Reviewer's report

Title: Data validity of routine electronic hospital admissions data with Scottish Stroke Care Audit (SSCA) data for identifying stroke in the Scottish population.

Version: 2 Date: 7 August 2015

Reviewer: Seo Young Park

Reviewer's report:

1. The title is poorly written. It is redundant ('data' was used 3 times) and does not represent what this manuscript is about. Consider:
   Validity of routine electronic hospital admissions data against SSCA data in identifying stroke in the Scottish population
   Or
   Agreement between routine electronic hospital admissions data and SSCA data in identifying stroke in the Scottish population

2. The motivation or background of this research is not really compelling. Why is it so important to identifying correct number of stroke cases in general? If the limitation of administrative hospital discharge coding data is known so well, why not just use the SSCA data? Is SMR01 data cheaper to obtain?

3. There is a section named 'reference population' under method, which is not appropriate. SSCA data are by no means 'population'. It is just another dataset. I suggest 'reference data'.

4. This paper says SSCA data or SMR01 data indentified XX% of stroke cases, but how do they know the denominator, which is the true number of stroke cases? Are they assuming all stroke cases were included either in SSCA and SMR01? That is probably unrealistic and untestable assumption. Given the data, all they can say is that SSCA data or SMR01 data indentified XX% or less of stroke cases, because there must be stroke cases that were not reported anywhere.

5. The numbers in the results section do not make sense, or I am not understanding them correctly because of poor explanation. For example, the authors said there were 22416 entries in SSCA and 22200 entries in SMR01, and when those two datasets were combined, there were 22601 entries. This can happen only when these two datasets have considerable overlap, but the next sentence says only 55.5% (n=12552) records were found in both datasets. How come? Diagrams for # of cases in each dataset will help.

6. In the discussion, the authors said the % of strokes identified in this study was an increase from previous linkage data. It would be great if there's any discussion why it increased.
7. The last sentence in the discussion section has nothing to do with this research.

8. Overall the quality of writing is poor. There are numerous errors and grammatical mistakes.

**Level of interest:** An article of insufficient interest to warrant publication in a scientific/medical journal

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I have nothing to disclose. I have no competing interests.