PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Volume-outcome relationship in transcatheter aortic valve implantations in Germany 2008-2014: A secondary data analysis of electronic health records
AUTHORS	Kaier, Klaus; Oettinger, Vera; Reinecke, H; Schmoor, Claudia; Frankenstein, L.; Vach, Werner; Hehn, Philip; von zur Mühlen, Constantin; Bode, Christoph; Zehender, Manfred; Reinöhl, Jochen

VERSION 1 – REVIEW

REVIEWER	David McAllister University of Glasgow, UK
REVIEW RETURNED	04-Dec-2017

GENERAL COMMENTS	This is the second version of the manuscript I have been asked to review. There are no new issues. I have therefore commented (IN CAPS) for each point of my review as to whether I agree with the authors' response. 1. In most settings anonymisation is not usually considered sufficient for a study not to require ethical review, or at least some consideration of the research by a body who has a remit to consider routine data research. It looks as though sufficient data is available in terms of the covariates and outcomes for individuals to be identified, and perhaps for disclosure to occur. Was some information governance, ethics or privacy review conducted? How does this compare to policies and practice in Germany for routine data research? We have added some details on the specifics of the patient privacy assurance process to the methods section. All measures were taken by DESTATIS prior to release of the data for scientific use and are in accordance with German data protection law. See first paragraph of the Methods section. FINE. 2. The discussion of some principles around organisation of health services is interesting. I had not previously heard of “economies of scope” We have edited the discussion of the “economies of scope” concept to be more concise. FINE. 3. + 4. The introduction includes a lot of material on the wider-issue of procedure volume and outcomes. I wonder if the introduction should focus more on the rationale and current research evidence for a relationship between procedure volume and outcomes for patients undergoing TAVI. The actual aims and objectives are not clear from the final paragraph of the introduction. We have edited the introduction for improved clarity. FINE

	5. If the methods section had sub-heading it might be easier to follow, eg Patients, Exposure, Outcomes and covariates would be useful. We have added subheadings to the Methods section. I STILL FEEL THAT THE METHODS SECTION IS THIN. THE CHARACTERISTCS OF THE SAMPLE IS FUNDAMENTAL TO UNDERSTANDING THE RESEARCH. 6. Additional key information would help the readers judge the quality of the research and reproduce the findings, such as a complete list of procedure codes, inclusion and exclusion criteria, sources of data for variables, missingess, any validation carried out for any of the variables etc. We added further details regarding the inclusion and exclusion criteria and a reference to a previous publication with an online appendix containing a complete list of procedure codes to address the concerns of missingness of variables and the overall validity of the data source. AS ABOVE. 7. Attempting to cover the items suggested in the STROBE guidelines for routine healthcare data (REporting of studies Conducted using Observational Routinely collected Data (RECORD)) would also be useful – although I appreciate that some of this information will not be available to the researchers. We have expanded the abstract and methods sections in accordance with the RECORD statement. FINE. 8. Some statements on the completeness of coverage of this dataset, relative to the German population who have undergone TAVI, would also be useful. Generally, all TAVI procedures performed (and reimbursed) in Germany are available in the dataset. This information was also added to the methods section). As added to the methods section (see above, question 6), patients with a baseline diagnosis of pure aortic regurgitation (main or secondary diagnosis other than I35.0, I35.2, I06.0, I06.2) and those with concomitant cardiac surgery or percutaneous coronary intervention were not included in this analysis. FINE. 9. I am not clear on what statistical analyses were performed. Several models at different levels appear to have been fitted. Including a more detailed description of the models as well as the analysis code in the supplementary appendix would be helpful. See response to the next question. We added a supplemental table detailing the process used to generate the results shown in Figure 1.10. I wonder why, when the researchers had access to individual-level patient data, a two-step approach was taken where (I believe) a model was fit at individual-patient level, then a subsequent model was performed at centre-level. Why not use a generalized-linear mixed model including patient-level covariates and centre-level covariates? Given that the between centre variation is the focus of this research, a hierarchical model seems more suitable than a GEE-type approach. We agree that it is also possible to analyse the data by one big model. However, we felt that it would be harder to follow the results, if there were no in-between results, which help the reader understand what we actually analyse. Hence we felt that a two-step approach would be clearer. We also agree with the author that hierarchical models would be the method of choice for understanding center heterogeneity. However, we are mainly interested in understanding how the relation of volume to outcome

	changes over time. Of course we cannot ask such a question without assuming heterogeneity, but it is not the heterogeneity in itself which is the object of interest. The question of the relation of volume to outcome is indeed a question from a marginal view, as we ask about the average outcome in dependence on center characteristics. In other sub-projects, we did analyze the data with hierarchical models, for example to obtain an understanding of variation across centres in their learning curves. However, we have to address subtle issues in such approaches, as the center specific covariates change over time and variances have to be modelled as time varying. So this is a further reason why we believe that the two-step approach is preferable here. We added a supplemental table detailing the process used to generate the results shown in Figure 1. WHILE IT WOULD NOT HAVE BEEN MY CHOICE (I WOULD HAVE PREFERED A HIERARCHICAL MODEL) THIS SEEMS REASONABLE. THE SUPPLEMNTARY TABLE IS HELPFUL. 11. Related to 4. Figures 1 and 2 are difficult to follow. It would be interesting to see spaghetti plots for the 113 centres (with some degree of transparency) or a representative sample of centres, with time as the x-axis and the adjusted mean on the y-axis. Regression-based estimates (with confidence intervals) for low, medium and high-volume centres across time, at the mean of the covariates, could be overlaid onto such a plot. As stated before (see response to major comment 1), we added some more details on practices in Germany regarding ethics or privacy review which were conducted by DESTATIS (see first paragraph of the Methods section). When conducting the study, we also considered adding some spaghetti plots for the outcomes of the 113 centers. The use of the anonymous, persistent "institute indicator of hospitals", however, is highly restricted by DESTATIS, making such spaghetti plots impossible. In practice, DESTATIS prohibits any analyses that allow drawing inferences about individual hospitals. Even results on the level of our number-of-procedures-per-year-and-center-categorization (i.e. n<50, 50=n<100, n=100) were provided only if more than 3 centers were included in a summary category. See Table 2 as an example of the highly restrictive censoring policy: DESTATIS censored the exact number of hospitals in the >=100 category in 2008 and 2009 in order to prevent traceability of individual hospitals. We added a supplemental table detailing the process used to generate the results shown in Figure 1. OK BUT THE STATEMENT THAT “DESTATIS prohibits any analyses that allow drawing inferences about individual hospitals” SHOULD THEN BE INCLUDED IN THE METHODS SECTION AS READERS SHOULD BE AWARE OF THIS LIMITATION OF THE DATASET. 12. Other potential confounding factors might account for the apparent effect of volume, including the experience of surgeons at each centre, or the selection of more complex/difficult cases by tertiary referral centres, or unmeasured patient-level confounding factors. If available some of these variables could be included in the model. Otherwise particular confounders, rather than the catch-all “we cannot guarantee that all parameters of relevance are included in the model”, could be mentioned in the discussion. We added two major limitations (see ‘STRENGTHS AND LIMITATIONS OF THIS STUDY’): “A major limitation is that the data source does not include information on the type of device used in individual TAVI procedures. In addition, information regarding the experience of surgeons at each centre would be highly relevant for the analysis but

	is unavailable." I THINK THAT THE LIMITATIONS STATEMENTS SHOULD BE STRENGHTENED TO STATE THE POTENTIAL CONSEQUENCES OF THE LIMITATIONS, EG CONFOUNDING. 13. Unless the covariates included in the model were distributed very similarly between centres with different procedure-volumes, Table 3 is potentially misleading. It would be more useful to present estimates for the adjusted analyses. As estimates for the adjusted analyses are shown in the Figures and discussed in the manuscript, we think it is still of relevance to also present the unadjusted outcomes. IF ONLY ONE SET OF ESTIMATES ARE TO BE PRESENTED FULLY IN A TABLE IT SHOULD BE THE ADJUSTED ESTIMATES. 14. It would be useful if the abstract and results section both reported the number of centres and the number of patients experiencing each outcome for the whole cohort. The number of centers performing the procedure each year is shown in Table 2. The number of patients experiencing each outcome may be derived from the procedure numbers (Table 2) and the unadjusted outcomes (Table 3). I WOULD STILL ARGUE THAT THIS INFORMATION SHOULD BE INCLUDED IN THE ABSTRACT. 15. Reporting hypothesis tests alone without effect measure estimates in the results section is uninformative, particularly for such a large dataset. We added details on the annual change of the volume effect. FINE.

REVIEWER	Eberhard Grube Bonn, Germany
REVIEW RETURNED	20-Jan-2018

GENERAL COMMENTS	The authors tried to address all issues raised by the reviewers. However, the major drawback is that data are limited to the time period until 2014 and, therefore, it is hard to draw any conclusions for daily clinical practice in 2018 (use of next generation trans catheter heart valves, decision of the "Gemeinsame Bundesausschuss" etc.).

REVIEWER	Ulrike Nimptsch Technische Universität Berlin, Germany
REVIEW RETURNED	15-Feb-2018

GENERAL COMMENTS	The authors study the volume-outcome relationship in TAVI procedures over time. This is an important issue. It is quite possible that the volume-outcome relation might have changed, as in Germany the use of TAVI strongly increased during the recent years. Thus, this study may add important findings to existing research. The authors defined criteria for inclusion and exclusion of cases. In general, it would be helpful, if the rationale for the selection of cases was stated in the methods section (e.g., why were patients with PCI excluded, why were patients with pure aortic regurgitation excluded?). Cases with concomitant cardiac surgery were excluded from the study. However, TAVI is usually a planned procedure which is predominantly performed in patients who are not eligible for cardiac surgery. Therefore, a cardiac surgery procedure during the same

	hospital stay might likely represent a complication following a TAVI procedure. This implies that patients who possibly have experienced a complication are excluded from the analysis, although these cases might contribute strongly to the volume-outcome relation. The authors should explain the rationale for this exclusion criteria in the methods section and discuss the potential bias occurring from this approach in the limitations section. Hospital volume was classified in three fixed categories (<50, 50-99, >=100). Since in the later years some hospitals perform more than 200 TAVI procedures per year, a possible effect related to very high volumes might be hidden in the analyzed group of patients treated in hospitals with >=100 cases per year. The authors should discuss this issue in the limitation section. The discussion and conclusion might be more balanced, following the study results and limitations. Existing research on the persistence of volume-outcome relationships (Reames BN et al. Ann Surg 2014;260:244-51) should be considered. Statistical review The authors studied risk-adjusted mortality as primary outcome measure. The use of cluster-robust standard errors is appropriate to account for the data structure, i.e., clustering of patients in hospitals. Risk-adjusted mortality was derived from a two-step approach where adjusted predictions at the means were subsequently derived from a first-step logistic model including patient characteristics, year of treatment, procedure volume of treating hospital, and an interaction term between year and procedure volume. This is an appropriate way to generate risk-adjusted measures, and when comparing unadjusted mortality and predicted probabilities the results seem plausible. However, this approach might appear uncommon to many readers, so it might be helpful if the authors would briefly explain in the methods section how to interpret the predicted probabilities. Meta-regression was appropriately used to quantify annual changes. I would recommend that the authors explain how the confounders considered in risk adjustment were chosen (e.g., why was transapical vs. transfemoral approach not considered?). As far as I understood, the same set of confounders was applied to adjust secondary outcomes. It would be interesting to have information on the models performances for the different outcomes. Minor comments References 1 and 20 are identical. In 2014 an additional sub-code for TAVI procedures was introduced to the German procedure coding system OPS (5-35a.02). Was this code considered within the case definition? In table 1 the proportions of patients with diagnoses of aortic valve stenosis and combined aortic valve stenosis do not sum up to 100%. However, as far as I understood, only patients with such diagnoses were included. The description of reimbursement data (page 18) should be moved to the methods section. It might be helpful, if a list of codes for definition of cases, outcomes and confounders was given in the appendix of the manuscript, rather than referring to a previous publication.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: David McAllister

This is the second version of the manuscript I have been asked to review. There are no new issues. I have therefore commented (IN CAPS) for each point of my review as to whether I agree with the authors' response.

For the sake of readability we have removed the issues that Dr. McAllister considered to have been appropriately addressed in the earlier round of revisions (marked as FINE).

5. If the methods section had sub-heading it might be easier to follow, eg Patients, Exposure, Outcomes and covariates would be useful.
Second revision: I STILL FEEL THAT THE METHODS SECTION IS THIN. THE CHARACTERISTCS OF THE SAMPLE IS FUNDAMENTAL TO UNDERSTANDING THE RESEARCH.

We substantially expanded the methods section according to the comments below.

6. Additional key information would help the readers judge the quality of the research and reproduce the findings, such as a complete list of procedure codes, inclusion and exclusion criteria, sources of data for variables, missingess, any validation carried out for any of the variables etc.

We added further details regarding the inclusion and exclusion criteria and a reference to a previous publication with an online appendix containing a complete list of procedure codes to address the concerns of missingness of variables and the overall validity of the data source.

Second revision: AS ABOVE.

See above.

9. I am not clear on what statistical analyses were performed. Several models at different levels appear to have been fitted. Including a more detailed description of the models as well as the analysis code in the supplementary appendix would be helpful.

See response to the next question. We added a supplemental table detailing the process used to generate the results shown in Figure 1.10.

I wonder why, when the researchers had access to individual-level patient data, a two-step approach was taken where (I believe) a model was fit at individual-patient level, then a subsequent model was performed at centre-level. Why not use a generalized-linear mixed model including patient-level covariates and centre-level covariates? Given that the between centre variation is the focus of this research, a hierarchical model seems more suitable than a GEE-type approach.

We agree that it is also possible to analyse the data by one big model. However, we felt that it would be harder to follow the results, if there were no in-between results, which help the reader understand what we actually analyse. Hence we felt that a two-step approach would be clearer. We also agree with the author that hierarchical models would be the method of choice for understanding center heterogeneity. However, we are mainly interested in understanding how the relation of volume to outcome changes over time. Of course we cannot ask such a question without assuming heterogeneity, but it is not the heterogeneity in itself which is the object of interest. The question of the relation of volume to outcome is indeed a question from a marginal view, as we ask about the average outcome in dependence on center characteristics. In other sub-projects, we did analyze the data with hierarchical models, for example to obtain an understanding of variation across centres in their learning curves. However, we have to address subtle issues in such approaches, as the center specific covariates change over time and variances have to be modelled as time varying. So this is a

further reason why we believe that the two-step approach is preferable here. We added a supplemental table detailing the process used to generate the results shown in Figure 1.

WHILE IT WOULD NOT HAVE BEEN MY CHOICE (I WOULD HAVE PREFERED A HIERARCHICAL MODEL) THIS SEEMS REASONABLE. THE SUPPLEMNTARY TABLE IS HELPFUL.

11. Related to 4. Figures 1 and 2 are difficult to follow. It would be interesting to see spaghetti plots for the 113 centres (with some degree of transparency) or a representative sample of centres, with time as the x-axis and the adjusted mean on the y-axis. Regression-based estimates (with confidence intervals) for low, medium and high-volume centres across time, at the mean of the covariates, could be overlaid onto such a plot.

As stated before (see response to major comment 1), we added some more details on practices in Germany regarding ethics or privacy review which were conducted by DESTATIS (see first paragraph of the Methods section). When conducting the study, we also considered adding some spaghetti plots for the outcomes of the 113 centers. The use of the anonymous, persistent "institute indicator of hospitals", however, is highly restricted by DESTATIS, making such spaghetti plots impossible. In practice, DESTATIS prohibits any analyses that allow drawing inferences about individual hospitals. Even results on the level of our number-of-procedures-per-year-and-center-categorization (i.e. n<50, 50=n<100, n=100) were provided only if more than 3 centers were included in a summary category. See Table 2 as an example of the highly restrictive censoring policy: DESTATIS censored the exact number of hospitals in the >=100 category in 2008 and 2009 in order to prevent traceability of individual hospitals. We added a supplemental table detailing the process used to generate the results shown in Figure 1.

OK BUT THE STATEMENT THAT “DESTATIS prohibits any analyses that allow drawing inferences about individual hospitals” SHOULD THEN BE INCLUDED IN THE METHODS SECTION AS READERS SHOULD BE AWARE OF THIS LIMITATION OF THE DATASET.

In the second paragraph of the Data section, the following is stated: “All summary results were anonymized by DESTATIS. In practice, this means that any information allowing the drawing of conclusions regarding a single patient or a specific hospital are censored by DESTATIS to guarantee data protection.”

12. Other potential confounding factors might account for the apparent effect of volume, including the experience of surgeons at each centre, or the selection of more complex/difficult cases by tertiary referral centres, or unmeasured patient-level confounding factors. If available some of these variables could be included in the model. Otherwise particular confounders, rather than the catch-all “we cannot guarantee that all parameters of relevance are included in the model”, could be mentioned in the discussion.

We added two major limitations (see ‘STRENGTHS AND LIMITATIONS OF THIS STUDY’):
“A major limitation is that the data source does not include information on the type of device used in individual TAVI procedures. In addition, information regarding the experience of surgeons at each centre would be highly relevant for the analysis but is unavailable.”

I THINK THAT THE LIMITATIONS STATEMENTS SHOULD BE STRENGHTENED TO STATE THE POTENTIAL CONSEQUENCES OF THE LIMITATIONS, EG CONFOUNDING.

We added some further details in the “STRENGTHS AND LIMITATIONS OF THIS STUDY”. In the methods section, we added some limitations regarding the exclusion of concomitant procedures and the limitations of risk-adjustment due to the lack of relevant clinical information.

13. Unless the covariates included in the model were distributed very similarly between centres with different procedure-volumes, Table 3 is potentially misleading. It would be more useful to present estimates for the adjusted analyses.

As estimates for the adjusted analyses are shown in the Figures and discussed in the manuscript, we think it is still of relevance to also present the unadjusted outcomes.

IF ONLY ONE SET OF ESTIMATES ARE TO BE PRESENTED FULLY IN A TABLE IT SHOULD BE THE ADJUSTED ESTIMATES.

We added Table S2 to Table S7 including all adjusted estimates. Accordingly, we now provide all estimates: Unadjusted estimates in Table 2, Adjusted estimates in the Figures and also in Table S2 to Table S7.

14. It would be useful if the abstract and results section both reported the number of centres and the number of patients experiencing each outcome for the whole cohort.

The number of centers performing the procedure each year is shown in Table 2. The number of patients experiencing each outcome may be derived from the procedure numbers (Table 2) and the unadjusted outcomes (Table 3).

I WOULD STILL ARGUE THAT THIS INFORMATION SHOULD BE INCLUDED IN THE ABSTRACT.

We have added the following sentence to the Abstract: “Between 2008 and 2014, a total of 43,996 TAVI procedures were performed in 113 different centers in Germany with a total of 2,532 cases of in-hospital mortality.”

Reviewer: Eberhard Grube
The authors tried to address all issues raised by the reviewers. However, the major drawback is that data are limited to the time period until 2014 and, therefore, it is hard to draw any conclusions for daily clinical practice in 2018 (use of next generation trans catheter heart valves, decision of the "Gemeinsame Bundesausschuss" etc.).

We fully agree that the generalizability of the results is subject to limitations, and the manuscript has been expanded substantially to better reflect this. We feel that our snapshot of the evolution of the volume-outcome relationship during an important period in the development of the TAVI technique is nevertheless of interest to researchers.

Reviewer 3: Ulrike Nimptsch
The authors defined criteria for inclusion and exclusion of cases. In general, it would be helpful, if the rationale for the selection of cases was stated in the methods section (e.g., why were patients with PCI excluded, why were patients with pure aortic regurgitation excluded?).

We have added the following rationale for the exclusion of concomitant procedures in the methods section: “Although some concomitant procedures might be informative (a cardiac surgery procedure during the same hospital stay as TAVI might likely represent a complication following a TAVI procedure), these cases may not be consistently identified in our dataset as, in many cases, concomitant procedures might have taken place in another center.”

Cases with concomitant cardiac surgery were excluded from the study. However, TAVI is usually a planned procedure which is predominantly performed in patients who are not eligible for cardiac

surgery. Therefore, a cardiac surgery procedure during the same hospital stay might likely represent a complication following a TAVI procedure. This implies that patients who possibly have experienced a complication are excluded from the analysis, although these cases might contribute strongly to the volume-outcome relation. The authors should explain the rationale for this exclusion criteria in the methods section and discuss the potential bias occurring from this approach in the limitations section.

We fully agree with you that this represents a major limitation of the dataset. The described cases (first and unsuccessfully TAVI, then conversion to cardiac surgery) represent a major complication from the TAVI procedure. Unfortunately, these cases may not be adequately identified in our dataset as, in many cases, the cardiac surgery takes place in another center. Including such cases in the analysis would put centers with cardiac surgery (where these cases are included) at disadvantage in comparison to centers without cardiac surgery (where patients are 'discharged' at the time the patients were transferred to cardiac surgery). Sometimes, it also could be that TAVI and cardiac surgery are conducted in the same hospital, which however has two IK-numbers for the respective departments.
We stated the following rationale for the exclusion of concomitant procedures in the methods section: "Although some concomitant procedures might be informative (a cardiac surgery procedure during the same hospital stay as TAVI might likely represent a complication following a TAVI procedure), these cases may not be consistently identified in our dataset as, in many cases, concomitant procedures might have taken place in another center."
Hospital volume was classified in three fixed categories (<50, 50-99, >=100). Since in the later years some hospitals perform more than 200 TAVI procedures per year, a possible effect related to very high volumes might be hidden in the analyzed group of patients treated in hospitals with >=100 cases per year. The authors should discuss this issue in the limitation section.
We added this detail to the section STRENGTHS AND LIMITATIONS OF THIS STUDY.

The discussion and conclusion might be more balanced, following the study results and limitations. Existing research on the persistence of volume-outcome relationships (Reames BN et al. Ann Surg 2014;260:244-51) should be considered.

We have included the study by Reames et al.

Risk-adjusted mortality was derived from a two-step approach where adjusted predictions at the means were subsequently derived from a first-step logistic model including patient characteristics, year of treatment, procedure volume of treating hospital, and an interaction term between year and procedure volume. This is an appropriate way to generate risk-adjusted measures, and when comparing unadjusted mortality and predicted probabilities the results seem plausible. However, this approach might appear uncommon to many readers, so it might be helpful if the authors would briefly explain in the methods section how to interpret the predicted probabilities.

We have added an explanation in the methods section.

I would recommend that the authors explain how the confounders considered in risk adjustment were chosen (e.g., why was transapical vs. transfemoral approach not considered?).

We added a sentence stating that the 21 baseline patient characteristics used for risk adjustment were previously identified by Reinöhl et al. to describe risk profiles between procedural groups.

As stated in the Strengths and Limitations section, "A major limitation is that the data source does not include information on the type of device used in individual TAVI procedures. In addition, information regarding the experience of surgeons at each centre would be highly relevant for the analysis but is unavailable". The German Operation and Procedure Classification system allows distinguishing

transapical and transfemoral procedures only. Distinguishing transapical and transfemoral procedures, however, shall not be seen as a proxy for the individual surgeons experience with a specific type of device and was therefore not used for further subdividing the patient population.

As far as I understood, the same set of confounders was applied to adjust secondary outcomes. It would be interesting to have information on the models performances for the different outcomes.

We added Tables S2 to S6, which include the regression details for all secondary outcomes.

References 1 and 20 are identical.

Reference 1 refers to the article in question, Reference 20 to the article's supplementary appendix. We split this into two references in order to improve findability of the supplementary data.

In 2014 an additional sub-code for TAVI procedures was introduced to the German procedure coding system OPS (5-35a.02). Was this code considered within the case definition?

We included all TAVI procedures with the OPS coding 5-35a.0, which includes this new category.

In table 1 the proportions of patients with diagnoses of aortic valve stenosis and combined aortic valve stenosis do not sum up to 100%. However, as far as I understood, only patients with such diagnoses were included.

In Table 1, the variables "Aortic valve stenosis" and "combined aortic valve diseases" refer to the respective coding as main diagnosis. This detail was added in Table 1.

The description of reimbursement data (page 18) should be moved to the methods section.

We have implemented this change and moved the description to the methods section.

It might be helpful, if a list of codes for definition of cases, outcomes and confounders was given in the appendix of the manuscript, rather than referring to a previous publication.

We added a list of codes for definition of cases, outcomes and confounders: Table S1.

VERSION 2 – REVIEW

REVIEWER	Ulrike Nimptsch Technische Universität Berlin, Germany
REVIEW RETURNED	22-Mar-2018

GENERAL COMMENTS	I still have some concerns about the exclusion criteria. I recommend to state in the limitations section, that the exclusion of patients with pure aortic regurgitation and those with cardiac surgery or PCI (approximately 1800 cases in 2014) possibly caused bias in the measurement of hospital volume and outcome. To my opinion, the conclusion is quite strong in view of the results. The statement that a disappearing volume-outcome relation over time "might be the case for other interventional procedures, too" is not supported by the data. To my knowledge, it is also not supported by existing research. In figure 1 the confidence interval of mortality in the high-volume group for the year 2008 is very narrow. Is it correctly displayed?

	(Sorry that I missed this point in my first review.) The authors stated in their response that “distinguishing transapical and transfemoral procedures, however, shall not be seen as a proxy for the individual surgeons experience with a specific type of device and was therefore not used for further subdividing the patient population”. I just wish to note that a transapical approach might be a proxy for the patient’s health state (e.g. severe peripheral artery disease). Transapical approach has also a higher procedural risk. Therefore, it might be a relevant factor in risk adjustment. One minor comment: On page 6 (last paragraph) the OPS code 5-35a.02 should be added to the text.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

I still have some concerns about the exclusion criteria. I recommend to state in the limitations section, that the exclusion of patients with pure aortic regurgitation and those with cardiac surgery or PCI (approximately 1800 cases in 2014) possibly caused bias in the measurement of hospital volume and outcome.

We have added this sentence to the new limitations discussion in the Conclusions section of the manuscript and refer to it in the Limitations section, which had to be shortened to meet editorial requirements (see responses to editorial comments above).

To my opinion, the conclusion is quite strong in view of the results. The statement that a disappearing volume-outcome relation over time “might be the case for other interventional procedures, too” is not supported by the data. To my knowledge, it is also not supported by existing research.

The phrase ‘might be the case for other interventional procedures, too’ was deleted in the Abstract and Discussion section.

In figure 1 the confidence interval of mortality in the high-volume group for the year 2008 is very narrow. Is it correctly displayed? (Sorry that I missed this point in my first review.)

We double checked the confidence interval of mortality in the high-volume group for the year 2008. It is correctly displayed.

The authors stated in their response that “distinguishing transapical and transfemoral procedures, however, shall not be seen as a proxy for the individual surgeons experience with a specific type of device and was therefore not used for further subdividing the patient population”. I just wish to note that a transapical approach might be a proxy for the patient’s health state (e.g. severe peripheral artery disease). Transapical approach has also a higher procedural risk. Therefore, it might be a relevant factor in risk adjustment.

We fully agree with you that the existence of peripheral artery disease is part of the patients’ health state and should be used for risk adjustment. The treatment decision between transapical and transfemoral procedures, however, is highly centre-specific and not solely dependent on the diagnosis of severe peripheral artery disease. As the center-level number of TAVI procedures is of relevance for discussing mandatory minimum thresholds, we aimed at determining the effect of such thresholds and ignored the access route. We however added a sentence pointing out that type of device and access route was not used for risk adjustment (see - briefly, due to editorial requirements - in the ‘STRENGTHS AND LIMITATIONS OF THIS STUDY’ and then again, in detail, in the limitations discussion at the end of the conclusions section).

One minor comment: On page 6 (last paragraph) the OPS code 5-35a.02 should be added to the text.

OPS code 5-35a.02 was added on page 6.