Reviewer's report

Title: Hepatitis C prevalence in Denmark - a capture-recapture estimate based on multiple national registers

Version: 1 Date: 26 March 2012

Reviewer: Ruth King

Reviewer's report:

Major compulsory revisions

1. Methods, Data sources: DANVIR:
   a. why do 4 of the 18 Danish laboratories not contribute to this database? It is also stated that this database is estimated to cover 85% of the Danish population – what is the “missing” 15% - is this regional?
   b. “we included all subjects who had a positive HCV-RNA test not followed by a negative HCV-RNA” – In the results section (sensitivity analysis) it states that “We excluded 749 patients from the laboratory register as they were initially HCVRNA positive but later became negative. However, 564 of these were registered with chronic hepatitis C in one or more of the other registers.” Please clarify if these individuals are only removed from the DANVIR register but may still be listed on the other registers (also see below point).

2. Methods, Data sources above Drug Treatment Register “Except the laboratory register it was not possible to exclude patients who cleared the infection from the source registers” – what is the “clearance” rate of chronic HCV? Could this lead to a potential source of bias (such potential issues are discussed in the Discussion) – it might be useful to add something to the current point of different classification of chronic HCV between the sources, this related issue.

3. Methods, two step procedure – Step 1
   a. “The selection of the best fitting models was based on the Schwartz and Akaike information criterion” – more statistical detail is needed here, including the set of possible log-linear models that are fitted to the data, the definition of “best fitting model” (was this solely on a single information criterion?) and why two information criterion appear to be used (and what the procedure is in the case of discrepancies between the information criterion used for different datasets).
   b. “stratified by gender, three age groups, five geographical regions and time period” – what are the age groups and geographical regions, and what is the rationale for these stratifications. In addition how do these stratifications compare with the 85% of the Danish population assumed to be covered by the DANVIR source? Note that taking this number of stratifications leads to relatively “sparse” contingency tables (a total of 60 strata leading to an average of only 116 individuals per contingency table with 16 cells per table).
   c. “Confidence intervals for the total estimate were derived from boot-strap
analysis of 1000 samples” – please clarify whether the bootstrap algorithm used fits only the best fitting model.

4. Methods, two step procedure – Step 2
a. “Assuming the same prevalence among the non-tested we calculated the total number of hepatitis C infected in the drug treatment register” – is this assumption reasonable? If individuals are showing some symptoms of HCV are they more likely to be tested for HCV than others who do not (in other words being tested is NOT independent of having HCV). If this is the case, the proportion of individuals with HCV who are tested will be higher than the proportion of individuals with HCV who are not tested.
b. “Assuming the same proportion of diagnosed hepatitis C infection outside the treatment register we calculated the total prevalence of hepatitis C” – again is this assumption reasonable? Is the proportion of diagnosed HCV drug users identified on the drug treatment register likely to be the same as the proportion of diagnosed HCV individuals?
c. This two-step procedure, in terms of applying a multiplicative factor to the estimated diagnosed number of chronic HCV individuals via the capture-recapture data analysis to obtain an estimate of the total chronic HCV population fails to incorporate any uncertainty with regard to this estimated multiplicative factor.

5. Results “The regional prevalence ranged from 0.15% in the North region to 0.28% in Copenhagen and the capital region represented 40% of all diagnosed cases – it would be useful to include these regional prevalence estimates (and population sizes for each region) in Table 2, so that it is easier to compare the different regions.

6. Results “Estimate of the undiagnosed population with chronic hepatitis C” – “In the laboratory register among all Danish patients with a positive anti-HCV and a test result for HCV-RNA, 62.2% (3,999/6,431) were positive” – where does this 3,999 come from? In Table 1 this figure appears to be 3960 (also see point 11 below regarding consistency of numbers).

7. Results “Estimate of the undiagnosed population with chronic hepatitis C” – “We assumed the same diagnostic coverage (54.3%) among hepatitis C patients outside the drug treatment register” – is this realistic? Will the probability of diagnosis be the same for those identified on the drug treatment register and the rest of the HCV population? (See point 4c regarding treating this estimate as known without error in applying the multiplicative factor for undiagnosed HCV infections).

8. Results “Sensitivity analysis” – “A simple model... gave an estimate of 7,012” – however this appears a little “unfair” as a comparison and in advance it would be predicted that the model will perform poorly, as it assumes independence between the sources (and removes strata). Perhaps a better sensitivity analysis would be to remove only the strata, and consider a range of log-linear models (as before), allowing for interactions between the sources. It is also not surprising
that considering only 2 sources typically underestimates the population size. For the use of three sources, it appears independence is again assumed between the sources – why not allow possible interactions? I do not see the point of removing the 749 patients from the simple analyses and rerunning – why not do this sensitivity analysis for the full data analysis (stratified and allowing log-linear interactions).

9. Results – it might be useful to summarise what could be regarded as the most important interactions between the sources for the different stratified analyses in terms of the interactions being identified in the majority of cases (e.g. “80% of the analyses across the strata identified an interaction between sources A and B; furthermore when the interaction was identified, it was positive in nature”).

10. Discussion para 3 starting “The assumption of independence…” – this whole paragraph is unclear.

11. Table 1 and Figure 1 – I cannot “marry” the numbers together provided in these. Using Figure 1, I obtain estimates of the number of individuals identified by each source (in order of Table 1) as 6191, 3225, 4834, 3100 (compared to 3960, 2890, 4484, 3065).

12. Discussion final sentence “Balancing the above stated possible bias and uncertainty we estimate that the total population with chronic hepatitis C is probably in the range 15,000 – 21,000 (0.35%-0.48%).” This seems a little ad-hoc to conclude the discussion with.

Minor essential revisions

1. Abstract, Results, “We estimated that 46% of the hepatitis C infected individual had not been diagnosed” – wording is poor – rephrase.

2. Abstract, Results “16,907 (“ – one of these brackets is not closed. Also “ 18,216), - 0.39%” – use of hyphen is unclear (it could look like a minus sign), perhaps rephrase to something along the lines of “18,216), this corresponds to 0.39%”.

3. Abstract, Conclusions “Half of the patients with chronic hepatitis C in Denmark have been identified” – the estimate is 6935/16907 (=41%) – which to me is less considerably than half.

4. Methods, Data sources: DANVIR “177.453” should read “177,453” (i.e. comma instead of a full stop).

5. Methods, Data sources: Communicable Diseases Register “The register is estimated to cover 35-40% of individuals diagnosed with hepatitis C” – please clarify – is this acute or chronic HCV or simply chronic HCV?

6. Results “Estimate of the undiagnosed population with chronic hepatitis C” – “16,906 (“ – one of these brackets is not closed.

7. Results “Sensitivity analysis” – “HCVRNA” -> “HCV-RNA”; and “age gender” ->
“age, gender”.

8. Discussion end of para 2 “it is likely that 62.2% (3064 patients) of this group had chronic HCV infection” – this is poorly worded statistically – the “probability” that exactly 3064 of these patients had chronic HCV is very small – however, it would be the “best point estimate” for the number of chronic HCV – ideally there should be a confidence interval on this estimate.

9. Discussion para 4 “found a 40% hepatitis C prevalence, identical to”… “treatment register were identical to the survey results” – poor use of the word “identical” – perhaps use “very similar” or “consistent with”.

10. Table 1 and 2 “Zealand” and “Sealand” – which is correct?

Discretionary revisions

1. Methods, Data sources: Drug Treatment Register – it would perhaps be useful to add in some specific numbers, such as the total number on the register, number identified or tests for chronic HCV.

2. Figure 1 – I found this very difficult to follow – perhaps split into 2 figures and provide some further explanation to the figures?

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I have answered no to all of the above - but I would note that I collaborate with one of the authors (Hay) in other academic publications - however I do not see this as a competing interest.