PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Representativeness of the dabigatran, apixaban, and rivaroxaban clinical trial populations to real-world atrial fibrillation patients in the United Kingdom: A cross-sectional analysis using the General Practice Research Database
AUTHORS	Monz, Brigitta; Lee, Sally; Clemens, Andreas; Brueckmann, Martina; Lip, Gregory

VERSION 1 - REVIEW

REVIEWER	Robby Nieuwlaat Assistant Professor McMaster University Canada No conflicts of interest.
REVIEW RETURNED	18-Jul-2012

GENERAL COMMENTS	Review comments paper bmjopen-2012-001768 ‘Representativeness of the dabigatran, apixaban, and rivaroxaban clinical trial populations to real-world atrial fibrillation patients in the United Kingdom: A cross-sectional analysis using the General Practice Research Database’ by Lee et al. Lee et al. analyzed what proportion of UK general practice patients meet the eligibility criteria of the RE-LY, ARISTOTLE and ROCKET-AF studies, in order to assess to what extent the results of these studies are generalizable. I have several comments: 1. The notion that ROCKET-AF represents a smaller fraction of the total AF population is not surprising considering the well-known higher risk inclusion criteria. It is more interesting to use these results to provide practical clinical implications of these findings: which therapies to apply in which patients, and to what extent can one can extrapolate results to non-RCT patients. 2. The AVERROES study adds to the generalizability of apixaban, please consider adding to analysis, or at least Discussion. 3. It has been reported that the positive predictive value of AF in the GPRD is 64.4%, which means that 35.6% are incorrectly classified as AF. The authors dismiss this notion and indicate that this will not bias results, but it is definitely possible that the ‘over-diagnosed’ group was different (post-operative AF, or other transient causes?). Please discuss how they could differ and potentially change the overall results. 4. As the authors indicate, the GPRD data will not contain as many details as would be required to address all possible RCT eligibility criteria. It seems more likely that data on exclusion criteria is harder to verify, discuss how this can affect results.

	5. The authors refer to both CHADS2 and CHA2DS2-VASc, but the latter was not yet known in the design stage of these RCTs. This has to be acknowledged. 6. It is stated that combinations of risk factors, rather than individual ones, account for the lower generalizability of ROCKET-AF. However, no numbers are shown for these combinations, it would be useful to know which combinations caused the differences. 7. There are other potential issues around generalizability of these RCT results besides eligibility criteria (included countries, TTR in the control group), add to Discussion. 8. Explain why the RE-LY study is the reference group for the statistical analysis. If you were looking for any difference among the 3 groups as stated in the objective, this should be clear and adjustment for multiple testing should be done. 9. Please clarify, page 7: artificial randomisation date was defined for 31st March 2008 to allow sufficient time for the application of prospective exclusion criteria. Not sure what you mean by this, since you did not randomize.

REVIEWER	Frank de Vries Utrecht University
REVIEW RETURNED	24-Aug-2012

THE STUDY	There are no stats at all. I have not seen a strobe checklist.
RESULTS & CONCLUSIONS	The research question has not been statistically evaluated
REPORTING & ETHICS	I have not seen a strobe checklist that was submitted with this paper.

REVIEWER	Irene Petersen, PhD Senior Lecturer in Epidemiolgy and Statistics. Department of Primary Care and Population Health, UCL London, UK
REVIEW RETURNED	05-Sep-2012

THE STUDY	It is an interesting question whether the trial populations differs from the 'real' world populations and important to understand. However, I am not convinced the study design really answer the question. A simple % of people from GPRD that would have been included in a potential trial may be a too crude way to assess the generalisability of the trials. A few questions and comments below that may help improve the study. It wasn't quite clear to me how missing data was dealth with, and some description of the proportion of missing data by study variables would be helpful. The end of the study period should be stated. In my experience patients often don't have a Read code for hypertension recorded in primary care databases, but from the blood pressure measurements recorded you can deduct whether they have hypertension. Did the study team look into this.

	A chi-square test comparing the proportion of eliglible patients to the three trials doesn't really answer the question. It just tells you that there is a relative difference between the three trials. Though with sample sizes like these p-values from a significant test provides little information.
RESULTS & CONCLUSIONS	I am sorry, but to me this study design seems slightly backward. A % of popopulation doesn't really tell you enough about generalisability. To evaluate the generalisability of the trial findings I would 'replicate' the trial in a 'real' world population e.g. in GPRD and then see if you reached same results as in the trial. Or 'replicate' the trial in the population not included in the trial and see if you still reach same results as in the trial.
REPORTING & ETHICS	I think it would be helpful to reader to know whether any of the authors have any conflicts of interest associated with the products of the trials that they seek to evaluate.

VERSION 1 – AUTHOR RESPONSE

1. The notion that ROCKET-AF represents a smaller fraction of the total AF population is not surprising considering the well-known higher risk inclusion criteria. It is more interesting to use these results to provide practical clinical implications of these findings: which therapies to apply in which patients, and to what extent can one can extrapolate results to non-RCT patients.

We agree with the peer-reviewer that these are important clinical questions. However, it is beyond the scope of this manuscript to answer these questions. We applied the trial inclusion and exclusion criteria to see how reflective the trial population would be of the real-life population. We don’t think it appropriate to derive from this research a recommendation on which patient should be treated with which product.

2. The AVERROES study adds to the generalizability of apixaban, please consider adding to analysis, or at least Discussion.

We have investigated the inclusion and exclusion criteria of studies that investigated any of the three new anticoagulants versus warfarin. The AVERROES study was explicitly conducted in patients unsuitable for warfarin (Connolly et al. 2011). We have clarified this point in the background section: ”Although these three RCTs have demonstrated that the three new anticoagulants are superior or non-inferior to warfarin in terms of stroke prevention, these studies applied specific inclusion and exclusion criteria that may have excluded patients who would otherwise be treated in real-life clinical practice, currently with warfarin.”

We added in the discussion section:
“The results of this analysis demonstrate that the warfarin-controlled pivotal trials for the novel oral anticoagulants dabigatran (RE-LY), apixaban (ARISTOTLE), and rivaroxaban (ROCKET-AF) vary in their representativeness of the AF population enrolled.”

3. It has been reported that the positive predictive value of AF in the GPRD is 64.4%, which means that 35.6% are incorrectly classified as AF. The authors dismiss this notion and indicate that this will not bias results, but it is definitely possible that the ‘over-diagnosed’ group was different (post-operative AF, or other transient causes?). Please discuss how they could differ and potentially change the overall results.

It is true that those patients incorrectly classified as having AF can introduce bias. However, there is no evidence to indicate that this would bias the results in favour of the percentage of patients from

GPRD that would have been eligible for the respective studies, as this misclassification would apply similarly to all three studies. We have clarified this in the article summary: “AF diagnosis in the GPRD may not always be accurate. However, the majority of AF cases were correctly coded according to a recent systematic review, and any errors would not be expected to systematically bias the findings of this research in favour of one study.”

4. As the authors indicate, the GPRD data will not contain as many details as would be required to address all possible RCT eligibility criteria. It seems more likely that data on exclusion criteria is harder to verify, discuss how this can affect results.

We have mentioned in the limitation section that, for example, planned major surgery was an exclusion criterion which could not be operationalized in GPRD. However, as this was a criterion used in all three clinical studies, there is no evidence to assume this to bias the results into a particular direction and we therefore did not expand the limitation section beyond what has already been stated.

5. The authors refer to both CHADS2 and CHA2DS2-VASc, but the latter was not yet known in the design stage of these RCTs. This has to be acknowledged.

We have added this to the background section: “The CHA2DS2-VASc scoring became available after the three clinical studies had been initiated.”

We additionally added one new paragraph on the ESC guidelines update that became available in August 2012 and recommends (only) the CHA2DS2-VASc as risk scoring algorithm.

6. It is stated that combinations of risk factors, rather than individual ones, account for the lower generalizability of ROCKET-AF. However, no numbers are shown for these combinations, it would be useful to know which combinations caused the differences.

We added the following to clarify this point: “Of note, the ROCKET-AF trial required patients to have a history of stroke, TIA, or systemic embolism (i.e. secondary prevention cohort) or had to have two of the following: age ≥75 years, congestive heart failure or ejection fraction ≤35%, diabetes, or hypertension.”

7. There are other potential issues around generalizability of these RCT results besides eligibility criteria (included countries, TTR in the control group), add to Discussion.

We added the following text to the limitation section: “A further limitation is that other study factors that can influence generalisability have not been investigated in this research, such as the countries participating in the studies or the quality of the warfarin arm as an indicator for the quality of patient care.“

8. Explain why the RE-LY study is the reference group for the statistical analysis. If you were looking for any difference among the 3 groups as stated in the objective, this should be clear and adjustment for multiple testing should be done.

The selection of a reference group is arbitrary in any case and we selected as reference group the one study that led to the highest percentage. Multiple testing does not apply as all analyses are exploratory and no confirmatory testing conducted. We added this statement to the methods section: “All analyses are descriptive and exploratory.”

9. Please clarify, page 7: artificial randomisation date was defined for 31st March 2008 to allow sufficient time for the application of prospective exclusion criteria. Not sure what you mean by this,

since you did not randomize.

We reworded this to “start date”.

The research question has not been statistically evaluated

We refer to the results section where we report p-values for the pair-wise comparisons of the proportions. All analyses are exploratory and no confirmatory testing conducted. We added this statement to the methods section: “All analyses are descriptive and exploratory.”

I have not seen a strobe checklist that was submitted with this paper.

We believe that a STROBE checklist is not applicable to our research as we did not investigate the association between an exposure/treatment and effect variables.

It wasn't quite clear to me how missing data was dealt with, and some description of the proportion of missing data by study variables would be helpful.

Handling of missing data in our type of research is of much less significance than in studies investigating the association of exposure variables to (relative) effects, as missingness on certain inclusion/exclusion criteria (e.g. age) would be applicable to all three studies similarly.

The end of the study period should be stated.

This is not really applicable. This was a cross-sectional application of inclusion and exclusion criteria, and the start date of March 31, 2008 was only instituted to handle the prospective study inclusion/exclusion criteria, For example, it is stated in one study that: “clinically significant GI bleeding within six months of randomisation” was an exclusion criterion (see supplementary data). Therefore, a study start date had to be determined to allow such time-dependent criteria to be observable within the database. The methods sections states: “The study design was a cross-sectional database analysis.” We furthermore added an example for the prospective exclusion criteria to clarify this to the reader (same example as in the answer above).

In my experience patients often don't have a Read code for hypertension recorded in primary care databases, but from the blood pressure measurements recorded you can deduct whether they have hypertension. Did the study team look into this.

We have consistently used Read codes for all variables of interest and have not explored alternative ways to identify them, including hypertension, in this analysis.

A chi-square test comparing the proportion of eligible patients to the three trials doesn't really answer the question. It just tells you that there is a relative difference between the three trials. Though with sample sizes like these p-values from a significant test provides little information.

We agree that p-values become quickly significant with such large sample sizes and would rather encourage the reader to interpret the magnitude of the proportion per study along the differences.

I am sorry, but to me this study design seems slightly backward. A % of population doesn't really tell you enough about generalisability.

To evaluate the generalisability of the trial findings I would 'replicate' the trial in a 'real' world population e.g. in GPRD and then see if you reached same results as in the trial. Or 'replicate' the trial in the population not included in the trial and see if you still reach same results as in the trial.

We thank the peer-reviewer for this comment. To undertake such real-life assessments takes typically several years as the drugs in question need to become used in routine care. Therefore, Health Technology Assessment (HTA) bodies (such as NICE in the UK) often request that evidence is presented to what extent a trial population is reflective of the population for which the coverage decision has to be taken and for which the drug is likely to be used in routine practice. Is a study population very different from the one for which the drug will be used in routine care, this will increase the uncertainty in such HTA decisions. Such assessment as ours therefore can serve as a first indication of generalisability. However, our research can not and did not intend to answer the question of whether the clinical trial results will be "repeated" under real-world conditions. As correctly stated, other study designs are required to accomplish this. We have amended the limitation section to make this aspect clearer and have reworded "generalisability" to "representativeness" and "applicability" in several instances to reflect the scope of our research better.
The following text was added:
"In order to answer the question of generalisibility, it would be necessary to compare clinical trial results with effectiveness and safety findings observed in routine care. However, to undertake such real-life assessments typically takes several years as the drugs in question need to become used widely. Therefore, Health Technology Assessment (HTA) bodies (such as NICE in the UK) often request that evidence is presented to what extent a trial population is reflective of the population for which the coverage decision has to be taken and for which the drug is likely to be used in routine practice. If a study population is very different from the one for which the drug will be used in routine care, this will increase the uncertainty in such HTA decisions. Such assessment as ours therefore can serve as a first indication of generalisability."

I think it would be helpful to reader to know whether any of the authors have any conflicts of interest associated with the products of the trials that they seek to evaluate.

We had already stated this on page 16 of the originally submitted manuscript. We have expanded this as follows: "SL, BM, AC, and MB are employees of Boehringer Ingelheim, the manufacturer of dabigatran."

VERSION 2 – REVIEW

REVIEWER	Irene Petersen, Senior lecturer UCL, UK I have no competing interest
REVIEW RETURNED	14-Oct-2012

THE STUDY	I don't feel the authors have taken the initial feedback on board.