Reviewer’s report

Title: Developing an algorithm to identify people with Chronic Obstructive Pulmonary Disease (COPD) using administrative data

Version: 1 Date: 6 February 2012

Reviewer: Ane Johannessen

Reviewer’s report:

The authors develop an algorithm for identifying COPD subjects in general practices, based on data on hospital admissions, medication use and lung function testing. Identification of COPD subjects is useful to help target patients in need of proactive and planned care. The best fitted algorithm is shown to identify between 30% and 97% of COPD patients, with increasing accuracy in older patients.

The manuscript is generally well-written, and the Tables and Figures are beautiful.

The topic is interesting and relevant. However, there are some considerations that should be addressed by the authors:

Discretionary Revisions:
1. There seems to be three different study populations in this paper:
   - Seven practices with 26 GPs in Aarhus identified all patients in their practices with a known diagnosis of COPD. N = 266.
   - Nine algorithms were evaluated, and the chosen one was used to identify patients with possible COPD in five different practices with 17 GPs in the Central Denmark region. N = 532.
   - Patients identified by the chosen algorithm in the Ringkoebing-Skjern and Ikast-Brande municipalities were asked to verify their diagnosis. N = 1984.

   It is not always clear which study population is being referred to, especially in the Discussion, and it would improve the manuscript if the authors could be clearer about this. Perhaps the different populations could be given specific names separating them from each other?

Major compulsory revisions:
2. Methods and material, the Analysis paragraph: "Hansen et al. found a 9% overall prevalence for COPD in the population aged 45 - 85 and this was used as the prevalence measure in order to identify patients with COPD already being treated in the healthcare system." In the background paragraph 3 it says that 14.3 % live with COPD in Denmark. Why is this prevalence rate not used instead? This should be elaborated upon and explained in the Discussion section.
3. Discussion, last paragraph: "The long-term prognosis of COPD may improve with early interventions and the earlier the patient is diagnosed the better the health gains (26). Organised efforts aimed at early identification for early and proactive intervention for COPD are especially important as COPD is increasingly recognised as a disease state which in principle is preventable and treatable (27)." This discussion point does not seem relevant for the present study and should probably be excluded. Since the algorithm includes hospitalization data, medication use data and spirometry registrations, it cannot be said to identify COPD patients early. As stated elsewhere in the Discussion page 9, first paragraph, the algorithm identifies "the sickest patients".

4. The relatively low identification in younger groups might be explained by the inclusion of many patients with asthma. This means that the algorithm identified groups might better be classified as the population with chronic lung diseases (COPD and asthma). It would improve the manuscript if the authors could focus even more on this in the Discussion. What are the implications? What does this mean for the identification of COPD patients? What can be done in the future to even better identify COPD patients?

Minor comments:

5. Material and methods, line 3 in the GP verification paragraph: "The GPs were asked to refer to the Gold Standard where the diagnosis of COPD is confirmed by means of a spirometry which shows a lung function < 70" should be "The GPs were asked to refer to the GOLD guidelines where the diagnosis of COPD is confirmed by means of a post-bronchodilator spirometry which shows a forced expiratory volume in one second (FEV1) / forced vital capacity (FVC) < 70."

6. In the Discussion, paragraph 5, it is stated: "A very high specificity of the algorithm is one of its strengths. This ensures that less healthy people will be suspected of having COPD unnecessarily." Is this a legitimate statement? Do we know anything about the COPD prevalence in all those who are NOT identified by the algorithm – was it not developed solely based on COPD subjects?

7. Discussion, paragraph 6: "a Gold Standard for COPD" should be replaced with "the GOLD guidelines for COPD".

8. Discussion, paragraph 6: "This could both mean that less and that more patients could have been identified." Could the authors elaborate on this? Next sentence: "Properly more considering the fact that suggestions have be made that the overall prevalence of people with COPD is 14.2%." should be "Probably more considering the fact that suggestions have be made that the overall prevalence of people with COPD is 14.2%.

9. Discussion, paragraph 6: "...their lung disease Thus..." The authors should include a punctuation mark between disease and Thus.

10. Page 13, list of abbreviations: COPD is listed as "Chronic Obstructive Disease", this should be corrected to "Chronic Obstructive Pulmonary Disease".
SD is listed as "Standard Derivation", this should be corrected to "Standard Deviation".

11. Table 3: It is not quite clear what the lines represent here – are they the 10-year age groups? I suggest adding an explanatory column at the far left to state the categories for each line.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

I declare that I have no competing interests