Reviewer's report

Title: Statistical estimates of absenteeism attributable to seasonal and pandemic influenza from the Canadian Labour Force Survey

Version: 1 Date: 25 November 2010

Reviewer: Mark Jit

Reviewer's report:

The authors apply a well established regression method that is used to determine the proportion of respiratory disease outcomes in time series data that are due to a particular organism (eg. influenza) using laboratory reports for that organism as explanatory variables. While there are variations in testing and reporting within and between seasons, in this case, these are taken into account by the use of indicator variables for both month of the year and influenza year. The novelty of the analysis is that it is applied instead to work-related absenteeism rather than disease outcomes or health care resource use. This application of the method appears legitimate, although there is the concern that it is quite a broad outcome measure (compared to eg. hospitalisations for respiratory outcomes).

Major compulsory revisions

1. My main concern is the comment that variables for non-influenza organisms were dropped due to “lack of statistical significance”. There could well be a substantial burden of disease in the relevant age groups due to organisms like RSV and S. pneumoniae. Indeed, since the outcome variable (work-related absenteeism) is not specific to respiratory disease outcomes, the explanatory organisms should not be restricted to respiratory organisms – other organisms (eg. gastrointestinal) also display seasonality and may lead to absenteeism. The authors need to provide more detail in order to justify this absence, including giving details of the variable selection algorithm, the significance cut-off, the organisms considered and (perhaps in an appendix) a graph of the time-series data for laboratory reports of each organism. The concern is that the variables dropped out of the regression simply because the data are aggregated into monthly counts (rather than eg. weekly or daily) and hence it is difficult to be sensitive to seasonal differences in organisms which peak around the same time (eg. influenza, RSV, S. pneumoniae, norovirus etc.)

2. There are a large number of coefficients in the regression equation (I counted 40) compared to the number of data points (around 150). While this looks necessary in order to take into account both within and between season trends, this means that the model is likely to fit data points well for this reason alone. Hence the results of the analysis need to be interpreted with some caution given that there is no way to independently validate them. This caveat probably needs to be mentioned in the discussion, unless the authors can think of a means of independent validation using a separate data source.
Minor essential revisions

1. Was the coefficient beta_5 (time off for non-influenza related ILI after the pandemic was announced) significant? I would be concerned if it was – while it is quite plausible that the pandemic situation would cause workers to take precautionary time off, it is difficult to believe that the amount of time was identical throughout the pandemic. Also what was used to decide when the pandemic ended?

2. Were all hospitalisations for respiratory causes tested for H1N1 during the pandemic?

3. In Table 1, it’s not clear why the % of hours lost attributable to influenza is much greater than the % of hours lost due to own illness or disability – surely the first is a subset of the second.

4. In Figure 3, why are the error bars present only for a few characteristics?

**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 declare that I have no competing interests