Comments by Reviewer 1

The authors addressed all my queries accordingly.

Response: Thank you for the careful review of our manuscript. Your comments were highly insightful and enabled us to greatly improve the quality of our manuscript.

Comments by Reviewer 2

I'd like to thanks the authors for addressing my comments but I still have some further queries.

1) In the revised text the authors mention the three surgeries were chosen because the number of female surgeons who performed these surgeries was sufficient for analysis – this is vague. What constitutes 'sufficient'? Was this some sort of sample size calculation required for the analysis?

Response: We thank you for your careful review and helpful comments. We would also like to apologise to the reviewer for the unclear description. The number of surgeries such as pancreaticoduodenectomy performed by female surgeons was too low for anonymous analysis. For example, the number of pancreaticoduodenectomies performed by female surgeons belonging to the age category of ≥21 years after medical licence registration was 108 during the 5-year study period (unpublished data). Therefore, experts in these surgical fields in Japan can identify them, and presenting the adjusted odds ratios would not be appropriate because it directly publicises their quality of surgery.
Therefore, we have revised our text accordingly to clearly indicate that the three surgeries were sufficient for analysing data anonymously, i.e., without the individual surgeon being identified by the readers (page 3, lines 110–112).
"These three procedures were chosen because the number of female surgeons who performed these surgeries was sufficient for analysis without the individual surgeon being identified."

2) In the main analysis the authors mention that continuous variable were categorised to account for non-linear relationships between the variable and outcome. I mentioned this in my previous comment – why did you not explore this and show the relationship – you could have used splines or polynomial regression to account for this non-linearity. I accept you did what planned to do and do later explore some of this in the newly added sensitivity analysis

Response: We completely agree that non-linear relationships should be analysed with appropriate methodologies. It was conventional to categorise continuous variables in previous research and the adjusted odds ratios of the categorical variables were easier to understand than that of splines where the other covariates needed to be fixed at a certain, sometimes arbitrary, value for obtaining the odds ratios of the splines. In addition, as we responded to the committee's comment, the categorisation of continuous variables enabled observation of the differences between male and female surgeons in greater detail for each category of the years after medical licence registration, as shown in Table 1. If only the median value had been presented, the unequal frequency distribution between male and female surgeons in each category could not have been elucidated. Finally, exploring non-linear relationships in depth was not the aim of our study. However, we admit that a more careful research plan should have been developed. We also agree that different methods for treating continuous variables may have yielded different results and they should have been explored with the given data for

analysis. We would like to thank the reviewer for accepting what we planned for analysing the data at the initiation stage of the study and what we did later as a sensitivity analysis. We thank the reviewer for their recommendation. We believe that the results of the sensitivity analysis made our conclusions more robust.
We have added the following text in the sensitivity analysis (page 5, lines 200–202):
"This analysis was included to explore confounding effects that might vary from previous studies depending on how to model the non-linear relationship between the variable and outcome."

3) However, later on the authors say non-linear relationship were assumed based on previous research on the volume-outcome relationship. This seems less satisfactory – you have data, why don't you explore the relationship in your data.

Response: As mentioned in our previous response, we agree that one should consider modelling non-linear relationships with appropriate methods such as splines and not with arbitrary categorisation extracted from previous research. We hope the added text above is sufficient to inform the readers about the less appropriate approach in the initial analysis.

4) Your additional sensitivity analysis are post-hoc analyses which are not described in your protocol – you should describe them as such. These analyses become more exploratory.

Response: We completely agree that this should be stated explicitly. We have revised the subtitle and the text of the sensitivity analysis section accordingly (page 5, lines 191 & 193–194)

5) In one of the tables you have a footnote to suggests that number of surgeons in each category does not add up to the total number of surgeons in the study population because some surgeons moved to a higher category (in terms of seniority) during the study period. So this is clearer why do you not have a rule and report the highest category achieved in the study timeframe.

Response: We appreciate your kind suggestion. We have revised the table so that the addition of the number of each category represents the total number of surgeons in the study population (Table 1).

6) I would check all tables and results for errors. E.g. table 2 – you need to specify the number of surgeries per year is reported as a median and IQR. I think the last category for the number of surgeries per year should also be >=50. Same applied for all other tables. Report all estimates to a consistent number of decimal places.

Response: We would like to thank the reviewer for pointing this out. We have checked through all tables and results, corrected errors, and revised vague expressions (Tables 1–4, Supplementary Figures 1–4, Supplementary Tables 1–7). The number of decimal places for p values may seem inconsistent, but we followed the convention as follows: report to two decimal places when p values are >0.01, report to three decimal places when p values are between 0.01 and 0.001, and report p<0.001 when p values are <0.001 [1]. We were not able to find specific guidance on decimal places for publication in the BMJ, but if there is one, we would like to follow it. Please let us know.

7) I think your Abstract needs to report some of the OR's and CI's you found from the main analysis. I wonder if all the OR's to the analyses performed would also be better in a Table as well as reported in the text.

Response: We would like to thank the reviewer for pointing this out. We agree that ORs and CIs would be informative for the readers, but because of the limited word count, we could not include them in the abstract. Now, we have added the ORs and CIs and removed some parts of the previous abstract to meet the word limit (page 2).
The ORs and CIs are presented in a table format in Figure 2. As per the recommendation of the other reviewer in the previous round, we have presented the results with a graphical representation.

Reference
1. Aguinis H, Vassar M, Wayant C. On reporting and interpreting statistical significance and p values in medical research. *BMJ Evid Based Med* 2021;**26**:39–42. doi:10.1136/bmjebm-2019-111264