Comments by the Committee

- It is unclear why you have categorised many of the adjustment variables or hospital case numbers - why not treat these as continuous? Some adjustment has been made for cases from different hospitals but the same hasn't been done at the surgeon level, i.e. many of the surgeons in the study are undertaking many surgeries

Response: Thank you for pointing this out. We admit that the richness of the data may have been lost since the continuous variables are categorised. However, we categorised the variables because we assumed that the effect of a one-unit-increase in those variables would not be equivalent for patients across the range for those variables. For example, a one-year-increase in age does not have the same effect for those in their 20s as for those in their 80s. Based on clinical experience and knowledge, we believe that it is generally accepted that risks do not increase linearly according to age. As for case volume, a downward trend or a hockey-stick like shape is often observed in the relationship between outcome and case volume, and for body mass index and laboratory data, a U-shaped relationship is often seen between an outcome and a variable. These relationships would not be accounted for if the variables were considered continuous and treated linearly; thus, we categorised the variables to account for a non-linear relationship. The categories that were assigned in our study were based on those used in previous studies, including mortality and morbidity prediction model development using the NCD (Hoshino et al., Gotoh et al.). The number of years after medical licence registration were grouped based on the following assumption: surgeons with an experience of 5 years or less were considered to not have completed the general surgery training program; those with an experience of 6-10 years were assumed to be board certified general surgeons; 11-15 years, board certified gastroenterological surgeons; 16-20, board certified trainers; and 21 years or more, directors (or a similar position) of surgical departments. The board certification system and hence, the years after medical licence registration, may differ in other countries, but the rationale behind this categorisation was to group surgeons according to surgical skill levels. We have added this explanation in the manuscript (page 4, lines 174–180).

Furthermore, owing to this categorisation, we were able to observe the differences between male and female surgeons in greater detail for each category, as shown in Table 1. If only the median value was considered, the unequal frequency distribution between male and female surgeons for the years after medical licence registration could not have been elucidated.

However, we agree that the categorisation might be arbitrary and hinders reader comprehension. We added summary statistics for the continuous variables in the tables (Table 1–4) and performed a sensitivity analysis that included years after medical licence registration, patient age, patient body mass index, and hospital case volume as continuous variables (page 5, lines 200–203) and confirmed that this new analysis did not affect our conclusion (supplementary Fig 1–3). Unfortunately, laboratory data such as haemoglobin concentration are not collected in a continuous manner; these data are provided only if they are abnormal. Regarding the adjustment for cases at the surgeon level, we assumed that case volume is a surrogate of surgical experience that significantly affects outcome. For hospitals, hospital case volume would reflect surgical experience. On the other hand, for surgeons, we considered years after licence registration to be a more accurate measure of surgical experience rather than the annual

case volume of the individual surgeon because the years after licence registration account for surgical experience during the entire professional career, not just for the surgical experience of that year. Furthermore, in a previous study using the NCD, hospital case volume had a stronger effect on improved operative mortality than surgeon case volume (Iwatsuki et al.). Lastly, surgeon case volume and hospital case volume are relatively highly correlated (correlation coefficient = 0.5). For better interpretability and computational efficiency, we did not include surgeon volume in the regression analysis. However, we agree that surgeon volume may be a candidate confounding factor, particularly for surgeons who perform a higher number of surgeries (as you pointed out). Thus, we conducted an additional regression analysis including surgeon volume (page 5, lines 204–211). Our findings are described in the results section (page 7, lines 328–335) and in the supplementary material (supplementary Fig 1–3).

- You have adjusted for a lot of covariates, many of which do not appear to be confounders. Therefore, the authors could expand why they adjusted for these factors. Should these confounders be adjusted for? 1) average annual surgery volume per surgeon (It's possible female surgeons generally have lower average volume). 2) SES of the patients. 3) Region of the hospital

Response: We would like to thank the committee for these valuable comments. We apologise that the rationale of the statistical methods was not clear in the manuscript. Our assumption was that just as preoperative risks could affect whether a female or male surgeon is assigned to the patient this could also affect the surgical outcome. We first believed that patients with high surgical risk would be more likely to be assigned to male surgeons. Hence, risk factors affecting the outcome were considered to be confounders in this analysis. The preoperative risk factors that we selected were chosen from subject matter knowledge and previous research, including risk prediction model development using the NCD (Hoshino et al., Gotoh et al.).
As for the additional candidate confounders, surgical volume was included in the additional analysis. Unfortunately, data regarding socioeconomic status (SES) were not available in our database. As far as we were able to find, in the literature, the regional average household income, which only partly captures the concept of SES, did not affect the outcome of cardiovascular surgery in Japan (Lee et al.). This finding may apply only to Japan and also may not be applicable to other surgical fields. We agree that the lack of data on SES is a limitation of this study; however, later in the manuscript, we included urban-rural status as an additional covariate in the regression analysis for the sensitivity analysis, and SES may be partly explained by this status. We included this explanation in the methods and discussion sections (page 6, lines 211–221; page 9, lines 431–432). Finally, we agree that the region of the hospital is also a candidate confounder. We adjusted for hospital-level characteristics and considered these to be stronger confounders than the region of the hospital variable because regionality is considered to be a characteristic of a higher level in a nested structure and the region would not have a direct effect on the choice of the gender of the surgeon; however, we also performed an additional regression analysis that included the urban-rural status of the hospital as a covariate (page 6, lines 211–221; supplementary table 4 & supplementary Fig 1–3).

- There are big differences between men and women in the types of surgeries they are performing, raising concerns about using MV models. Perhaps they should be stratified by surgical approach? Moreover, shouldn't hierarchical models be used (or at least GEE) to account for the clustering of patients by surgeon/hospital? Do surgeons operate at multiple hospitals?

Response: We apologise for the unclear description of the methods and the flow diagram. The study population was recruited from among patients who underwent three different surgical procedures, i.e., distal gastrectomy, total gastrectomy, and low anterior resection. We conducted a regression analysis separately for each surgical procedure. A multilevel (or hierarchical) model was used to account for hospital level characteristics. We have revised our manuscript and flow diagram to clarify this part (pages 4–5, lines 155–159; Fig 1). Surgeons may operate at multiple hospitals; therefore, it may not be reasonable to assume that surgeons are nested in hospitals. Surgeon-level characteristics were not treated as random effects. We used the year after medical licence registration as a fixed effect to account for surgeon-level characteristics. In the additional sensitivity analysis, surgeon case volume was added as a variable to further account for surgeon characteristics (page 5, lines 204–211; supplementary table 4 & supplementary Fig 1–3).

- Competing risk was not discussed.

Response: We apologise that our original text may have mislead readers; however, our study was an ordinary regression analysis with a binary outcome, not a survival analysis. In accordance with the comments made by the reviewers, the methods section was revised for enhanced clarity and readability; we hope that our revisions will aid reader comprehension (page 4, lines 136–137; page 4, lines 155–160). Please let us know if further clarification is needed. We would be more than willing to address any further concerns.

- Complete case analysis was done, but some ideas about the extent of missing data will be essential.

Response: Thank you for pointing this out. We agree that information regarding the extent of missing data was lacking in our manuscript. Thus, we provided additional summary tables to present information regarding patients with missing values for each surgical procedure and conducted an additional analysis with multiple imputation (page 5, lines 196–199; supplementary tables 1–3 & supplementary Fig 1–3). We were able to confirm that the adjusted odds ratios changed only minimally and our conclusion did not change based on the results of this analysis.

-More context is needed to understand how a surgeon is assigned to a procedure. Are there patients' preference? Some other factors? Or is it kind of random? This will also help us understand if simple adjustments are adequate in dealing with a potential selection bias.

Response: We appreciate your valuable suggestions. Generally, in Japan, patients cannot nominate a primary surgeon; they are assigned to each surgery at random or at the discretion of the department head. We have accordingly made the following additions to the main text (page 8, lines 403–406).
"Generally, in Japan, patients cannot nominate a primary surgeon, and primary surgeons are assigned to each surgery at random or at the discretion of the department head discretion; thus, the process for case assignment to female surgeons by supervisors is essential in the training process for female surgeons."

- *Findings for Female surgeons with <=5 years of experience are concerning. Will that affect their employability and trust/confidence among the patients?*

Response: We are grateful for your valuable suggestion. The mortality and postoperative complication rates of DG performed by female surgeons with an experience of ≤5 years and LAR by females with an experience of 16-20 years were statistically higher than those by male surgeons with the same years of experience, respectively; however, female surgeons with 6-10 years of experience showed better outcomes in DG, and female surgeons in other subgroups tended to have the same surgical outcomes as their male counterparts. Furthermore, the rate of leakage in LAR performed by female surgeons with ≤5 years of experience was lower than that in males. We believe that the slightly worse outcomes of DG and LAR in a few subgroups may not be enough to affect the employment of female surgeons or the trust and confidence of their patients. Further, because there are far fewer female surgeons than male surgeons, the statistical analysis is susceptible to female surgeons who represent extreme outliers, and our findings must be interpreted cautiously. In order to clearly show that there are fewer female surgeons than male surgeons and that the distribution of years since medical licence registration differs considerably between the two groups, the characteristics of surgeons have been separately listed in Table 1.
This information was included in the discussion section as follows (pages 7–8, lines 362–370):
"For DG performed by female surgeons with an experience of ≤5 years post-registration, the adjusted odds ratio for 'surgical mortality' and 'surgical mortality with a complication grade of CD-3 or higher' were statistically higher than those for male surgeons of the same category. For LAR, females with an experience of 16-20 years had a statistically higher adjusted risk for 'surgical mortality' than males with the same surgical experience. However, the adjusted risks for 'surgical mortality or a complication grade of CD-3 or higher' in DG performed by female surgeons with 6-10 years of experience was lower than those for males, and the rate of leakage in LAR performed by female surgeons with ≤5 years of experience was lower than that for males. Female surgeons in other subgroups in DG and LAR and in all subgroups in TG tended to have comparable surgical outcomes to male counterparts."

- *Could you include operations in other surgical fields (but maybe this is beyond the scope of this study?*

Response: Thank you for your excellent suggestion. We believe that it is essential to promote such studies in the future. We are a research team consisting mainly of gastrointestinal surgeons. It would,

thus, have been challenging to discuss the validity of the analysis of postoperative complications for surgeries other than gastrointestinal surgical procedures. Similar studies in other subspecialties of surgery are awaited in the future.

- There is not a lot of detail on how the data were collected nor how complete they are. "Post operative" could cover quite a time period.

Response: We completely agree with you; the methods section lacked details regarding the data entry system of NCD. We provided more details in the manuscript as follows (pages 3, lines 107–109):
"The NCD data entry system does not allow missing values except for laboratory data that were not taken from the patient. Validity of the data entries is evaluated through site visits and audits every year and has been proven to be high."
The postoperative time period was defined as that within 30 days after surgery (page 4, lines 142–144).

- It should be clarified in main text that only surgeries for gastric cancer/rectal cancer were included (flow diagram).

Response: Thank you again for pointing this out. As mentioned earlier, we revised our manuscript and flow diagram to include this information (page 3, lines 124–125; Fig 1).

- Error of number: table 4, OR for Surgical mortality or severe complications of DG: 1.28 (0.93, 1.14).

Response: We apologise for this error. The true point estimate was 1.03; we have corrected this in the manuscript (page 6, line 294; Fig 2).

- We don't see a role for PPI here as it would turn into another research question. You could declare this and how they plan to disseminate it.

Response: Thank you for your helpful advice. The PPI section in the manuscript was modified as follows (page 5, lines 235–240):

"Although patients and the public were not involved in the conception, design, or implementation of this study, we wish to publicise the study results among patients and the public to raise awareness regarding the surgical outcomes of female surgeons being comparable to those of their male counterparts. In the Japanese society, it has been a concern that women spend more time engaged in housework and childcare, making it difficult for them to work in a profession such as surgery. We would like to widely publicise these results through the media and public symposiums to encourage women's participation in professional fields, including surgery."

- *To make this paper better you could stress these 3 surgeries are only representative but that across medicine equity and in training, inclusion, mentoring and practice produce better medicine.*

Response: Thank you for your helpful advice. We have added this information in the manuscript as follows:
Study design and data source section (Page 3, lines 114–115).
"Other procedures among the aforementioned eight were considered difficult to analyse because fewer female surgeons performed these procedures."
Discussion section (page 8, lines 422–423)
"The three surgical procedures we analysed are only representative, but we believe that equality in training, inclusion, mentoring, and practice across the genders would produce better outcomes in medicine."

- *Please submit the study protocol.*

Response: We have included a study protocol file along with the revised version of the manuscript.

References

1 Hoshino N, Endo H, Hida K, et al. Emergency surgery for gastrointestinal cancer: A nationwide study in Japan based on the National Clinical Database. *Ann Gastroenterol Surg* 2020;4:549–61.
2 Gotoh M, Miyata H, Hashimoto H, et al. National Clinical Database feedback implementation for quality improvement of cancer treatment in Japan: from good to great through transparency. *Surg Today* 2016;46:38–47.
3 Lee SL, Hashimoto H, Kohro T, et al. Influence of municipality-level mean income on access to aortic valve surgery: a cross-sectional observational study under Japan's universal health-care coverage. *PLoS ONE* 2014;9:e111071.
4 Iwatsuki M, Yamamoto H, Miyata H, et al. Effect of hospital and surgeon volume on postoperative outcomes after distal gastrectomy for gastric cancer based on data from 145,523 Japanese patients collected from a nationwide web-based data entry system. *Gastric Cancer* 2019;22:190–201.

Comments by Reviewer 1

1. Severe postoperative outcomes: define if this is in-hospital? Who classified the postoperative complications (e.g. CDC)?

Response: We apologise that we did not provide sufficient details regarding this outcome variable. Severe postoperative complications were defined as those that occurred within 30 days after surgery. We revised our manuscript accordingly (page 4, lines 142–147). The Clavien–Dindo (CD) classification, which was proposed by Dindo et al., was used for categorising the complications (Reference #15: Dindo et al.).

2. Have the authors considered using any postoperative complication as an outcome of interest?

Response: While we considered using any complication as an outcome of interest, herein, we only considered complications with a CD grade of ≥3 because there was concern regarding underreporting of non-severe complications. Further, we considered severe complications that require extensive treatment as being of greater interest. The definition of complications with a CD grade of ≥3 is described in the methods section (page 4, lines 144–147):
"The CD classification was proposed by Dindo et al. for evaluating postoperative complications and comparing them among different hospitals, and a CD grade of ≥3 indicates that surgical, endoscopic, or radiological procedures are required for the treatment of the complication."

3. Statistical analysis section: It would be helpful to indicate what an OR >1 indicate.

Response: Thank you for your helpful advice. We have added the relevant information in the methods section as follows (page 4, lines 159–160):
"An adjusted odds ratio (OR) of >1 indicated a higher risk and an adjusted OR of <0 indicated a lower risk of the analysed outcome."

4. Page 4 (Line 57). 'A total of 14,0971 eligible DG surgeries…' Please note typo on the total number of DG surgeries.

Response: We apologise for this typographical error; this was corrected in the revised manuscript (page 6, line 249).

5. Adjusted variables. Its noticeable that all adjusted variables were dichotomised. It would be important to describe the method used for this? And also the reason for not using it as a continuous variable. Perhaps even an ad hoc analysis included as an online supplementary material would be helpful.

Response: Thank you for this comment. We performed an ad hoc analysis including the number of years after medical licence registration, patient's age, patient's body mass index, and hospital volume as continuous variables (page 5, lines 200–203; supplementary Fig 1–3).

6. Tables 4-6. The non-adjusted ORs should be added. The footer should also include the variables that were used for the adjustment. Is it possible to include another column and have the analysis done for overall cohort (e.g. DG + TG + LAR as gastrointestinal surgeries)? This would address some of the limitations reported in the discussion section.

Response: We completely agree that non-adjusted ORs should be presented along with the adjusted ORs. In order to demonstrate how risk adjustment worked for the ORs, we created figures instead of tables (Fig 2–5). In addition, the main analysis was repeated for the overall cohort, including the three procedures (page 5, lines 222–226; supplementary Fig 4). As you pointed out, this would alleviate the bias created due to the small number of female surgeons who had a strong effect on the result.

7. 'Surgical mortality or complications with a CD classification of 3 or higher'. Could this be changed to severe postoperative complication? As mortality is classified as CDC V (i.e. redundant).

Response: We apologise for the unclear definition of surgical mortality provided in the original manuscript. Surgical morality was defined as in-hospital death that occurred within 90 days postoperatively and any death up to 30 days postoperatively, whereas the time frame of severe postoperative complications was defined as 'occurring within 30 days after surgery'. CD class 5 only includes death within 30 days after surgery. The rationale of extending the time frame for death was that death can occur long after operation during index hospitalisation. The number of in-hospital deaths nearly doubles if the time frame is extended from 30 days to 90 days. Since death is the most severe type of complication, it should be properly evaluated, with a sufficient time period. The definition of surgical mortality used herein is commonly used in previous studies based on the NCD (Gotoh et al.), and we believe that this definition is well-accepted. We apologise for not explaining this in the text. The relevant information has been added in the methods section as follows (page 4, lines 137–144):
"In this study, surgical mortality was defined as all-cause death up to 30 days postoperatively, including death that occurred after discharge, and deaths that occurred within 90 days postoperatively after the index hospitalisation. The extended time frame for mortality after index hospitalisation was intended to provide sufficient time for the outcome to be captured, because nearly the same number of patients die between 30 and 90 days after surgery as those within 30 days. This measure has been commonly used in previous NCD-based research to evaluate surgical outcomes. Severe postoperative complications were defined as any postoperative surgical and medical complications with a Clavien–Dindo (CD) classification of ≥3 that occurred within 30 days postoperatively."

8. Page 5 (Line 18): 'Female surgeons performed surgeries on relatively high-risk patients.' It would be interesting to compare outcomes for high-risk patients / low risk patients only between female x male surgeons.

Response: Thank you for your valuable suggestion. We agree that it is an interesting comparison. We performed the relevant analysis and described this in the text (page 5, lines 227–230); the results are presented in supplementary tables 5–7.

Best wishes!

References

1. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: A new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg* 2004;240:205–13.
2. Gotoh M, Miyata H, Hashimoto H, et al. National Clinical Database feedback implementation for quality improvement of cancer treatment in Japan: from good to great through transparency. *Surg Today* 2016;46:38–47.

Comments by Reviewer 2

This study compared the overall 30-day mortality rate and complication rates between female and male surgeons in Japan for elective distal gastrectomy, total gastrectomy, and low-anterior colon resections between 2013-2017 in the Japanese National Clinical Database. There was no difference in mortality rates, complication rates, pancreatic fistulas, or length of hospital stay between female and male surgeons, even though the female surgeons performed more open surgical cases and had higher risk patients.
This study has the usual criticisms that can be leveled against any study that is examining outcomes from administrative data. The large number of cases studied should have resulted in some statistical differences which were not seen.
This is a valuable study for the gender comparison with these operations

Response: Thank you for your feedback. This study showed no statistically significant difference in surgical outcomes between male and female surgeons, despite some limitations. This report encourages more women to participate in the field of surgery and, through their efforts, contribute to improving surgical skills and patient outcomes.

Comments by Reviewer 3

1) What is the rationale for choosing just three surgeries – this could be seen as a highly selective cohort. Why did you not look at all surgeries – some rationale is needed. You could look at surgery type as a factor rather than looking at them separately

Response: Thank you for pointing this out. While this study could have been more valuable if all types of surgeries were investigated, unfortunately, we only had access to a database of gastroenterological surgeries, and very few other procedures performed by female surgeons were registered in the database. In addition, for low-risk surgery, such as inguinal hernia surgery and laparoscopic cholecystectomy, data on preoperative risk factors and outcomes are not included in the database. These surgical procedures may be of interest in terms of elucidating other aspects of the gender gap; however, we aimed to focus on whether there is a substantial difference in the surgical quality of male and female surgeons. This could only be investigated for surgical procedures that a sufficient number of female surgeons had performed and for which a sufficient number of outcomes of interest occurred and were collected in the database, for valid statistical analysis. We have added this information in the main text (page 4, lines 112–115).
Further, we agree that the three procedures could be analysed together. The main analysis was repeated for the overall cohort, including the three procedures (page 5, lines 222–226; supplementary Fig 4).

2) Can you also justify the date range, 2013 to 2017 is a bit dated – is what is presented still reflective of current practice – again this could be seen as highly selective.

Response: Herein, we used the data that were available at the inception of study design. It took us a while to begin the study because data on gender and years after medical licence registration of the doctors could not be used initially, as these data were personal information of doctors: we needed time so that doctors could be provided the opportunity to refuse to their personal information being used in this study. The gender gap is a sensitive issue in Japan; thus, we were required to first build consensus among surgeons for conducting such a study. Because until now, no dramatic changes have been made to improve the working environment of female doctors, we believe that our results still reflect the current situation.

3) There is some missing data in this data set which is not discussed in much detail – a complete case analysis was in fact performed, can the missingness not be investigated further. Can you describe the case mix of those with missing data for example, are they reflective of those used in the full analysis.

Response: Thank you for pointing this out. We agree that information regarding the missingness in our data was lacking in our manuscript. Thus, we provided additional summary tables to present information regarding patients with missing values for each surgical procedure and conducted an additional analysis with multiple imputation (page 5, lines 196–199; supplementary tables 1–3 & supplementary Fig 1–3). We were able to confirm that the adjusted odds ratios changed only minimally and our conclusion did not change based on the results of this analysis.

4) The time point of analysis for all outcomes also needs justifying, this can be important here on outcome. Surgical mortality was defined as in-hospital deaths up to 90 days post-op but any death up to 30 days post-hoc – why the inconsistency? Which was used in the mortality analysis?

Response: We apologise for the unclear definition that was included in the original manuscript. Surgical morality was defined as any death up to 30 days postoperatively, including those who were discharged from hospital, and death that occurred within 90 days postoperatively during index hospitalisation. The time frame for in-hospital deaths was extended because death can occur long after operation during index hospitalisation. The number of in-hospital deaths nearly doubles if the time frame is extended from 30 days to 90 days. Since death is the most severe type of complication, it should be properly evaluated over a sufficient period. The definition of surgical mortality is used commonly in previous studies based on the NCD (Gotoh et al.) and we believe that this definition is or can be well accepted among surgeons. We apologise for not explaining this in the text. The relevant information has been added in the methods section as follows (page 4, lines 137–144):
"In this study, surgical mortality was defined as all-cause death up to 30 days postoperatively, including death that occurred after discharge, and deaths that occurred within 90 days postoperatively after the index hospitalisation. The extended time frame for mortality after index hospitalisation was intended to provide sufficient time for the outcome to be captured, because nearly the same number of patients die between 30 and 90 days after surgery as those within 30 days. This measure has been commonly used in previous NCD-based research to evaluate surgical outcomes. Severe postoperative complications were defined as any postoperative surgical and medical complications with a Clavien–Dindo (CD) classification of ≥3 that occurred within 30 days postoperatively."

5) I am unclear why so many continuous variables were categorised – it makes more sense to treat these as continuous ad would make the paper easier to read.

Response: Thank you for your valuable comments and suggestions. However, we categorised the variables because we assumed that the effect of a one-unit-increase in those variables would not be equivalent for patients across the range for those variables. For example, a one-year-increase in age does not have the same effect for those in the 20s as for those in the 80s. Based on clinical experience and knowledge, we believe that it is generally accepted that risks do not increase linearly according to age. As for case volume, a downward trend or a hockey-stick like shape is often observed in the relationship between outcome and case volume, and for body mass index and laboratory data, a U-shaped relationship is often seen between an outcome and a variable. These relationships would not be accounted for if the variables were considered continuous and treated linearly; thus, we categorised the variables to account for a non-linear relationship. The categories that were assigned in our study were based on those used in previous studies, including mortality and morbidity prediction model development using the NCD (Hoshino et al., Gotoh et al.).
However, we agree that the categorisation might be arbitrary and hinders reader comprehension. We added summary statistics for the continuous variables in the tables (Table 1–4) and performed a sensitivity analysis that included years after medical licence registration, patient age, patient body mass

index, and hospital case volume as continuous variables (page 5, lines 200–203) and confirmed that this new analysis did not affect our conclusion (supplementary Fig 1–3). Unfortunately, laboratory data such as haemoglobin concentration are not collected in a continuous manner; these data are provided only if they are abnormal.

6) The authors need to describe in full the statistical models that were used, how variables were selected for inclusion, the assumptions and the methods of estimation. I think hospital level characteristics are accounted for but what about the surgeon case mix. The surgeons in this data set are conducting many surgeries but how is this accounted for in the modelling? The correlation here may be important. Based on the final model, can we graphically see the profiles of the female vs male surgeons. We can also see from the univariate analyses that males are considerably more experienced and females are operating on higher risk, older and patients with more morbidity factors – how is this adjusted for in the model. All these things will affect outcome. This links to model selection. This needs describing better in the methods.

Response: Thank you for pointing this out. We admit that the methods for statistical analysis lacked sufficient details regarding our data. In terms of surgeon-level characteristics, we added surgeon case volume as an additional covariate in the regression model (page 5, lines 204–211; supplementary Fig 1–3). We also added a graphical presentation of the unadjusted and adjusted ORs to aid reader understanding of how the adjustment worked in the final model (Fig 2–5).

References

1. Hoshino N, Endo H, Hida K, et al. Emergency surgery for gastrointestinal cancer: A nationwide study in Japan based on the National Clinical Database. *Ann Gastroenterol Surg* 2020;4:549–61.
2. Gotoh M, Miyata H, Hashimoto H, et al. National Clinical Database feedback implementation for quality improvement of cancer treatment in Japan: from good to great through transparency. *Surg Today* 2016;46:38–47.

Comments by Reviewer 4

1. It is pretty difficult to read / understand the overall data in the first paragraph of the results. ie how many were eligible but then xx (%) opted out and of these, how many were performed by male and how many by female surgeons (and %). I would suggest a simple Table.

Response: We agree that the first paragraph of the results was not reader-friendly. After reviewing our manuscript, we considered that the flow diagram was not helpful for understanding the selection of eligible patients. We made three separate flow diagrams, each representing one surgical procedure (Fig

1). As per your suggestion we also presented all data regarding the characteristics of the surgeons in Table 1. We hope that our changes will aid reader comprehension.

2. Because the dataset is large even small differences will be statistically significant. Eg the operating time differences between 259 and 261 minutes - a time difference of 2 minutes is really not clinically significant. You need to highlight this for all the operations. And everything else that does not make much clinical difference.

Response: We completely agree with you that results from large datasets should be discussed in light of clinical significance. We removed the findings that were statistically significant but clinically insignificant from the results section.
The following sentence was omitted:
"The operation time of surgeries performed by female surgeons was longer for DG (261 vs. 259 min) and LAR (269 vs. 265 min) than that of surgeries performed by male surgeons but not significantly different for TG (282 vs. 279 min)."

3. You need to run some multi-variate analyses. I think you need to highlight in the discussion that some of the outcome differences could be accounted by the fact that there are very small numbers of female surgeons, even one adverse event is likely to cause a large effect whereas the same does not apply to male surgeons. Additionally, the female surgeons are operating on sicker patients who are older, more likely to be anaemic, malnourished, and have comorbid conditions. Therefore, the seemingly higher morbidity rates, could be easily accounted for because of these reasons, rather than a genuinely clinically important difference for female surgeons < 5 years of experience. I think this needs to be flashed out in discussion more.
The message in Table 6 can be damaging when you are comparing with female surgeons who contribute to a very small workload. This really needs to be highlighted as a limitation.

Response: Thank you for your suggestion. We admit that the methods section was unclear. We conducted a multilevel, multivariable regression analysis to account for the differences in patient, surgeon, and hospital level characteristics. We ensured that the statistical analysis section was more detailed regarding this part, for clarity (page 4, lines 155–185). In addition, the discussion section was revised (page 7, lines 352–353).
As you pointed out, a major limitation of the study is that a small number of surgeries performed by a small number of female surgeons were investigated. This point was already discussed as a limitation in the discussion section; however, we also modified the text to state that that careful consideration is required when interpreting our findings as follows (page 9, lines 433–435):
“When interpreting the results, it is important to note that because there are so few female surgeons, a single adverse event can significantly impact the entire result; this is not the case for male surgeons.”

4. I think that you need to clarify if these data are for public vs private hospitals and what proportion of females are in more regional vs major teaching hospitals. Eg private hospitals may have a greater proportion of male surgeons and they may have also have patients with earlier stage disease. Or major teaching hospitals may be better equipped for minimally invasive techniques compared to more regional centres where there may be more female surgeons. I think these are important aspects of the discussion which has not been mentioned at all in the discussion.

Response: Thank you for your valuable suggestions. Unfortunately, the NCD hospital identifiers do not differentiate between public, private, or medical school hospitals. However, unmeasured hospital level characteristics were adjusted using the multilevel model. The region of the hospital could be identified using the NCD, and it was added as a covariate in the sensitivity analysis. We agree that more information regarding hospital characteristics may shed light on where and how female surgeons work. The added sensitivity analysis revealed that female surgeons are more likely to work in urban areas and experience less surgical cases than male surgeons (page 7, lines 328–329). Further, as described in the original manuscript, the distribution of female surgeons was lower in hospitals with very high case numbers (page 6, lines 260–264; Tables 2–4). The characteristics of patients in public, private, and medical school hospitals, such as early tumour stage and undergoing minimally invasive surgery (laparoscopic surgery), were included as risk-adjustment variables in the regression analysis. Therefore, we believe that our regression analyses were well-controlled in terms of confounding factors.

5. I think the authors need to also ask themselves why they feel the need to compare outcomes between male and female surgeons - both will have professional qualifications, both will be equally trained - and as we all know, there are training biases such that male trainees and surgeons are more likely to have more training, go away for fellowship (again - not discussed in manuscript), so, we need to ask ourselves, what is the reason for these fundamental differences. Is it not the doing of the society as a whole because of the glass ceiling that these female trainees / surgeons face? I agree that it is important to demonstrate that the female surgeons are not inferior, they are if anything, possibly better in view of the sicker patients they manage and this also needs stronger discussion in your manuscript.

Response: Thank you for this comment. As you have pointed out, changes are required in the surgical field and in Japanese society to address the gender gap. In traditional Japanese culture, women have often been considered unsuitable for surgery and unwelcome. We believed that if we could show no differences in the results of surgical procedures performed by men and women, this would facilitate the improvement of the work environment for women in surgery. This explanation was included in the manuscript as follows (page 8; lines 410–413).

“In traditional Japanese culture, women have often been considered unsuitable for performing surgery and are unwelcome in the field. We believed that showing that there were no differences in the results of surgical procedures performed by men and women would make it easier for women to be accepted as surgeons and professionals.”

6. The authors really need to make stronger statements in their discussion - eg surgeons performing LAR with > 20 years experience. Patients of female surgeons have a 11 times increased risk of dying. The 95% CI however goes from 1.05 to 118 - this is a massive CI and suggests that the estimate is so imprecise and yet this is not mentioned in discussion. The small numbers of female surgeons needs to be discussed as a limitation.

Response: We deeply appreciate this comment, thanks to which, we were able to identify a miscalculation in the subgroup analysis. The estimates for the interaction terms were miscalculated. We sincerely apologise for this mistake, which should have been noticed due to the extremely wide CI. The true point estimate was 2.01 with a 95% CI of 0.57 to 7.10 for female surgeons with >20 years of experience in LAR. Further, in several other subgroups, the 95% CI now included a value of 1.0; therefore, the significant differences between male and female surgeons disappeared. The revised numbers are presented in Fig 3–5. Statistically significant differences between male and female surgeons remain in several subgroups; however, these results should be interpreted with caution because of the small number of surgeries performed by female surgeons (page 9, lines 431–435).

7. Despite the comprehensive nature of NCD, the authors need to include in their discussion that they lose over 1/4 of their cases from exclusions. This a large number of cases that have been excluded. This needs to be acknowledged as a limitation. And please see point 1.

Response: Thank you for pointing this out. We agree that a fair number of patients were excluded from the study. Non-members of the Japanese Society of Gastroenterological Surgery were excluded because they were assumed to be doctors who specialise in other surgical fields, such as cardiovascular surgery. In Japan, these doctors need to complete a general surgery program, which includes performing gastroenterological surgery, to enter a subspeciality program. Therefore, they are considered to be separate from doctors who specialise in gastroenterological surgery, and the effect on the outcome was also considered to be different for surgeries performed by these doctors. In order to improve comparability of the outcome of surgeries, the background of patients and doctors needed to be as similar as possible, except for the doctor's gender. We believe that this was a necessary exclusion. Additionally, we aimed to assess the quality of surgery performed as standard or major procedures, which was considered to improve comparability, because non-standard procedures may have complicated confounders such as the treatment preferences of patients and doctors. Hence, emergency patients and those with metastases were excluded. These exclusions might be considered limitations which negative affect generalisability; however, the exclusions were necessary to ensure comparability. The methods were updated to contain the aforementioned information (page 3–4, lines 119–129). We have also included the loss of generalisability as a limitation in the discussion (page 9, lines 436–439).

- And I would dare say that this needs to be presented at a Japanese meeting and perhaps, published in some way in a Japanese journal.
If needed, I will be happy to be involved in an editorial for this paper if it gets published.

Response: We sincerely appreciate your kind comments. We would like to present our results at symposiums of Japanese surgical societies. Further, we would consider publication in a Japanese journal if the BMJ permits to do so. If this paper is published in the BMJ, we strongly hope that you will editorialise it and stimulate discussion.