Reviewer’s report

Title: Identifying optimal indicators and purposes of population segmentation through engagement of key stakeholders: a qualitative study

Version: 0 Date: 11 Sep 2019

Reviewer: Richard Wood

Reviewer's report:

This paper is essentially of two parts: the first concerning the assessment of "indicators" used to describe/define segments of the population, in accordance to some desired "outcomes", or purposes (the second part). The authors conduct these assessments by means of qualitative surveying techniques of a range of experts from across government and healthcare authorities. Finally, the authors attempt to define a "hybrid" approach to segmentation involving a combination of data-driven and expert-defined components.

Firstly, I am pleased to see a paper address the issue of the purpose of population segmentation. I believe there is a gap in the literature in this question, and that a discussion of its varied utility from a diverse group of system stakeholders is welcome. I also believe that a comprehensive assessment of the "indicators" is useful content which, again, has not been broached. This will help practitioners and researchers in this field.

I do however have some concerns about this manuscript in its current state, as follows.

Determination of "domains"

It is not clear how the 8 indicator and 6 outcome domains have been determined. Same too for the "categories". Were these prescribed from the outset, or were they somehow derived from the stakeholder interviews. The authors state (page 5, line 7-9) that "This allowed for subthemes to be derived from the data instead of pre-existing ideas from existing literature or frameworks". Is this the case? I think there at the very least needs to be more information in the "Data analysis" sub-section of "Methods" that details specifically how these themes/sub-themes described relates to the domains/categories within the "Results" section. If the domain/categories are indeed pre-defined then where have they come from? If that is the case then there will need to be something in the Discussion section to explain further the limitations (i.e. how were they validated? Did you and stakeholders interviewed agree with all of them? How did you know that none others were missing?). In the Discussion (page 7, line 18) it mentions "Our findings revealed eight broad domains…" - from this it appears to suggest that these were not pre-defined. Again, this leaves me a little confused, and some clarification is needed.

Use of word "Outcome/s"

While "indicators" could do with better-defining (i.e. its first mention on page 3, line 48 should come with a definition), I think the use of the word "outcomes" needs to be more closely looked at. What you mean by this is "purpose" (as you state page 3, line 51-53) - i.e. the objective of segmentation whether it be to understand differences in healthcare demand, need, cost - so why not just call it this. The use of
the word outcome is confusing since one of the other "outcomes" is, as you say (page 8, line 29), to study health outcomes, i.e. in the clinical, patient-related sense as opposed to the totally different meaning as an "objective" or "purpose".

Hybrid model

While there is some reference in the introduction to two types of segmentation approach including expert and data-driven constructions (page 3, line 30), I feel that mention of a "hybrid model" (page 9, line 7) is not within the narrative of this paper, and does not logically flow from the findings presented. Combining expert and data within a segmentation "model" and tailoring interventions accordingly is nothing new and I would urge the authors to study and cite the relevant literature, including grey literature, in which making use of both elements is considered. I do not necessarily think that reference to this "hybrid model" firmly requires removing from the "Discussion" (although removal of it I don't think would detract from the value of the contribution of this manuscript), but perhaps more thought should be given to its point and placement, as I am unsure how this logically follows from what is quite an objective review of indicators/outcomes. If it is to stay then its introduction (page 9, lines 1-5) need to be more clear and convincing: the first sentence is a tenuous link-in as I believe from the literature over many years it should be clear that segmenting on one variable type is insufficient (even Bridges to Health implicitly includes many) - i.e. it is not just "our findings" that "reveal" this - maybe change "reveal to "support" and cite the other studies that include other variables. And the second sentence (lines 2-5) would need to be developed further (e.g. "often omitted" - I don't understand this, "omitted" from what?). Finally, perhaps the most crucial complexity in segmentation is given only a cursory mention: "Second, individuals are grouped into homogenous clusters based on the selected indicators of health" (line 11) - determining these rules/criteria is far from a trivial undertaking.

Paper structure

While the paper reads well I believe the structure could be slightly improved. Perhaps breaking down the "Results" section to make it clear the divide between the two parts/areas of the paper (indicators/outcomes). This too could be considered in the "Discussion" where it may benefit from an "Indicators" sub-heading at page 7, line 17, an "Outcomes/(Purpose)" sub-heading at page 8, line 25, and a "Strengths and limitations" sub-heading at page 9, line 25.

Literature review

There is no substantial literature review section in this paper. This is to expectation since there is not much previous effort in this field (hence part of the reason why I think this could be a valuable submission). However, I believe the recent findings of Wood et al, 2019 "A comparison of population segmentation methods" would be worthy to mention and discuss at some points. For instance, in explaining the range of data-driven methods (page 3, line 39) and also on the pros/cons of data vs expert methods on the grounds of discrimination, segment interpretability, etc (perhaps in the "Discussion"). The range of variables/"indicators" considered in this work could also be raised in the discussion on page 7, lines 31-34). And discussion on the utility of age could be raised where, in comparing methods side-by-side, age has (itself) not been found to be that informative (chronic conditions is the main driver).
Direct vs indirect benefit

It would benefit from being a bit clearer over what is a direct output of population segmentation and what are many steps away. For instance, on page 4, lines 30-32 it is mentioned "[an] ageing population would entail a significant increase in healthcare needs. Population segmentation can be utilised to anticipate and effectively meet these needs". While I don't dispute this, segmentation is only an ingredient to this, one would also need to be able to temporally forecast drift in and out of each segment and arguably this is the more difficult ask. I also refer to page 6, lines 56-58: here, "identifying high healthcare user segments" is an immediate output from segmentation, whereas "addressing their care needs" is many steps beyond.

Other points

A few times it is mentioned "as population…" (e.g. page 3, line 7). Would this not be "as a population…".

Page 3, line 18 - "Segmentation is a concept to group patients and healthy people into segments with relatively similar needs or characteristics". For most of the time in my experience, yes, but given it could be used for many other things (e.g. to segment based on outcomes, experience, etc) and that you are in your paper reviewing these "purposes" would it not make sense to be a little more vague here, i.e. "segmentation is a concept typically used to group patients and…".

Page 3, line 55 - "..suffer from being idiosyncratic and piecemeal" - please explain why being bespoke is necessarily a bad thing; could it not also be valuable to be bespoke?

Page 3, line 56 - "..existing expert-defined or data-driven frameworks may not necessarily encapsulate the characteristics and needs of a heterogeneous population" - why not? Needs some further explanation.

Page 5, line 24 - explain what is meant by "data saturation".

Page 5, line 44 - "mentioned the categories" - in what respect, in a positive way, in a negative way? Or just a mere mention. This could affect legitimacy of results surely? Needs more explanation.

Page 5, line 48 - is "race" a legal/appropriate variable that can be used. May need to be careful as to what end, or at least flag that this could be a controversial indicator.

Page 6, lines 45+ - it might be interesting to break down (or at least discuss) desired "purposes" of segmentation by the type of stakeholder (i.e. whether govt, clinician, social worker).

Page 6, line 4 - "...segmentation could help health authorities to identify the poorest strata in the society". If you have "financial status" then you can surely identify these anyway, without having to "segment" per se.

Page 8, line 12-23 - "financial status" - very interesting, maybe require some discussion on i) how this would be useful in a private vs public health economy, and ii) how this information may be acquired (in the UK this info would not be readily available, or possible governance-wise to link at record level).

Page 8, line 29 - "Health outcomes across the care pathway" - also what about off the care pathway, i.e.
those that need care but aren't receiving it need to be considered too.

Discussion, on indicators - the indicators are discussed individually but little mention is made of the need to combine them in a credible segmentation approach. I feel this is probably beyond the scope of this study but some mention could be made (i.e. further / in replace of that on page 9, line 11).

**Level of interest**
Please indicate how interesting you found the manuscript:

An article of importance in its field

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

None

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments
which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal