type: none"> <li>1. Education level, migration background, type of household, income, receipt of social welfare, stage at HIV diagnosis, use of mental health care, and use of antidepressants were included in the initial multivariable model for MSM. Age was forced into the model.</li> <li>2. Education level, migration background, type of household, and income were included in the initial multivariable model for cisgender heterosexual men. Age was forced into the model.</li> <li>3. Migration background, year of HIV diagnosis, and use of antidepressants were included in the initial</li> </ol>

	multivariable model for women. Age was forced into the model.
	<b>Revised text, footnote Figure 1</b>
	<i>For MSM, age, education level, migration background, type of household, income, receipt of social welfare, year of HIV diagnosis, stage at HIV diagnosis, use of antidepressants, and use of anti-psychotic medication were included in the initial multivariable model. For cisgender heterosexual men, age, education level, migration background, type of household, income, year of HIV diagnosis, and use of antidepressants were included in the initial multivariable model. For women, age, migration background, type of household, and income were included in the initial multivariable model.</i>

6. The split into one model that includes only socio-demographic and economic variables and the other one including only health-related variables seems a bit arbitrary. Also, the second model does not even include some basic information such as age (that is usually included almost in any regression analysis as a consequence).

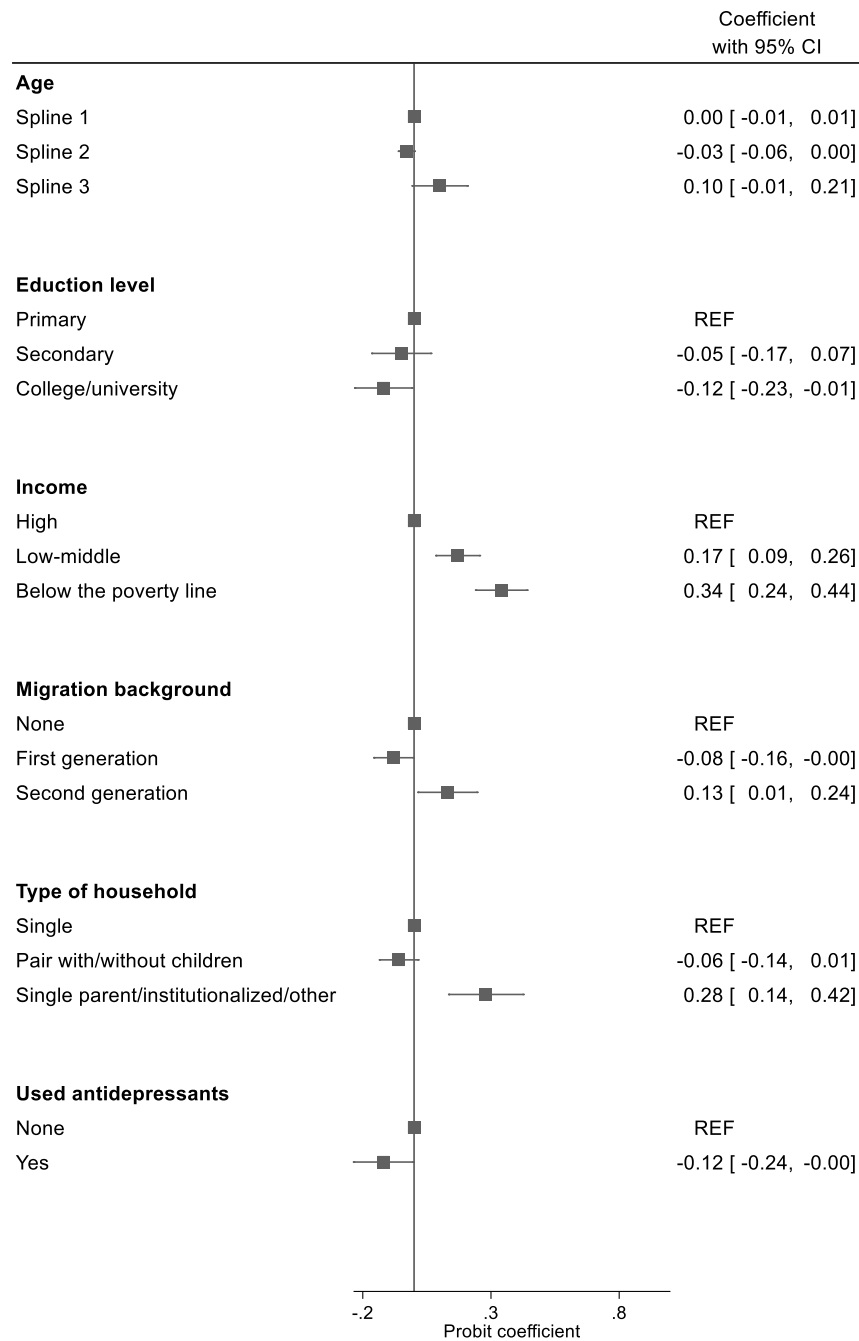
**Re. We appreciate this comment. In our initial analyses (Bayesian logistic regression), the model had excessive autocorrelation between runs of the MCMC for obtaining a posterior ORs. To avoid this issue, we decided to assess the different sets of variables in two distinct models that (1) could represent different pathways through which determinants may influence outcomes and (2) resulted in no autocorrelation.**

**Given the various suggestions from the Reviewers, we used a different approach that made it no longer necessary to separate determinants in different models. We now report univariable and multivariable models including all covariates.**

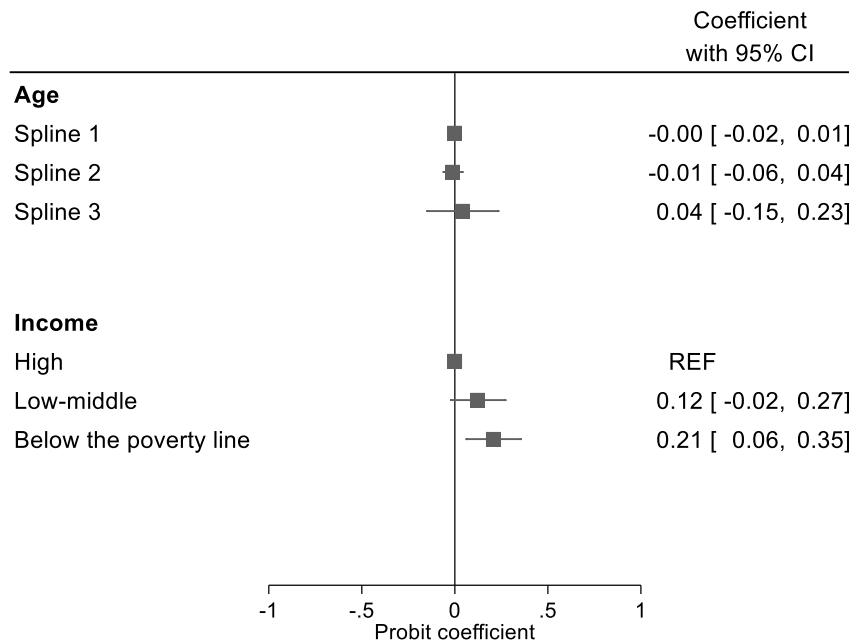
7. Why was age included linearly? Did you also consider the use of splines of age categories?

**Re. Thank you for this suggestion. We have included age as cubic splines (with 4 knots at the 20<sup>th</sup>, 40<sup>th</sup>, 60<sup>th</sup>, and 80th percentile) to assess its impact on the multivariable model. As shown in the Figures below, the coefficients of the other variables barely changed with cubic-spline transformed age. Although we understand the importance of improving statistical efficiency with the use of splines, it does hamper interpretability of the parameter estimate. Taken together, we opted to report estimates from the model without age as cubic splines, but rather age as a categorical variable.**

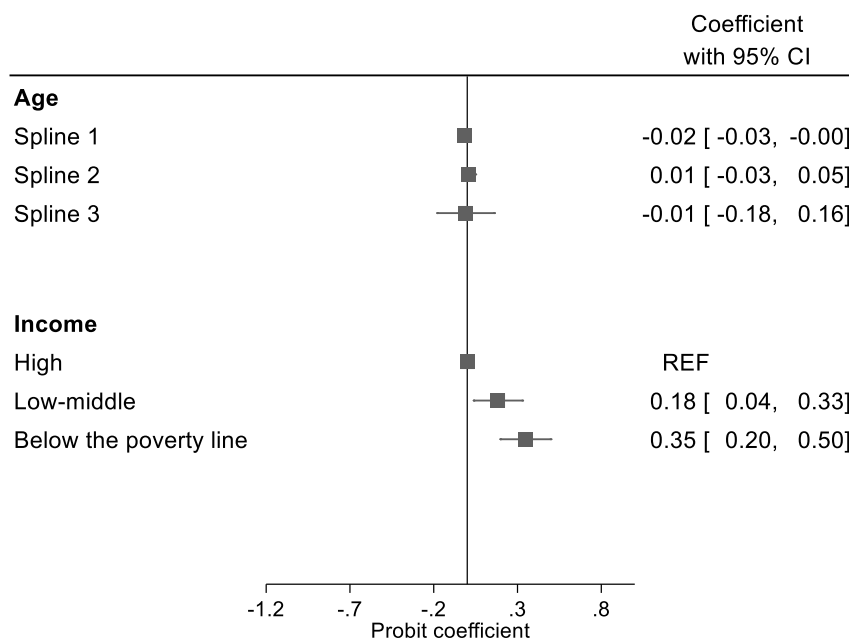
## Men who have sex with men



## Women



## Cis-gender heterosexual men



**Rebuttal Figure 2. Socio-demographic, economic, and health related determinants of having a detectable viral load for MSM (A), women (B), and cisgender heterosexual men (C). Age was included as cubic splines with 4 knots.**

8. I am not sure it is novel enough to be published in Nature Communication but would see it more appropriate for a more specialized journal.

Re. We understand the Reviewer's comment. Lower socioeconomic status and poorer health outcomes go hand in hand and most of the results are not entirely surprising in this regard.

Most of the literature linking socioeconomic status to HIV care-related outcomes has relied on questionnaire data, which involve an extensive amount of selection and response biases. In our cohort, 98% of individuals with HIV in care in the Netherlands are represented in our cohort, minimizing selection bias. We do admit, however, that individuals who are not in care might be less represented in our cohort. Nevertheless, the nationwide stretch of the ATHENA cohort coupled with data from a governmental source very capable of collecting extensive data on the Dutch population makes this likely the most comprehensive, population-based study available. We further minimized the bias from failed linkage with more advance models. As a result, we are able to provide estimates that are the closest approximation to what we would expect in the population (i.e., estimates with the least possible bias).

The study is not only focused on socioeconomic determinants, but also utilization of other health care services prior to HIV diagnosis. Our data provide strong evidence for the public health implications on broadening STI and HIV prevention for the key populations using these services.

Furthermore, social inequalities has played an increasingly important role in our approach to HIV elimination. We require less biased and more exhaustive data to figure out which needs should be more urgently addressed. As mentioned by the first reviewer, others have been linking governmental and municipal data to their cohorts to address this issue.<sup>3</sup> The example given here could help others develop a strategy for evaluating social inequalities for HIV care.

Finally, we are addressing the call for the Special Collection (Social determinants of infectious disease) on understanding and addressing social drivers of infectious diseases. Given what we have presented and with the help of your comments, we feel our study makes a meaningful contribution to the literature and fits within the scope of this Collection.



## References

1. Broyles LN, Luo R, Boeras D, Vojnov L. The risk of sexual transmission of HIV in individuals with low-level HIV viraemia: a systematic review. *The Lancet* 2023; **402**(10400): 464-71.
2. van Sighem A.I., Wit F.W.N.M., Boyd A.C., Smit C., Jongen V.W., T.S. B. Monitoring Report 2024. Human Immunodeficiency Virus (HIV) Infection in the Netherlands. <https://www.hiv-monitoring.nl/nl/resources/monitoring-report-2024>, 2024.
3. Budu MO, Kooij KW, Heath K, et al. Cohort Profile Update: Reflecting back and looking ahead: Updating the Comparative Outcomes and Service Utilization Trends (COAST) Study to include 28 years of linked data from people with and without HIV in British Columbia, Canada. *Int J Popul Data Sci* 2025; **10**(1): 2496.