Dear Dr. Godlee,

We thank you and your expert reviewers for the positive feedback, constructive review, and the offer to reconsider a revised version of our manuscript entitled:

**Manuscript ID BMJ 2015 027294: Risk Prediction Models for Contrast Induced Nephropathy: A Systematic Review**

for publication as a “Research Article” in *The BMJ*. Appended please find a detailed point-by-point response to all issues raised by the reviewers and the manuscript committee, highlighting the changes to and location of these changes in the manuscript. Changes in the manuscript have been highlighted using tracked changes, and page numbers refer to the tracked changes version. We have also included a clean copy of the revised manuscript. We hope that this manuscript meets with your approval.

Sincerely,

Samuel Silver

Ziv Harel
Itemized Responses to Reviewer and Committee Comments

**Reviewer 1:**
We are pleased that the Reviewer found our study to be novel, well-designed, and addresses an important and common clinical problem of interest to a general medical readership. We appreciate his or her comments to strengthen the presentation of our results and its applications.

1. The manuscript has a clearly defined research question. However, it is in the main text answered in too general terms. A more precise presentation of each model would be preferred in the result section as a part of the main text and not just in table and appendix table format.

**Response:** Thank you for this suggestion. We have added specific references to each prediction model throughout the Results section in order to highlight the strengths and weaknesses of individual models. We also added several paragraphs that explicitly compared the different models to one another.

*Change 1 (page 13, paragraph 2):* Five models had good discriminative ability (c-statistic > 0.80) upon validation, with the number of predictors in each risk model ranging from 3 to 15. These higher performing models included the following risk factors: pre-existing kidney disease (all models), age (4/5 models), diabetes (4/5 models), heart failure/impaired ejection fraction (4/5 models), and hypotension/shock (3/5 models). Only one of these models included contrast volume (Victor et al.). Of these five models, only Maioli et al. was validated externally (c-statistic=0.82), although the external cohort was from the same hospital as the derivation cohort.

*Change 2 (page 13, paragraph 3):* Three models had moderate discriminative ability (c-statistic 0.70-0.80). These models included the following risk factors: pre-existing kidney disease (all models), age (2/3 models), diabetes (1/3 models), heart failure/impaired ejection fraction (2/3 models), and hypotension/shock (1/3 models). All of these models included contrast volume as a risk factor.

*Change 3 (page 14, paragraph 3):* Three additional studies included electronic calculators to facilitate knowledge translation: Gurm et al. (online calculator: https://bmc2.org/calculators/cin), Mehran et al. (smartphone application: http://www.qxmd.com/calculate-online/nephrology/contrast-nephropathy-post-pci), and Victor et al. (Excel® spreadsheet: available with the article’s supplementary data).

2. Moreover, a thorough discussion of each model is missing. Is one model better validated than other? What could be the reason for differences in c-statistics for the models validated in more than one study?

**Response:** We have added a discussion of each model to the Discussion section, outlining which models were better validated. We also explained why Tziakas et al., a high performing model that was validated in more than one study, had different c-statistics reported.

*Change (page 16, paragraph 2):* Eight of the models had moderate to good discrimination based upon the c-statistic. However, this is only the first step in the development of a clinical prediction rule. The next steps involve testing the rule in a separate population (external validation) and measuring its effect on clinical outcomes (impact analysis). From these eight models, two were
externally validated in the same hospital as the derivation cohort (Fu et al. and Maioli et al.), which limits their generalizability. Tziakas et al. was externally validated in two different settings, once in the same hospital as the derivation cohort and once as part of a multi-centre study. The c-statistic is a function of the sensitivity and specificity of a given risk score, and both sensitivity and specificity are influenced by case mix, disease severity, or risk factors for disease. Therefore, it was not unexpected that there were variations in the c-statistic for the same model among different studies because of differences in population characteristics in each of the externally validated studies.

Unfortunately, some model characteristics make their clinical application challenging. Both Fu et al. and Tziakas et al. included contrast volume in their models, however the volume of contrast needed is usually not known until the procedure has been performed. Since decisions on diagnostic testing and prophylactic therapies are usually made before the procedure, inclusion of contrast volume as a model variable limits model usefulness. This criticism applies to other promising models, such as Gao et al. and Victor et al. Maioli et al. did not include contrast volume as a variable, and so this model may be a reasonable starting point for further refinement.

3. What are the precise differences between the models? The authors use general terms such as “heterogeneity in patient populations and clinical scenarios” (page 15 line 13).

Response: We have made the precise differences between the models more explicit in both the Results and Discussion (refer to Major Points 1 and 2). We also specifically described the reasons for heterogeneity between studies.

Change (page 18, paragraph 1): Heterogeneity also precluded the performance of meta-analyses. Contributors to study heterogeneity included different clinical settings, cointerventions, the type of contrast administered, the timing of creatinine measurement to ascertain CIN, and the method used to define baseline creatinine.

4. Could one model potentially be recommended for clinical use since no good alternatives exist at the moment? What should the clinician consider when making a decision in practice?

Response: We attempted to highlight the higher performing models in the Discussion, while at the same time not overstating their predictive ability (refer to Major Points 1 and 2). We also extended our section on the clinical applications of our study to assist the physician in daily practice.

Change (page 23, paragraph 2): While we await these results, clinicians should consider using one of the higher performing scores that do not include contrast volume to estimate a patient’s risk of CIN (Maioli et al., Chen et al., Gurm et al., Liu et al.). Potential applications are numerous. First, predictive models can inform patient-centered decision-making, whereby high risk patients may choose alternative imaging methods or opt out of further investigation. Second, they allow for the selective use of the few pre-procedural maneuvers that are likely to be helpful (e.g., holding diuretics to prevent intravascular volume depletion, fluid hydration, etc.) to potentially mitigate the risk of CIN. Third, they allow clinical trials and quality improvement interventions to target patients most likely to benefit from these complex efforts.
Minor Comments
1. Page 11, line 10: Variables such as age, DM, and contrast volume is included in the majority of the models. Please include this.

Response: Thank you for this suggestion. We have now explicitly listed in the text any variables included in the majority of the risk models (7 of 12 models).

Change (page 12, paragraph 3): Other common measures included in the majority of risk models were age (8/12 models), diabetes (8/12 models), heart failure/impaired ejection fraction (8/12 models), and contrast volume (7/12 models).

2. Table 2: Add number of patients included in each study

Response: The number of patients included in each study was listed in Table 1, but we have now also included these data at the top of Table 2.

3. Table 3: In the part “derivative cohort” a column with “number of included patients” and “population” would provide a better overview for the reader.

Response: The “number of included patients” is already presented in Table 3 under “sample size.” The “population” of each derivation cohort is described in Table 1, but we have now also included these data in Table 3 to provide a better overview for the reader.

4. Table 4: The authors should consider not including “clinical usefulness” in a judgment of risk of bias. In general the table in hard to understand and does not with the current layout add anything useful to the manuscript.

Response: Thank you for this comment. We have tried to clarify Table 4 by adding an explanatory Table to the Supplementary Material, adapted from Hayden et al. and Tangri et al. We feel it is important to keep Table 4 in the manuscript, since the inclusion of methodologically weak studies can threaten the internal validity of a systematic review. We have also clarified the term “clinical usefulness” in the text and Supplementary Appendix to clarify is meaning and relevance to risk prediction.

Change 1: Supplementary Appendix 2

Criteria for measuring risk of bias and clinical usefulness

<table>
<thead>
<tr>
<th>Bias</th>
<th>Areas to Consider</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study participation</td>
<td>• Sample population matches the population of interest</td>
</tr>
<tr>
<td></td>
<td>• Inclusion and exclusion criteria</td>
</tr>
<tr>
<td></td>
<td>• Sampling frame and recruitment strategy, including methods to identify the study sample and the period/place of recruitment</td>
</tr>
<tr>
<td></td>
<td>• Adequate participation in the study by eligible patients</td>
</tr>
<tr>
<td></td>
<td>• Characteristics of baseline study sample</td>
</tr>
<tr>
<td>Study attrition</td>
<td>• Proportion of the study sample who complete the study</td>
</tr>
<tr>
<td></td>
<td>• Attempts to collect data on patients who withdraw from the study</td>
</tr>
<tr>
<td></td>
<td>• Reasons for loss to follow-up</td>
</tr>
<tr>
<td>Characteristics of patients loss to follow-up</td>
<td>Differences in characteristics and outcomes between patients who complete the study and patients lost to follow-up</td>
</tr>
<tr>
<td>Prognostic factor selection</td>
<td>Clear definitions of candidate predictors</td>
</tr>
<tr>
<td>Prognostic factor measurement</td>
<td>Clearly defined method of measurement for each predictor</td>
</tr>
<tr>
<td>Outcome measurement</td>
<td>Clearly defined outcomes</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>Sufficient data presentation to assess the results</td>
</tr>
<tr>
<td>Reporting of model performance</td>
<td>Discrimination</td>
</tr>
</tbody>
</table>

**Clinical Usefulness**

| Utility | Risk categories provided that would trigger a diagnostic or therapeutic decision | Clinical trial data to demonstrate an effect on decision-making or clinical outcomes |
| Usability | Simple risk calculator or nomogram to facilitate use at the bedside | Electronic or web-based calculator to facilitate use at the bedside |

**Change 2 (page 9, paragraph 3):** Similar to a previous systematic review, we also assessed the clinical usefulness of each study, which was defined as the combination of clinical utility and usability. For clinical utility (the effect on a clinical decision linked to a risk category or threshold), we assessed whether authors linked their models to specific risk categories and discussed how the risk categories would aid diagnostic evaluations. For usability (the availability of a clinical decision aid), we noted whether authors included a calculator or risk score that would facilitate knowledge translation and use at the bedside.

5. Please discuss whether it is at all possible to develop a predictive model of CIN which can be used in different clinical settings and populations. How should we design future studies when aiming to develop a potentially robust model for predicting CIN?

**Response:** We have summarized the next steps in model derivation at the end of the Discussion in order to develop a robust CIN model that can be applied in different clinical settings.

**Change (page 23, paragraph 1):** Many of the existing models have been adequately derived, but newly derived models should ensure standardized definitions are used to select and measure prognostic factors in order to reduce heterogeneity and misclassification bias. Subsequent steps involve external validation in multi-center cohorts and integrating risk assessment with diagnostic or therapeutic decisions. This second step will require experts to suggest clinically
relevant risk thresholds for CIN in the AKI clinical practice guidelines, which are currently lacking. Lastly, there is a need for model derivation that includes patients who undergo intravenous CT procedures, CT angiography and non-coronary angiography as these procedures constitute the majority of contrast procedures and the risk associated with such is likely different than with arterial contrast for coronary angiography.

6. Reference 2 (McCullough PA et al. 1997) and 3 (Nash K et al. 2002): Please consider substituting with more up-to-date references.

Response: Thank you for pointing this out. These references have been updated, with the text slightly modified to reflect the new information.

Change 1 (page 4, paragraph 1): McCullough PA et al. 1997 has been replaced by Rsu RK et al. 2013. No change to text required.

Change 2 (page 4, paragraph 1): Nash K et al 2002 has been replaced by Waikar SS et al. 2008. This part of the manuscript now reads: “CIN may be as high as the third most common cause of AKI in hospitalized patients, after ischemic and drug-induced injury.”

Reviewer 2:
We thank the Reviewer for his or her constructive comments, and are pleased that he or she found our systematic review to contribute valuable findings towards an increasing problem given the frequency with which diagnostic procedures that require radiocontrast are performed. We have clarified his or her questions regarding the Results section below.

1. The selection process in this paper is clearly explained and depicted in Figure 1 and in Supplemental Appendix 1. However, on page 10 in the first Results section 3509 citations were excluded by screening of title and abstract. Please state the main reasons for exclusions.

Response: We thank the reviewer for his or her comment. We have added the main reasons for exclusion of the 3509 papers in our text.

Change (page 11, paragraph 1): Our search strategy yielded 3567 unique citations. We excluded 3509 citations based on screening of title and abstract mainly due to non-relevant outcomes, animal studies and review articles, leaving 58 articles for full-text review. We subsequently excluded 42 studies that did not fulfill our inclusion criteria as they consisted of unpublished abstracts (n=17), models which did not have an associated CIN risk score (n=13), letters to the editor and narrative reviews (n=6), duplicate publications (n=3), studies with non-CIN related outcomes (n=2), and non-validated risk scores (n=1). This yielded, yielding 16 studies comprising 12 unique risk prediction models (Figure 1).

2. In a next step 42 studies were excluded, I guess this time after full text evaluation, please include information in a sentence or two why these studies did not fulfill the inclusion criteria.

Response: We thank the reviewer for his or her comment and have included this information in our Results section. Please refer to the prior comment for the full explanation.
3. The authors state that they assessed study quality (bias, Table 4) using a modification of criteria recommended by Hayden et al, Ann Intern Med 2006, and also that they assessed clinical usefulness of the prediction models as suggested by Tangri et al., Ann Intern Med 2013. Despite these references it is not clear to the reader how the authors conclude the size of bias in Table 4 (low, high, ?), it seems somehow arbitrary, which is in contrast to a "systematic" review. Please clarify and extend this paragraph and also Table 4, so that the reader can follow your rationale.

Response: Thank you for this suggestion. We have clarified the criteria used to assess bias and clinical usefulness in the text and Supplementary Appendix (see Reviewer 1, Minor Comment #4). We agree with the Reviewer that the criteria are limited, and have added these to our Discussion. We have also tried to more clearly specify the criteria in the Table 4 legend. However, we believe it is important to keep Table 4 in the manuscript, since this is one of the only published frameworks for reporting the quality of studies included in a systematic review. Without this reporting, the inclusion of methodologically weak studies can threaten the internal validity of a systematic review.

Change 1 (page 21, paragraph 1): Fourth, the criteria for risk of bias and clinical usefulness were adapted from the existing literature on clinical risk prediction and have not been extensively studied or prospectively validated.

Change 2 (Table 4): Bias was evaluated as low risk, high risk, or unknown risk (?). Clinical usefulness was evaluated as Yes or No. Two authors (ZH and SS) evaluated the studies on these criteria independently. Discrepancies were resolved by consensus. Refer to Supplementary Appendix for more details.

4. Further, in Table 4 nine of twelve risk predictors are termed "clinically useful" which is in sharp contrast to the Discussion section, in which the authors explain all the (correct) limitations of the studies. This is puzzling and contradictory, please clarify and include in the Discussion section.

Response: Thank you for pointing this out. This discrepancy arose because we did not separate clinical usefulness into utility and usability, as originally defined by Tangri et al. Since many of the models provided risk scores, they technically satisfied criteria for clinical utility even though the risk scores were not linked to diagnostic or therapeutic decisions. We have resolved this by evaluating the models for clinical utility and clinical usability separately (refer to Reviewer 1, Minor Point 4). Table 4 now demonstrates that few models meet both criteria for clinical usefulness. We have also elaborated upon these explanations in the Discussion.

Change 1 (Table 4): Table 4 separated based upon “clinical utility” and “clinical usability.”

Change 2 (page 14, paragraph 3): Most models (9/12 models) stratified their cohorts into risk categories (low, moderate, high risk). However, none of the models explained how assignment to a risk category would affect diagnostic or therapeutic decisions. Simple risk calculators or nomograms were provided by four studies. Three additional studies included electronic calculators to facilitate knowledge translation: Gurum et al. (online calculator: https://bmc2.org/calculators/cin), Mehran et al. (smartphone application: http://www.qxmd.com/calculate-
online/nephrology/contrast-nephropathy-post-pci), and Victor et al. (Excel® spreadsheet: available with the article’s supplementary data).

Change 3 (page 19, paragraph 1): Even though prediction models for CIN perform similarly to some other popular prediction models (Framingham Risk Score, QRISK2) based upon the c-statistic, there are a number of important differences that have limited their use in clinical practice. First, these aforementioned models have been externally validated in multicenter studies. Second, many predictive models, including the Framingham Risk Score and QRISK2, clearly outline how assignment to a risk category affects diagnostic or therapeutic decisions. Finally, electronic risk calculators exist for both the Framingham Risk Score and QRISK2. None of the predictive models for CIN satisfy all three of these elements, and only Mehran et al. satisfy two of three elements, but with a median c-statistic of 0.57 on external validation.

5. Finally, the authors correctly state that this systematic review illustrates the heterogeneity of the studies per se, and particularly of discovery and validation cohorts. This is a very valuable finding in particular because all of these studies evaluated risk prediction in the same setting: coronary angiography and PCI. No intravenous CT scans! When risk prediction models perform so poorly in studies evaluating the same intervention, how can we be able to define risk prediction models for CIN in general (PCI, peripheral angiography, venous for CT, venous PTA, etc...)?

Response: This suggestion is similar to Reviewer 1 (Minor Point 5). We recommend building upon existing high performing models, as well as developing new models using standardized definitions endorsed by clinical practice guidelines. Subsequent steps beyond model derivation are also outlined in the new Discussion paragraph (refer to Reviewer 1, Minor Point 5).

Manuscript Committee:
We are pleased that the committee agreed that the prediction of contrast induced nephropathy is a timely research topic that affects many medical specialties, and our systematic review was well-designed to summarize the available evidence. We appreciate the suggestions made to enhance the presentation of Results and the Discussion, so as to highlight the important clinical applications and research implications of our work.

General Committee Comments
1. We agreed with reviewers that it would be helpful to put the findings in more clinical context. For example, might you say which of the models is the most likely candidate for further development? Might you comment on how the C statistic and performance of some of these models compare to other risk prediction models (in other therapeutic areas) that are in clinical use despite not being perfect?

Response: Thank you for these suggestions to put the manuscript’s results into context. We have taken the Committee’s advice and highlighted the higher performing prediction models (also refer to Reviewer 1, Major Points 1 and 2), as well as compared CIN models to prediction models from other therapeutic areas.
Change 1 (page 23, paragraph 2): While we await these results, clinicians should consider using one of the higher performing scores that do not include contrast volume to determine a patient’s risk of CIN (Maioli et al., Chen et al., Gurm et al., Liu et al).

Change 2 (page 19, paragraph 1): Even though prediction models for CIN perform similarly to some other popular prediction models (Framingham Risk Score, QRISK2) based upon the c-statistic, there are a number of important differences that have limited their use in clinical practice. First, these aforementioned models have been externally validated in multicenter studies. Second, many predictive models, including the Framingham Risk Score and QRISK 2, clearly outline how assignment to a risk category affects diagnostic or therapeutic decisions. Finally, electronic risk calculators exist for both the Framingham Risk Score and QRISK2. None of the predictive models for CIN satisfy all three of these elements, and only Mehran et al. satisfy two of three elements, but with a median c-statistic of 0.57 on external

2. We also wondered whether some results and findings might be presented graphically, since this sort of visual presentation is often very interesting for readers.

Response: Thank you for this suggestion. We discussed the presentation of results amongst the authorship team, but could not think of a graphical presentation since the data were too heterogeneous to conduct meta-analyses.

3. Can you examine or speculate about the source(s) of heterogeneity? Many of these models seem to have been evaluated in the setting of coronary angiography, so is it the case that their utility should be examined more narrowly, in particular patients or populations? You might say more about how they perform in the predominant situations in which they are used.

Response: We have clarified the potential sources of heterogeneity in the Discussion (refer to Reviewer 1, Major Point 3), which mostly stem from issues with prognostic factor measurement, patient selection, and operational definitions of CIN and creatinine measurement. This makes the need to derive the models more narrowly less important. In fact, we are in need of a model applicable to both coronary angiography, as well as intravenous CT scans. We also suggested how current and future models could be derived to minimize the issue of heterogeneity going forward.

Change 1 (abstract): There was significant heterogeneity among the included studies, as a result of different clinical settings, cointerventions, and the timing of creatinine measurement to define CIN.

Change 2 (page 23, paragraph 1): Many of the existing models have been adequately derived, but newly derived models should ensure standardized definitions are used to select and measure prognostic factors in order to reduce heterogeneity and misclassification bias.

4. Since existing models do not perform well, we thought it would be useful to expand your discussion of what the next generation of research should look like. For example, do we need a trial showing whether using these models changes patient outcomes?
Response: Thank you for this suggestion. We have now included several paragraphs on higher performing models in the Results and Discussion. These changes are outlined in our responses to Reviewer 1 (Major Points 2 and 4). We also made changes to the abstract and conclusion.

Change 1 (Abstract): The majority of higher performing models included measures of pre-existing chronic kidney disease, age, diabetes, heart failure/impaired ejection fraction, and hypotension/shock.

Change 2 (Conclusion): While higher performing models usually include pre-existing CKD, age, diabetes, heart failure/impaired ejection fraction, and hypotension/shock, most of these published models have limited predictive ability when validated externally and are not relevant to individuals receiving intravenous contrast or non-coronary angiography.

6. The C statistics are on average, moderate, and in some cases (the Victor study) look quite good indeed (c/f Framingham, QRISK2, etc). What’s missing is a clinical look at the utility of these tools, not unlike our discussion at last week’s MM about pre-eclampsia tools. Are these tools used in coronary angiography? What are the implications of being scored as high risk in terms of prevention?

Response: Thank you for these excellent suggestions. We have added a paragraph comparing CIN prediction models to the popular models suggested by the Committee, which also highlights why CIN prediction models may not be commonly used in clinical practice (refer to Reviewer 2 Major Point 4 and General Committee Item 1). We have now explicitly pointed out the limitations of some models with a high c-statistic (such as the Victor study, refer to Reviewer 1 Major Point 2). Lastly, we have included two paragraphs on the utility of CIN risk prediction,
including the implications of being scored as high risk by using a clinical trial and quality improvement intervention as practical examples for the researcher and clinician respectively.

**Change 1 (page 19, paragraph 1):** Predictive models for CIN have been available for clinical use for almost ten years. However, uptake by cardiologists and radiologists has been low judging by their omission from recent clinical practice guidelines and survey studies.

**Change 2 (page 22, paragraph 1):** The Prevention of Serious Adverse Events following Angiography (PRESERVE) trial is underway, with the plan to enroll 8,680 patients to compare the effectiveness of isotonic sodium bicarbonate versus isotonic saline and oral NAC versus oral placebo (NCT01467466). Dr. Chertow is on the Steering Committee of PRESERVE. Another effective prevention strategy for CIN may be system-based quality improvement (QI) efforts. Brown et al. reduced the rate of CIN by 20% in consecutive PCI patients at multiple centers through a multi-faceted intervention that included: standardized fluid orders, loosening NPO restrictions prior to a procedure, cessation of nephrotoxic medications, self-expansion of the extracellular fluid volume, mandatory procedure delays to ensure adequate volume status and team training.

Predictive risk scores for CIN were not used in either of these strategies, representing a potential missed opportunity. For the PRESERVE trial, a robust risk score might help target ideal candidates for enrollment to enrich CIN event rates and ensure a beneficial effect is not missed. For the QI intervention, a risk score would allow the clinical team to concentrate their system-changes on the highest risk patients. This would help protect against improvement fatigue given the multiple quality initiatives that now exist. In fact, patients with an eGFR < 60mL/min/1.73m² seemed to benefit more from the intervention (28% reduction), suggesting integration of a risk score to target high risk patients may achieve an even greater reduction in the rate of CIN.

**Change 3 (page 23, paragraph 2):** While we await these results, clinicians should consider using one of the higher performing scores that do not include contrast volume to determine a patient’s risk of CIN (Maoli et al., Chen et al., Gurm et al., Liu et al.). Potential applications are numerous. First, predictive models can inform patient-centered decision-making, whereby high risk patients may choose alternative imaging methods or opt out of further investigation. Second, they allow for the selective use of the few pre-procedural maneuvers that are likely to be helpful (e.g., holding diuretics to prevent intravascular volume depletion, fluid hydration, etc.) to potentially mitigate the risk of CIN. Third, they allow clinical trials and quality improvement interventions to target patients most likely to benefit from these complex efforts.

7. The paper needs a bit of tightening up, the authors could loosen up a little and try and take some positives from their work.

**Response:** Agreed. We have added several sections on the higher performing models to the Results and Discussion, as well as a section outlining how current CIN prediction models could be used to enhance a recently published quality improvement intervention. These Changes are highlighted in previous sections of this response letter (Reviewer 1: Major Points 2 and 4; and General Committee Comments: Item 6).

**Specific Committee Comments**
Tiago Villanueva
1. It would have been nice to have an idea if these risk prediction tools are actually used widely or not in clinical practice, since they've been around for 10 years, which is too much time for these tools to have been used inappropriately.

Response: Our clinical impression is that these are not widely used. We have included the relevant references in the Discussion, along with a rationale for why CIN prediction models are not widely used. This led nicely into the next section on the absence of prediction models for intravenous contrast and non-coronary angiography.

Change (page 19, paragraph 1): Predictive models for CIN have been available for clinical use for almost ten years. However, uptake by cardiologists and radiologists has been low judging by their omission from recent clinical practice guidelines and survey studies. Even though prediction models for CIN perform similarly to some other popular prediction models (Framingham Risk Score, QRISK2) based upon the c-statistic, there are a number of important differences that have limited their use in clinical practice. First, these aforementioned models have been externally validated in multicenter studies. Second, many predictive models, including the Framingham Risk Score and QRISK 2, clearly outline how assignment to a risk category affects diagnostic or therapeutic decisions. Finally, electronic risk calculators exist for both the Framingham Risk Score and QRISK2. None of the predictive models for CIN satisfy all three of these elements, and only Mehran et al. satisfy two of three elements, but with a median c-statistic of 0.57 on external validation

Another reason for the low clinical uptake of predictive models for CIN is that they have focused exclusively on populations receiving intra-arterial contrast for coronary angiographic procedures…

Wim Weber
1. The conclusion that ‘more research is needed’, is probably true, and I initially wondered if that is enough for us? But their cover letter is quite persuasive; it is a topic that affects many disciplines as almost all will order these investigations, and this now shows that we cannot predict who will be affected.

Response: We agree with the Committee member that CIN is relevant to virtually every medical discipline. We have tried to extend our conclusion beyond “more research is needed” by including two new paragraphs on the clinical utility of CIN prediction for the researcher and clinician (Refer to General Committee Comments: Item 6).

Rebecca Burch
1. I agree with the concerns about clinical usefulness that Busch (Reviewer 1) raised. In the real world some kind of prediction has to be used, and while it's helpful to know that more work needs to be done I found myself wondering how helpful this was for current practice.

Response: Thank you for these comments. We have extended our conclusion beyond “more research is needed” by including two new paragraphs on the clinical utility of CIN prediction for the researcher and clinician (Refer to General Committee Comments: Item 6).
2. From looking through table 3, the prediction score that looks the best is Maioli, which was externally validated in a separate cohort (different institution? unclear) and performed fairly well. Could the authors do more to recommend what to do now?

Response: We have described the model by Maoli et al. in more detail in both the Results and Discussion (outlining its strengths and limitations, refer to Reviewer 1 Major Points 1 and 2), as well as described the next steps from both a clinical and research perspective (refer to Reviewer 1 Major Point 4 and General Committee Items 4 and 6).

Change (page 17, paragraph 2): Maoli et al. did not include contrast volume as a variable, and so this model may be a reasonable starting point for further refinement.

Alison Tonks
1. The paper reads well and maintained by interest. As others have said, the authors could do a better job comparing and contrasting the real value of these models in clinical situations. Which ones look better validated? Better performing? Should future researchers build on those and discard the others?

Response: We thank the Reviewer for his or her laudatory comment. As other Reviewers and Committee members have suggested, we have added several sections in the Results and Discussion to compare the models and highlight the better performing models (Reviewer 1: Major Points 1 and 2). We have also added sections to the Discussion to emphasize the value of these models to the clinician and researcher, including future research directions (Refer to General Committee Comments: Item 4 and 6).

2. They could also point researchers where to go next. Top of the list would be prospective "trials" to see if using any of the better performing models actually improved patient outcomes. This final step is often missing in prediction modelling, and the authors should emphasize the need for this kind of evaluation. Without it, developing models is a dry statistical exercise that may or may not actually help patients.

Response: We agree with the Committee Member on all points, especially regarding the need for prospective trials to determine if risk prediction has any effect on patient outcomes. This is one of the final steps in prediction modelling, and we have tried to outline in the manuscript the steps that are needed for CIN prediction models to reach this stage (refer to General Committee Item 4).

Change 1 (page 21, paragraph 1): Finally, none of the internally or externally validated models were prospectively evaluated in clinical practice to determine their effect on clinical decision-making and patient outcomes.

Change 2 (conclusion): Given the increasing incidence of CIN and the multiple clinical applications of CIN risk prediction, it is necessary to build upon current models in order to develop a clinically useful and generalizable prediction model for CIN that demonstrates an impact on clinical decision-making and patient outcomes.
3. Currently, their definition of clinical utility is very broad. A model had utility if the authors discussed how it might help decision making. Surely just a measure of the authors own faith in their model.

Response: This is an excellent point and was also raised by Reviewer 1 (Minor Point 4) and Reviewer 2 (Major Point 3). We have clarified the criteria by Tangri et al. used to assess clinical utility and usability in the Methods and Supplementary Appendix so it is not as broadly defined.

Rafael Perera
1. I would have liked to see some figures summarizing the data but in principle there is not much wrong with the article.

Response: Thank you for this suggestion. We discussed the presentation of results amongst the authorship team, but could not think of a graphical presentation since the data were too heterogeneous to conduct meta-analyses.

2. There is an issue as to how these results could be made more useful. It could be through specific recommendations on which models appear to have the most potential stratified by setting.

Response: Thank you for this important comment. We have significantly revised the manuscript to highlight the higher performing models (Reviewer 1: Major Points 1 and 2) and describe the applications of our findings for both the clinician and researcher (Refer to General Committee Comments: Item 4 and 6).

Other Revision Requests:
1. BMJ PICO template

Response: This has been completed and emailed to papersadmin@bmjgroup.com, as requested.

2. What this paper adds/what is already known box

Response: See below

What This Paper Adds
What is already known on this subject:
- Contrast induced nephropathy is associated with cardiovascular complications, dialysis and death
- Over 80 million iodinated contrast studies are performed each year, ordered by a wide-range of medical specialties
- Several models exist to predict contrast induced nephropathy, but their impact on decision-making and patient outcomes is unclear

What this study adds:
• Most of the models reviewed have limited predictive ability in external populations and are only relevant to individuals undergoing coronary angiography
• Future research should involve the external validation of models, preferably across multiple sites with diverse populations, and efforts to integrate risk assessment with diagnostic or therapeutic decisions