Integrated data analysis for early warning of lung failure

Geisinger Health Collider Project: Stage 1

The Outliers: Rebecca Barter and Shamindra Shrotriya

November 17, 2015

1 Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality worldwidee [Lozano et al., 2012], however many of those with COPD remain undiagnosed [NIH, 2010]. Individuals with undiagnosed COPD typically experience related adverse health effects resulting in increased utilization of health care services. One of the primary reasons for hospitalizations amongst undiagnosed COPD patients is pneumonia; amongst patients with a secondary diagnosis of acute exacerbation of Chronic obstructive pulmonary disease (COPD), pneumonia was the primary reason for hospitalization for 22.3 percent of admissions. For this project, our aim is to develop methods capable of effectively predicting cases of undiagnosed COPD amongst those whose primary reason for hospitalization was pneumonia. Most existing algorithmic approaches to this prediction problem focus only on utilizing clinical information, however we aim to incorporate external data sources primarily related to relevant socioeconomic factors that are not captured by the clinical records using a process called "data blending".
One of the most common causes of COPD is cigarette smoking, however as many of 10-20% of COPD patients have never smoked and only a fraction of smokers develop COPD. These insights are suggestive of alternate risk factors such as genetic and environmental, particularly the ingestion of biomass fuel combustion. With this in mind, our goal is to "blend" external environmental data as well as data containing smoking information and medical claims information with the clinical data using a novel blending approach with the hopes of increasing our ability to predict undiagnosed COPD amongst those hospitalized with pneumonia. .

2 Data

The clinical data we will use in this project will include data collected from patients who are a part of the Geisinger Health System. The dataset consists of data from a total population of 88,000, among whom, 1,033 have a "discharge diagnosis of COPD". The clinical dataset contains approximately 80 unique variables sourced primarily from medical records with a few variables sourced from billing records. We will be comparing two sub-populations

1. The COPD subpopulation: all patients who have been diagnosed with COPD.
2. The non-COPD pneumonia subpopulation: those who are diagnosed at a single visit with pneumonia and subsequently recover. We do not include in this subpopulation those who are re-admitted at multiple visits with pneumonia, since this may be indicative of undiagnosed COPD.

2.1 External data sources

As a major focus of this project is to blend external data sources with the clinical data described above. This task is poised with a number of challenges. Firstly to ensure that the external

data was suitable for the task at hand, we aimed to ensure that a number of requirements were met: the data must be publicly available, obtained from a trustworthy source, be de-identified, be representative of the population of interest and contain some variables common to the clinical dataset.
Although these requirements seem fairly straighforward; in practice, it turned out to be extremely difficult to find relevant publicly available individual-level data with variables common to the clinical dataset. As a result, we decided to take two alternative approaches.
First, we first propose to use data collected within the Geisinger health system, but not included in the provided COPD clinical dataset. The primary reason for this approach is three-fold: first of all, as noted above, it is extremely difficult to find individual-level data that is both publicly available AND from a trusted source. Secondly, using data collected by Geisinger ensures that the underlying population mostly matches the population from which the COPD data is drawn, a key concern for the validity of the conclusions drawn from our subsequent analyses. Finally, for the publicly available individual-level data that we could find, it quickly became clear that is is extremely difficult to find adequate variables to match on (where the matching is to match individuals in the external dataset to similar individuals in the clinical dataset). For example, in most cases, the best we could do is matching on age, gender and approximate location, however if we are aiming to impute a variable such as smoking or exposure to biomass fuel combustion, the imputed values is likely to be extremely noisy and not particularly trustworthy (for example, it is highly unlikely that all females, aged 27 who live in Danville, Pennsylvania have the same smoking status; we need more common information between the two datasets).
Our second approach to identifying external data is to use the widely available location-based data (such as weather data, air pollution data or flu trends data), which cannot be imputed at an individual-level. Values of variables imputed from these types of less-granular data sources will be the same for all patients who attend the same Geisinger medical clinic, with the presumption that these patients live within the general proximity of the medical clinic.
The following sections describe the data that we intend to blend (using various methods that will be described below), with the clinical data.

2.1.1 Smoking data from the Geisinger psychiatric diseases dataset

As mentioned above, one of the most common causes of COPD is smoking. Although the COPD dataset provided by Geisinger, does not contain the smoking status attribute, the psychiatric diseases dataset does. Assuming that it is not possible to obtain the smoking status of the COPD patients in our dataset, we could use the common attributes of the psychiatric diseases and COPD datasets to predict/impute smoking status. This is an ideal dataset for matching, since (1) the data are sourced from the same population as the patients in the COPD dataset, (2) the data are sourced from a trusted source, and (3) there will be a large number of overlap between the variables. Further, if there are common patients in each dataset, then we could use these patients to devise a supervised learning algorithm which could predict the smoking status of the non-overlapped patients in the COPD dataset (although the method we discuss in our blending section below does not assume any overlap in the two datasets).

2.1.2 Insurance data from Geisinger health systems

One of the primary causes of COPD is the inhalation of biomass fuels. In order to infer the likelihood of being exposed to biomass fuels a key indicator would be the employment history of the patient. A reliable source for this information would be the insurance history of the patient under the Geisinger system which would record the employer of the patient and as such allow us to identify the associated industry and likelihood of exposure to biomass fuels. The employment zipcode would also allow for an additional longitudinal location statistic to blend.
Furthermore by accessing the deidentified insurance data of the patient we could also get a reliable longitudinal history of the previous ailments/ treatments undertaken by the patient. By

combining this with the ontology of diseases data a clustering analysis could be undertaken to better understand which type of past ailment types were more closely linked to COPD.
We understand that this data may be sensitive, however if it is cleaned and deidentified for each patient to mitigate privacy risks it becomes a rich and reliable longitudinal data source on patient health/ employment history.

2.1.3 Weather data from Pennsylvania

Although there is no direct established link between cold/ wet weather causing increased risk of pneumonia, studies have shown that cold and flu infection cases are generally higher in Fall and Winter months (which could lead to pneumonia). A likely reason for this is that people are often contained indoors during such weather conditions and have a higher risk of transferring airborne diseases. By longitudinally blending recent average temperature and rainfall data by patient housing zipcode and patient employment zipcode (from the insurance data above) could assist in classifiying true pneumonia cases [McCoy and MS, 2008].

2.1.4 Air pollution data from Pennsylvania

A leading identified cause of COPD is the inhalation of biomass fuels. Often this can be due to working/ living in areas which are affected by heavy air pollution - particularly soot and particulate matter [WHO, 2014]. This longitudinal air pollution data is availible from the Pennsylvania government.

3 Data blending

In this section, we will describe our intended methodology for data blending. Given that we have external data of various different levels of granularity, we will describe each situation individually.

3.1 Individual-level data blending

For the datasets for which we can perform individual-level matching, such as the Geisinger psychiatric diseases dataset and the insurance data, we have developed a two-stage blending method. The first stage is a matching stage which involves, for each individual in our clinical dataset, identifying individuals in the external dataset that are most similar in terms of the covariates common to both datasets. In order to effectively blend data sources, there need to be exist common variables using which the datasets can be combined. We propose the following variables to be the minimum required individual-level variables for blending external data sources to the COPD dataset provided: age, gender, zipcode (or hospital zipcode) and the date of patient treatment.
The second stage is an imputation stage which involves imputing or predicting the external variables of interest for the clinical patients by aggregating the values obtained by the matched external individuals. To set up some notation, for observation i in the clinical dataset, we can define a vector of the p observed variables by

$$\mathbf{y}_i = (y_{i1}, ..., y_{ip})^T, \qquad i = 1, ..., n$$

similarly, for the jth observation in the external dataset, we can define a vector of the q observed variables by

$$\mathbf{z}_j = (z_{j1}, ..., z_{jq})^T, \qquad j = 1, ..., m$$

Figure 1 describes this process for the first individual in the clinical dataset in a general setting. Suppose that we can reorder these variables so that the first $k \leq \min(p, q)$ variables are common to both datasets, then we can re-write our observations as

$$\text{clinical observation } i\text{:} \quad \mathbf{y}_i = (x^c_{i1}, x^c_{i2}, ..., x^c_{ik}, y_{i(k+1)}..., y_{ip})^T, \qquad i = 1, ..., n$$

$$\text{external observation } j: \quad \mathbf{z}_j = (x^e_{j1}, x^e_{j2}, ..., x^e_{jk}, z_{j(k+1)}..., z_{jq})^T, \qquad j = 1, ..., m$$

where $x_{i1}, ..., x_{ik}$ correspond to the common variables, $y_{j(k+1)}, ..., y_{jp}$ correspond to the variables that are in the clinical dataset but not in the external dataset, and $z_{i(k+1)}, ..., z_{ip}$ correspond to the variables that are in the external dataset but are not in the clinical dataset.
Our goal is to predict the unobserved values of the variables $z_{k+1}, ..., z_q$ for the patients in the clinical dataset. As mentioned above, this process is done in two steps: matching and imputation.

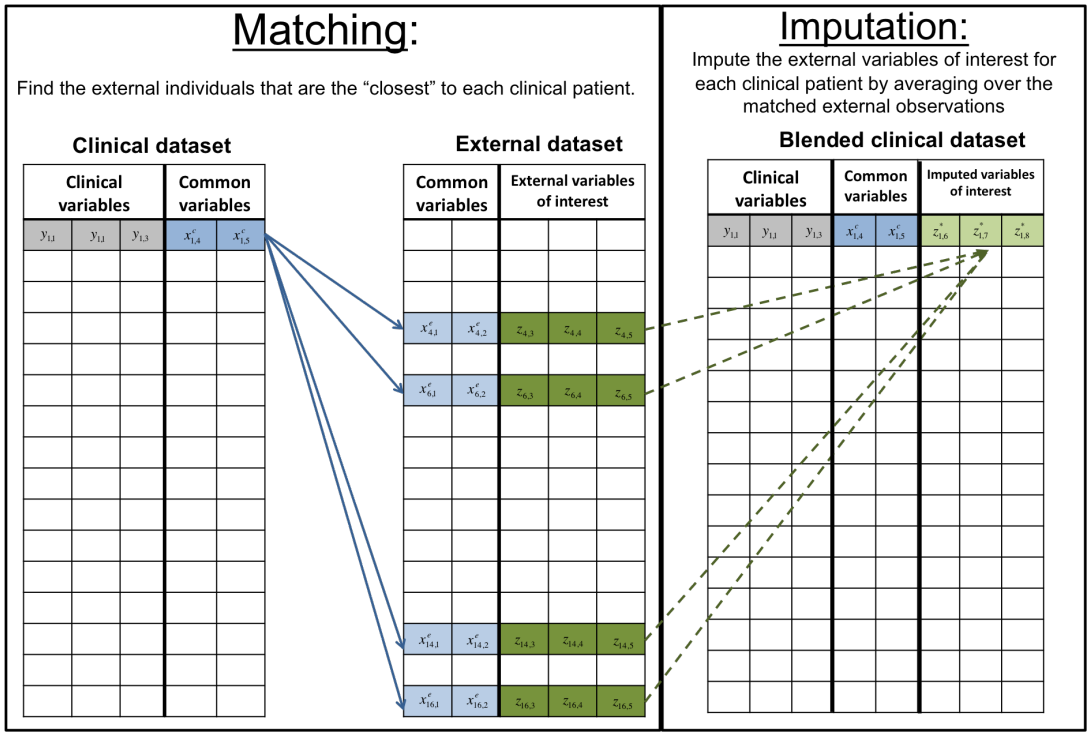

Figure 1: The data blending process for individual-level blending. For each individual in the clinical dataset, we must identify "matched" individuals in the external dataset that are the most similar based on the common variables for both datasets. Then the external variables of interest are imputed for the clinical dataset based on the matched individuals in the external dataset

3.1.1 The matching step

To perform matching, we must be able to define a measure of distance between observations in the clinical and external datasets based on the variables that are common to both. We can define the common variables for an observation in the clinical dataset by $\mathbf{x}^c_i = (x^c_{i1}, ..., x^c_{ik})^T$ and the common variables for an observation in the external dataset by$\mathbf{x}^e_j = (x^e_{j1}, ..., x^e_{jk})^T$. The distance between an observation i in the clinical dataset and an observation j in the external dataset, will be defined based on the common variables only, that is

$$d(\mathbf{y}_i, \mathbf{z}_j) = d(\mathbf{x}^c_i, \mathbf{x}^e_j)$$

where we can define the distance metric, d, in a number of ways, such as the euclidean distance metric:

$$d_{euclidean}(\mathbf{x}^c_i, \mathbf{x}^e_j) = \sum_{l=1}^{k} (x^c_{ij} - x^e_{jl})^2,$$

the Mahalanobis distance metric:

$$d_{mahalanobis}(\mathbf{x}_i^c, \mathbf{x}_j^e) = \sqrt{(\mathbf{x}_i^c - \mathbf{x}_j^e)^T S^{-1} (\mathbf{x}_i^c - \mathbf{x}_j^e)},$$

where $S = cov(\mathbf{x}_i^c, \mathbf{x}_j^e)$ is the covariance matrix, the correlation metric:

$$d_{corr}(\mathbf{x}_i^c, \mathbf{x}_j^e) = 1 - corr(\mathbf{x}_i^c, \mathbf{x}_j^e)$$

or any of the large number of other distance metrics.
For a clinical observation, $\mathbf{y}_i$, we define a *matched* external observation, to be any $\mathbf{z}_j$ such that $d(\mathbf{y}_i, \mathbf{z}_j) = d(\mathbf{x}_i^c, \mathbf{x}_j^e) < \lambda$ for some threshold, λ, which could be chosen by a number of methods such as cross-validation to find the value of λ that yields the lowest prediction error.

3.1.2 The imputation step

Next, having found matched external observations for each clinical observation, we can impute the values of the external variables of interest that were not present in the clinical dataset. More specifically, suppose that for a clinical observation $\mathbf{y}_i$, we have t matched external observations, $\mathbf{z}_1, ..., \mathbf{z}_t$. We could impute the external observations of interest for the clinical observation by defining $z^*_{i(k+1)}, ..., z^*_{iq}$ by

$$z^*_{ir} = \frac{1}{t} \sum_{j=1}^{t} z_{jr}$$

for continuous variables, or by majority vote of the z_{jr} for $j = 1, ..., t$ for categorical variables. Thus, we now have a blended clinical dataset consisting of the blended clinical observations;

$$\mathbf{x}_i^* = (x^c_{i1}, x^c_{i2}, ..., x^c_{ik}, y_{i(k+1)} ..., y_{ip}, z^*_{i(k+1)}, ..., z^*_{iq})^T$$

which contains all of the original observations in $\mathbf{x}_i$ along with the imputed observations based on the external matched observations.
An alternative approach which does not involve deciding on a threshold, λ, for matched observations, is to perform *weighted imputation*. To perform weighted imputation, instead of simply averaging over the external observations most similar to the clinical observation of interest, we calculate a weighted average over all of the external observations, where the external observations "closest" to the clinical observation get the highest weights and those furthest get the lowest weights:

$$z^*_{ir} = \frac{1}{m} \sum_{j=1}^{m} \frac{z_{jr}}{d(\mathbf{y}_{ir}, \mathbf{z}_j)}$$

3.1.3 Assessing reliability

Having performed the data blending, how can we tell that it is (a) stable (do not have extremely large variance) and (b) meaningful (i.e. close to the values of the external covariates that the clinical observations would have obtained). We here propose a number of tests that can be used to assess the reliability of the imputed values. First, we could compare the values imputed using a similar, but different dataset (for example, if we had used to psychiatric diseases dataset, we could compare with the values we obtained if we had imputed using the obesity dataset). Next, we could assess how the blending method performed when imputing a common variable, and assessing how close the imputed values are to the true values. We could also compare the results of different distance metrics, as well as different matching threshold values. Further, we could assess stability by randomly witholding subsets of the data and assessing how much the imputed values and predictions change (ideally, there will be little difference).

3.2 Region-level data blending

For the air-quality and weather datasets, we have only data at the zip-code level. As a result, the blending is simple: for each clinical observation we impute the values of the air-quality/weather variables to be the values that correspond to the zip-code region in which they live or which contains the medical center that they attended.

4 Prediction

Recall that our primary goal is to be able to predict undiagnosed COPD amongst patients who are diagnosed with pneumonia. The first thing that we woudl like to identify is whether there are subpopulations whose COPD status can be effectively predicted using the clinical data alone. Our approach will focus on, not improving the prediction for the overall population, but specifically idenftifying whether the incorporation of external information can improve the prediction accuracy on those whose COPD status cannot be predicted using clinical data alone.

4.1 Prediction methods

Our prediction approach, comes from two different angles. The first is to use standard prediction approaches (such as random forest, support vector machines, linear discriminant analysis) using the extra external variables as additional predictors. However, this approach does not utilize the longitudinal information contained in the dataset at all; particularly the fact that the dataset contains repeated observations from the same individuals over time. Further, these individuals are nested within hospitals.

Thus, the second approach that we would like to take is somewhat more complex because we would like to utilize the longitudinal features of the dataset. For instance, we might like to fit a logistic regression model where the coefficients are fit using generalized estimating equations (GEE) if we are interested in a population interpretation of the model coefficients, or mixed effect models if we are interested in an individual-level interpretation of the model coefficients. The primary advantage of using these longitudinal methods is that we can specify, and thus make us of, the correlation structure that exists between the observations (at the individual and the hospital level).

Further, the utilization of longitudinal approaches allows us to identify changes over time that might contribute to evidence that one has COPD that is not captured by considering a cross-sectional view of the data. There exist a number of other classification methods for longitudinal data that we would also like to try, such as extensions of functional data analysis for random trajectories to binary regression models for longitudinal data [Müller, 2005] as well as quadratic inference function classifiers [Wang and Qu, 2014].

4.2 Feature selection

One of our primary interests in the identification of which external variables are useful in predicting undiagnosed COPD. We propose to perform various feature selection methods (to identify the most informative features, rather than simply throwing in all available features). These include stepwise (sequential) feature selection approaches for classification [**?**], which can be combined with cross-validation to identify the set of features that minimize the predictive accuracy.

4.3 Evaluation metrics

Prior to our prediction analysis, we intend to withold a subset of the data to act as a test set. Once we have decided on our final model(s) which were build using the remainder of the data (the training set) we will evaluate its effectiveness based on how well it predicts COPD on the witheld test set based on a number of measures including the overall classification accuracy, the within-class accuracy (the average of the accuracy for each class), the log-loss to guage a sense

of entropy of classification and help tune model to minimise cross entropy, and the Area Under Curve (AUC) and Receiver Operating Characteristic (ROC).

5 Conclusion

For this project, our aim is to develop methods capable of effectively predicting cases of undiagnosed COPD amongst those whose primary reason for hospitalization was pneumonia. We aim to incorporate external data sources primarily related to relevant socioeconomic and climate factors that are not captured by the clinical records using a process called "data blending".
Our proposed requirements for the external datasets are that they are publicly availible to Geisinger, obtained from a trustworthy source, contain no personally identifiable information (PII) on patients and are representative of the Pennsylvania population of interest.
For blending purposes, the external data must contain combinations of the common variables - age of person, gender, zipcode and date. We have propose a two-stage blending method. The first stage is a matching stage which involves, for each individual in our clinical dataset, identifying individuals in the external dataset that are most similar in terms of the covariates common to both datasets. The second stage is an imputation stage which involves imputing or predicting the external variables of interest for the clinical patients by aggregating the values obtained by the matched external individuals.
In order to ensure that the data is relevant for modeling purposes we intend to blend features which will help identify the underlying causes of COPD/ pneumonia. We propose using publicly availible datasets such as weather and pollution data which can be combined using zipcode and date. Furthermore we propose that we utilise deidentified Geisinger patient-level insurance data to obtain a rich longitudinal dataset caputuring socioeconomic indicators which may be relevant for COPD classification.
For predictive modeling we aim ti utilise a combination of modelling techniques which employ supervised machine learning (logistic regression, random forests etc) and also established longitudinal data analysis methods to leverage the temporal component of the panel data collected on patients. The evaluation of these models will be undertaken by utilising several common classification metrics (confusion matrix , classifcation rate etc) as well as a stepwise approach to assessing the impact of the features included from the external data sources.

Appendix: External data sources

Weather data from Pennsylvania
Data source: `http://www.ncdc.noaa.gov/data-access`

Air pollution data from Pennsylvania
Data source: `http://www.dep.pa.gov/Pages/default.aspx#.Vkte0a6rSso`. However the access is restricted for the general public. Geisinger may be able to request access to this from the Pennsylvania government for use in this project.

Appendix: Other external data ideas

Despite these challenges this we searched for several external data sources that might fit the identified requirements above. The most trusted publicly available data sources we could identify are significantly less granular than required given the above issues faced. They are listed below with brief description and key issues in utilising them for the blending process:

- IRS Income Statistics Data by Zipcode
  - source: `https://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-ZIP-Code-Data`

  - This is usefult to analyse change in income levels over time. Potentially useful as an indicator of stress (e.g. low income high unemployment trend) which could be a contributing factor to COPD.

- Pennsylvania Smoking Rankings by Zipcode
  - source: `http://www.countyhealthrankings.org/app/pennsylvania/2015/measure/factors/9/data`
  - This is only at zipcode level
  - Potentially useful to help distinguish patients with higher risk of COPD driven by smoking.
  - This is potentially useful if we blend using patient-level zipcode to get the maximum level of variation from this metric in our classification methodology

- Additional Income Level Data by Zipcode from Qubit Consulting
  - source: `https://www.incomebyzipcode.com/search/`
  - This is publicly availible but difficult to verify credibility. It may be a useful cross-validation against the IRS income statistics data as identified above

- Housing Vacancy Data in Pennsylvania
  - This is split by Gender and zipcode separately unfortunately, not as a composite gender-zipcode summary
  - There is only an overall gender distribution by zipcode provided which may be useful to help redistribute other zipcode-level metrics (e.g. population metrics) by gender.

- Climate and Temperature Data
  - This is split by Zip code and has a Temporal component for blending (by month)
  - This could be a useful to identify longitudinal trends in temperature and linking them to identifying any true pneumonia (non-COPD) cases. The hypothesis here is that colder temperatures may lead patients to suffer from pneumonia and help refine the classification
  - To source individual months may be a bit difficult and scraping would need to be done given the lack of API for downloading the data

- Census Fact Finder
  - source: `http://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml`
  - Has thorough census information which can be searched by PA zipcode
  - There is also a distribution by Gender which could be potentially be used to re-distribute other metrics collected at the zipcode level by gender

- Flu Trends Data
  - source: `https://public.tableau.com/views/s14_15_20150522/Forecasts?%3Aembed=y&%3AshowTabs=y&%3Adisplay_count=no&%3AshowVizHome=no`
  - This is only availible for main PA cities i.e. Philadelphia, Pittsburgh and State College
  - Previously part of the Google Flu Trends Project (now discontinued by Google). Could be potentially used as a crude indicator of the likelihood of pneumonia alone occurring (pneumonia can occur as a complication of the flu virus).

References

[Lozano et al., 2012] Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S. Y., Alvarado, M., Anderson, H. R., Anderson, L. M., Andrews, K. G., Atkinson, C., Baddour, L. M., Barker-Collo, S., Bartels, D. H., Bell, M. L., Benjamin, E. J., Bennett, D., Bhalla, K., Bikbov, B., Bin Abdulhak, A., Birbeck, G., Blyth, F., Bolliger, I., Boufous, S., Bucello, C., Burch, M., Burney, P., Carapetis, J., Chen, H., Chou, D., Chugh, S. S., Coffeng, L. E., Colan, S. D., Colquhoun, S., Colson, K. E., Condon, J., Connor, M. D., Cooper, L. T., Corriere, M., Cortinovis, M., de Vaccaro, K. C., Couser, W., Cowie, B. C., Criqui, M. H., Cross, M., Dabhadkar, K. C., Dahodwala, N., De Leo, D., Degenhardt, L., Delossantos, A., Denenberg, J., Des Jarlais, D. C., Dharmaratne, S. D., Dorsey, E. R., Driscoll, T., Duber, H., Ebel, B., Erwin, P. J., Espindola, P., Ezzati, M., Feigin, V., Flaxman, A. D., Forouzanfar, M. H., Fowkes, F. G. R., Franklin, R., Fransen, M., Freeman, M. K., Gabriel, S. E., Gakidou, E., Gaspari, F., Gillum, R. F., Gonzalez-Medina, D., Halasa, Y. A., Haring, D., Harrison, J. E., Havmoeller, R., Hay, R. J., Hoen, B., Hotez, P. J., Hoy, D., Jacobsen, K. H., James, S. L., Jasrasaria, R., Jayaraman, S., Johns, N., Karthikeyan, G., Kassebaum, N., Keren, A., Khoo, J.-P., Knowlton, L. M., Kobusingye, O., Koranteng, A., Krishnamurthi, R., Lipnick, M., Lipshultz, S. E., Ohno, S. L., Mabweijano, J., MacIntyre, M. F., Mallinger, L., March, L., Marks, G. B., Marks, R., Matsumori, A., Matzopoulos, R., Mayosi, B. M., McAnulty, J. H., McDermott, M. M., McGrath, J., Mensah, G. A., Merriman, T. R., Michaud, C., Miller, M., Miller, T. R., Mock, C., Mocumbi, A. O., Mokdad, A. A., Moran, A., Mulholland, K., Nair, M. N., Naldi, L., Narayan, K. M. V., Nasseri, K., Norman, P., O'Donnell, M., Omer, S. B., Ortblad, K., Osborne, R., Ozgediz, D., Pahari, B., Pandian, J. D., Rivero, A. P., Padilla, R. P., Perez-Ruiz, F., Perico, N., Phillips, D., Pierce, K., Pope, C. A., Porrini, E., Pourmalek, F., Raju, M., Ranganathan, D., Rehm, J. T., Rein, D. B., Remuzzi, G., Rivara, F. P., Roberts, T., De León, F. R., Rosenfeld, L. C., Rushton, L., Sacco, R. L., Salomon, J. A., Sampson, U., Sanman, E., Schwebel, D. C., Segui-Gomez, M., Shepard, D. S., Singh, D., Singleton, J., Sliwa, K., Smith, E., Steer, A., Taylor, J. A., Thomas, B., Tleyjeh, I. M., Towbin, J. A., Truelsen, T., Undurraga, E. A., Venketasubramanian, N., Vijayakumar, L., Vos, T., Wagner, G. R., Wang, M., Wang, W., Watt, K., Weinstock, M. A., Weintraub, R., Wilkinson, J. D., Woolf, A. D., Wulf, S., Yeh, P.-H., Yip, P., Zabetian, A., Zheng, Z.-J., Lopez, A. D., Murray, C. J. L., AlMazroa, M. A., and Memish, Z. A. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet (London, England)*, 380(9859):2095–2128.

[McCoy and MS, 2008] McCoy, K. and MS (2008). Health Myths Center.

[Müller, 2005] Müller, H.-G. (2005). Functional Modelling and Classification of Longitudinal Data*. *Scandinavian Journal of Statistics*, 32(2):223–240.

[NIH, 2010] NIH (2010). Chronic Obstructive Pulmonary Disease. Technical report.

[Wang and Qu, 2014] Wang, X. and Qu, A. (2014). Efficient classification for longitudinal data. *Computational Statistics & Data Analysis*, 78:119–134.

[WHO, 2014] WHO (2014). WHO | Household air pollution and health.