Reviewer's report

Title: Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium

Version: 2 Date: 9 October 2008

Reviewer: ANN Jolly

Reviewer's report:

Review of “Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium”

This paper describes the collection, using diaries, of the number, type and frequency of contacts made by a random sample of people of all ages in Belgium during the work week and on weekends. The authors also analyse their data for associations between intimacy or number of contacts and types of contacts such as household contacts, work or school associates.

This paper has the potential to deliver some indispensable parameters for mathematicians and others interested in simulating infection spread in populations. However, they are clouded in a long paper with too many goals, new and foreign methods which need much more explanation for this audience, and vague discussions and conclusions.

First, there is no doubt that more data on social contact patterns are needed, but the implications of different contact characteristics may be different for different diseases. For example, Mycobacterium tuberculosis usually is transmitted during more intimate contact over a long period of time, so duration of contact, and intimacy of contact, matter. In ‘flu, where coughing may produce aerosolised virus, less intimate contacts with limited frequency may be sufficient. This is all the more important for scholars to realise, as the two are linked; of course a person is unlikely to have highly frequent contact with some of those most intimate in her life, as she spends a lot of time with them, so the intensity of contact is made at the expense of frequency. The research which is presented here is very much a “fishing expedition” where the researchers have gone to great lengths to collect valuable findings, and analyse them with a number of sophisticated techniques only to find in the end, things like, that people who were intimate were more likely to touch each other, and spend longer times together - not unexpected. This is where I think that hypothesis testing is a good approach to the problem, at least as an heuristic device to clearly articulate what the research questions to be answered are, and what the answers may mean.

Second, the paper is too long. I suggest it be broken into at least two if not three papers; one describing clearly the data sought, the rationale for the project in specific detail, for example, whether the data is to be used in modelling pandemic ‘flu, bioweapons, TB, or norovirus, the specific parameters sought for each,
including a statement about why they are thought to be important. The data collection methods should then be clearly explained, without frequent references to the POLYMOD research as that may be unfamiliar to readers. Included in the methods section should also be some explanation of the probabilistic sampling methods, and why a random representative sample should be expected to deliver useful results, when we know from recent experience of SARS and HIV that people with extremely high numbers of contacts may play an important role. Then one of the analyses could be brought in to further elucidate the results, such as the associate rules analysis, which should be described here in clear, simple language. This is where the English in which the paper is written definitely needs improvement; phrases like; “...various interestingness measures could be studied...” It should be understood by the authors that to an epidemiologist or a reader of medical studies, “associations” are usually the result of some sort of regression analyses, and predictions of associations are usually accomplished in discriminant function analysis. This seems similar to what the authors may be trying to do, though with terms such as “support”; “confidence”, “lift” it is very difficult to tell. Again the classification trees seem to resemble cluster analysis, but any description of what the analysis is intended to reveal and how it does that is deficient. After the methods section the results should show a table of descriptors of individuals surveyed and perhaps the means or medians and IQR of the frequencies and durations of different types of contacts. Again, certain results could be emphasised here on their own merit, such as the number and shape of distribution of contacts, the same with the durations of contacts, and some idea of both together, i.e. how long do people spend with their spouses, daughters, and intimate friends? All of these questions are extremely important in determining disease spread, especially as the routes and infectious doses of some pathogens still need to be estimated. Last, obviously the results could include those of one of the analysis which further elucidates on one of the central themes of the paper, for example, what determines whether people have extremely large numbers of contacts? Or how many intimate friends do people have outside their homes? Such questions have immediate implications for disease control, such as which people may spread infection to the largest numbers of others? Or how important are non household intimate contacts in exposure, given that many infections require prophylaxis of household contacts. In the discussion, limitations of the sampling should be mentioned as well as any attempts to mitigate these; such as using RDD on land lines when it is not at all obvious how many people may not have land lines, and use cell phones only, or who have no access to phones. Recently, TB, HIV and community acquired MRSA have been spread in low income, minority, inner city populations, in which there is a lower proportion of people with telephones. Last the discussion should clarify the implications of the results for public health practitioners, and mathematical modellers. The use of the analytic methods should also be clearly rationalised in terms of what it adds to our knowledge.

Third, the language of the papers needs to be simplified, and would benefit from review by an English scholar familiar with these methods.

I think this paper holds critical information needed by epidemiologists, public
health staff, emergency planners, and infectious disease modellers, but the presentation obscures the value. Shortening the papers to one analysis, will help greatly in following one theme through from beginning to end, without having to constantly refer back to the methods of a very complex analysis.

I recommend that this paper (s) be resubmitted after compulsory major revision.

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

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

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

I declare I have no competing interests