PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	STI epidemic re-emergence, socio-epidemiological clusters characterisation, and HIV coinfection in Catalonia, Spain, during 2017–2019: a retrospective population-based cohort study
AUTHORS	Sentís, Alexis; Montoro-Fernandez, Marcos; Lopez-Corbeto, E; Egea-Cortés, Laia; Nomah, DK; Díaz, Yesika; Garcia de Olalla, Patricia; Mercuriali, Lilas; Borrell, Núria; Reyes-Urueña, Juliana; Casabona, Jordi

VERSION 1 – REVIEW

REVIEWER	Van Gerwen, Olivia The University of Alabama at Birmingham Division of Infectious Diseases
REVIEW RETURNED	28-May-2021

GENERAL COMMENTS	In this article, the authors aim to describe the epidemiology of incident STI cases, factors associated with STI-HIV coinfection, and socio-epidemiologic clusters in Catalonia, Spain through a population-based retrospective cohort study. The STIs of interest included syphilis, gonorrhea, chlamydia, LGV, and HIV from a regional registry. They describe increasing STIs in their region among women and young people and identified STI socio-epidemiological clusters using a novel methodology of k-means clustering. This paper describes the state of STIs in Catalonia from 2017-2019, showing the number of incident cases of STIs for a variety of infections and outlining the dramatic increase seen in these over a short two-year period. The clustering methodology is interesting, but not described in enough detail. Therefore, the results of this portion of the manuscript are difficult to understand. I think a deeper dive into explaining this methodology in general and how the clusters were created will enhance the value and impact of these data. There are several grammatical shortcomings throughout this manuscript, so would suggest extensive copy editing prior to resubmission. After reading through the whole paper, it is clear that there is epidemiologic value in these data that could inform further studies and prevention efforts, but the lack of clarity from a grammatical and syntax perspective dampen its impact. General Comments • Several times (in the abstract, strengths and limitations section) the authors mention increases in STI, but it's unclear in those sections what the increase is in relation to- does this refer to increases over time? If so, please clarify the time frame. If this is in comparison to another group, please also clarify that. This

	happens at several points throughout the manuscript. I have detailed a few specific instances, but please make sure to clarify this throughout the manuscript. • The authors refer to “rates” of STIs several times throughout the manuscript- please make sure to clarify whether this refers to incidence, prevalence or some other epidemiologic measure. The term “rate” is too nonspecific. Introduction Page 6, Lines 20-25: The reported increasing percentages of these STIs is confusing the way it is written here. Have cases risen by the noted percentage? Please clarify. Page 6, Lines 27-32: The way this sentence is structured is confusing. Would be better to say that from 2000 to 2017, STI cases have increased 10-fold, with 23,975 cases in 2017 alone.” Would also make sure that it is clear that these are incident cases. Page 8: Authors begin talking about “spatiotemporal clustering of cases” and “STI socio-epidemiological clusters” at the end of the introduction and do not explain what composes such an entity. Suggest adding a sentence or two defining this. Page 8, Line 17: Is this really a “hidden” epidemic? The authors have made a strong argument up to this point that STIs and HIV are rampant. Would take out this adjective. Methods Page 7, Line 38: The hyphens enclosing the names of the STIs are unnecessary. Please remove. Also this first sentence in the paragraph is very long and confusing. Please break it up into multiple sentences for clarity. Page 7, Line 47: I would like to know more about the case definitions used as all readers may not be familiar with the ECDC (i.e., this is an international journal with an international audience). Perhaps this would be a good thing to include in a supplementary reference like a table. Page 7-8, line 59: It’s unclear what the basic health area deprivation index is and what it is used for in the study design. Please clarify. You describe what this is later on in the statistical analysis section—would suggest defining this earlier since you start to talk about it at the beginning of the methods section. Would define your clusters here as “A, B, and C” here in the methods. Then when the reader gets to the results, they have a frame of reference for what these clusters are. I would place your map of the clusters in the methods portion of the manuscript. It’s also confusing in the methods as to how exactly the clusters were chosen. Were ABSs that were similar in the parameters you listed figured into a cluster? Results Page 10, Line 44: It’s still unclear what a deprivation index is so these results are difficult to understand. Page 10: You defined “reinfection” earlier in the methods, but it would be useful to discuss if these people were treated

	appropriately or not if you have that data. Distinguishing between reinfection and untreated infection is important, if those data are available. Discussion Page 14, Line 16: would take out the phrase “blood borne diseases”—unnecessarily vague Page 14, Line 21: “Last years” doesn’t make sense here, may be a typo?

REVIEWER	Thinh, Vu Toan CUNY Graduate School of Public Health and Health Policy, Center for Innovation in Mental Health
REVIEW RETURNED	31-May-2021

GENERAL COMMENTS	The manuscript deals with a very important topic: STI and HIV surveillance systems. The method used is a retrospective cohort to describe incidence rates of STI cases stratified by socio-epidemiological characteristics and their trends as well as correlates of HIV/STI co-infection in Catalonia between 2017 and 2019. Generally, the manuscript will substantially benefit from proofreading by a native speaker so as to correct typos as well as grammar. Please consider the following comments: ABSTRACT: - Line 6-13: in the epidemiological aspect, the authors could combine the 1st and 3rd objectives without any harm. If the authors highlight the K-means algorithm on purpose, separating objectives are acceptable, however, the authors should re-order the objectives. - Line 39, page 3: the authors stated that “…. The increase in chlamydia and gonorrhea in women…”. Looking at table 1, males accounted for more than 80% of gonorrhea cases. Please explain the discrepancy. - Line 46, page 3: “… aged 30-60…were associated with an increased risk of HIV coinfection”. Given table 2, the odds of having co-infection in 20+ year-old group are “statistically significant” higher than those who are less than 20. Please revise the conclusion and its respective parts. - Line 55, page 3: “A) similar distribution-values”. Please clarify this confusing sentence by summarizing main findings in this cluster. - Line 3, page 4: “C) higher incidence rates for all STI”. Is that true? Cluster C occupied 38% of Chlamydia cases which is much smaller than the overall rate at 55% (Table 3). - Conclusion: please tell us what do the findings imply instead of repeating the objectives. More importantly, the authors mentioned the “key populations” which is not really the case for this paper. Up to 80% of those who self-reported sexual orientation are heterosexual (Table 1). INTRODUCTION - Please shorten the introduction part to summarize key rationales and could brief some information of the Catalan HIV/STI registry of Catalonia here. - Line 25, page 6: “(LGV) have been increased 50%, 36% and 69%, respectively, from 2014 to 2018 [5]”. What did the authors mean? The incidence rates increased “by/to” 50%, 36%, and 69%. Please clarify the rates. Additionally, the reference [5]

	referred to the Annual Epidemiological Report for 2017, how could the authors have data in 2018? Did the authors reference the cited papers which are not original? METHODS - Surveillance systems often set up age limitations to monitoring prevalence, incidence, and their trend. The authors should describe some eligibilities of surveillance participants. - Outcome measurement: it would be great if the authors could provide kinds of tests used to diagnose STI and HIV. - In this manuscript, the authors merged data from the surveillance systems, how the authors confirmed the same person between two systems (using social security number, fingerprint, etc), please clarify. If possible, please describe some step findings such as numbers of records duplicates within each surveillance system, numbers of records merged, matched, etc. - Line 20, page 8: what kinds of statistical tests did the authors apply to evaluate the final model: goodness-of-fit, multi-collinearity, or observational influences (Table 2)? If the authors simply put “significant variables” in the simple/binary logistic regression model into the final model as described in line 3, page 9, please provide the tests so as to assess the final model. - Line 3, page 9: importantly noted that the results are bivariate, not univariate logistic regression. Please make sure to use the appropriate terminologies. - Line 22, page 10: the supplemental materials are mistakenly cited. The sexual preference is found in table 1 while the education attainment belongs to table S4. RESULTS - Table S3: the authors should provide the test for trend (P trend) when mentioning the incidence rate changes over the study period. Also, the authors did not reveal the number of populations at risk (denominators) for the years 2017-2019, so this challenges audiences in comparing rates. - Line 28-41, page 10: the authors should review the numbers so as to make solid conclusion. The detailed comments follow: o Line 28, page 10: the authors stated that “the STI episodes in men were significantly higher than in women”. This sentence could be quite subjective without any statistical tests. o Line 33-34, page 10: the authors have given wrong data, based on data the authors mentioned, there are about 55% out of participants with sexual orientation information self-reported WSM. In addition, please confirm “in men, 51% were MSM”. o Line 37-41, page 10: Chlamydia is prevalent among those younger than 40 in lieu of those aged 30 and less because the 30-39-year-old group occupied a high proportion of 21%. Likewise, the 20-49-year-old bracket covers up to 80% of syphilis cases. - Table 1: o Regarding choices of categories, having 5 categories for exposure (e.g., age group) are well chosen to reflect the size of exposure effect expected across categories. Nevertheless, the authors divided the age bracket into 6 categories, of which one open-ended category (60+). This can result in residual confounding. o Moreover, the authors implemented quintile method to create boundaries for deprivation index. This could lead to diluted effect if deprivation is a strong confounder or is unevenly distributed. Please discuss these limitations at the end of discussion section.

	o Should the authors count missing values in order to calculate percentages in table 1 while the research team only used denominators with complete values (Lines 32-34, page10 and lines 1-8, page 11)? - Line 11-29, page 11: please be consistent when using abbreviations such as adjusted Odds Ration (aOR in lieu of ORa), confidence interval (CI instead of IC) in both text and tables. Also, importantly note that the results are multivariable, not multivariate logistic regression. - The fourth paragraph: why did the authors decide to compare each cluster to the overall? Is there any way to make it more sense such as comparing between three clusters? - Please rewrite the whole confusing paragraph (Line 11-29) so as to reflect the main findings based on respective tables. - Table 2: o The authors must explain how to get both crude and adjusted odds ratios for missing value category regarding deprivation index. What kinds of missing value do the authors detect (missing completely at random, missing at random, or missing not at random)? And how did the authors handle missing values in order to get ORs for this category? And how to interpret ORs for missing data? o Please explain the discrepancy between Tables 1 and 2. In table 1, there remain 2,433 HIV/STI co-infected cases while in table 2, there are only 1,376 cases. - Again, please rewrite the confusing paragraph (Line 18-27). Picking data for presentation plays a key role in determining the main conclusion, however, it should reflect the fact. Having said that the authors should not combine numbers in a subjective and/or random way. For example, 2nd and 4th quintile for deprivation index accounted for approximately 70% of STI cases in cluster A (instead of 4th and 5th quintile as the author mentioned). Likewise, more than 70% of STI cases in cluster B belonged to 3rd and 4th quintiles (in lieu of 1st to 3rd). - Table 3, Table S3, and Table S4: please put sample size for each column title. - Figure 1: It would be greatly appreciated if the authors please put the figure title on the same page of the map. Is there a possibility of change the chart type so as to see geographic areas and incidence rates simultaneously? DISCUSSION - The first paragraph seems to give a conclusion that repeats the main results. Please shorten the paragraph and make comparison to previous studies. Also, the authors are encouraged to paraphrase the sentence, do not use exactly the long sentence from other parts of the manuscript. - Line 13-22, page 13: the authors used old data from 2010 to confirm the current trend, please find out the most updated data to elaborate the main findings. - Line 11-29, page 14: this paragraph is used to discuss the 2nd objective, the authors should spend more space incorporating the evidence from previous studies to discuss the key findings. - Limitation paragraph, page 15: the authors should discuss the reasons for a high volume of missing value, it is due to the quality of data collection or discrimination on sexual orientation. - The authors consider referencing the paper “Recent trends in sexually transmitted infections among adolescents, Catalonia, Spain, 2012–2017” (https://doi.org/10.1177/0956462420940911)

	which might be relevant for discussing incidence rates and their trend. - The questionnaire collected behavioral data (Table S1), yet the authors excluded these individual risk behaviors for STI and HIV co-infection for logistic models. Please provide insights. REFERENCES - References #5 and #22 are duplicate.

REVIEWER	Obiri-Yeboah, Dorcas University of Cape Coast
REVIEW RETURNED	07-Jun-2021

GENERAL COMMENTS	The STI re-emergence in Catalonia (2017-2019): epidemic characterization, socio-epidemiological clustering approach, and HIV co-infection associated factors. General Well structured, well written generally. Very important and relevant study topic using retrospective data. Abstract: clear and contain all relevant information. Data collection and analysis well described and appropriate. Study limitations have been clearly stated and discussed. Below are some specific relatively minor comments to help improve the manuscript. Specific comments 1. Page 6, line 9: you have “The epidemic of sexually transmitted infections (STI) is a major public health concern in high-income-, middle-, and low-income countries”. This basically imply the situation is global so can you consider stating just that concisely? 2. Sexually transmitted infections (STI)- usually to use it as plural, STIs is used. This is used correctly in some places in the manuscript and not used when needed in other places. Please check 3. Page 7 lines 2- 8: you have “Some studies have identified the social determinants of health, discrimination, and inequities as main factors associated with the appearance of STI spatiotemporal clustering of cases”. Please rephrase especially towards the last part, it gets unclear. 4. Gonorrhoea has been spelt in both British and American ways. Be consistent 5. Page 8 line 8: you have “(total number of episodes due to any STI that had the same person during the study period)”, rephrase. I believe you mean the total number of episodes of any STIs that the same person had during the study period? 6. Page 10 lines 26-32: You have “proportionally, the STI episodes in men were significantly higher than in women, for gonorrhoea, syphilis, and LGV, but less frequent for chlamydia (80%, 87%, 99%, and 38% were in men, respectively)”. I find it confusing as I don’t see the proportions in women. Consider rephrasing. 7. Page 10, lines 55-60: this sentence is not clear and needs to be revised---The STI episodes in HIV-positive counted 6% from the overall, however, with higher proportion in syphilis and LGV (13% and 25%, respectively) and the lowest in chlamydia (2%) 8. Page 14 lines 32-35: please rephrase the sentence “Last years, the K-means clustering methodology has proven its potential in classifying and grouping health related outputs in different study fields” 9. Page 15, line 43: rephrase this “While the HIV trend to decrease, mainly because the wider and earlier use of ARV,……”

	10. Check the formatting for table 3 11. Though secondary data was used, this is a sensitive topic, what ethical considerations did the researchers have? There is no comment on ethics at all as far as I saw. 12. Some grammatical and typographical errors to correct e.g. check line 13 (page 5), 48 (page 6), etc.

REVIEWER	Makuza, Jean Damascene Rwanda Biomedical Center, Institute of HIV, Diseases Prevention and Control
REVIEW RETURNED	08-Jun-2021

GENERAL COMMENTS	Comments on the article This is a good article exploring a sensible subject to help most people in need. The article is well written and has sufficient information on STIs epidemics in the study population however, there are some errors to be addressed like strengthening the abstract, avoid repetitions, abbreviations to spell, … 1) Abstract Some abbreviations were spelled on the first-time user like STIs, HIV, LGV, and others, please spell them on their 1st use. Results: line 37: you said the number of STIs-cases doubled, what is the basic number or from which period to which period? It is better also to show the figures (numbers or proportion of cases in women and in young people). Conclusion: Your conclusion need to be linked with figures from results and these are not described in the results. In addition to that, there is a need for recommendations in this part based on your findings, please try to suggest any of them. Strengths and limitations There are repetitions between points 1 and 4 and these need to be summarized together 2) Introduction Lines 20-25: This incidence is in which population? General population or specific population? Methods For study participants, do you have inclusion and exclusion criteria, if yes can give them? You cited different variables, which are the dependent and how are they described or defined (continuous or categorical), which are independent? What is the study site and why did you choose it? For statistical analysis, page 8 lines 44-48, this sentence describing ABS could go into the study population. Page 9, line 4, I think this is not a univariate analysis, it could be the bivariate analysis, please do the correction. What was the cut-off of the p-value for consideration of variables in the final model (multivariate)? What other criteria did you consider for variables to be included in the final model? 4) Results Page 10, line 15, please precise the % of STIs for Barcelona which is the highest, and for Alt Pirineu I Aran which has the lowest as you did for urban and rural areas. Page 10, line 34, spell abbreviation WSM. Discussion Paragraph 1, page 12-13, line 36-60 then 1-10, this is a pure repetition of the results, can summarize it in other few words? Or

	delete it definitely as its contents come again in the following paragraphs. Page 15, line 3-10, you said missing data could lead to biased results, which bias do you expect? Better to precise the types of biases expected. Do not you think that your study was prone to confounders? Please tell me something about that could confound your results. Page 15, lines 48-55. There could be a typo error "….These populations need to be considered a priority for the preventive strategies…… Please bring the correction. 5) References References 3, 10, 15, 20, and 21 seem to be aged, better to use references published no lesser than 2010. Reference 28 is not well presented, please correct it. 6) Additional Please check the English used

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr. Olivia Van Gerwen, The University of Alabama at Birmingham Division of Infectious Diseases

General comment: In this article, the authors aim to describe the epidemiology of incident STI cases, factors associated with STI-HIV coinfection, and socio-epidemiologic clusters in Catalonia, Spain through a population-based retrospective cohort study. The STIs of interest included syphilis, gonorrhoea, chlamydia, LGV, and HIV from a regional registry. They describe increasing STIs in their region among women and young people and identified STI socio-epidemiological clusters using a novel methodology of k-means clustering. This paper describes the state of STIs in Catalonia from 2017-2019, showing the number of incident cases of STIs for a variety of infections and outlining the dramatic increase seen in these over a short two-year period

Reviewer 1: comment	**Response**
1. The clustering methodology is interesting, but not described in enough detail. Therefore, the results of this portion of the manuscript are difficult to understand. I think a deeper dive into explaining this methodology in general and how the clusters were created will enhance the value and impact of these data.	We appreciate your valuable suggestion. We have described the methodology in more detail in a separate subsection with the heading 'K-means clustering of STIs'. In particular, we elaborated on the principles underlying the clustering algorithm procedure and the associated validation processes. Please see the new subsection on Page 7, line 19.
2. There are several grammatical shortcomings throughout this manuscript, so would suggest extensive copy editing prior to resubmission. After reading through the whole paper, it is clear that	Thank you for the suggestion. The revised manuscript has been copyedited by a native English speaker for language and grammatical errors, and to improve readability.

Reviewer 1: comment	Response
there is epidemiologic value in these data that could inform further studies and prevention efforts, but the lack of clarity from a grammatical and syntax perspective dampen its impact.	
3. Several times (in the abstract, strengths and limitations section) the authors mention increases in STI, but it's unclear in those sections what the increase is in relation to- does this refer to increases over time? If so, please clarify the time frame. If this is in comparison to another group, please also clarify that. This happens at several points throughout the manuscript. I have detailed a few specific instances, but please make sure to clarify this throughout the manuscript.	Thank you for highlighting these ambiguities. We have corrected this throughout the report by indicating timeframes where appropriate, and specifying where trends refer to rates or number of cases etc.
4. The authors refer to "rates" of STIs several times throughout the manuscript- please make sure to clarify whether this refers to incidence, prevalence or some other epidemiologic measure. The term "rate" is too nonspecific.	Thank you for highlighting these ambiguities. We have corrected this throughout the report.
Introduction	
5. Page 6, Lines 20-25: The reported increasing percentages of these STIs is confusing the way it is written here. Have cases risen by the noted percentage? Please clarify.	Thank you. Yes, the cases increased from 2014 to 2018 by the specified percentages. We have rewritten that sentence to clarify. Please refer to page 5, line 6–8.
6. Page 6, Lines 27-32: The way this sentence is structured is confusing. Would be better to say that from2000 to 2017, STI cases have increased 10-fold, with 23,975 cases in 2017 alone." Would also make sure that it is clear that these are incident cases.	Thank you. We have rewritten this so that it is clear these are new STI cases. Please refer to page 5, line 9–11. 'This trend is reflected in Spain where new STI cases have been reported to increase 10-fold from 2000 to 2017, with 23,975 cases of gonorrhoea, syphilis, chlamydia and LGV reported in 2017 alone.'

Reviewer 1: comment	Response
7. Page 8: Authors begin talking about "spatiotemporal clustering of cases" and "STI socio-epidemiological clusters" at the end of the introduction and do not explain what composes such an entity. Suggest adding a sentence or two defining this.	Thank you. We have elaborated as below, with reference to Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010;31:651–66: 'While spatiotemporal clustering may be useful in grouping events or cases, other methodologies including k-means clustering allow grouping of different geographical units by common characteristics such as sociological and epidemiological factors.' Please refer to page 5, line 24–27.
8. Page 8, Line 17: Is this really a "hidden" epidemic? The authors have made a strong argument up to this point that STIs and HIV are rampant. Would take out this adjective	Thank you. This has been deleted.
Methods	
9. Page 7, Line 38: The hyphens enclosing the names of the STIs are unnecessary. Please remove. Also this first sentence in the paragraph is very long and confusing. Please break it up into multiple sentences for clarity	Thank you. We have rewritten the methods section substantially for clarity. Please refer to page 6, line 6–21.
10. Page 7, Line 47: I would like to know more about the case definitions used as all readers may not be familiar with the ECDC (i.e., this is an international journal with an international audience). Perhaps this would be a good thing to include in a supplementary reference like a table	Thank you for the suggestion. Further information regarding the ECDC case definitions that are used in Catalonia have been included as supplementary material (table S2).
11. Page 7-8, line 59: It's unclear what the basic health area deprivation index is and what it is used for in the study design. Please clarify. You describe what this is later on in the statistical analysis section—would suggest defining this	Thank you. We have rewritten the methods section substantially to describe basic health areas more clearly and detailed how the deprivation indices were determined and used. Please see below and page 7, line 1–10. 'A Basic Health Area (BHA; Àrea Bàsica de Salut [ABS], in Catalan) is a territorial unit of coverage served by a primary healthcare team. Each BHA typically serves a population of approximately

Reviewer 1: comment	Response
earlier since you start to talk about it at the beginning of the methods section.	5,000–25,000 people. The socioeconomic level of the BHAs were classified according to a deprivation index (calculated by the Agency of Health Quality and Assessment of Catalonia) which was attributed to each individual according to their residential address. The deprivation index is a composite measure based on indicators such as proportion of residents with low educational level, proportion of manual workers, proportion of residents with an annual income below a specified amount and rate of premature mortality. Deprivation indices were categorized in quintiles, with the first quintile being the least deprived.'
12. Would define your clusters here as "A, B, and C" here in the methods. Then when the reader gets to the results, they have a frame of reference for what these clusters are. I would place your map of the clusters in the methods portion of the manuscript. It's also confusing in the methods as to how exactly the clusters were chosen. Were ABSs that were similar in the parameters you listed figured into a cluster?	Thank you. We have introduced the three clusters upfront in the methods section (page 8, line 4–5), where the map is also referred to. Further details around how the clusters were grouped are now supplied in the new subsection 'K-means clustering of STIs'. Please see from page 7, line 19.
Results	
13. Page 10, Line 44: It's still unclear what a deprivation index is so these results are difficult to understand	As per comment #11 above, we have rewritten the methods section substantially to detail how the deprivation indices were determined and used. Please see below and page 7, line 1–10. We hope this is satisfactory.
14. Page 10: You defined "reinfection" earlier in the methods, but it would be useful to discuss if these people were treated appropriately or not if you have that data. Distinguishing between reinfection and untreated infection is important, if those data are available	Thank you for raising this important point. We aligned our definition of reinfection with guidance from the CDC ie an episode of the same STI detected after a defined period, which differs for each STI, following the previously recorded infection in the same individual, which guarantees that the previous episode was cured. To clarify, we added the following sentence (page 7, 16–18): 'As information regarding treatment response was not available, episodes occurring outside of the

Reviewer 1: comment	Response
	specific timeframes for each STI were assumed not to be a persistent infection resulting from treatment failure'
Discussion	
15. Page 14, Line 16: would take out the phrase "blood borne diseases"— unnecessarily vague	Thank you. This has been replaced with 'viral hepatitis' on page 13, line 20.
16. Page 14, Line 21: "Last years" doesn't make sense here, may be a typo?	Thank you. This has been rectified.

Reviewer 2: Mr. Vu Toan Thinh, CUNY Graduate School of Public Health and Health Policy

General comment: The manuscript deals with a very important topic: STI and HIV surveillance systems. The method used is a retrospective cohort to describe incidence rates of STI cases stratified by socioepidemiological characteristics and their trends as well as correlates of HIV/STI co-infection in Catalonia between 2017 and 2019.

Generally, the manuscript will substantially benefit from proofreading by a native speaker so as to correct typos as well as grammar. Please consider the following comments:

Reviewer 2: comment	Response
Abstract	
1. Line 6-13: in the epidemiological aspect, the authors could combine the 1st and 3rd objectives without any harm. If the authors highlight the K-means algorithm on purpose, separating objectives are acceptable, however, the authors should re-order the objective	Thank you. We have reordered as suggested in the abstract (page 3, line 2–4) and introduction section (page 6, line 1–3.
2. Line 39, page 3: the authors stated that "…. The increase in chlamydia and gonorrhoea in women…". Looking at table 1, males accounted for more than 80% of gonorrhoea cases. Please explain the discrepancy.	Chlamydia is the STI that has shown greatest increase from 2017 to 2019, both in total numbers and in incidence rate (table S4), and accounts for more than 50% of all STI cases in 2019 in Catalonia. More than 60% of chlamydia cases occurred in women and more than 65% in people under 30 years of ages (see table 1). Gonorrhoea showed the second biggest increase during the study period and 47% were in people under 30 years of age. In the introduction, we discussed that a separate study showed that in Catalonia, during almost the same period (2018–2019), a proportionally higher increase was observed in young adults,

Reviewer 2: comment	Response
	particularly women, particularly for chlamydia but also gonorrhoea. This is consistent with our findings (in 2019, the ratio men: women decrease from 6 to 3.69 which means that proportionally increase was higher in women, page 19-29: https://canalsalut.gencat.cat/web/.content/_A-Z/S/sida/enllasos/anual_ITS.pdf, reference 8). We truly believe that both diseases, chlamydia and gonorrhoea, mainly in these two populations at higher risk, women and young adults, are those that have contributed most to the STI re-emergence. In the revised manuscript, we have endeavoured to discuss this more clearly in the results (page 9, line 18–32) and discussion (page 12, line 11–13, page 12-13, line 32-5). Besides, we have added STI incidence rates by sex, age, and sex&age in current table 2.
3. Line 46, page 3: "… aged 30-60…were associated with an increased risk of HIV coinfection". Given table 2, the odds of having co-infection in 20+ year-old group are "statistically significant" higher than those who are less than 20. Please revise the conclusion and its respective parts.	Thank you very much for your comment. We agree and indeed individuals above the age of 20 years are at higher risk those below 20 years. We have rectified this throughout the manuscript, in particular in the abstract (page 3, line 20), and the results (page 10-11, line 32-2) discussion (page 13, line 20–22) sections.
4. Line 55, page 3: "A) similar distribution-values". Please clarify this confusing sentence by summarizing main findings in this cluster	Thank you. We have rewritten the main findings of the cluster analysis so that it is a summary of the main findings. Please see page 3, line 24–27.
5. Line 3, page 4: "C) higher incidence rates for all STI". Is that true? Cluster C occupied 38% of Chlamydia cases which is much smaller than the overall rate at 55% (Table 3).	Thank you for the query. As in comment #4, we have rewritten the main findings of the cluster analysis so that it is a summary of the main findings. By 'higher incidence rates for all STIs', we mean that the overall incidence rate of STIs, which was the highest among all groups (721.0 versus 160.6 per 100,000 population in the overall group). We have clarified this in the results section (page 11, line 30). Indeed, all 4,359 STI cases in Cluster C were reported in BHAs in the highest quintile of STI

Reviewer 2: comment	Response
	incidence rate. We have made this clear on page 12, line 5–6.
6. Conclusion: please tell us what do the findings imply instead of repeating the objectives. More importantly, the authors mentioned the "key populations" which is not really the case for this paper. Up to 80% of those who self-reported sexual orientation are heterosexual (Table 1).	Thank you for the suggestion. We have rewritten the conclusion (page 3, line 28–30) as follows: "We recommend socioepidemiological identification and characterization of STI clusters and factors associated with HIV coinfection to identify at-risk populations at a health area level to design effective interventions."
Introduction	
7. Please shorten the introduction part to summarize key rationales and could brief some information of the Catalan HIV/STI registry of Catalonia here.	Thank you for the suggestion. The introduction section has been rewritten substantially and we hope your feedback has been addressed appropriately.
8. Line 25, page 6: "(LGV) have been increased 50%, 36% and 69%, respectively, from 2014 to 2018 [5]". What did the authors mean? The incidence rates increased "by/to" 50%, 36%, and 69%. Please clarify the rates. Additionally, the reference [5] referred to the Annual Epidemiological Report for 2017, how could the authors have data in 2018? Did the authors reference the cited papers which are not original?	The increase refers to increase in incidence from 2014 to 2018 by the specified percentages. We have rewritten that sentence as below (page 5, line 6–9): 'Across Europe, incidence of STIs continue to be on the rise with confirmed cases reported in national surveillance systems increasing by 50% for gonorrhoea, 36% for syphilis, 68% for lymphogranuloma venereum (LGV), and 0.6% for chlamydia from 2014 to 2018.' Additionally, we have rectified the citation error by citing the primary references, which are the ECDC annual reports (see references 3–6).
Methods	
9. Surveillance systems often set up age limitations to monitoring prevalence, incidence, and their trend. The authors should describe some eligibilities of surveillance participants.	Thank you. Our data source was the Catalan HIV/STI Registry, which uses case definitions that are aligned with standardized case definitions established by the ECDC, of which neither age nor any other personal characteristics are specified. These case definitions are also aligned with the CDC case definitions for surveillance

Reviewer 2: comment	Response
	(https://www.cdc.gov/std/statistics/2019/case-definitions.htm). We aimed to clarify this by reporting this in the methods section (page 6, line 18–21), with further information detailed in supplementary table S2.
10. Outcome measurement: it would be great if the authors could provide kinds of tests used to diagnose STI and HIV.	Thank you. Unfortunately, this information is not collected in the surveillance system. Nonetheless, all notified cases in the surveillance system are confirmed cases as per the established case definitions established by the ECDC.
11. In this manuscript, the authors merged data from the surveillance systems, how the authors confirmed the same person between two systems (using social security number, fingerprint, etc), please clarify. If possible, please describe some step findings such as numbers of records duplicates within each surveillance system, numbers of records merged, matched, etc.	Thank you for the helpful suggestion. This is described on page 6, line 24–31: 'All individuals who had experienced at least one STI episode during the study period were linked, through the Spanish healthcare system personal identification code (CIP), to the Catalan HIV/STI Registry to identify HIV coinfections either before or after the recorded STI episode. In addition to the CIP, Catalan HIV/STI Registry surveillance team performs duplicate checks at least twice annually using a unique STI episode number (assigned to each notification and disease), name and date of birth. For our analysis, a deduplicated, HIV/STI-linked and anonymized version was provided.' Unfortunately, information other than those described above was not available as the dataset provided to our investigation team was a de-duplicated, anonymized version.
12. Line 20, page 8: what kinds of statistical tests did the authors apply to evaluate the final model: goodness-of-fit, multi-collinearity, or observational influences (Table 2)? If the authors simply put "significant variables" in the simple/binary logistic regression model into the final model as described in line 3, page 9, please provide the tests so as to assess the final model.	Thank you. Sexual preference, country of birth and education level were excluded from the models because more than 50% of values were missing. We used backward stepwise elimination regression to include all analysed variables that showed statistical significance ($P<0.05$) by the Wald test in the final multivariable logistic regression model. This is now described on page 8, line 25–28.

Reviewer 2: comment	Response
13. Line 3, page 9: importantly noted that the results are bivariate, not univariate logistic regression. Please make sure to use the appropriate terminologies	Thank you for pointing out the error, which has been rectified.
14. Line 22, page 10: the supplemental materials are mistakenly cited. The sexual preference is found in table 1 while the education attainment belongs to table S4.	Thank you for pointing out the error, which has been rectified.
Results	
15. Table S3: the authors should provide the test for trend (P trend) when mentioning the incidence rate changes over the study period. Also, the authors did not reveal the number of populations at risk (denominators) for the years 2017-2019, so this challenges audiences in comparing rates	Thank you for the suggestion. Incidence trends were analysed using the $\chi 2$ test for linear trend, which is now described in the methods section (page 8, line 17). Further, we revised the table (now table 2) to include P-values for all comparisons, and specified the P-values in the main text where appropriate. The number of populations at risk (denominators) are provided in table S4. Besides, we have added STI incidence rates by sex, age, and sex&age in table 2.
Line 28-41, page 10: the authors should review the numbers to make solid conclusion. The detailed comments follow:	
16. Line 28, page 10: the authors stated that "the STI episodes in men were significantly higher than in women". This sentence could be quite subjective without any statistical tests.	Thank you. We have rewritten as follows (page 10, line 1–2): 'The vast majority of reported cases occurred in men for all STI types except chlamydia, of which 61.9% occurred in women (table 1).'
17. Line 33-34, page 10: the authors have given wrong data, based on data the authors mentioned, there are about 55% out of participants with sexual orientation information self-reported WSM. In addition, please confirm "in men, 51% were MSM".	Thank you. We have rewritten this to reflect the representation of MSM and MSW in reported cases instead (page 10, line 5–7). 'Among the 15,023 (35 .5%) reported STI cases for which information regarding sexual preference was available, half (54.5%) were reported in women who have sex with men (WSM), 21.8% in MSM and 21.0% in MSW (table 1).'
18. Line 37-41, page 10: Chlamydia is prevalent among those younger than 40 in lieu of those aged 30 and less because the 30-39-	Our intention was to highlight that chlamydia was most common among individuals below the age of 30 years while syphilis was most common in those above the age of 30 years.

Reviewer 2: comment	Response
year-old group occupied a high proportion of 21%. Likewise, the 20-49-year-old bracket covers up to 80% of syphilis cases.	(Please note that the age range of 20–49 years also covers up to 80% of gonorrhoea and LGV cases). We have rewritten as follows (page 10, line 3–5): 'Chlamydia was reported most frequently among individuals below 30 years of age (66 .1%) while syphilis occurred most in those above 30 years of age (77.1%).'
19. TABLE 1: Regarding choices of categories, having 5 categories for exposure (e.g., age group) are well chosen to reflect the size of exposure effect expected across categories. Nevertheless, the authors divided the age bracket into 6 categories, of which one open-ended category (60+). This can result in residual confounding. Moreover, the authors implemented quintile method to create boundaries for deprivation index. This could lead to diluted effect if deprivation is a strong confounder or is unevenly distributed. Please discuss these limitations at the end of discussion section.	Thank you for the helpful comment. We have discussed this these as limitations in the discussion section (page 14, line 14–19).
20. TABLE 1: Should the authors count missing values in order to calculate percentages in table 1 while the research team only used denominators with complete values (Lines 32-34, page10 and lines 1-8, page 11)?	Thank you. The issue of missing data is stated as a limitation of this study upfront under 'strengths and limitations' and described extensively in the discussion section (page 14, line 6–19) as below. We believe that showing missing data in table 1 is a good practice of transparency. However, we also wanted to highlight the percentage among those with available data only for some variables such as sexual preference due to its relevance as a potential risk factor. We hope you can share this point of view. 'A key limitation of this study is the high proportion of missing data around sociodemographic and lifestyle characteristics, a common phenomenon in population-based epidemiological studies where questionnaires are used. This may have potentially introduced information bias or inaccurate representation of the true situation when describing high-risk

Reviewer 2: comment	Response
	populations. Although not formally assessed, we classify these missing data as missing completely at random due to time constraints in completion of the epidemiological questionnaires by surveillance officers and healthcare professionals who notified the diseases to the surveillance systems. Nonetheless, our findings are similar to those reported in previous analyses.'
21. Line 11-29, page 11: please be consistent when using abbreviations such as adjusted Odds Ration (aOR in lieu of ORa), confidence interval (CI instead of IC) in both text and tables. Also, importantly note that the results are multivariable, not multivariate logistic regression	Thank you for the comment. These have been corrected throughout the manuscript.
22. The fourth paragraph: why did the authors decide to compare each cluster to the overall? Is there any way to make it more sense such as comparing between three clusters?	Thank you for the query. Our intention is to demonstrate the value of having results for these smaller socioepidemiological clusters as opposed to the pooled results that, in some instances, may not be relevant to the local context. We explained this in the discussion section on page 13 (line 32-35) and page 15 (line 1–8).
23. Please rewrite the whole confusing paragraph (Line 11-29) so as to reflect the main findings based on respective tables.	Thank you. We have rewritten this under the heading 'Factors associated with HIV coinfection among individuals with STIs'. Please see page 10-11, line 29–3.
24. TABLE 2: The authors must explain how to get both crude and adjusted odds ratios for missing value category regarding deprivation index. What kinds of missing value do the authors detect (missing completely at random, missing at random, or missing not at random)? And how did the authors handle missing values in order to get ORs for this category? And how to interpret ORs for missing data?	Thank you for your comment. Although not formally assessed, we classify these missing data as missing completely at random due to time constraints in completion of the epidemiological questionnaires by surveillance officers and healthcare professionals who notified the diseases to the surveillance systems. The issue of missing data is stated as a limitation of this study upfront under 'strengths and limitations' and described extensively in the discussion section (page 14, line 6–13). Missing data for deprivation index were handled similarly. The data were included because the sample size in this category is proportionally quite relevant so we tried to get some information about it. Unfortunately, the values

Reviewer 2: comment	Response
	are quite similar in most categories. If the missing data category was more similar to any other specific category, we could attribute that most of the missing data were coming from that specific category. We believe that we could probably say the opposite ie that probably the missing data category is fed by all the other categories. As this discussion does not add any particular value and can be deduced from the table, we have decided to focus the discussion on interpretation of the main findings instead.
25. TABLE 2: Please explain the discrepancy between Tables 1 and 2. In table 1, there remain 2,433 HIV/STI co-infected cases while in table 2, there are only 1,376 cases	Thank you for the query. Table 1 describes total number of cases (42,283) and table 3 reports number of affected individuals (34,600). In assessing risk factors associated with HIV coinfection (table 3), individuals with more than one STI episode were counted once (first episode), and successive episodes in the same individual were grouped in a variable that considers the number of episodes, and included in the models. This is described in the methods section (page 8, line 22–24). Based on your query, we have endeavoured to make the distinction between 'cases' and 'individuals' clearer throughout the manuscript where appropriate.
26. Again, please rewrite the confusing paragraph (Line 18-27). Picking data for presentation plays a key role in determining the main conclusion, however, it should reflect the fact. Having said that the authors should not combine numbers in a subjective and/or random way. For example, 2nd and 4th quintile for deprivation index accounted for approximately 70% of STI cases in cluster A (instead of 4th and 5th quintile as the author mentioned). Likewise, more than 70% of STI cases in cluster B belonged to 3rd and 4th quintiles (in lieu of 1st to 3rd).	Thank you for your comment. The deprivation index quintiles were not chosen at random. We aimed to describe extreme incidence rates and we believe that the way they are described currently gives an idea of the proportion for extreme values in each cluster: higher proportion of very high incidence rates (4th and 5th quintile) are more frequent in Clusters A and C, while the proportion of lower incidence rates are more frequent in Cluster B. In table 4 you can see the proportion of number of episodes and the proportion of ABS in each cluster, and understand that Clusters A and C had higher incidence rates. To further clarify this, we have added the following explanation on page 12, line 6–9: 'This correlated well with the fact the number of STI cases per BHA was higher in Clusters A and C (105.8 and 544.9 cases per BHA, respectively) than in the total (97.4 cases per

Reviewer 2: comment	Response
	BHA), which indicates higher proportion of high incidence rates (table 4 and Figure 1).' We hope we have clarified your query satisfactorily and that you share our point of view.
27. Table 3, Table S3, and Table S4: please put sample size for each column title.	Thank you. Sample sizes have been included in all tables. Please note that tables S3 and S4 are now tables 2 and S5 in the revised manuscript.
28. It would be greatly appreciated if the authors please put the figure title on the same page of the map. Is there a possibility of change the chart type so as to see geographic areas and incidence rates simultaneously?	We believe that in the final proof, the title and legend will be placed directly under the figure. Thank you for the suggestion regarding chart type. Respectfully, we prefer to present the information this way as our intention is to allow a general view of incidence rates and clusters in each ABS independently. Nevertheless, when describing the findings in the main text, it will be helpful to combine both, as we have done on page 12, line 3-6: 'Almost 60% of STI cases in Cluster B occurred in BHAs in the three lowest quintiles of STI incidence rates, while more than 60% in Cluster A occurred in areas of high STI incidence rates (fourth and fifth quintiles). All 4,359 STI cases in Cluster C were reported in BHAs in the highest quintile of STI incidence rate.'
Discussion	
29. The first paragraph seems to give a conclusion that repeats the main results. Please shorten the paragraph and make comparison to previous studies. Also, the authors are encouraged to paraphrase the sentence, do not use exactly the long sentence from other parts of the manuscript.	Thank you. The discussion section has been reorganized and rewritten substantially to address the reviewers' comments and the revised manuscript has been copyedited by a native English speaker for language and grammatical errors, and to improve readability. We would appreciate it if you could review the section in the revised manuscript.
30. Line 13-22, page 13: the authors used old data from 2010 to confirm the current trend, please find out the most updated data to elaborate the main findings.	Thank you for your comment. We have updated the discussion to reflect more recent data, referring to publications from 2015–2020. Please see page 12, line 23 onwards.
31. Line 11-29, page 14: this paragraph is used to discuss the 2nd objective, the authors	We appreciate and agree with this comment. We endeavour to discuss this in our paper, but unfortunately data around socioepidemiological

Reviewer 2: comment	Response
should spend more space incorporating the evidence from previous studies to discuss the key findings.	clusters are limited. Interestingly, since we received your comments, a new paper has been published discussing the use of cluster analysis in the HIV/STI field (Blondeel K et al. BMJ Open 2021;11:e33290). We have incorporated this in the discussion section (page 13-14, line 35, 3): 'In a recent study of STI risk among MSMs, hierarchical cluster analysis, another machine learning methodology, identified factors other than behaviour, such as sexual networks and risk perception, that influence the vulnerability to STIs and HIV infections.'
32. Limitation paragraph, page 15: the authors should discuss the reasons for a high volume of missing value, it is due to the quality of data collection or discrimination on sexual orientation	Thank you. The issue of missing data is stated as a limitation of this study upfront under 'strengths and limitations' and described extensively in the discussion section (page 14, line 6–14) as below. 'A key limitation of this study is the high proportion of missing data around sociodemographic and lifestyle characteristics, a common phenomenon in population-based epidemiological studies where questionnaires are used. This may have potentially introduced information bias or inaccurate representation of the true situation when describing high-risk populations. Although not formally assessed, we classify these missing data as missing completely at random due to time constraints in completion of the epidemiological questionnaires by surveillance officers and healthcare professionals who notified the diseases to the surveillance systems. Nonetheless, our findings are similar to those reported in previous analyses.'
33. The authors consider referencing the paper "Recent trends in sexually transmitted infections among adolescents, Catalonia, Spain, 2012–2017" (https://doi.org/10.1177/0956462420940911) which might be relevant for discussing incidence rates and their trend.	Thank you for the suggestion. We agree that this aligns well with our discussion and have described this as follows (page 13, line 9–11): 'Our findings are consistent with earlier studies of STIs Catalonia in 2007–2015, 2012-2017 and 2018-2019, showing a proportionally higher increase in young adults, mostly women, especially for chlamydia but also for gonorrhoea.'
34. The questionnaire collected behavioral data (Table S1), yet the authors excluded these	Thank you for the query. Sexual preference, country of birth and education level were excluded from the logistic regression models

Reviewer 2: comment	Response
individual risk behaviors for STI and HIV co-infection for logistic models. Please provide insights.	because more than 50% of values were missing. This is stated in the methods section (page 8, line 25–26).
References	
35. References #5 and #22 are duplicate	Thank you for pointing out the error, which has been rectified.

Reviewer 3: Dr. Dorcas Obiri-Yeboah, University of Cape Coast

General comments: Well structured, well written generally. Very important and relevant study topic using retrospective data. Abstract: clear and contain all relevant information. Data collection and analysis well described and appropriate. Study limitations have been clearly stated and discussed. Below are some specific relatively minor comments to help improve the manuscript**.**

Reviewer 3: comment	Response
1. Page 6, line 9: you have "The epidemic of sexually transmitted infections (STI) is a major public health concern in high-income-, middle-, and low-income countries". This basically imply the situation is global so can you consider stating just that concisely?	Thank you. We have revised the sentence to reflect your suggestion (page 5, line 2–3): 'The epidemic of sexually transmitted infections (STIs) continues to be a major concern and threat to global public health.'
2. Sexually transmitted infections (STI)- usually to use it as plural, STIs is used. This is used correctly in some places in the manuscript and not used when needed in other places. Please check	Thank you for pointing out the inconsistencies. We have checked and ensured that the right subject-verb agreement is used throughout the manuscript.
3. Page 7 lines 2- 8: you have "Some studies have identified the social determinants of health, discrimination, and inequities as main factors associated with the appearance of STI spatiotemporal clustering of cases". Please rephrase especially towards the last part, it gets unclear.	Thank you for the comment. We have rewritten it as follows (page 5, line 22–24): 'Some studies have described social determinants of health, discrimination and inequalities as the main factors associated with the spatiotemporal clustering of STI case'
4. Gonorrhoea has been spelt in both British and American ways. Be consistent	Thank you for pointing out the inconsistencies. We have ensured the British spelling is used throughout.
5. Page 8 line 8: you have "(total number of episodes due to any STI that had the same person during the study period)", rephrase. I	Thank you. Yes, we indeed meant to say that, so have rephrased as follows (page 7, line 14–16):

Reviewer 3: comment	Response
believe you mean the total number of episodes of any STIs that the same person had during the study period?	'Multiple STI episodes were defined as total number of episodes of any STI reported for the individual during the study period'
6. Page 10 lines 26-32: You have "proportionally, the STI episodes in men were significantly higher than in women, for gonorrhoea, syphilis, and LGV, but less frequent for chlamydia (80%, 87%, 99%, and 38% were in men, respectively)". I find it confusing as I don't see the proportions in women. Consider rephrasing.	Thank you. We have rewritten as follows (page 10, line 1–2): 'The vast majority of reported cases occurred in men for all STI types except chlamydia, of which 61.9% occurred in women (table 1).'
7. Page 10, lines 55-60: this sentence is not clear and needs to be revised---The STI episodes in HIV-positive counted 6% from the overall, however, with higher proportion in syphilis and LGV (13% and 25%, respectively) and the lowest in chlamydia (2%)	We have rephrased as follows (page 10, line 26–28): 'In total, 6% of the STIs episodes affected HIV-positive people, however, higher proportion was observed in syphilis and LGV (13% and 25%, respectively) and the lowest in chlamydia (2%) (table 1).'
8. Page 14 lines 32-35: please rephrase the sentence "Last years, the K-means clustering methodology has proven its potential in classifying and grouping health related outputs in different study fields"	Thank you. We have rephrased as follows (page 13, line 26–27): 'The k-means clustering methodology is a machine learning approach that has proven its utility and potential in classifying and grouping health-related outcomes.'
9. Page 15, line 43: rephrase this "While the HIV trend to decrease, mainly because the wider and earlier use of ARV……"	We have rewritten the phrase as follows (page 14, line 32–34): 'While declines in HIV infection has been observed in the last decade in Catalonia, as in many other regions in Europe, primarily due to the success of wider and earlier use of antiretroviral therapies.'
10. Check the formatting for table 3	Thank you for the suggestion. We have reformatted the tables to hopefully improve the presentation of data.
11. Though secondary data was used, this is a sensitive topic, what ethical considerations did the researchers have? There is no comment on ethics at all as far as I saw.	Thank you for the comment. Please refer to the 'Ethics approval statement' and 'Patient and public involvement' subsections of the methods section on page 8-9.

Reviewer 3: comment	Response
12. Some grammatical and typographical errors to correct e.g. check line 13 (page 5), 48 (page 6), etc.	The revised manuscript has been copyedited by a native English speaker for language and grammatical errors, and to improve readability.

Reviewer 4: Dr. Jean Damascene Makuza, Rwanda Biomedical Center

General comments: This is a good article exploring a sensible subject to help most people in need. The article is well written and has sufficient information on STIs epidemics in the study population however, there are some errors to be addressed like strengthening the abstract, avoid repetitions, abbreviations to spell, …

Reviewer 4: comment	Response
Abstract	
1. Some abbreviations were spelled on the first-time user like STIs, HIV, LGV, and others, please spell them on their 1st use.	Thank you for the comment. We have checked the manuscript to ensure abbreviations are defined at first mention.
2. Results: line 37: you said the number of STIs-cases doubled, what is the basic number or from which period to which period? It is better also to show the figures (numbers or proportion of cases in women and in young people).	Thank you for the query. We have rewritten this in the abstract and the main manuscript so that it is clearer (page 9, line 15-17) 'The number of STI cases doubled from 9,687 in 2017 to 18,872 in 2019 (table 2). The incidence rate of STIs increased by 91.3% from 128.2 cases per 100,000 population in 2017 to 248.9 cases per 100,000 population in 2019.'
3. Conclusion: Your conclusion need to be linked with figures from results and these are not described in the results. In addition to that, there is a need for recommendations in this part based on your findings, please try to suggest any of them.	Thank you for the suggestion. We have rewritten the conclusion (page 3, line 28–30) as follows: "We recommend socioepidemiological identification and characterization of STI clusters and factors associated with HIV coinfection to identify at-risk populations at a health area level to design effective interventions."
4. Strengths and limitations: There are repetitions between points 1 and 4 and these need to be summarized together	Thank you for the helpful suggestion. We have restructured the strengths and limitations section to make the points more succinct.

Reviewer 4: comment	Response
Introduction	
5. Lines 20-25: This incidence is in which population? General population or specific population?	Thank you for the query. The statement refers to increase in incidence reported in national surveillance systems. We have rewritten that sentence as below (page 5, line 6–9): 'Across Europe, incidence of STIs continue to be on the rise with confirmed cases reported in national surveillance systems increasing by 50% for gonorrhoea, 36% for syphilis, 68% for lymphogranuloma venereum (LGV), and 0.6% for chlamydia from 2014 to 2018.'
Methods	
6. For study participants, do you have inclusion and exclusion criteria, if yes can give them?	Our data source was the Catalan AIDS/HIV/STI Surveillance System, which includes all confirmed case per case definitions that are aligned with those established by the ECDC. These case definitions are also aligned with the CDC case definitions for surveillance (https://www.cdc.gov/std/statistics/2019/case-definitions.htm). We aimed to clarify this by reporting this in the methods section (page 6, line 18–21), with further information detailed in supplementary table S2.
7. You cited different variables, which are the dependent and how are they described or defined (continuous or categorical), which are independent? What is the study site and why did you choose it?	Thank you. Table 2 outlines all categories of independent variables, with HIV coinfection as the dependent variable. We believe that this information is clear from the table and in the methods section. This was a retrospective population-based cohort study of the Catalonia region in Spain. We have endeavoured to clarify the site of analysis throughout the manuscript where appropriate.
8. For statistical analysis, page 8 lines 44-48, this sentence describing ABS could go into the study population.	Thank you for the suggestion. We have described ABS (now abbreviated in English as Basic Health Area [BHA]) under a new heading 'Analysis variables' (page 7, line 1–10).
9. Page 9, line 4, I think this is not a univariate analysis, it could be the bivariate analysis, please do the correction. What was the cut-off of the p-value for consideration of variables in the final model (multivariate)? What other criteria did you consider	Thank you. We have rewritten the paragraph to reflect your queries as follows (page 8, line 25–28): 'Sexual preference, country of birth and education level were excluded from the models because more than 50% of values were missing. We used backward stepwise elimination regression to include all analysed variables that showed statistical significance (P<0.05) by the Wald test in the final multivariable logistic regression model.'

Reviewer 4: comment	Response
for variables to be included in the final model?	
Results	
10. Page 10, line 15, please precise the % of STIs for Barcelona which is the highest, and for Alt Pirineu I Aran which has the lowest as you did for urban and rural areas.	Thank you. We have rewritten this part as follows (page 10, line 18–22): Barcelona reported the highest incidence rate of STIs while Alt Pirineu i Aran recorded the lowest consistently throughout the study period (table 2). In 2019, the incidence rate of STIs was 307.8 cases per 100,000 population in Barcelona and 45.7 cases per 100,000 population in Alt Pirineu i Aran.'
11. Page 10, line 34, spell abbreviation WSM.	Thank you for the comment. We have checked the manuscript to ensure abbreviations are defined at first mention.
Discussion	
12. Paragraph 1, page 12-13, line 36-60 then 1-10, this is a pure repetition of the results, can summarize it in other few words? Or delete it definitely as its contents come again in the following paragraphs.	Thank you. While we agree that some elements are repeated, we would respectfully like to start our discussion with an overview of the general findings. We have rewritten it such that it presents a broad summary. Please refer to the first paragraph in the discussion section on page 12.
13. Page 15, line 3-10, you said missing data could lead to biased results, which bias do you expect? Better to precise the types of biases expected.	Thank you for this comment. We believe that the statement that missing data can lead to many kinds of biases remains debatable (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643276/). As detailed discussion of this topic is outside the scope of this paper, we highlight the most relevant one (https://catalogofbias.org/biases/information-bias/) as follows (page 14, line 6–10): 'A key limitation of this study is the high proportion of missing data around sociodemographic and lifestyle characteristics, a common phenomenon in population-based epidemiological studies where questionnaires are used. This may have potentially introduced information bias or inaccurate representation of the true situation when describing high-risk populations.'
14. Do not you think that your study was prone to confounders? Please tell me something about	Thank you for your comment. We believe that most of our variables (e.g. sex, age, deprivation index) may be associated with higher exposure and vulnerability to STIs so they indeed could be cofounders. These are adjusted for in

Reviewer 4: comment	Response
that could confound your results.	the logistic regression model, which we believe would reduce cofounding. In any case, we thought it would be relevant to comment on this as part of discussion around limitations (page 14, line 14-19): 'The age category above 60 years old may contribute to residual cofounding although the risk is minimal because it is the smallest group and the range is larger than for other age categories. Categorisation of the deprivation indices by quintiles could have diluted the findings if deprivation was a strong confounder or unevenly distributed, although we do not believe either event to be the case in our analysis.'
15. Page 15, lines 48-55. There could be a typo error "….These populations need to be considered a priority for the preventive strategies…… Please bring the correction.	Thank you. We have double checked the manuscript to ensure there are no typos throughout.
References	
16. References 3, 10, 15, 20, and 21 seem to be aged, better to use references published no lesser than 2010. Reference 28 is not well presented, please correct it.	Thank you. We have removed these references as suggested and replaced, where appropriate, with more recent publications.
Additional	
17. Please check the English used	The revised manuscript has been copyedited by a native English speaker for language and grammatical errors, and to improve readability.

VERSION 2 – REVIEW

REVIEWER	Thinh, Vu Toan CUNY Graduate School of Public Health and Health Policy, Center for Innovation in Mental Health
REVIEW RETURNED	05-Oct-2021

GENERAL COMMENTS	The authors made substantial edits and have resolved comments. Yet, there are points needed to take into consideration: 1) The authors implemented the predictive model so as to identify factors associated with HIV co-infection, however, the revised manuscript has yet to answer the previous question regarding what statistical tests are used to assess model calibration as well

	as predictive accuracy. Stepwise regression is ruled out by an automatic procedure and discards the least statistically significant variables. That being said, all covariates in the final models are significant but it doesn't mean that this is a good model to explain the variability of HIV co-infection. 2) Sample size: is there any possibility of creating a data flowchart? a. Table 1: the denominator is based on cases (42,283 episodes); i. What if a person has all STIs, their socio-demographic data will be counted more than one. Is that understanding true? ii. It will make more sense if we could know that among a total of 34,600 participants, how many people experienced at least one, two, three or all out of 4 STI types, but not "cases". b. Table 2: it would be great if authors could provide person-time information, it should be counted based on "cases" or "invidivuals" c. Table 4: among 35,831 cases. i. What is different from the 42,283 cases? Please clarify the sample. 3) Additional comments: a. Please be coherent by not separating paragraphs if it contains one or two sentences. b. There remain some typos and the author needs to review the manuscript again. E.g., Table 1 contains many redundant brackets (line 57, page 22).

REVIEWER	Obiri-Yeboah, Dorcas University of Cape Coast
REVIEW RETURNED	31-Aug-2021

GENERAL COMMENTS	Extensive revision has been made to the manuscript. Most of the comments have been addressed satisfactorily, Thanks you. However, consider the following: 1. I feel the title is too long. Consider this "STI epidemic re-emergence, socio-epidemiological clusters characterisation, and HIV coinfection in Catalonia, Spain, during 2017–2019: a retrospective population-based cohort study". I don't think you lose significant content with this title. The study objective will then include the additional details 2. socioepidemiological should be socio-epidemiological 3. abstract under participants, insert WHO so it reads "42,283 confirmed syphilis, gonorrhoea, chlamydia, and lymphogranuloma venereum (LGV) cases among 34,600 individuals who reported to the Catalan HIV/STI Registry in 2017–2019" 4. under strengths and limitations, the bullet number 3 does not fit. It states "• MSM, heterosexual women and young adults should be considered priority target populations for preventative strategies of STI and HIV, taking into account structural and social determinants that were identified as crucial in this analysis". It is neither a strength nor limitation of this study 5. extensive revision was made so please read through once again for some grammatical and typographical errors

REVIEWER	Makuza, Jean Damascene Rwanda Biomedical Center, Institute of HIV, Diseases Prevention and Control

REVIEW RETURNED	31-Aug-2021

GENERAL COMMENTS	Thank you for considering reviews, currently, the manuscript looks good. All my suggestions and comments were considered.

VERSION 2 – AUTHOR RESPONSE

Answer to Reviewer 2 Answer to:
Mr. Vu Toan Thinh, CUNY Graduate School of Public Health and Health Policy

Comments to the Author:
The authors made substantial edits and have resolved comments. Yet, there are points needed to take into consideration:
We appreciate your valuable improvement suggestions in both rounds' revisions. In spite of this, on this occasion we have to disagree in some of your suggestions, please see below our explanations. We hope you understand our reasons.

1) The authors implemented the predictive model so as to identify factors associated with HIV co-infection, however, the revised manuscript has yet to answer the previous question regarding what statistical tests are used to assess model calibration as well as predictive accuracy. Stepwise regression is ruled out by an automatic procedure and discards the least statistically significant variables. That being said, all covariates in the final models are significant but it doesn't mean that this is a good model to explain the variability of HIV co-infection.
Thank you for your comment. We did not evaluate the model itself, since is not a predictive model but an explanatory model. In this case, it is not necessary to assess the accuracy.
As mentioned in the introduction, all the variables in the epidemiological questionnaire and those "additional" variables analysed such is the case of deprivation index have been described as potential risk factors for HIV coinfection. Keeping this is mind and our specific objective ("determine factors associated with HIV coinfection in Catalonia, Spain, during 2017–2019"), we want to disentangle which of these potential risk factors, among the patients in our cohort, were more strongly associated or not show association with HIV coinfection and the magnitude of this association. We believe that the model is good to provide and answer to this objective which we think is not exactly the same than trying to explain the variability of HIV coinfection.

2) Sample size: is there any possibility of creating a data flowchart?
Thank you for your suggestion. As we have used centralise data from surveillance systems, which used the presented case definitions provided in the supplementary material, we think that there is no any additional exclusion which could be presented in a flowchart. In other words, the sample size, as explained in the manuscript are all the number of confirmed reported cases in the STI surveillance system in Catalonia.

a. Table 1: the denominator is based on cases (42,283 episodes);
i. What if a person has all STIs, their socio-demographic data will be counted more than one. Is that understanding true?
ii. It will make more sense if we could know that among a total of 34,600 participants, how many people experienced at least one, two, three or all out of 4 STI types, but not "cases".
Thank you for your comment. We have largely discussed if it was more appropriate present the first descriptive table by person as you suggest. But as this manuscript is a piece of work about communicable disease surveillance, we thought essential to include a table with all the episodes as it is presented in table 1 because actually is a very common analysis displayed in the yearly surveillance reports delivered by many public health agencies.

We think that table 1 provides the big picture of the magnitude of the STI epidemics. Besides as people with reinfection count more times in their categories, the results highlight the categories of the more vulnerable individuals which in a person-base table will become more diluted.
Nevertheless, in table 2 is displayed the proportion of different ranges of number of STI episodes by person but, we believe that information by disease in table 1 is really needed for surveillance (more than the analysis of potential combination of different STI in the same individual). We think that most of the analysis you are suggesting are very relevant to be studied but our manuscript is already very dense in terms of results and these new suggestions need to be tackle in further studies.

b. Table 2: it would be great if authors could provide person-time information, it should be counted based on “cases” or “invidivuals”
Thank you for your suggestion but the incidence rate as it provided is the usual way to display the results in surveillance since the main objective is to monitor the episodes per year. We truly believe that maintained the usual way to provide results is the most appropriate for specialists in surveillance read, interpret and analyse our manuscript. Besides, since is very complicate or impossible to determine the real time of exposure by person and giving the added complexity to interpret some person time outputs (e.g., person days), we truly believe that our table 2 works better as it is now displayed.

c. Table 4: among 35,831 cases.
i. What is different from the 42,283 cases? Please clarify the sample.
As mentioned in page 11 line 5-7: “Of the 373 Catalan BHAs, five (Garraf rural, Polinyà-Sentmenat, Ribes-Olivella, Roquetes-Canyelles and Viladecans 3) were excluded from the K-means clustering analysis because their delimitations and populations changed during the study period.” Besides we had 5,773 episodes with no information available about BHA of residence. These facts reduced the sample size for the cluster analysis from 42,283 to 35,831 STI cases. In order to clarify this, point we have added the following sentence (page 11 line 7-11): “In these five BHAs 679 episodes were reported during the three years of the study period. This fact and having 5,773 episodes with no information available about BHA of residence reduced the sample size for the cluster analysis from 42,283 to 35,831 STI cases”

3) Additional comments:
a. Please be coherent by not separating paragraphs if it contains one or two sentences.
Thank you very much for you suggestion, we have now solved this issue.

b. There remain some typos and the author needs to review the manuscript again. E.g., Table 1 contains many redundant brackets (line 57, page 22).
Thank you for your remark, we have reviewed the manuscript again.

Answer to Reviewer 3 Answer to:

Dr. Dorcas Obiri-Yeboah, University of Cape Coast

Comments to the Author:
Extensive revision has been made to the manuscript. Most of the comments have been addressed satisfactorily, Thanks you. However, consider the following:
1. I feel the title is too long. Consider this "STI epidemic re-emergence, socio-epidemiological clusters characterisation, and HIV coinfection in Catalonia, Spain, during 2017–2019: a retrospective population-based cohort study". I don't think you lose significant content with this title. The study objective will then include the additional details
Thank you for your remark, the tittle has been changed following your suggestion.

2. socioepidemiological should be socio-epidemiological
Thank you for your remark, we have modified the word in all the text.

3. abstract under participants, insert WHO so it reads "42,283 confirmed syphilis, gonorrhoea, chlamydia, and lymphogranuloma venereum (LGV) cases among 34,600 individuals who reported to the Catalan HIV/STI Registry in 2017–2019"
Thank your comment. We have added the word in the abstract.

4. under strengths and limitations, the bullet number 3 does not fit. It states "• MSM, heterosexual women and young adults should be considered priority target populations for preventative strategies of STI and HIV, taking into account structural and social determinants that were identified as crucial in this analysis". It is neither a strength nor limitation of this study
Thank you for your comment. Following your suggestion, we have removed it.

5. extensive revision was made so please read through once again for some grammatical and typographical errors
Thank you for your remark, we have reviewed your previously highlighted words-sentences in yellow in the manuscript again.

Answer to Reviewer 4 Answer to:

Dr. Jean Damascene Makuza, Rwanda Biomedical Center, The University of British Columbia School of Population and Public Health

Comments to the Author:
Thank you for considering reviews, currently, the manuscript looks good. All my suggestions and comments were considered.

Thank you very much for your review.