Author's response to reviews

Title: Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales in the San Francisco Bay Area

Authors:

Michelle L Kirian (michellekirian@sfdph.org)
June M Weintraub (june.weintraub@sfdph.org)

Version: 2 Date: 2 July 2010

Author's response to reviews: see over
July 2, 2010

Natalie Pafitis MSc,
The BioMed Central Editorial Team
editorial@biomedcentral.com

Dear Ms. Pafitis:

Thank you for your email of 7 June 2010 regarding our manuscript (1185628764366661) previously titled "Time series analysis of the relationship of readily available diarrheal remedy over-the-counter drug sales data and gastrointestinal disease in the San Francisco Bay Area". My co-author and I are grateful for the reviewers’ careful consideration and thoughtful suggestions to improve our manuscript.

On behalf of my co-author, I am pleased to submit the attached revised manuscript addressing the reviewers’ comments. On the following pages, I copied each comment in italics, followed by an explanation of my response.

The manuscript is now 3257 words, and has 36 references, 3 tables, and 2 figures. The abstract is 196 words.

Thank you for the opportunity to revise this manuscript.

Sincerely,

Michelle Kirian, MPH
Referee: 1

1. “With one exception, the statistical methods are clearly described. The exception is the simulation of “random alerts” (pp. 4-5) which is not described.”

   - Dates were randomly chosen and not simulated. The text, “To further test the ability of the over-the-counter drug sales model to predict outbreaks, random alerts were simulated and sensitivity and specificity calculations were performed again.” was changed to,

   **P 7, Paragraph 2:** “To further test the ability of the Diarrheal Remedy Sales model to predict outbreaks, sensitivity and specificity calculations were also performed for three sets of randomly chosen dates.”

2. “The only other comment that I have is that the description of the ARIMA models that were fit (now on p. 4) belongs in the Results section.”

   - The description of the ARIMA models was moved to the first paragraph of the Results section (**P8, Paragraph 1**).

Referee: 2

1. The title is awkward and should be more descriptive of the text, which says little about time series analysis.

   - The title, “Time series analysis of the relationship of readily available diarrheal remedy over-the-counter drug sales data and gastrointestinal disease in the San Francisco Bay Area” was changed to,

     **Title:** Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales in the San Francisco Bay Area

2. The abstract needs some modification.

   - The abstract was modified to reflect the changes in terminology in the rest of the text. Please see P 2 track changes for more details.

3. Data utility, and significance of findings: It is important to make the distinction between evaluation of utility of NRDM data and of OTC data in general. On p. 2,line 3, the authors overstate when they say they “evaluated the ability of readily available diarrhea remedy over-the-counter drug sales data to predict”. Their later statements are clearer on this point, especially the first full paragraph on page 7, beginning with “Our study was not able to assess
whether improvements in over-the-counter drug sales reporting systems might improve the performance of this type of syndromic surveillance. Systems that include more complete information on retailer participation and more detailed sales data might perform better than the system that was available to us…”. From my own projects and from talking to users at conferences, I am familiar with the weaknesses of the NRDM dataset. What I am asking is that the authors make this distinction more prominent in the early part of the paper and also discuss it in the Limitations section. Indeed, the Public Health Agency of Canada has reported better results with their own OTC data.

More specifically, the statement that “over-the-counter drug sales did not coincide with diarrhea due to other [beside Norovirus] etiologies” reminded me of my experience evaluating records of prescription drug sales. In that project, no one could demonstrate any correlation between GI-related prescriptions and illness outbreaks until a physician investigated the products included and removed many of them for improved specificity, and then correlations were found. Purchasers of NRDM data cannot modify the data in this way.

I agree with their statement that “the ability of over-the-counter syndromic surveillance to enhance the detection of waterborne disease outbreaks has not been adequately demonstrated”. However, they also point out that “Our study was not able to assess whether improvements in over-the-counter drug sales reporting systems might improve the performance of this type of syndromic surveillance”, and they should not elsewhere extrapolate their findings to OTC data in general. Judicious reorganization and some rewording are needed.

- We agree that this analysis should in no way be considered an evaluation all syndromic surveillance systems using over-the-counter drug sales but instead an evaluation of the predictive ability of one data source in particular. To make this point more obvious the following have been changed or added:

- “The authors evaluated the ability of sales of over-the-counter diarrheal remedies to predict endemic and epidemic gastrointestinal disease in the San Francisco Bay Area”, was changed to,

**P 2. Paragraph 1 (Abstract):** “The authors evaluated the ability of sales of over-the-counter diarrheal remedies available from the National Retail Data Monitor to predict endemic and epidemic gastrointestinal disease in the San Francisco Bay Area.”

- “However, similarly, the low percentage of store participation by NRDM enrolled stores and the lack of information on the participating stores such as market coverage or drugs included in each category indicate that, like other over-the-counter drug data sources, serious limitations may inhibit the use of this data for syndromic surveillance. Our study was not able to assess whether improvements in over-the-counter drug sales reporting systems might improve the performance of this type of syndromic surveillance. Systems that include more complete information on retailer participation and more detailed sales data might perform better than the system that was available to us for this analysis.” was added to and moved to the limitations section,
P 13, Paragraph 3: “Over-the-counter drug sales records as provided by the NRDM have several limitations. The usability of these data could be improved if participation by enrolled stores was increased, or if meta-information on participating stores such as market coverage and on the drugs included in each were made available. While we did not find any association between gastrointestinal disease and purchases of diarrheal remedies in general, it is possible that one product or a subset of products included in this category might have coincided with known disease. Furthermore, our study was not able to assess whether improvements in over-the-counter drug sales reporting systems might enhance the performance of this type of syndromic surveillance. The use of over-the-counter drugs sales for surveillance may be prohibitive due to the cost and logistics of data collection, or the proprietary and secret nature of the data.”

- “This work does not support the implementation of syndromic surveillance with over-the-counter diarrheal remedy sales for gastrointestinal disease.” was changed to,

P 2, Paragraph 4 (Abstract): “This work does not support the implementation of syndromic surveillance for gastrointestinal disease with data available through the National Retail Data Monitor.”

- “To clarify the validity and representativeness of over-the-counter drug sales data for prospective outbreak detection, we sought to determine if these data are related to known outbreaks of infectious gastrointestinal illness in the drinking water service area.” was changed to,

P 3, Paragraph 2: “To clarify the validity and representativeness of sales of over-the-counter diarrheal remedies available through the National Retail Data Monitor (NRDM) for prospective outbreak detection, we sought to determine if these data are related to known outbreaks of infectious gastrointestinal illness in the drinking water service area.”

4. More detailed discussion of results and evaluation: The authors’ discussion seems to confuse forecasting and detection. On page 2 is the statement, “Signals generated by forecasting with the drug sales model did not perform better than signals generated randomly.” How do signals “perform”? Do the authors mean that random signals match the reported outbreaks as well as time series anomalies based on forecasting? To me, this finding means that the time series collected are not what should be monitored. The related discussion on page 6 also needs clarification. Please explain.

- In using the term “perform” the authors intended to say that the signals generated by the model did not give a better indication that an outbreak was occurring than did signals which were chosen randomly. We agree that this means that the time series we studied—NRDM diarrheal remedy sales—is not a good indicator for advanced monitoring of gastrointestinal disease. To make this point clearer, changes were made to the text as stated below. Further details on the analysis findings are given in the results section (P9-10, Paragraphs 3, 4 & 1), “No significant correlation at any lag was found between Diarrheal Remedy Sales and diarrhea cases (Figure 2). Furthermore, regression analysis of the Diarrheal Remedy Sales univariate model residuals did not reveal an
association between the weekly number of outbreaks or outbreak-associated cases and Diarrheal Remedy Sales when all outbreaks data were included or when restricted to Norovirus and/or non-institutional outbreaks.

Four signals were generated by the Diarrheal Remedy Sales model (6/11/06, 1/29/06, 10/15/06 and 6/10/07). Four of the 20 outbreaks with 50 or more and one of three with 100 or more cases started during or lasted through a week with a signal. The two outbreaks with 100 or more cases without signals were both non-institutional Norovirus outbreaks with steep epidemic curves. Sensitivity for all outbreaks and for outbreaks with 50 or more or 100 or more cases was low and specificity high (Table 3). Three sets of four random signals had identical sensitivity and specificity to those generated by the model, further supporting the conclusion that any relationship between Diarrheal Remedy Sales and gastrointestinal illness is spurious.

- “Signals generated by forecasting with the drug sales model did not perform better than signals generated randomly.” was replaced with,

**P 2, Paragraph 3, Abstract:** “Signals generated by forecasting with the diarrheal remedy sales model did not coincide with weeks with outbreaks more reliably than signals chosen randomly.”

- “NRDM diarrheal remedy over-the-counter sales data did not predict outbreaks of gastrointestinal disease or correlate with individual cases of diarrhea illness. Signals generated by the diarrheal remedy sales model did not perform better than signals generated at random in predicting gastrointestinal outbreaks.” was changed to read,

**P 10, Paragraph 2:** “NRDM Diarrheal Remedy Sales did not predict outbreaks of gastrointestinal disease or correlate with individual cases of diarrhea illness. Signals generated by the Diarrheal Remedy Sales model did not coincide with weeks with outbreaks more reliably than signals chosen randomly.”

5. Explain how sensitivity and specificity were defined to clarify the statement “The sensitivity and specificity of these alerts were calculated by week”

- The following was added,

**P 7, Paragraph 2:** “Model sensitivity was calculated as the number of weeks where there was a signal and at least one outbreak divided by the total number of weeks with an outbreak. Specificity was calculated as the total number of weeks without a signal and no detected outbreaks divided by the total number of weeks without an outbreak.”

6. I had to read the statement “Forecasting with the univariate over-the-counter drug sales model produced four signals” several times. I think that it means that there were 4 instances where the observed value exceeded the forecast value by more than the upper 95% confidence limit. Is this true? These explanations should not require much space.

- The methods section states that, (P 4, Paragraph 6) “Signals were generated when actual observations exceeded the upper 95% confidence limit.” To clarify the number of signals generated, “Forecasting with the univariate over-the-counter drug sales model produced four signals for the weeks beginning 6/11/06, 1/29/06, 10/15/06 and 6/10/07.” was replaced with,
**P 9, Paragraph 4:** “Four signals were generated by the Diarrheal Remedy Sales model (6/11/06, 1/29/06, 10/15/06 and 6/10/07).”

7. **p. 2, title:** the phrase “diarrheal remedy over-the-counter drug sales data” is awkward in the title and in the text. One suggestion is “over-the-counter sales records of diarrheal remedies”.

   - We appreciate this suggestion and replaced this phrasing as follows:

   - “Time series analysis of the relationship of readily available diarrheal remedy over-the-counter drug sales data and gastrointestinal disease in the San Francisco Bay Area” was replaced by,

**P1, Title:** “Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales in the San Francisco Bay Area”

   - “The authors evaluated the ability of readily available diarrhea remedy over-the-counter drug sales data to predict endemic and epidemic gastrointestinal disease in the San Francisco Bay Area.” was replaced by,

**P 2, Abstract:** “The authors evaluated the ability of sales of over-the-counter diarrheal remedies available from the National Retail Data Monitor to predict endemic and epidemic gastrointestinal disease in the San Francisco Bay Area.”

   - Throughout the text, “Diarrheal Remedy Sales” was substituted for all instances referring to the analysis variable as defined by: “the proportion sales of non-promotional diarrhea remedy to sales of non-promotional drugs for all categories combined”. Where the text was discussing anti-diarrheal drug sales in another context the phrases “over-the-counter anti-diarrheal drug sales” or “Over-the-counter drug sales records”.

8. **p. 3, “may include as few as one or two cases”**

   - The most liberal definitions of outbreaks require at least two cases—please modify the text to read “as few as two cases”.

   - Done.

9. **The statement “Data were received from three adjacent counties and transmitted in electronic formats” should be rewritten as “Data were transmitted in electronic formats from three adjacent counties”. If this is not correct, then a better explanation is needed.**

   - Done

10. **“Only cases with residence in the respective county were sent by each county...”**

    - Wouldn’t it be more correct to say “were requested from each county...”?”
“Reports of cases of gastrointestinal disease from 2001-2007 were requested from each of the county health departments in the drinking water service area. … Only cases with residence in the respective county were sent by each county to avoid duplicate case records.” was changed to,

**P4, Paragraph 2:** “Reports of cases of gastrointestinal disease among residents from 2001-2007 were requested from each of the county health departments in the drinking water service area. … (second sentence deleted)”

11. Please change “receives outbreaks records” to “receives outbreak records” or “receives records of outbreaks.”

“Electronic records of outbreak data for all three participating counties were provided by the California Department of Public Health which receives outbreaks records” was changed to,

**P 4, Paragraph 3:** “Electronic records of outbreak data for all three participating counties were provided by the California Department of Public Health which receives outbreak reports…”

12. In the data descriptions, the general reference to “data” is too vague and should be clarified with references to data records and their fields. In particular, please replace the clause “gastrointestinal case and outbreak data were aggregated into a single dataset” with more detail.

Details on how data (cases, outbreak and sales) were aggregated are provided in,

**P6, Paragraph 2:** “Over-the-counter drug sales, and gastrointestinal case and outbreak data were aggregated by week for analysis. Diarrheal Remedy Sales were aggregated by week of sale, cases by week of report to the health department and outbreaks by week of onset of first outbreak-associated case. Data were divided into three parts for model building, model validation, and forecasting.”

“Data included etiology, date of report to the health department, gender, age, city and county.” was replaced with,

**P 4, Paragraph 2:** “Reports for each case included etiology, date of report to the health department, gender, age, city and county.”

“Outbreak data included information on etiology, number of cases, date of symptoms onset for the first and last cases, affected counties, and whether the outbreak occurred in an institutional setting such as a nursing home.” was replaced with,
P4, Paragraph 3: “For each outbreak, information on etiology, number of cases, date of symptoms onset for the first and last cases, affected counties, and whether the outbreak occurred in an institutional setting such as a nursing home was provided.”

- “Over-the-counter drug sales data were purchased from the National Retail Data Monitor (NRDM). Data for the years 2005-2007 were provided as an electronic file. Data for years 2003-2004 were downloaded using the NRDM web interface. NRDM over-the-counter drug sales data” was replaced with,

P5, Paragraph 2: “Over-the-counter drug sales records were purchased from the NRDM. Records for the years 2005-2007 were provided as an electronic file. Records for years 2003-2004 were downloaded using the NRDM web interface. NRDM over-the-counter drug sales records…”

13. “drug sales data are divided into 18 categories based on common use, form and whether intended for adult or pediatric populations…”
--Does this description mean that the categories are such as “adult/pills”, “child/liquid”? Please give examples. This sentence needs to be split into 2 sentences.

- The 18 NRDM sales categories are listed in the text and category names include basic information identifying form (liquid or tablet) and population (adult/pediatric) where appropriate and available. (P 5, Paragraph 2) “NRDM over-the-counter drug sales records are divided into 18 categories based on common use, form and whether intended for adult or pediatric populations. NRDM drug categories are: diarrhea remedies, anti-fever adult, anti-fever pediatric, bronchial remedies, baby/child electrolytes, chest rubs, cold relief adult liquid, cold relief adult tablet, cold relief pediatric liquid, cold relief pediatric tablet, cough syrup adult liquid, cough adult tablet, cough syrup pediatric liquid, cough/cold, hydrocortisones, nasal product internal, throat lozenges, and thermometers.”

- Examples of medications which are included in the NRDM diarrheal remedy category are also included. (P6, Paragraph 1) “Diarrheal remedies are products taken for the relief of diarrhea and include bismuth, attapulgite, subsalicylate, and loperamide hydrochloride products.”

14. ”Data distinguish between promotional and non-promotional sales”
-- How do the data make this distinction?

- NRDM provides units sales counts by category as total and as total unit sales minus unit sales for products for which discounts or other promotions were offered for during the reporting period. No additional information detailing the necessary criteria for promotional labeling is available. “Data distinguish between promotional and non-promotional sales.” was replaced with,

P 5, Paragraph 2: “Daily total sales are available for both all units sold by category and units sold by category excluding units for which discounts or other promotions were offered during the reporting period.”
15. p.5, Please add the missing word in “From July 2003 through December 2007, the proportion diarrheal remedy sales …“

- Done

16. p. 6, As in the title, I suggest replacing “NRDM diarrheal remedy over-the-counter sales data” with “NRDM over-the-counter sales data for diarrheal remedies”

- “NRDM diarrheal remedy over-the-counter sales data did not predict outbreaks of gastrointestinal disease or correlate with individual cases of diarrhea illness.” was replaced with,

P10, Paragraph 2: “NRDM Diarrheal Remedy Sales did not predict outbreaks of gastrointestinal disease or correlate with individual cases of diarrhea illness.”

17. p. 7, Please correct the grammar in “the incidence of respiratory illnesses… is likely to greatly exceeded that of diarrheal illnesses”

- Done.

18. p. 8 The Conclusions section says that “we cannot exclude the possibility that this system may be able to detect larger outbreaks”. No system has been described. What do the authors mean?

- This conclusion is meant to provide a more concise summary of the point in made in the first paragraph of the limitation section, “First, there were no large regional outbreaks in our dataset and the high data variability of diarrhea remedy sales may make it difficult to discern changes resultant from relatively small increases in illness.” We have changed the sentence to read: “However, we cannot exclude the possibility that NRDM data may be useful for detecting large outbreaks.”

19. p. 7 “The number of surveillance systems and lack of published reports on over-the-counter drug sales monitoring systems, and NRDM specifically, suggests publication bias may be present.”

To me, this reference to “publication bias” suggests that more successful applications exist but are not published. Is this interpretation correct? I would have expected bias in the other direction, with the successful surveillance systems more likely to publish. Having spoken to multiple users, I find it much more likely that the combination of understaffed user departments and weaknesses in the data has resulted in a reduced number of evaluations, successful experiences, even regular usage. The authors may wish to revise the statement based on experience.

- Publication Bias is a form of selection bias where the results from studies or work are less likely to be submitted by authors or accepted by reviewers for publication based on the direction or strength of the findings. It is sometimes referred to as “positive outcome bias” because significant, positive findings are more likely to be published than null findings,
although the bias could certainly go in the other direction. The statement in question is intended as a conclusion to the entire review of syndromic surveillance studies which starts on P6, Paragraph 3 and not only the paragraph it is in. To make this clearer, “The number of surveillance systems and lack of published reports on over-the-counter drug sales monitoring systems, and NRDM specifically, suggests publication bias may be present.” was changed to,

**P 13, Paragraph 1:** “Although our literature review did identify a number of reports suggesting that syndromic surveillance with over-the-counter anti-diarrheal drug sales could enhance traditional disease control activities, the widespread adoption of syndromic surveillance systems and the paucity of published reports on over-the-counter drug sales monitoring systems, and NRDM specifically, suggest publication bias may be present.”