Author's response to reviews

Title: Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis

Authors:

Caryn Roth (carynroth@gmail.com)
Randi E Foraker (rforaker@cph.osu.edu)
Philip RO Payne (philip.payne@osumc.edu)
Peter J Embi (peter.embi@osumc.edu)

Version: 2
Date: 23 April 2014

Author's response to reviews: see over
Dear Prof. Duftschmid,

Please consider the attached revised manuscript, “Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis” for publication as an original research article in *BioMed Central Medical Informatics and Decision Making*.

We are grateful to have had the opportunity to revise and resubmit our manuscript for your consideration. We have addressed the reviewer’s comments point-by-point below.

Our work describes the secondary use of electronic health record data, integrated with socioeconomic and obesogenic community-level data, to study obesity and overweight in a diverse population. We leveraged current health informatics advances for new applications and demonstrated the benefit of this methodology for future use to reduce chronic disease and improve health.

All authors have read and approve of the manuscript as written. We have no disclosures.

Thank you,

Caryn Roth
Responses to Associate Editor – Dr. Duftschmid

Editor’s Comments:

As you will see from the attached reviews, the 2 reviewers have come to heterogeneous assessments of your submission.

Reviewer Maldonado criticizes that the paper does not fit in a medical informatics journal, as it does not cite enough medical informatics references. In accordance with the Section Editor I differ from this view insofar as I see the integration of GIS data with EHR data as an important trend in medical informatics and there are several GIS-related papers cited in your submission.

However, I recommend another extension of your reference list: I found 2 conference publications of a subgroup of the current authors, which seem to describe results from the same project that is presented in the current submission:


Please add citations of these preceding conference publications in your submission and make explicit how the current submission extends the conference papers.

Concerning the second point of criticism (not enough insights concerning the problem of heterogeneous clinical data sources) I recommend you add a discussion on the general problems of integrating data from different sources and to what extent you were
confronted with them. Your situation was probably alleviated by the fact that you had to integrate on the coarse level of ZIP-codes only instead on an integration on patient level.

Reviewer Grossmann only recommends a discretionary revision. If you follow his recommended split of the data into 3 sets, the split should be done by zip code (not just randomly selecting it among the patients), as from a knowledge discovery perspective the real unit of analysis here is the zip code.

**Author Response**

Thank you for your interest in this manuscript and your kind acknowledgement of its relevance to medical informatics, as it integrates GIS and EHR data in a novel way to advance health. We appreciate your encouragement to include our previous publications related to this work. Of note, both of these are one-page abstracts, published in conference proceedings when this work was at a preliminary stage. They are both much less detailed than the current manuscript, and the research has since evolved due to feedback received while presenting at these conferences.

The first abstract, “Augmenting EHR-Derived Clinical Data with Geographic-Level Public Health Information to Develop Research Hypotheses for Population Obesity Rates” described preliminary results and statistical methods that differ from what we ultimately used and report in this manuscript. For this reason, we prefer not to cite this abstract and confuse readers. However, we have added the second conference proceeding abstract “Integrating population- and patient-level data for secondary use of electronic
health records to study overweight and obesity” as a reference in the current manuscript, since it describes some data mining applications and challenges of using these data.

Thank you for your recommendation to add a discussion on the general problems of integrating data from different sources the extent to which these issues applied to our work. This is certainly a relevant and important topic in biomedical informatics, as there are many difficulties and open research questions surrounding the integration of data for patient-centered research. As an example, the Electronic Data Methods (EDM) Forum recently commissioned multiple articles for a supplement in Medical Care addressing exactly these topics: http://journals.lww.com/lww-medicalcare/toc/2013/08001. As mentioned below in the response to reviewer #2, we have added this section:

“Additional challenges arise from linking data from multiple sources. While some data will match according to the linker in each dataset, other entries will remain unmatched.[1] We were confronted with this issue when we queried the PrimeLocation database for community-level data describing zip codes within Franklin County, which did not in fact cover all the zip codes that comprised our patient cohort. This mismatch was due to the fact that zip codes and counties do not share perfect overlap, so the PrimeLocation classification system did not include some of the zip codes which were not fully contained in the county. This forced us to exclude patients residing in those districts from our analysis, which may have induced a selection bias. Furthermore, each distinct data source contained data from one point in time, but these do not necessarily match. For example in our analysis, our patient data were from 2010-2011, and the
PrimeLocation data were from 2011. However, we could not get 2011 data on number of establishments from the US Census, so we had to settle for older data, adding possible misclassification into the analysis. Further integration challenges likely would have arisen had we been able to acquire this type of socioeconomic and community resource data at the more granular level of the individual patient.”

Responses to reviewer 1 – Dr. Grossmann

Reviewer's report:
The paper considers an interesting approach for combining data from EHR with community level factors for finding risk factors for obesity at the community level.

As statistical method logistic regression is used.

As statistician I cannot evaluate the medical implications of the results but in my opinion the approach making use of different data sources in retrospective observational studies is important and of interest in many areas of evidence based medicine and by no means limited to the considered question of obesity Overall the paper is well written and describes data preparation and the method for model development quite clear. Limitations of the approach are clearly stated.

However, I would propose to go one step further in the analysis and put it into the framework of data mining. This would mean that the authors use for model assessment the method of splitting the data into a training set, a validation set and a test set.

(See for example the textbook Hastie et al. : The Elements of Statistical
The frequently recommended sizes are 50% for training, 25% for validation and 25% for testing. Besides the training and the validation step, which was combined in the paper by using the method of variable selection for the logistic model you can apply the model to an independent test set. Application of the model for independent test data would give a better idea about the generalization power of the model. Due to the large amount of data available this approach should work quite well in this case.

**Author Response**

Thank you for your interest in our work and your insightful suggestion to repeat the analysis with a data mining framework. While we appreciate this suggestion, we feel that the predictive model answers a different question than our current explanatory model seeks to answer. Please see Shmueli, 2010 (1) for a thorough description of the different purposes and methods for these two diverse modeling approaches, including different decisions regarding data selection and analysis. There are also additional biases that arise with the use of data mining methods, separate from those we encountered using our current statistical methodology, and we would need to address these in any future data mining applications of this work. (2) Since our analysis aims to study the associations between obesity and community level factors, rather than predict obesity, we fit an explanatory model to these data.

We have clarified this point by adding it in our limitations and future directions section:
“We also will explore creating a predictive model with these data, in order to see if we can better predict a patient’s BMI category using community-level data.”

References:


Responses to reviewer 2 – Dr. Maldonado

Reviewer's report:

The overall scope of the paper is relevant since it is aligned with one current hot topic in (bio)medical informatics: the reuse of EHR data in research. The paper is well written, the structured is adequate and the level of detail is sufficient. Nevertheless and in my opinion, the paper is not suitable for publication in a medical informatics journal such as BMC Medical Informatics & Decision Making. My main concern is that the “informatics” part is missed. The paper fits in a medical/public health/epidemiology/… journal rather than in a medical informatics one. Actually only 3 out of 43 references in
the paper are from medical informatics journals. The paper describes a use case and it seems that data was collected and integrated manually and then analyzed. There is no reference to any informatics tool or methodology used to seamlessly integrate and augment EHR with community-level data. The reuse of EHR in clinical research is a difficult topic. An important problem is the heterogeneity of clinical data sources, which may differ in data models, schemas, naming conventions, semantics, and degree of detail used to represent similar data. Furthermore, clinical research very often require data at a level of abstraction higher than raw clinical data, a problem known as the ‘impedance mismatch’. Any paper about this topic to be published in a medical informatics journal should give some insights or solutions about the aforementioned problems.

**Author’s Response**

Thank you for your comments and suggestions to make this manuscript more applicable for an informatics audience. In addition to the reuse of EHR data for research, use of GIS is also increasing in informatics, and the integration of diverse data sources is an ongoing point of research in the informatics literature. Of note, this paper is an example of applied clinical informatics, rather than an informatics basic science paper, so we believe it does fit in well with an informatics journal such as BMC Medical Informatics and Decision Making. Please see this article for a nice discussion of applied informatics as an integral component of the informatics life cycle:

We have added additional citations to address your concern that there were not enough citations from informatics journals. These citations include:


Thank you for discerning the lack of mention of any informatics tool or methodology used to seamlessly integrate and augment our EHR with community-level data. This was certainly a challenge in our analysis, and doing the data integration manually is not a long term solution that we endorse. We have considered ways that informatics methodologies could be used to do this in a more systematic and sustainable way, and looked toward others who are attempting to do such work. We have expanded the challenges section in our discussion to address this point:

“Our approach for combining and analyzing these data was semi-manual, but we are exploring options for better integrating and sustaining these data sources long-term. For example, by incorporating community-level data into EHRs directly, we would not only attain more granular data at the patient level, but we would be better able to
validate and analyze it in a longitudinal fashion. Physicians could use this additional patient context to make more relevant health recommendations, which might improve compliance and outcomes. Unfortunately, adding external data into the EHR presents its own set of challenges related to accuracy and timeliness, which may not make this option viable on a large-scale. However, efforts to integrate community-level data into enterprise data warehouses are underway, and these novel methods present a promising approach to systematically integrate these data to answer such research questions. Our current work makes the case for the importance of such efforts and informatics solutions to facilitate this type of research.”

In response to your comment about the reuse of EHR in clinical research, we have expanded the challenges section of our manuscript and added the following paragraph:

“Beyond erroneous values, EHR-derived data present additional challenges when used for secondary purposes such as research. These issues include missing data, inconsistent data entry practices over time and between physicians, and data stored in text fields, which are difficult to access and use for large-scale research. Even data stored in an easily-accessible structured format may be unreliable or unusable for research, due to its initial intended purpose, for example billing. Furthermore, unknown provenance and insufficient granularity may limit the data from answering the research question of interest. Many are working to address such issues, for example tackling the free text issue by developing sophisticated Natural Language Processing (NLP) techniques. As the larger informatics community works to address such obstacles in the
secondary use of EHR data, we can still utilize this rich data source, keeping cognizant of its limitations.”

We have also added an additional paragraph about challenges of integrating the types of data we used in this analysis. All of these expanded discussion sections include some of the additional references mentioned above.

“Additional challenges arise from linking data from multiple sources. While some data will match according to the linker in each dataset, other entries will remain unmatched.[1] We were confronted with this issue when we queried the PrimeLocation database for community-level data describing zip codes within Franklin County, which did not in fact cover all the zip codes that comprised our patient cohort. This mismatch was due to the fact that zip codes and counties do not share perfect overlap, so the PrimeLocation classification system did not include some of the zip codes which were not fully contained in the county. This forced us to exclude patients residing in those districts from our analysis, which may have induced a selection bias. Furthermore, each distinct data source contained data from one point in time, but these do not necessarily match. For example in our analysis, our patient data were from 2010-2011, and the PrimeLocation data were from 2011. However, we could not get 2011 data on number of establishments from the US Census, so we had to settle for older data, adding possible misclassification into the analysis. Further integration challenges likely would have arisen had we been able to acquire this type of socioeconomic and community resource data at the more granular level of the individual patient.”