PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	NURSING WORKLOAD, PATIENT SAFETY INCIDENTS AND MORTALITY – AN OBSERVATIONAL STUDY FROM FINLAND
AUTHORS	Fagerström, Lisbeth; Kinnunen, Marina; Saarela, Jan

VERSION 1 – REVIEW

REVIEWER	Peter Griffiths University of Southampton
REVIEW RETURNED	24-Mar-2017

GENERAL COMMENTS	Thank you for the opportunity to review this interesting study. It provides another 'take' on exploring the association between nurse staffing and patient outcomes with the potential to refine our understanding. I think it potentially merits publication as it has a distinctive contribution. However, there are a number of issues that the authors should consider. I would certainly want to see revisions/clarifications and responses before I could offer a definitive judgement. 1) (minor) The background needs to say a little more about the RAFAELA system so that it is clearer what it is – both in terms of the nature of the data gathered and the measure it provides – for example does it give a direct measure of the required nurse staffing level in some way? More direct reference to the extensive evidence base behind the system would be appropriate as well although I appreciate the need to be concise at this point. The meaning of the term 'optimal' in this context needs to be clarified. I realise more detail is given later but I think the nature of the tool needs to be established in the background. 2) (minor) The account of the system (Raphaela) and the measures used in it (PAONCIL, OPCq) in the methods is confusing to me – who is not uninitiated – I suspect it would be much worse for those not already familiar. I may appear to be contradicting myself when I say suggest that much of this be abbreviated by addressing this as a study which 'simply' uses a measure of nursing workload that is based on a system designed to determine nursing workload based on patient classification – which then gives us a measure of nursing workload… and contrast that to studies where the implied nursing workload is uniform across all patients (and hence invariant for any given nurse to patient ratio). If you could do this and give appropriate reference to the system itself you would be offering more detail than Needleman's study! If you avoid the extensive detail you can focus on reporting your variables more clearly. 3) (major) The aim of the study is not clearly described. This

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	relates to point 2 above. It is stated that “The aim of this observational study was to investigate whether a staffing model based on optimal NWL, measured daily using the RAFAELA Nursing Intensity and Staffing system, can positively affect patient safety incidents (incidents) and mortality. Also, we want to compare staffing models based on the RAFAELA system with those based on patient-to-nurse ratios.” I do not think this study explores the effect of the staffing models (indeed it can’t) – rather it is using a different measure of nursing workload, derived from a staffing model / system compared to the classic approach using nurse to patient ratios. 4) (major) I realise that modelling these associations is complex and the models can be difficult to describe. However, for that reason it is all the more important to be absolutely clear because nothing can be assumed. Can I ask the authors to clarify the level on which the outcome is measured? I presume from the account that it is at the level of the day – so for each days staffing data there is a count of adverse outcomes (presumably 0 on most days?). Alternatively if it is the patient level there is (potentially) a count or a binary outcome (0/1) for the event. If this is the case how is the exposure of the patient to variable staffing levels taken into account? 5) (major) The comparison of models and associations between different approaches to estimating nursing workload is certainly an interesting feature of this paper although the approach taken is somewhat unsystematic. It is not aided by the fact that the different approaches to measuring workload are not treated in the same way without a common basis for forming categorisations (0ver and above optimal vs tertiles etc). This renders direct comparison rather difficult. Comparisons are (at times) made between levels of statistical significance (not ideal). Model fit indices are not presented. I think this needs to be addressed to provide a sounder basis for making comparison between the models and to make this conclusion more convincing (if it is correct) 6) (minor / discretionary) The fact that results of the OPCCq based workload measure and the nurse patient ratio measure become more similar when ward effects are introduced is striking – one might take this one step further and look at deviations in staffing (nurse to patient ratio and OPCq) from the norm for that ward rather than from the overall means. Of course the Rafaela system means this is, in effect, already done for the OPSq based ratings unless I am mistaken (it is not entirely clear). 7) (major) A major question for me is how a staffing level can be defined as ‘optimal’ when improvements on it are associated with significant reductions in adverse events. To me this suggests that the ‘optimal’ staffing level is simply not optimal. This does not fundamentally undermine the study as one where dependency based criteria are used to assess nursing workload rather than nurse patient ratios. However, it does rather call into question the particular system that is being used to define the “optimal” workload. The relationships presented here suggest it might yet provide a valid measure but I would question ‘optimality’. This needs to be addressed in the discussion 8) (minor) The authors offer a conclusion that associations between nurse staffing levels (based on ratios) and outcomes can

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	be challenged – but their results in fact largely confirm these associations although they do suggest potential refinements to methods (although as it stands I am not convinced that differences in significance full make the case). More minor points 1) The conclusions (final lines) do not arise from the study 2) Causal language in aim (abstract and elsewhere) – the nature of the study makes this inappropriate although I appreciate that it is the ultimate goal of this line of enquiry. There needs to be a tighter alignment between the actual analyses of the paper and the stated objectives… 3) I think the abstract would be difficult to follow for someone not versed in the research field or indeed already familiar with the RAFAELA system as

REVIEWER	Walter SERMEUS KU Leuven, Belgium
REVIEW RETURNED	02-Apr-2017

GENERAL COMMENTS	The study is interesting and is addressing shortcoming in the literature on nurse staffing and patient outcomes that most of research is cross-sectional, observational and is mostly performed on the hospital level instead of nursing unit level what is closer to the nurse-patient relation. This study is still observational, but is performed on the unit level during 1 year. But the study design has several flaws. 1. A first one is that the design is still observational what means that the objective (abstracts) and aim of study (introduction) should be rewritten: Now it is written "to investigate if a staffing model .... can affect patient safety accidents". It should be rewritten as ".... correlates with ..." 2. Secondly, PSI are measured using a PSI reporting system (HaiPro). Research by Classen et al. (Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, Whittington JC, Frankel A, Seger A, James BC. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood). 2011 Apr;30(4):581-9.) has shown that by using a trigger tool 10x more adverse events are found than by using indicators. In the same article, Classen et al. report that 100x more PSIs are found by the trigger tool than by PSI reporting. PSI reporting has shown to be an unreliable source of information for this type of research. In the manuscript, we don't find any data how these Hair data are collected, how much PSIs are reported, the reliability of the data etc... 3. The authors also report mortality data, coming from mortality registers, what is a good and independent source. But I don't see how they have linked the mortality data to unit staffing data. Most patients stayed during hospitalisation in different wards and not the date, time and place is not enough to explain. Again here the data might be cross-sectional again: if the patient dies on day x at ward y, we have to be careful to directly link nurse staffing on ward y and day x with mortality (time difference of cause to effect). 4. All analyses have been done on a aggregated unit level. It means that difference in nurse staffing on the unit/day level are evaluated on their impact on PSI and mortality rates on that unit/day instead of

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	on individual patients. Individual characteristics of patients (age, sex, disease, ...) are not taking into account. It might be that the case mix is different from day to day, explaining some of the variability. The day rates are not independent because some patients are staying several days at a nursing wards. My suggestion is that this should be explained more as a limitation of the study or the analysis should takes these elements into account. 5. Table 1 is too large and the legend is not fully explanatory. Table 1 should give an overview of descriptive data (not per unit, but globally). I would like to know how patients are in de study, with a short description of the main demographic and diagnostic data. I would like to know the nursing staffing (average and variability). Number of days in optimal staffing (and upper/lower). Number of PSI (total - variability). Number of deaths 6. Table 2 is too complex. The scenarios are not so different. Best is to make a choice which one to present, saying that you have evaluated different scenarios. It is interesting to see that if you have too many groups, the results become inconsistent. Based on the results it is difficult to conclude what are optimal nurse staffing ratios. If you have more nurses ("above optimum") you have lesser incidents and lesser deaths what seems to me the right optimum. I don't understand the number of events at the bottom line in the table (p20 line 11). You report 1367 incidents: 848 near misses, 400 adverse events (=1248), 246 >1 incident (Total= 1494). Table 2, 3 and 4 could be combined in one table. Unsure about conclusions. I want to see the new tables and data first

REVIEWER	Thomas Debray Julius Center for Health Sciences and Primary Care, The Netherlands
REVIEW RETURNED	22-May-2017

GENERAL COMMENTS	In this study, the authors investigate the association between nursing workload (measured by the RAFAELA system) and patient safety incidents using data from 4 Finnish hospitals. Alhough the research question certainly sounds relevant, I found it rather difficult to critically appraise and interpret the research findings, mostly because key information on the design and analysis was often missing or not adequately reported. Some examples are given below * Study participants are not sufficiently described. Which types of nurses were considered eligible, besides the fact that they were using the RAFAELA system? * All primary outcomes are prognostic, yet, no clear description is provided about the time frame in which outcomes were assessed (1 year?). For instance, when did incidents have to occur to be counted as such? Similarly, for mortality no description is given as which deaths were counted as events. Are patients who deceased after 5 days treated equally as patients who deceased after 1 year? Further, since the outcome is prognostic, why not adopt survival models to analyze the associations of interest (e.g. Cox regression)? * The authors are interested in the effects of nursing workload

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	(NWL), but use several reference (so called "optimal") values throughout the paper. It is not clear to me (1) why these optimal levels are considered relevant (as the paper aims to investigate associations between NWL levels and PSI) and (2) what the exact definitions are of NWL and "optimal" NWL. Why not simply treat NWL as a continuous covariate, or, alternatively, as a categorized covariate with clearly defined thresholds? * Variables used in the statistical models are not clearly described. In particular, which variables have been considered to adjust for confounding? Also, the exact definition of the different variables are missing (e.g. nursing intensity) * The data of the study is clustered within units of several Finnish hospitals. For this reason, it would be appropriate to adopt statistical models that account for clustering, as otherwise standard errors and confidence intervals are likely too narrow. This can, for instance, be achieved using stratified (and/or random) intercept terms. The authors mention the use of unit-fixed effects, perhaps these are the models that account for clustering? (as confidence intervals are indeed smaller in this case). If so, what methods were used to decide on the complexity of these models? In general, it is not clear what statistical models have been used in the manuscript, it would therefore be helpful to describe them in an appendix. Finally, I think it would be appropriate to report that the RAFAELA system was proposed by the primary author of this paper. It is not clear whether use of this system involves any commercial elements (e.g. licensing or support fees), but if so it should be reported as a potential conflict of interest. Minor comments * Abstract: "Main outcome measures were PSI, and mortality data" - Please mention the type (and timing) of mortality, rather than the source of data. E.g. "Primary and secondary outcome measures: PSI and all-cause mortality within XXXX months." * Study setting: Please describe the RAFAELA system or provide a reference. The system is discussed in more detail later on, but it would be helpful to provide some information early on. * Results & Discussion - The authors mention the use of unstandardized and standardized models. This terminology is rather confusing, as standardization often refers to transformation of covariates or outcomes (yet no description of the type of standardization seems to be provided). When adjusting statistical models for confounders, it is more common to use terms like "adjusted" or "multivarible". Vice versa, when estimating crude associations, it is common to use terms like "unadjusted" or "univarable". * Page 9 - "[...] associations between the NWL in relation to optimal ratio and the occurrence of incidents remained almost the same". What is the definition of "the optimal ratio". What outcome does it represent? Or rather, did the authors aim to describe the behaviour of differences in NWL levels on PSI? * Page 11 - "An assumption is that when NWL is very high only some incidents are reported". Where does this assumption apply? Which results should be interpreted according to this assumption? * Page 11 - The authors state that "we find evidence that a staffing model based on [...] can predict incidents and mortality rates better

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	than a patient-to-nurse model". I dont fully agree with this statement as the authors did not investigate the predictive performance of these models. * Table 2: Please indicate that the odds ratios are unadjusted (i.e. they represent crude associations) * Table 3 and 4, and Figure 1: Please indicate that presented odds ratios are adjusted, and describe the covariates of adjustment. Current descriptions such as "unit-specific effects" and "effects of weekday, holiday and season" are not clear, and obfuscates whether the model involves random effects distributions or additional (e.g. linear) covariates. * Figure 1 represents odds ratios, not risks Spelling/grammar * The term multivariate is commonly used to address multiple outcomes, rather than multiple predictors. Please rephrase as "multivarable" throughout the manuscript. (e.g. in absract "Using multivariate logistic regression analyses" should read as "Using multivariable logistic regression analyses" * Abstract: The sentences "The odds for one PSI" and "The odds for death" are somewhat odd, please consider rephrasing as "The odds for PSI occurrence" and, respectively, by "The odds for all-cause mortality within ... months" * Study setting: "In that the study received approval from the chief administrative physicians of all four hospitals involved, no further ethical approval was necessary" - Please consider rephrasing as "The current study received approval from the chief administrative physicians of all four hospitals involved, and therefore no further ethical approval was necessary" * Page 11 - "We find evidence that a staffing model based on daily measurement of ..." - I think past tense would be more appropriate here: "We found evidence that ..."

VERSION 1 – AUTHOR RESPONSE

Dear Editor,
We are grateful for the insightful comments on our submission. They made us realise that we were somewhat unclear on some central points. We have therefore now revised the paper in accordance with the recommendations. Below follows a point-to-point response to each of the comments raised.
EDITOR

- Please edit the title so that it is not declarative. It should also contain the country.

Response: The title has now been revised to be non-declarative and to include the country.

- Please leave all recommendations on the STROBE checklist - on the second page these have been deleted.

Response: The checklist has been corrected.

- The document needs to be portrait formatted, not landscape.

Response: The document is now portrait formatted.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

- Please ensure that your manuscript is proofread by a native English speaker prior to resubmission, to check for any language errors.

Response: The manuscript has been proofread prior to resubmission.

REVIEWER 1

1) (minor) The background needs to say a little more about the RAFAELA system so that it is clearer what it is – both in terms of the nature of the data gathered and the measure it provides – for example does it give a direct measure of the required nurse staffing level in some way? More direct reference to the extensive evidence base behind the system would be appropriate as well although I appreciate the need to be concise at this point. The meaning of the term 'optimal' in this context needs to be clarified. I realise more detail is given later but I think the nature of the tool needs to be established in the background.

Response: A short description of the RAFAELA system has been added. The term 'optimal' has been defined. We hope that these additions will clarify the text, and make it easier for the reader to grasp the RAFAELA system and logic of the study.

2) (minor) The account of the system (Raphaela) and the measures used in it (PAONCIL, OPCq) in the methods is confusing to me – who is not uninitiated – I suspect it would be much worse for those not already familiar. I may appear to be contradicting myself when I say suggest that much of this be abbreviated by addressing this as a study which 'simply' uses a measure of nursing workload that is based on a system designed to determine nursing workload based on patient classification – which then gives us a measure of nursing workload… and contrast that to studies where the implied nursing workload is uniform across all patients (and hence invariant for any given nurse to patient ratio). If you could do this and give appropriate reference to the system itself you would be offering more detail than Needleman's study! If you avoid the extensive detail you can focus on reporting your variables more clearly.

Response: We have now simplified the description of the RAFAELA system as a comprehensive system that produces detailed data on daily level.

3) (major) The aim of the study is not clearly described. This relates to point 2 above. It is stated that "The aim of this observational study was to investigate whether a staffing model based on optimal NWL, measured daily using the RAFAELA Nursing Intensity and Staffing system, can positively affect patient safety incidents (incidents) and mortality. Also, we want to compare staffing models based on the RAFAELA system with those based on patient-to-nurse ratios." I do not think this study explores the effect of the staffing models (indeed it can't) – rather it is using a different measure of nursing workload, derived from a staffing model / system compared to the classic approach using nurse to patient ratios.

Response: The aim of the study is rewritten: The aim of this observational study was to investigate whether using the optimal NWL as a measure based on the RAFAELA system, correlates with patient safety incidents and mortality. Also, we wanted to compare the use of optimal NWL as a measure with those based on standard patient-to-nurse ratios.

4) (major) I realise that modelling these associations is complex and the models can be difficult to describe. However, for that reason it is all the more important to be absolutely clear because nothing can be assumed. Can I ask the authors to clarify the level on which the outcome is measured? I presume from the account that it is at the level of the day – so for each days staffing data there is a

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

count of adverse outcomes (presumably 0 on most days?). Alternatively if it is the patient level there is (potentially) a count or a binary outcome (0/1) for the event. If this is the case how is the exposure of the patient to variable staffing levels taken into account?

Response: Yes, each outcome is measured at the level of the day. This is now better pointed out througout the text.

5) (major) The comparison of models and associations between different approaches to estimating nursing workload is certainly an interesting feature of this paper although the approach taken is somewhat unsystematic. It is not aided by the fact that the different approaches to measuring workload are not treated in the same way without a common basis for forming categorisations (0ver and above optimal vs tertiles etc). This renders direct comparison rather difficult. Comparisons are (at times) made between levels of statistical significance (not ideal). Model fit indices are not presented. I think this needs to be addressed to provide a sounder basis for making comparison between the models and to make this conclusion more convincing (if it is correct)

Response: We now give also model fit indices in the results table, and agree that they provide a sounder basis for making comparison between the models.

6) (minor / discretionary) The fact that results of the OPCCq based workload measure and the nurse patient ratio measure become more similar when ward effects are introduced is striking – one might take this one step further and look at deviations in staffing (nurse to patient ratio and OPCq) from the norm for that ward rather than from the overall means. Of course the Rafaela system means this is, in effect, already done for the OPSq based ratings unless I am mistaken (it is not entirely clear).

Response: Deviation from the norm of each ward is in fact what the Rafaela system measures.

7) (major) A major question for me is how a staffing level can be defined as ‘optimal’ when improvements on it are associated with significant reductions in adverse events. To me this suggests that the ‘optimal’ staffing level is simply not optimal. This does not fundamentally undermine the study as one where dependency based criteria are used to assess nursing workload rather than nurse patient ratios. However, it does rather call into question the particular system that is being used to define the “optimal” workload. The relationships presented here suggest it might yet provide a valid measure but I would question ‘optimality’. This needs to be addressed in the discussion

Response: The optimal nursing intensity level is determined by the PAONCIL method at each unit. We have now used the term ’recommended optimal level’ (for example 22-30 OPC points per nurse; see page 5) throughout the text. The validity and reliability of the PAONCIL method has been tested in earlier studies, both in Finland and Norway. If the workload per nurse is lower than 22 OPC/nurse, then you can say that the nurse has extra time per each patient. The question of ’optimality’ has been adressed in the discussion (in the beginning and at the end).

8) (minor) The authors offer a conclusion that associations between nurse staffing levels (based on ratios) and outcomes can be challenged – but their results in fact largely confirm these associations although they do suggest potential refinements to methods (although as it stands I am not convinced that differences in significance full make the case).

Response: The now reported model fit indices provide firmer evidence for the argumentation.

More minor points

1) The conclusions (final lines) do not arise from the study

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

Response: Both the discussion and the conlcusions have been rewritten.

2) Causal language in aim (abstract and elsewhere) – the nature of the study makes this inappropriate although I appreciate that it is the ultimate goal of this line of enquiry. There needs to be a tighter alignment between the actual analyses of the paper and the stated objectives…

Response: We hope that our corrections have clarified the entire text.

3)I think the abstract would be difficult to follow for someone not versed in the research field or indeed already familiar with the RAFAELA system as

Response: The abstract has been rewritten.

REVIEWER 2

1. A first one is that the design is still observational what means that the objective (abstracts) and aim of study (introduction) should be rewritten: Now it is written "to investigate if a staffing model .... can affect patient safety accidents". It should be rewritten as ".... correlates with ..."

Response: The aim is rewritten in accordance with the reviewer's recommendations.

2. Secondly, PSI are measured using a PSI reporting system (HaiPro). Research by Classen et al. (Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, Whittington JC, Frankel A, Seger A, James BC. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood). 2011 Apr;30(4):581-9.) has shown that by using a trigger tool 10x more adverse events are found than by using indicators. In the same article, Classen et al. report that 100x more PSIs are found by the trigger tool than by PSI reporting. PSI reporting has shown to be an unreliable source of information for this type of research. In the manuscript, we don't find any data how these Hair data are collected, how much PSIs are reported, the reliability of the data etc...

Response: We have added a new reference (Holmstrøm 2017) that refers to GGT and the research of Classen et al. However, we want to point out that GGT collects triggers and patient safety incidents from treatment periods, not on daily basis, whereas data on incidents collected from HaiPro can be targeted to certain days. Two of the authors are involved in a comprehensive patient safety study at Vaasa central hospital. In this study, associations between nursing intensity and incidents will be analyzed also by using GTT. The GTT is only looking for adverse events, but volunteering reporting on hazards also reveals near-miss situations, and thus significantly improves staff safety awareness.

3. The authors also report mortality data, coming from mortality registers, what is a good and independent source. But I don't see how they have linked the mortality data to unit staffing data. Most patients stayed during hospitalisation in different wards and not the date, time and place is not enough to explain. Again here the data might be cross-sectional again: if the patient dies on day x at ward y, we have to be careful to directly link nurse staffing on ward y and day x with mortality (time difference of cause to effect).

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

Response: Both the mortality data and the nurse staffing data is from the same unit and has been collected on daily basis. The assumption is that high workload decreases the nurses' time for caring and observations, and hereby the mortality risk during these days may raise.

4. All analyses have been done on a aggregated unit level. It means that difference in nurse staffing on the unit/day level are evaluated on their impact on PSI and mortality rates on that unit/day instead of on individual patients. Individual characteristics of patients (age, sex, disease, ...) are not taking into account. It might be that the case mix is different from day to day, explaining some of the variability. The day rates are not independent because some patients are staying several days at a nursing wards. My suggestion is that this should be explained more as a limitation of the study or the analysis should takes these elements into account.

Response:.Earlier studies have showed that OPC instrument identifies patients' individual characteristics such as functional ability, symptoms of diseases, and the effect on nursing intensity of the most central patient characteristics. Hence, the measurement by the OPC covers the actual patient case mix for each day. However the effect of these aspects, especially age and gender, are motivated to analyze more in detail in further studies.

5. Table 1 is too large and the legend is not fully explanatory. Table 1 should give an overview of descriptive data (not per unit, but globally). I would like to know how patients are in de study, with a short description of the main demographic and diagnostic data. I would like to know the nursing staffing (average and variability). Number of days in optimal staffing (and upper/lower). Number of PSI (total - variability). Number of deaths

Response: We have added in the end of Table 1 a description of the medical specialization of participating units. Regarding nurse staffing data, Patients per nurse, OPC per patients, OPC per nurse – that can be found in the table are data that give an overall picture of patients' aquity level and nurs staffing. Table 1 is already comprehensive, and more additional information may perhaps not make the table better.

6. Table 2 is too complex. The scenarios are not so different. Best is to make a choice which one to present, saying that you have evaluated different scenarios. It is interesting to see that if you have too many groups, the results become inconsistent. Based on the results it is difficult to conclude what are optimal nurse staffing ratios. If you have more nurses ("above optimum") you have lesser incidents and lesser deaths what seems to me the right optimum.
I don't understand the number of events at the bottom line in the table (p20 line 11). You report 1367 incidents: 848 near misses, 400 adverse events (=1248), 246 >1 incident (Total= 1494).

Response: We have now simplified, clarified and reconstructed the results table(s). The number of incidents were correct, but the types of incidents are now better described in the text in order to avoid confusion.

Table 2, 3 and 4 could be combined in one table.

Response: We have now combined the previous Tables 2-4 into one table (Table 2). Rest of the results are now commented upon in the text, and available upon request.

REVIEWER 3

* Study participants are not sufficiently described. Which types of nurses were considered eligible, besides the fact that they were using the RAFAELA system?

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

Response: The daily nursing intensity on each unit is assessed by all the responsible registered nurses on each day. One registered nurse may usually classify 1 to 6 patients per day. The assessment is done by classifying the patient's care needs by the Oulu Patient Classification (OPCq) instrument.

* All primary outcomes are prognostic, yet, no clear description is provided about the time frame in which outcomes were assessed (1 year?). For instance, when did incidents have to occur to be counted as such? Similarly, for mortality no description is given as which deaths were counted as events. Are patients who deceased after 5 days treated equally as patients who deceased after 1 year? Further, since the outcome is prognostic, why not adopt survival models to analyze the associations of interest (e.g. Cox regression)?

Response: The information is at level of the day at each ward studied, and on an individual level of each patient. Thus we do not have time-to-event setting that is needed for survival models.

* The authors are interested in the effects of nursing workload (NWL), but use several reference (so called "optimal") values throughout the paper. It is not clear to me (1) why these optimal levels are considered relevant (as the paper aims to investigate associations between NWL levels and PSI) and (2) what the exact definitions are of NWL and "optimal" NWL. Why not simply treat NWL as a continuous covariate, or, alternatively, as a categorized covariate with clearly defined thresholds?

Response: A short description of the RAFAELA system has been added and the optimal NWL is a central concept of the system. The term 'optimal' has been defined. We hope that these additions will clarify the text, and make it easier for the reader to grasp the RAFAELA system, the use of optimal NWL and the logic of the study.

* Variables used in the statistical models are not clearly described. In particular, which variables have been considered to adjust for confounding? Also, the exact definition of the different variables are missing (e.g. nursing intensity)

Response: The variables are now described in the notes of Table 2, and how they are used in the analyses are now better discussed in the text.

* The data of the study is clustered within units of several Finnish hospitals. For this reason, it would be appropriate to adopt statistical models that account for clustering, as otherwise standard errors and confidence intervals are likely too narrow. This can, for instance, be achieved using stratified (and/or random) intercept terms. The authors mention the use of unit-fixed effects, perhaps these are the models that account for clustering? (as confidence intervals are indeed smaller in this case). If so, what methods were used to decide on the complexity of these models? In general, it is not clear what statistical models have been used in the manuscript, it would therefore be helpful to describe them in an appendix.

Response: Yes, days in the data are clustered within each ward. We have chosen to report results of models with fixed ward-effects, which is now better described in the text. Unlike the random effects model, which rests on the assumption that the 'ward residual' must be normally distributed and that it must be uncorrelated with the characteristics represented by the covariates, no assumptions of that kind need to be made about the 'ward residual' in the fixed effect model . Our estimator is based on dummy variables for the ward effects (which is equivalent to using a fixed effect estimator in which the ward effects are 'differenced' out). We consequently obtain estimates of the ward effects directly from the estimation. Since the number of observations within each ward is large (177-365) and the within-ward variance is not large relative to the between-ward variance, we do not have any problem with an

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

unreliable estimator, and thus not with extremely small or extremely large estimates (which generally occur because of sampling variability and thus a small number of observations within each unit studied). However, since the number of observations within each ward is large, the fixed effects estimator and the random effects estimator will not produce wildly different results.

*Finally, I think it would be appropriate to report that the RAFAELA system was proposed by the primary author of this paper. It is not clear whether use of this system involves any commercial elements (e.g. licensing or support fees), but if so it should be reported as a potential conflict of interest.

Response: None of the authors have any conflict of interest, such as licensing fees.

Minor comments

* Abstract: "Main outcome measures were PSI, and mortality data" - Please mention the type (and timing) of mortality, rather than the source of data. E.g. "Primary and secondary outcome measures: PSI and all-cause mortality within XXXX months."

Response: Each outcome of interest is measured at the daily level of each ward during a period of one year, which means that we have 365 observations per ward, unless a ward has closed down during some period of the calendar year.

* Study setting: Please describe the RAFAELA system or provide a reference. The system is discussed in more detail later on, but it would be helpful to provide some information early on.

Response: The RAFAELA system is now better described.

* Results & Discussion - The authors mention the use of unstandardized and standardized models. This terminology is rather confusing, as standardization often refers to transformation of covariates or outcomes (yet no description of the type of standardization seems to be provided). When adjusting statistical models for confounders, it is more common to use terms like "adjusted" or "multivarible". Vice versa, when estimating crude associations, it is common to use terms like "unadjusted" or "univarable".

Response: We agree, and now refer to unadjusted and adjusted models.

* Page 9 - "[...] associations between the NWL in relation to optimal ratio and the occurrence of incidents remained almost the same". What is the definition of "the optimal ratio". What outcome does it represent? Or rather, did the authors aim to describe the behaviour of differences in NWL levels on PSI?

Response: The text has been rewritten, so we hope that this problem has been solved.

* Page 11 - "An assumption is that when NWL is very high only some incidents are reported". Where does this assumption apply? Which results should be interpreted according to this assumption?

Response: This phenomenon is about the problem: high stress – less time for reporting. The text has been corrected. ’An assumption is that when NWL is very high, that is a working situation when the nurse staff resources are too low and the nurses may not prioritize the registration of adverse events due to high workload, resulting in incidents connected to high NWL being underreported.’

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

* Page 11 - The authors state that "we find evidence that a staffing model based on [...] can predict incidents and mortality rates better than a patient-to-nurse model". I dont fully agree with this statement as the authors did not investigate the predictive performance of these models.

Response: As requested by Reviewer 1, we now report also model fit indices.

* Table 2: Please indicate that the odds ratios are unadjusted (i.e. they represent crude associations)

Response: The tables have now been reconstructed, and this comment has therefore been met.

* Table 3 and 4, and Figure 1: Please indicate that presented odds ratios are adjusted, and describe the covariates of adjustment. Current descriptions such as "unit-specific effects" and "effects of weekday, holiday and season" are not clear, and obfuscates whether the model involves random effects distributions or additional (e.g. linear) covariates.

Response: We have now rephrased and better define what we are doing.

* Figure 1 represents odds ratios, not risks

Response: The figure have been deleted.

* The term multivariate is commonly used to address multiple outcomes, rather than multiple predictors. Please rephrase as "multivarable" throughout the manuscript. (e.g. in absract "Using multivariate logistic regression analyses" should read as "Using multivariable logistic regression analyses"

Response: We have rephrased the terms.

* Abstract: The sentences "The odds for one PSI" and "The odds for death" are somewhat odd, please consider rephrasing as "The odds for PSI occurrence" and, respectively, by "The odds for all-cause mortality within ... months"

Response: The text has been checked.

* Study setting: "In that the study received approval from the chief administrative physicians of all four hospitals involved, no further ethical approval was necessary" - Please consider rephrasing as "The current study received approval from the chief administrative physicians of all four hospitals involved, and therefore no further ethical approval was necessary"

Response: The text has been rewritten on this point.

* Page 11 - "We find evidence that a staffing model based on daily measurement of ..." - I think past tense would be more appropriate here: "We found evidence that ..."

Response: This has been checked.

VERSION 2 – REVIEW

REVIEWER	Peter Griffiths University of Southampton UK
REVIEW RETURNED	27-Jun-2017

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

GENERAL COMMENTS	This paper is much improved and the current presentation is much clearer. I have a number of minor points and one very significant point. Major points The major reservation about publication (as before) relates to the claims made about 'optimal' nursing workload and the system that this measure derives from. This arises troughout the manuscript. "Strengths and limitations of this study" has the statement: "The study is the first to assess the relationship between optimal nursing workload and outcomes based on data obtained on a daily basis" As per comments on previous revision, this study does not measure 'optimal' nursing workload – this needs to be revised to remove confusion around the category that is defined as optimal by the system. P4. Has the statement "When the actual NWL is at the optimal level, the resources are considered to be allocated appropriately." – the results of this study demonstrate that this is not the case. While this may not need revising here it is worth pointing out or emphasising that this is an assumption. Conclusions (p11) States "By using the recommended optimal NWL as a tool and golden standard for allocation of nursing staff, the nurse managers can optimize the resources and ensure patient outcomes." This conclusion simply does not follow from this study. This study has quite convincingly demonstrated that i) This approach to measuring staffing is probably superior to a nurse to patient ratio ii) Variation in this workload measure is associated with variation in outcomes iii) That the recommended 'optimal' staffing levels of the system may in fact be wrong because improvements in outcome are seen what staffing is increased above it I cannot help but think that the author's investment in the system makes them reluctant to offer this conclusion. This seems unfortunate because it opens up an interesting and important line of research in order to investigate whether there is in fact a 'tipping point' where there is diminishing return from further improvements in staffing which could lead to a recalibration of the system. COI – I think the involvement of the authors in developing the tool should be declared as a potential conflict of interest for the sake of transparency. Minor points The statement is made "This study received approval from the chief administrative physicians of all four hospitals involved, and therefore no further ethical approval was necessary. " The need for no further approval does not follow without further explanation. I presume that the next sentence is that explanation but this needs to be linked with 'because'. I would also suggest including a statement that this is in accord with the regulatory regime for the conduct of health research in Finland. Para running over p 9 / 10 – this begins by acknowledging potential underreporting of incidents in the face of low staffing. However, it ends with "The OPC per nurse calculation is therefore more detailed…" This is a non sequitur and needs to be revised. What are the consequences of under reporting in the face of high workload? In my view this would tend to attenuate any observed relationship and so it does not undermine you conclusions, but you need to say this! P10 para 1 – I am confused by reference to the GTT which is not used in this study – the use of future tense causes further confusion.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	P10 para 2 “We found evidence that a staffing model based on daily measurement of individual patient care needs and optimal NWL…” As previously – I do not believe that this study is testing a staffing model – rather a staffing measure. Limitations – death caused by low staffing on a ward on one day may not occur of the same day or ward.

REVIEWER	Walter Sermeus Leuven Institute for Healthcare Policy, KU Leuven
REVIEW RETURNED	05-Jul-2017

GENERAL COMMENTS	The authors have really reworked the manuscript very well. We however didn't received a document with answers to the comments of the reviewers and/or what changes have been made. But using a different colour to indicate the changed content was helpful. Table 1 could be simplified by not giving the information on the level of the nursing unit, but rather on the level of department (cfr. Legend). This would be more appropriate. explanation for each of the variables should be added in the Legend Table 2 should be simplified as well. I should present OPC/nurse unadjusted and only 1 adjusted model and for patients/nurse unadjusted and only 1 adjusted model. You can drop "at optimum" for OPC/nurse and "2nd group" in patients/nurse. I would drop -2 log likelihood as statistic but would add significant * or NS. In both variables Below/above optimum and 1st group/3rd group should be defined in the legend. Note: in the legend Weekdays are all days. There should be a difference between week and weekend/holidays. As result, there is a difference in incidents reported but not in harm or death for OPC/nurse. There is no difference in patient/nurse. So the only difference in both nurse staffing methods (using plain statistics such as patient-to-nurse ratio or a more complicated one in using patient classification systems) is in the number of incidents reported. I don't agree with the statement on p10 l34 that a staffing model based on an optimal NWL can better predict mortality rates than a patient to nurse model. Yes for incidents, no for mortality and harm. I don't agree that reporting incidents is a reliable method of collecting adverse events (p10 l18). I refer to the study of Classen et al. (2011) who are showing than only 1/100 adverse events is reported by incidence reporting compared to GTT. You maybe be confident that the HAiPro database is a good reporting system (compared to other reporting systems) but the evidence is there that reporting is a weak form of collecting data on adverse events. (Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, Whittington JC, Frankel A, Seger A, James BC. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood). 2011 Apr;30(4):581-9.)

REVIEWER	Thomas Debray Julius Center for Health Sciences and Primary Care

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

REVIEW RETURNED	21-Aug-2017

GENERAL COMMENTS	Comment 1: Statistical models are still not clearly described. In particular, more explaination is needed for model 1, which was "adjusted for ward-specific effects". As I understand from the reviewer comments, this model treats "Ward" as a dummy variable in the model, and thus basically allows for heterogeneity in intercept term (without requiring the Normality assumption). It is not clear how odds ratios were modeled, but it appears that some interaction with a dummy term was involved ("Estimates for ward-specific effects and effects of weekday, holiday and season are not displayed here."). In this is indeed the case, the authors need to report how corresponding odds ratios were pooled across wards. If a fixed (i.e. common) effect was assumed, this needs to be supported, e.g. using prediction intervals demonstrating that estimates of across-ward variability are relatively small. Furhter, the paper needs to be more clear about the specificiation of the various models, e.g. the use of dummy factors to adjust for clustering is currently not clear from the main text. Finally, I strongly recommend to provide full details (i.e. model specification) of the analyses in an appendix, and to provide source code if possible. References * Riley, R. D., J. P. T. Higgins, and J. J. Deeks. “Interpretation of Random Effects Meta-Analyses.” BMJ 342, no. feb10 2 (February 10, 2011): d549–d549. doi:10.1136/bmj.d549. * Gardiner, Joseph C., Zhehui Luo, and Lee Anne Roman. “Fixed Effects, Random Effects and GEE: What Are the Differences?” Statistics in Medicine 28, no. 2 (January 30, 2009): 221-39. doi:10.1002/sim.3478. * Robinson, G K. “That BLUP Is a Good Thing: The Estimation of Random Effects.” Statistical Science 6, no. 1 (February 1991): 15-32. Comment 2: I still do not agree with the statement that "we find evidence that a staffing model based on [...] can predict incidents and mortality rates better than a patient-to-nurse model" because the authors did not perform comparative analyses that investigate the relative performance of these models. Reporting log likelihoods of individual models is simply not sufficient, and evidence is needed on their statistical difference (e.g. using AIC), as well as the difference in their discrimination and calibration performance. References * Steyerberg, Ewout W., Michael J. Pencina, Hester F. Lingsma, Michael W. Kattan, Andrew J. Vickers, and Ben Van Calster. “Assessing the Incremental Value of Diagnostic and Prognostic Markers: A Review and Illustration.” European Journal of Clinical Investigation 42, no. 2 (February 2012): 216–28. doi:10.1111/j.1365-2362.2011.02562.x.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	Comment 3: In line with previous comment, I do not agree with the conclusions that "Models estimated on basis of the RAFAELA classification system generally provided [...] better model fit than those based on the standard patients-to-nurse classification" and that "Hence, our analyses of the data suggest that, when it comes to predicting occurrence of patient incidents and mortality, measuring nursing workload according to the RAFAELA system is to be preferred above the standard patients-to-nurse approach." As illustrated in the Tables, and indicated by the authors "the difference was [in model fit was] not very large." In fact, the differences have not statistically been tested and appear rather neglegible. Further, as mentioned above, comparisons in terms of predictive performance (discrimination and calibration) are missing and should be included to make formal statements on predictive performance. Finally, to infer on the "preferred" approach, one should also investigate the (relative) impact of the different models, e.g. using decision curve analysis. References * Vickers, Andrew J., and Elena B. Elkin. “Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.” Medical Decision Making: An International Journal of the Society for Medical Decision Making 26, no. 6 (December 2006): 565–74. doi:10.1177/0272989X06295361. Comment 4: It appears that the RAFAELA system is owned by the Association of Finnish Local and Regional Authorities, and that its use is managed by the FCG Finnish Consulting Group Ltd. Although the author does not report any conflict of interest, I still think it would be appropriate to mention the (scientific) relation between the primary author and the RAFAELA system, although this could perhaps simply be done in the introduction. I leave this decision with the editors of BMJ open. Minor 1. Abstract: "To investigate whether the recommended daily workload per nurse (OPC/nurse) as measured by the RAFAELA system correlates with different types of patient safety incidents and with patient mortality, and to compare this interrelation with that based on the standard patients-to-nurse classification." I would indicate that the aim is to assess the magnitude of this correlation (or better, association), rather than its mere presence. 2. Outcomes: I would indicate that "whether at least one incident, of any type, occurred" within the available follow-up of 365 days (or less where relevant).

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

VERSION 2 – AUTHOR RESPONSE

Response to comments on bmjopen-2017-016367.R1

We appreciate the additional comments received, which we think were legitimate. We have now tried our best in responding to them all and revised the document accordingly. Below follows a point-by-point response to each comment.

Reviewer 1

Major points

The major reservation about publication (as before) relates to the claims made about 'optimal' nursing workload and the system that this measure derives from. This arises troughout the manuscript.

Response: We fully understand your point, and have now used the word 'assumed optimal' through out the text. Additionally, we have further described semantically the meaning of 'optimal' on page 3.

"Strengths and limitations of this study" has the statement: "The study is the first to assess the relationship between optimal nursing workload and outcomes based on data obtained on a daily basis" As per comments on previous revision, this study does not measure 'optimal' nursing workload – this needs to be revised to remove confusion around the category that is defined as optimal by the system.

Response: This sentence have been corrected.

P4. Has the statement "When the actual NWL is at the optimal level, the resources are considered to be allocated appropriately." – the results of this study demonstrate that this is not the case. While this may not need revising here it is worth pointing out or emphasising that this is an assumption.

Response: The sentence have been corrected.

Conclusions (p11) States "By using the recommended optimal NWL as a tool and golden standard for allocation of nursing staff, the nurse managers can optimize the resources and ensure patient outcomes." This conclusion simply does not follow from this study. This study has quite convincingly demonstrated that

i) This approach to measuring staffing is probably superior to a nurse to patient ratio
ii) Variation in this workload measure is associated with variation in outcomes
iii) That the recommended 'optimal' staffing levels of the system may in fact be wrong because improvements in outcome are seen what staffing is increased above it

I cannot help but think that the author's investment in the system makes them reluctant to offer this conclusion. This seems unfortunate because it opens up an interesting and important line of research in order to investigate whether there is in fact a 'tipping point' where there is diminishing return from further improvements in staffing which could lead to a recalibration of the system.
COI – I think the involvement of the authors in developing the tool should be declared as a potential conflict of interest for the sake of transparency.

Response: We have removed 'golden' based on the facts that further research and evidence is needed, before such a statement. However, after 15-20 years of daily use of the RAFAELA system, many nurse managers can agree that it is a tool for optimizing the staff resources. Part of the text (conclusions) have been rewritten. The authors still declare no potential conflicts of interest with

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

respect to the research, authorship, and/or publication of this article. However, we declare that the first author has been involved in developing the RAFAELA system.
Minor points

The statement is made “This study received approval from the chief administrative physicians of all four hospitals involved, and therefore no further ethical approval was necessary. ”The need for no further approval does not follow without further explanation. I presume that the next sentence is that explanation but this needs to be linked with ‘because’. I would also suggest including a statement that this is in accord with the regulatory regime for the conduct of health research in Finland.

Response: This section has been clarified in accordance with the suggestions.

Para running over p 9 / 10 – this begins by acknowledging potential underreporting of incidents in the face of low staffing. However, it ends with “The OPC per nurse calculation is therefore more detailed…” This is a non sequitur and needs to be revised. What are the consequences of under reporting in the face of high workload? In my view this would tend to attenuate any observed relationship and so it does not undermine you conclusions, but you need to say this!

Response: The sentence “The OPC per nurse calculation is therefore more detailed…” has been moved up, and the text regarding the consequences of under reporting has been rewritten.

P10 para 1 – I am confused by reference to the GTT which is not used in this study – the use of future tense causes further confusion.

Response: This section has been removed.

P10 para 2 “We found evidence that a staffing model based on daily measurement of individual patient care needs and optimal NWL…” As previously – I do not believe that this study is testing a staffing model – rather a staffing measure.

Response: You are correct. We have revised these statements in the text.

Limitations – death caused by low staffing on a ward on one day may not occur of the same day or ward.

Response: Two new sentences has been added according to this issue.

Reviewer 2

Table 1 could be simplified by not giving the information on the level of the nursing unit, but rather on the level of department (cfr. Legend). This would be more appropriate. explanation for each of the variables should be added in the Legend

Response: This comment is probably due to a misunderstanding. The information provided in the table is at the department (ward) level, which is legitimate when considering the setup of data and analyses. In the footnotes we now mention that the data are described in detail in the text. A detailed description in the footnotes of the table would, in our opinion, take too much space, particularly since the concepts are described better within the main text.

Table 2 should be simplified as well. I should present OPC/nurse unadjusted and only 1 adjusted model and for patients/nurse unadjusted and only 1 adjusted model. You can drop "at optimum" for OPC/nurse and "2nd group" in patients/nurse. I would drop -2 log likelihood as statistic but would add

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

significant * or NS. In both variables Below/above optimum and 1st group/3rd group should be defined in the legend. Note: in the legend Weekdays are all days. There should be a difference between week and weekend/holidays.

Response: We have removed the previous 'Adjusted model type 1' (as we agree that it is not necessary) and now include the unadjusted and fully adjusted models. We could not drop the indices of model fit, since they are concerned with the comments raised by Referee 3. In the footnotes of the table, and to some extent in the main text, we have rewritten part of the description of the variables. Also, full description of all models estimated and their estimates are now provided in the supplementary electronic files.

As result, there is a difference in incidents reported but not in harm or death for OPC/nurse. There is no difference in patient/nurse. So the only difference in both nurse staffing methods (using plain statistics such as patient-to-nurse ratio or a more complicated one in using patient classification systems) is in the number of incidents reported.

Response: Unfortunately, we do not fully understand this comment, but we have rewritten parts of the results section to be more clear about the interpretation of the findings.

I don't agree with the statement on p10 l34 that a staffing model based on an optimal NWL can better predict mortality rates than a patient to nurse model. Yes for incidents, no for mortality and harm.

Response: Please see the above comment.

I don't agree that reporting incidents is a reliable method of collecting adverse events (p10 l18). I refer to the study of Classen et al. (2011) who are showing than only 1/100 adverse events is reported by incidence reporting compared to GTT. You maybe be confident that the HAiPro database is a good reporting system (compared to other reporting systems) but the evidence is there that reporting is a weak form of collecting data on adverse events.

Response: We have added a sentence that 'we cannot guarantee that no reports are missing', and the reference,

Reviewer 3

Statistical models are still not clearly described. In particular, more explaination is needed for model 1, which was "adjusted for ward-specific effects". As I understand from the reviewer comments, this model treats "Ward" as a dummy variable in the model, and thus basically allows for heterogeneity in intercept term (without requiring the Normality assumption). It is not clear how odds ratios were modeled, but it appears that some interaction with a dummy term was involved ("Estimates for ward-specific effects and effects of weekday, holiday and season are not displayed here."). In this is indeed the case, the authors need to report how corresponding odds ratios were pooled across wards. If a fixed (i.e. common) effect was assumed, this needs to be supported, e.g. using prediction intervals demonstrating that estimates of across-ward variability are relatively small. Furhter, the paper needs to be more clear about the specificiation of the various models, e.g. the use of dummy factors to adjust for clustering is currently not clear from the main text. Finally, I strongly recommend to provide full details (i.e. model specification) of the analyses in an appendix, and to provide source code if possible.

Response: We are sorry about this unnecessary confusion. In the text we now describe the statistical models better. Yes, the wards are treated as dummies, and odds ratios are modelled as fixed effects for each dummy (no interactions). This is supported by the data, since across-ward variability is

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

modest. We now provide supplementary electronic files that contain description of all models estimated and complete results of all regressions.

I still do not agree with the statement that "we find evidence that a staffing model based on [...] can predict incidents and mortality rates better than a patient-to-nurse model" because the authors did not perform comparative analyses that investigate the relative performance of these models. Reporting log likelihoods of individual models is simply not sufficient, and evidence is needed on their statistical difference (e.g. using AIC), as well as the difference in their discrimination and calibration performance.

Response: Apart from the log likelihood, we now provide also the AIC and the Nagelkerke, and if needed we can provide also other measures. Conclusions are the same as before; the OPC/nurse measure provides better fit then the patients/nurse measure. However, we now explicitly point out that the difference is not very large.

In line with previous comment, I do not agree with the conclusions that "Models estimated on basis of the RAFAELA classification system generally provided [...] better model fit than those based on the standard patients-to-nurse classification" and that "Hence, our analyses of the data suggest that, when it comes to predicting occurrence of patient incidents and mortality, measuring nursing workload according to the RAFAELA system is to be preferred above the standard patients-to-nurse approach."

Response: We have rewritten (toned down) this argument.

As illustrated in the Tables, and indicated by the authors "the difference was [in model fit was] not very large." In fact, the differences have not statistically been tested and appear rather neglegible. Further, as mentioned above, comparisons in terms of predictive performance (discrimination and calibration) are missing and should be included to make formal statements on predictive performance. Finally, to infer on the "preferred" approach, one should also investigate the (relative) impact of the different models, e.g. using decision curve analysis.

Response: Apart from what relates to the comments and responses above, we have now performed decision curve analyses, according to the methodology suggested by Vickers and Elkin (2006). The results, which are summarised in Figure 1 and commented upon in the text, indicate that, when it comes to issues other than predictive accuracy, there is no clear evidence to suggest that one measure of nursing workload should be preferred above the other. Hence net benefit of the models using the OPC/nurse measure is not unequivocally higher than that of the models using the patients/nurse measure.

Comment 4:

It appears that the RAFAELA system is owned by the Association of Finnish Local and Regional Authorities, and that its use is managed by the FCG Finnish Consulting Group Ltd. Although the author does not report any conflict of interest, I still think it would be appropriate to mention the (scientific) relation between the primary author and the RAFAELA system, although this could perhaps simply be done in the introduction. I leave this decision with the editors of BMJ open.

Response: Yes, we have now pointed out this fact in the Declaration of conflict. ‘However, we want to declare that the first author has been involved in developing the RAFAELA system.’

Minor

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

1. Abstract: "To investigate whether the recommended daily workload per nurse (OPC/nurse) as measured by the RAFAELA system correlates with different types of patient safety incidents and with patient mortality, and to compare this interrelation with that based on the standard patients-to-nurse classification." I would indicate that the aim is to assess the magnitude of this correlation (or better, association), rather than its mere presence.

Response: We have decided to keep our aim, but some corections have been made in accordance with all the reviewers' commnets. 'The aim of this observational study was therefore to investigate whether the assumed optimal NWL, as a measure based on the RAFAELA system, correlates with patient safety incidents and patient mortality, using data collected on a daily basis. Also, we want to compare the estimates with those based on the standard patients-to-nurse ratio.'

2. Outcomes: I would indicate that "whether at least one incident, of any type, occurred" within the available follow-up of 365 days (or less where relevant).

Response: We have rewritten this section.

--

Additional responses to requested changes (email from the Editorial Office 18.09.201)
1. Kinnunen, Marina affiliations don't match from main document to scholar one.
This has been changed.

2. Aside from the marked copy, please also provide a clean copy of your manuscript without any highlights or tracked changes and upload it as your 'main document'.
YES, one clean and one with highlights has been uploaded.

3. Please ensure that the Competing Interests and Data sharing statement in your main document and Scholar One submission system are the same.
YES, CHANGES HAS BEEN MADE.

4. Please remove all your figures in your main document and upload each of them separately under file designation 'image'(except tables). NOTE: They can be in TIFF or JPG format and make sure that they have a resolution of at least 300 dpi. Figures in PDF, DOCUMENT, EXCEL and POWER POINT format are not acceptable.
THIS HAS BEEN DONE.

5. Please ensure that the embedded Tables are on editable format.
YES

6. The in text citation for Reference 18 is missing on your main document file. Please amend accordingly.
A MISTAKE, THE REFERENCE 18 HAS BEEN CHANGED TO 17,

7. We have noticed that you have uploaded a file under 'supplementary file'. However, we can't see any citation for this file within the main text. If this file needs to be published as supplementary file, please cite it as 'supplementary file' in the main text. Otherwise, you can change the file designation to 'Supplementary file for editors only'
WE HAVE CITATIONS FOR THIS FILE ON PAGES 7,8, AND 9.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

VERSION 3 – REVIEW

REVIEWER	Walter Sermeus KU Leuven, Belgium
REVIEW RETURNED	03-Nov-2017

GENERAL COMMENTS	It is the third time that the manuscript is presented and that the authors want to have it finished asap. But sending in a total number of 2282 pages is not helpful for the review process. The authors are meeting the comments of the reviewers in this version, what is good but there is always a reservation. They have to be more clear ad straightforward in their changes. I give a few examples: Conflict of interest: The authors declared no potential conflicts or interest.... . However, we want to declare that the first author.... This statement remains double for the first author "is there yes/no a conflict of interest". It would be better to say " LF was one of the developers of the RAFAELA system ... something about the commercial use of the RAFAELA system ... and if she is still involved in the company yes/no ... ; MK is ... (is there any conflict of interest); JS is .... (is there any conflict of interest)" Second example: one of reviewers is making a comment that the RAFAELA System is not calculating the "optimal" NWL. The authors have changed this in the "assumed optimal NWL". This change is not made in the title, in the tables but just a few times in the text. This limited change is not taking this comment seriously. The point is that the staffing level calculated by the RAFAELA system are in line with how the system was calibrated and validated but indeed not optimal. How to explain that the authors are showing that there is a relationship between OPC/nurse and patient mortality and that you can be below the optimum and still patients are dying. It should be recommended to talk about OPC-level instead of optimum/ below and above optimum. The manuscript should win a lot of credibility if this shift would be made across the manuscript (title, tables, text,..). One of the reviewers made the comment that the measurement of patient safety incidents is of concern. They added the reference to the paper by Classen et al. . The authors miss the point. It has been showed in literature that incident reporting is far from reality, because lack of time, lack of confidence in what to report. The evidence is showing that only 1% of adverse events is reported by incident reports. It is quite amazing that they even see some relationships between nurse staffing levels and patient incidents given the unreliability of the measure they are using. I would like to see that they discuss this. I still don't agree with the statement on p.9 (line 11) : "For the same outcome and adjusted model as discussed above, for instance, estimates based on the patients/nurse approach were not statistically significant at the five per cent level, while those based on the OPC/nurse were". I refer to table 2: in the OPC adjusted model: only the impact of "above optimum" on incidents (1.08-1.42) and deaths (1.18-1.73) is significant, while it is on the "below optimum" levels on incidents (0.67-0.93) and patients affected (0.64-0.96). For patient/nurse ratio it is indeed on none of the variables. The text give the idea that OPC/nurse is significant on all dependent variables which is not. Possible explanations are the definition of the split in the above/optimum/below in OPC/nurse that is not equal to the 3 groups split for the patient/nurse ratio. I want to know the effect of the

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	original variable (OPC/nurse or patient to nurse ratio that are both numerical variables. Another explanation is the dependent variable being incident reporting which is more variable than using indicators or other measures. The added value of the RAFAELA system is that it allows to measure patient intensity as a refined measure in the equation. It might be an option to calculate a patient intensity score on top of the patient/nurse ratio score and to evaluate if the explanatory power to patient safety is increased.

REVIEWER	Peter Griffiths University of Southampton, UK
REVIEW RETURNED	06-Nov-2017

GENERAL COMMENTS	In general, the authors have done a good job in responding to review comments. By and large, the strengths and limitations of the research are properly conveyed and a reader would be able to identify the weight that should be given to conclusions based on the results. The one residual issue of some significance relates (still) to the use of the term 'optimal' and conclusions that follow relating to optimality. These should now be easily addressed. I think it is important to do so although if the editors disagree I am happy to defer. I certainly see no reason to review the manuscript again as an informed reader would. I believe, be fully able to judge the appropriate conclusions. I offer the following specific comments on sections of the paper which address this issue and one or two other areas when clarity might be enhanced: [I have reviewed the word file as the pdf supplied ran to over 2800 pages. There are no page numbers on the word file] Abstract: Objective To investigate whether recommended daily workload per nurse… should read as daily workload per nurse? Results There is a typo on the first sentence ? “that” should be “than” but the sentence remains unclear…? (this is replicated in the main results). Introduction New insertion “While certain realities such as economic restraints cannot be entirely disregarded, the use of RAFAELA provides a measurement whereby an optimum situation can be assessed and achieved, with resources properly dedicated to the reduction or elimination of adverse events.” This reads like a marketing claim. I suggest that you add ‘aims to’ after “use of RFAELA” The section beginning “In the Rafaela system….” Could be usefully supported by a reference to where these methods are documented. “We have found only two studies on the relationship between the recommended optimal NWL and patient outcomes….” This is still misleading. Needleman’s study simply looked at what happened when staffing fell below that which was assessed as necessary. It did not consider optimality as defined here. I suggest you modify this

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	sentence “We have found only two studies on the relationship between nursing workload based on assessed requirements for care (as opposed to nurse patient ratios or equivalent measures) and patient outcomes….” Discussion Final para – stray word “one” at the end. Conclusion New insertion: “..by using the recommended optimal NWL…. Nurse managers can optimise the staff resources.” This conclusion does not follow and the issue has been pointed out on previous revisions – given that when staffing is above the system defined “optimal” outcomes are improved it is not clear what the basis for this recommendation is. One could equally (and perhaps more unequivocally) conclude – “by staffing at levels above the recommended optimal…..” . The problem is that none of the findings here actually allow us to determine an optimal solution without introducing other criteria (cost, benefit for example). I suggest this sentence is simply deleted.

REVIEWER	Thomas Debray Julius Center for Health Sciences and Primary Care
REVIEW RETURNED	26-Nov-2017

GENERAL COMMENTS	Thanks for addressing my previous comments. I only have one minor suggestions left. 1. Figure 1-5: Please add the reference lines as recommended by Vickers et al. One line is then based on a 'model' predicting outcome presence (e.g. death) for all patients, and the other one is based on a 'model' predicting "no event" (e.g. survival) for all patients.

VERSION 3 – AUTHOR RESPONSE

Response to comments on bmjopen-2017-016367.R2

We appreciate the additional comments received and have attempted to be highly perceptive and make revisions accordingly. Below follows a point-by-point response to each comment. We have uploaded Table A1 as a Supplementary File for Editors only.

EDITOR:

If possible, please reduce the number of pages you include as supplementary material. 2282 pages is very long for a manuscript submission, and we have concerns that it will lose reader interest.

Response: We have now reduced the supplementary material to approximate one tenth of the original size. We can certainly reduce it further, but with the risk that some information needed to understand the setup of the models might be lost, and that the full results of the regressions consequently may become difficult to understand.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

REWIEVER 2:

It is the third time that the manuscript is presented and that the authors want to have it finished asap. But sending in a total number of 2282 pages is not helpful for the review process.

The authors are meeting the comments of the reviewers in this version, what is good but there is always a reservation. They have to be more clear ad straightforward in their changes. I give a few examples:

Conflict of interest: The authors declared no potential conflicts or interest.... . However, we want to declare that the first author.... This statement remains double for the first author "is there yes/no a conflict of interest". It would be better to say

" LF was one of the developers of the RAFAELA system ... something about the commercial use of the RAFAELA system ... and if she is still involved in the company yes/no ... ; MK is ... (is there any conflict of interest); JS is .... (is there any conflict of interest)"

Second example: one of reviewers is making a comment that the RAFAELA System is not calculating the "optimal" NWL. The authors have changed this in the "assumed optimal NWL". This change is not made in the title, in the tables but just a few times in the text. This limited change is not taking this comment seriously. The point is that the staffing level calculated by the RAFAELA system are in line with how the system was calibrated and validated but indeed not optimal. How to explain that the authors are showing that there is a relationship between OPC/nurse and patient mortality and that you can be below the optimum and still patients are dying. It should be recommended to talk about OPC-level instead of optimum/ below and above optimum. The manuscript should win a lot of credibility if this shift would be made across the manuscript (title, tables, text,..).

Response: The supplementary material has been reduced considerably in size; please see our comment to the Editor above. The conflict of interest statement has been rewritten. Where possible, the use of the term ‘optimal’ has been changed or excluded and the text revised accordingly. However, since this terminology explicitly refers to that used by the RAFAELA classification system, we cannot exclude it entirely from the text and tables without the description and referral becoming too vague or even ambiguous.

One of the reviewers made the comment that the measurement of patient safety incidents is of concern. They added the reference to the paper by Classen et al. The authors miss the point. It has been showed in literature that incident reporting is far from reality, because lack of time, lack of confidence in what to report. The evidence is showing that only 1% of adverse events is reported by incident reports. It is quite amazing that they even see some relationships between nurse staffing levels and patient incidents given the unreliability of the measure they are using. I would like to see that they discuss this.

Response: We agree with these comments, and have now made several clarifications in the text.

I still don't agree with the statement on p.9 (line 11) : "For the same outcome and adjusted model as discussed above, for instance, estimates based on the patients/nurse approach were not statistically significant at the five per cent level, while those based on the OPC/nurse were". I refer to table 2: in the OPC adjusted model: only the impact of "above optimum" on incidents (1.08-1.42) and deaths

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

(1.18-1.73) is significant, while it is on the "below optimum" levels on incidents (0.67-0.93) and patients affected (0.64-0.96).

For patient/nurse ratio it is indeed on none of the variables. The text give the idea that OPC/nurse is significant on all dependent variables which is not.

Response: We are sorry about this confusion and have now made clarifications in the text.

Possible explanations are the definition of the split in the above/optimum/below in OPC/nurse that is not equal to the 3 groups split for the patient/nurse ratio. I want to know the effect of the original variable (OPC/nurse or patient to nurse ratio that are both numerical variables. Another explanation is the dependent variable being incident reporting which is more variable than using indicators or other measures.

The added value of the RAFAELA system is that it allows to measure patient intensity as a refined measure in the equation. It might be an option to calculate a patient intensity score on top of the patient/nurse ratio score and to evaluate if the explanatory power to patient safety is increased.

Response: As requested, we now provide an additional table, referred to as Table A1 (uploaded as a Supplementary File for Editors only). This table is constructed in a similar manner as the part of Table 2 that refers to adjusted models, but has used continuous measures of OPC/nurse and patients/nurse. Additionally, we estimate models that include both these continuous measures. Conclusions from these results are similar as those based on the categorical variables (in Table 2). OPC/nurse provides better model fit than patients/nurse for all outcomes, and in terms of all measures of fit (log likelihood, Aikake and R Square). We also see that the inclusion of OPC/nurse in addition to patients/nurse slightly improves the model fit. We are somewhat reluctant to include this new table into the final manuscript since it might give the impression that patients/nurse has a stronger effect (which is simply because the scale used is different from that of OPC/nurse). Furthermore, since OPC/nurse and patients/nurse are highly correlated and, thus, should be used as substitutes, the estimated effect of patients/nurse is reduced while the estimated effect of OPC/nurse is attenuated when both are included into the same model. Upon request we can of course include the table into the final document.

REVIEWER 1:

The one residual issue of some significance relates (still) to the use of the term 'optimal' and conclusions that follow relating to optimality. These should now be easily addressed. I think it is important to do so although if the editors disagree I am happy to defer.

Response: Please see our response to Reviewer 2 above.

Abstract: Objective. To investigate whether recommended daily workload per nurse… should read as daily workload per nurse?

Results. There is a typo on the first sentence ? “that” should be “than” but the sentence remains unclear…? (this is replicated in the main results).

Response: The sentences have been revised.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

Introduction. New insertion “While certain realities such as economic restraints cannot be entirely disregarded, the use of RAFAELA provides a measurement whereby an optimum situation can be assessed and achieved, with resources properly dedicated to the reduction or elimination of adverse events.” This reads like a marketing claim. I suggest that you add ‘aims to’ after “use of RFAELA”

Response: The sentence has been revised.

The section beginning “In the Rafaela system….” Could be usefully supported by a reference to where these methods are documented.

Response: References 16, 19, 21 have been added.

“We have found only two studies on the relationship between the recommended optimal NWL and patient outcomes….” This is still misleading. Needleman’s study simply looked at what happened when staffing fell below that which was assessed as necessary. It did not consider optimality as defined here. I suggest you modify this sentence “We have found only two studies on the relationship between nursing workload based on assessed requirements for care (as opposed to nurse patient ratios or equivalent measures) and patient outcomes….”

Response: The sentence has been rewritten.

Discussion. Final para – stray word “one” at the end.

Response: Corrected.

Conclusion. New insertion: “..by using the recommended optimal NWL…. Nurse managers can optimise the staff resources.”

This conclusion does not follow and the issue has been pointed out on previous revisions – given that when staffing is above the system defined “optimal” outcomes are improved it is not clear what the basis for this recommendation is. One could equally (and perhaps more unequivocally) conclude – “by staffing at levels above the recommended optimal…..” . The problem is that none of the findings here actually allow us to determine an optimal solution without introducing other criteria (cost, benefit for example). I suggest this sentence is simply deleted.

Response: The sentence has been deleted.

REVIEWER 3:

Figure 1-5: Please add the reference lines as recommended by Vickers et al. One line is then based on a 'model' predicting outcome presence (e.g. death) for all patients, and the other one is based on a 'model' predicting "no event" (e.g. survival) for all patients.

Response: We have now inserted the two lines in each figure.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

VERSION 4 – REVIEW

REVIEWER	Walter Sermeus KU Leuven Institute for Healthcare Policy
REVIEW RETURNED	23-Jan-2018

GENERAL COMMENTS	I would have hoped to be able to fully accept the manuscript, but still some minor issues are left that should be solved. Firstly in the abstract, the patient mortality figures should be reversed. "corresponding estimates for patient mortality are 1.43 () and 0.79 ()" The results on p8 are still misleading. Confidence intervals are given for patient safety incidents (CI: 1.13-1.45) and patient mortality (CI:1.19-1.69), but not for the 3 other outcome measures (patients affected, harm to patient, >1 incident). For these measures only the average ratio is given (1.13, 1.16, 1.25). But these ratios are NOT significant. It is also the case for the below optimum staffing as well for the unadjusted and adjusted models. This makes the result section very misleading and should be rewritten to make this clear. The authors added table 1 in a more readable format than in a previous version. But at the same time this raises many questions. We have limited information about the 36 nursing units that were included in the study: are they general med-surgical units, intensive care units, specialised units, oncology units? It might be useful to have this type of information as we can see that the optimal load (upper/low bounds) are varying from 8,90/12 up to 25.6/42.1. What is explaining this difference? In the table we see that the OPC/nurse is for some units above the upper bound (e.g. unit D4) and some other units below the lower bound (e.g. B11). How is this information taken into the analysis? I assume that the level of analysis is a unit at 1 day. During that day the OPC/nurse can be higher/lower than the optimal bound. How is this measured: binary such as 0,1, -1, number by calculating the difference? It should be clearly reported in the manuscript how the variable was operationalised in the model. The patients/nurse is quite different operationalised than in other research that is referenced. Usually, it is a number of patients per nurses during one shift. For 24h, an average is calculated. The numbers vary for med-surg wards from 5 patients per nurse in USA up to 17 patients per nurse in some other countries such as Spain, Belgium, Greece etc. They can be even lower on ICU (1 or 2 patients per nurse). In this study, the number is varying from 0,49 (Unit B11) to 2,85 (unit A3). So I don't understand the real metric. Please explain concretely how the measure was operationalised. An alternative for patient-to-nurse ratios in the literature are "nursing hours per patient day" (see recent review by Driscoll et al., European Journal of Cardiovascular Nursing, 2017). For both measures, the critique of the authors that they don't take nursing intensity into account, is valid. But the used measures should be comparable to allow comparison and critique. The report of incidents, patients affected, harm to patient and >1 incident is not clear. Based on the data, I presume that there is a connection between the first 3 indicators. If there is an incident, the patient might be affected or not, and when affected it might cause harm or not. In the table, it is unclear what it in the nominator and denominator (patients, days,...). For the indicator >1 incident, it also unclear (level of patients, days). For most indicators, I'm not sure

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

	what the distribution is, but it is unlikely that it a normal distribution. We expect more patients/days with low prevalence of incidents and less patients/days with high numbers. When the distribution is not normal, it might be that the regression model might not be longer valid. The last outcome variable "death" is more troublesome. The variable is retrieved from the mortality register of each hospital. That's good. But we are looking here for mortality that is not caused by the patient condition but by latent conditions such as nurse staffing and workload levels. We don't have any information about the patient characteristics (age, sex, medical diagnosis, co-morbidities,...) that are showing that patients have higher/not higher risks of mortality. There are some indexes such as Charlson / elixhauser that are used for adjusting the mortality rates for these types of risks. This is not done here. I see in table 1 that mortality rates are varying from 0,00 up to 0.18 (Unit D8). In 10 out of 36 units, no patients have died. Most of the numbers are quite low with a few outliers: D7:018; B5:0.16; C2:0.15. It might be interesting to know if these units would be ICUs, oncology units,... with a normal expected higher rate of mortality because of patient's condition. What is the impact of the outliers. A sensitivity analysis would help. Again, a clear description of this indicator (nominator; denominator) would also help. The distribution needs to be checked to see if the regression model is appropriate. The decision-analytic analysis was not clear to me (p.7). It is unclear what exact analysis has been performed, what the added value is of the performed analysis, if the results (p.9, line 40-55) are significant. I also don't see the added value of Figure 1 to 5. To conclude, I'm positive about the general aim of the study to explore if measuring workload/nursing intensity through a patient classification system gives more precise results in the relation between nurse staffing and patient outcomes. But I'm still not sure about the data you have available and the validity of the relations showed sofar. Explaining table 1 and the related variables is key to me to understand the work that is performed by the authors.

VERSION 4 – AUTHOR RESPONSE

Response to comments on bmjopen-2017-016367.R3

To the editor

We have now made what we hope are the final changes to the manuscript. First, we want to emphasize that the issues raised by the referee are not in fact highly critical points, since the major items seem to be based on misunderstandings of what we are doing from a methodological point of view. To adhere to the questions, we now describe the data structure and methods even more carefully. With regard to the comment on the decision-analytic analysis, we would like to stress that this part was included because it was requested by another reviewer (Reviewer 3), who obviously was pleased with our revision of the manuscript on this point. The main conclusion based on the decision-analytic analysis is that, we cannot unambiguously decide which measure is to prefer in

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

terms of net benefit values. Below follows a point-to-point response to the remaining comments Reviewer 2.

Response to Reviewer 2

Firstly in the abstract, the patient mortality figures should be reversed. "corresponding estimates for patient mortality are 1.43 () and 0.79 ()"

Response: We apologize for this mistake. This has now been corrected.

The results on p8 are still misleading. Confidence intervals are given for patient safety incidents (CI: 1.13-1.45) and patient mortality (CI:1.19-1.69), but not for the 3 other outcome measures (patients affected, harm to patient, >1 incident). For these measures only the average ratio is given (1.13, 1.16, 1.25). But these ratios are NOT significant. It is also the case for the below optimum staffing as well for the unadjusted and adjusted models. This makes the result section very misleading and should be rewritten to make this clear.

Response: We have now inserted the requested confidence intervals into the text.

The authors added table 1 in a more readable format than in a previous version. But at the same time this raises many questions. We have limited information about the 36 nursing units that were included in the study: are they general med-surgical units, intensive care units, specialized units, oncology units? It might be useful to have this type of information as we can see that the optimal load (upper/low bounds) are varying from 8,90/12 up to 25.6/42.1. What is explaining this difference?

In the table we see that the OPC/nurse is for some units above the upper bound (e.g. unit D4) and some other units below the lower bound (e.g. B11). How is this information taken into the analysis? I assume that the level of analysis is a unit at 1 day. During that day the OPC/nurse can be higher/lower than the optimal bound. How is this measured: binary such as 0,1, -1, number by calculating the difference? It should be clearly reported in the manuscript how the variable was operationalized in the model.

Response: The outcome per day, for each type of incident and for mortality, is consistently coded as 0 or 1. Hence, if there was an event in one day, say patient affected, the outcome variable is coded as 1, otherwise zero. The binary outcome, and the fact that the data are at the daily level, are also the reason to why we use logistic regression models, which easily can handle non-normal distributions. This is now pointed out in the text. We have added to Table 1 the type of specialty, and also added this text on page 4. ‘The following specialties were included in the data material: internal medicine (8 units), surgical (8 units), pediatrics (5 units), gynecology (4 units), maternity (2 units), neurology (2 units), orthopedics (2 units), oncology (1 unit), rehabilitation (1 unit), lung (1 unit), and otology (1 unit).’

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

The patients/nurse is quite different operationalized than in other research that is referenced. Usually, it is a number of patients per nurses during one shift. For 24h, an average is calculated. The numbers vary for med-surg wards from 5 patients per nurse in USA up to 17 patients per nurse in some other countries such as Spain, Belgium, Greece etc. They can be even lower on ICU (1 or 2 patients per nurse). In this study, the number is varying from 0,49 (Unit B11) to 2,85 (unit A3). So I don't understand the real metric. Please explain concretely how the measure was operationalized. An alternative for patient-to-nurse ratios in the literature are "nursing hours per patient day" (see recent review by Driscoll et al., European Journal of Cardiovascular Nursing, 2017). For both measures, the critique of the authors that they don't take nursing intensity into account, is valid. But the used measures should be comparable to allow comparison and critique.

Response: Please see the above response. We want to clarify how the classification of patients' nursing intensity is done by the OPC. The daily nursing intensity of each unit is based on daily classification of all patients on the unit. The registered nurses classify all the patients' nursing intensity by the OPC and every day (see page 5). The nurses' workload is calculated by dividing the daily total amount of nursing intensity points on the unit, e.g. 350, with the number of nurses who take care of patients, e.g. 12, during the same 24 hours. In this example, the patient-related NWL will then be 29.2 OPC points per nurse. Therefore, the measure 'patients per day' varies based on the exactly numbers of classified patients. As we all know, the number of patients varies from day to day and clear differences can be seen between different specialties. The example B11 is a pediatric unit, and the number of nurses is usually very high on Finnish pediatric units and A3 was a gynecology unit.

The report of incidents, patients affected, harm to patient and >1 incident is not clear. Based on the data, I presume that there is a connection between the first 3 indicators. If there is an incident, the patient might be affected or not, and when affected it might cause harm or not. In the table, it is unclear what it in the nominator and denominator (patients, days,...). For the indicator >1 incident, it also unclear (level of patients, days). For most indicators, I'm not sure what the distribution is, but it is unlikely that it a normal distribution. We expect more patients/days with low prevalence of incidents and less patients/days with high numbers. When the distribution is not normal, it might be that the regression model might not be longer valid.

Response: As stated in the third response above, and now also pointed out in the text, the outcome per day, for each type of incident and for mortality, is consistently coded as 0 or 1. The logistic regression models can easily handle non-normal distributions, which is in fact a major reason behind the use of logistic regression models in general. We now also mention in the text that the different types of incidents are coded as to roughly reflect the severity of an event.

The last outcome variable "death" is more troublesome. The variable is retrieved from the mortality register of each hospital. That's good. But we are looking here for mortality that is not caused by the patient condition but by latent conditions such as nurse staffing and workload levels. We don't have any information about the patient characteristics (age, sex, medical diagnosis, co-morbidities,...) that are showing that patients have higher/not higher risks of mortality. There are some indexes such as Charlson / elixhauser that are used for adjusting the mortality rates for these types of risks. This is not done here. I see in table 1 that mortality rates are varying from 0,00 up to 0.18 (Unit D8). In 10 out of 36 units, no patients have died. Most of the numbers are quite low with a few outliers: D7:018; B5:0.16; C2:0.15. It might be interesting to know if these units would be ICUs, oncology units,... with a normal expected higher rate of mortality because of patient's condition. What is the impact of the outliers. A sensitivity analysis would help. Again, a clear description of this indicator (nominator;

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.

denominator) would also help. The distribution needs to be checked to see if the regression model is appropriate.

Response: Since we find it is essential to provide estimates that refer to the same data, that is the same wards, we include wards with zero number of deaths; otherwise the results would not be fully comparable across columns. Since we in the adjusted models control for ward-specific effects (as pointed out in the text), and we can see that estimates from unadjusted and adjusted models do not differ largely, excluding wards with zero deaths (the so-called outliers) does nevertheless not impact on the mortality results reported to any noteworthy degree. This is now briefly pointed out in the text. Excluding wards with zero deaths from analyses of the four types of patient safety incidents would additionally mean that we would have a selected sample and not a complete data set.

The decision-analytic analysis was not clear to me (p.7). It is unclear what exact analysis has been performed, what the added value is of the performed analysis, if the results (p.9, line 40-55) are significant. I also don't see the added value of Figure 1 to 5.

Response: The approach undertaken for these analyses was suggested by Reviewer 3, and follow the methodology described by Vickers et al (which is in the reference list). For our data, the primary conclusion is that it is not evident which measure is to be preferred when it comes to decision making. Figures 1 to 5 summarizes net benefit values calculated based on the models estimated for each type of patient safety incident and patient mortality, respectively. Models based on the OPC/nurse measure and the patients/nurse measure are to be compared by looking at the net benefit values. The one with higher net benefit values is to be preferred above the other. As shown by the figures, there is no clear discrepancy. For some threshold probabilities, the OPC/nurse measure lies above the patients/nurse measure, while for others, the situation is the opposite. In addition, for each event (type of incident and mortality), the two curves are essentially overlapping, and in most instances the difference in net benefit values is rather modest. In terms of the magnitude of the benefit for patients, it is consequently not evident which measure of nursing workload is to be preferred.

To conclude, I'm positive about the general aim of the study to explore if measuring workload/nursing intensity through a patient classification system gives more precise results in the relation between nurse staffing and patient outcomes. But I'm still not sure about the data you have available and the validity of the relations showed sofar. Explaining table 1 and the related variables is key to me to understand the work that is performed by the authors.

Response: In the above comments and the revised version of the text, we now explain the data structure and methodology better, and as argued, we see no reasons to suspect any serious validity problems that would affect the conclusions drawn.

VERSION 5 – REVIEW

REVIEWER	Walter Sermeus Katholieke Universiteit Leuven, Belgium
REVIEW RETURNED	04-Feb-2018

GENERAL COMMENTS	The authors addressed most of the comments raised by the reviewers.

BMJ Open: first published as 10.1136/bmjopen-2017-016367 on 24 April 2018. Downloaded from http://bmjopen.bmj.com/ on July 13, 2024 by guest. Protected by copyright.