Reviewer’s report

Title: Water and sewage systems, socio-demographics, and duration of residence associated with endemic intestinal infectious diseases: A cohort study

Version: 2 Date: 4 November 2010

Reviewer: Marek Brabec

Reviewer’s report:

The paper is interesting to read. It is the “The unit of analysis was person-day …” statement that is highly problematic. If this means that the analyses were computed as if the person-day data were totally independent from each other (given the explanatory variables), then it is very likely that the significance of the tests is vastly exaggerated (p-values underestimated, width of confidence intervals underestimated) due to (positive) intraclass correlation induced by the similarity within the same individual, within the same family, within the same community, etc. It is likely that there is family, individual and spatial autocorrelation structure that should be taken into account (either directly via spatial and individual structure of random effects or by pragmatic adjustment of standard errors as in the GEE approach). Without this adjustment, the analyses are very hard to defend. The authors mention that they did perform GEE somehow (“… logistic regression with generalized estimating equations (to account for repeated measures on subjects) were also considered …”), but only in addition to the standard logistic regression (PROC LOGISTIC) whose results are reported predominantly in the paper. This is wrong. Situation is less than trivial here and the statistical analysis should reflect that carefully in order to be plausible and realistic.

The authors mention openly right at the beginning of the paper: “The results did not always follow prior expectations based on studies examining outbreaks and single systems, and underscore the importance of studying factors associated with endemic disease across water and sewage system types.”, this is certainly not surprising. Tendency to find counterfactuals might be related to problems well known in the context of phenomena termed as “ecological regression fallacy”, “Yule-Simpson effect”. The problematic side of the observation study like that reported in this paper should be acknowledged more explicitly in the text. Nevertheless, the approach has its advantages as well (ideally, the authors should state their view of pros and cons of performing a study in the way they did, in comparison with the “single systems studies”).

The study limitation stated as “The administrative data on disease events were provided by physicians, not laboratories, therefore clinical suspicion was used to assign general rather than specific disease coding, preventing analyses of
individual diseases such as giardiasis or cryptosporidiosis.” makes the analysis results much harder to interpret and use. There might be strong physician-effects superimposed on the studied process (can it happen that some physicians over-report and some under-report some of the gastrointestinal outbreaks of interest? Can it be that some of this is not explainable by the environmental factors but rather by subjective features?). Probably, some systematic testing for these effects and/or formalized data cleaning procedures might be needed before publishing the results.

If analyzed and presented carefully (including the changes in the statistical methodology indicated previously), the data might suggest interesting and important ideas, e.g. as those mentioned in the Conclusions (“The results also convey cautions: they support previous research indicating that surface water sources, even with protected watersheds, deserve special attention; and present new data suggesting municipal sewer systems may require more scrutiny than previously thought.”).