

Nat. Hazards Earth Syst. Sci. Discuss., author comment AC1
<https://doi.org/10.5194/nhess-2021-389-AC1>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply to RC1

Lennart Marien et al.

Author comment on "Machine learning models to predict myocardial infarctions from past climatic and environmental conditions" by Lennart Marien et al., Nat. Hazards Earth Syst. Sci. Discuss., <https://doi.org/10.5194/nhess-2021-389-AC1>, 2022

Comment 1:

This reviewer is a statistician familiar with the usual epidemiological methods for investigating variation of occurrences of adverse health outcomes over time (usually days) and their application, but with limited familiarity with machine learning methods.

This paper applies machine learning (ML) methods to obtain a prediction model for MI daily incidence counts from measures of weather and air pollution on the same and three previous days, and demographic data for that year. There is a very substantial literature on dependence on weather and air pollution of variation of occurrences of adverse health outcomes over time (usually days) and on a variety of methods for doing so. However, I have no reason to doubt the authors statement that very few of these have applied ML methods. I had only seen the Zhang 2014 paper cited in this paper, and was eager to see another application.

I learned things from the paper, and was impressed with several aspects of the work. I was unpersuaded that it demonstrated ML was indeed useful in this context, at least from the current analysis and its description, but perhaps that's irrelevant. However, I do believe that my first and most important comment needs addressing before almost any of the results can be usefully.

Response: We thank the reviewer for taking the time to do this thorough review, and for these positive comments, and welcome the suggestions made. We address these one by one, below.

Comment 2:

Most important issue: In all the studies I know of environmental predictors of variation in health outcome over time, control for seasonal and other long term temporal trends is included in the model. This is because these are typically strong predictors, even if demographic changes are allowed for, and otherwise confound the association between environmental variables and outcome. (This includes some publications with several of the same authors as this paper.) This paper appears not to have done so (sorry if I missed it). Assuming not, I could not know how much the importance of the predictors included merely reflected their association with trend or season. For example, given the steep trend in MI counts over the duration of the study (figure 1) I wondered for example if the apparent importance of air pollution might in part be due to a trend in pollution concentration (true of many pollutants in most western European locations).

Perhaps the usual methods for control do not fit easily into the ML framework, but I notice that Zhang (2014) managed by initially discounting the expected counts based on season and year ("To define the generally expected level of daily mortality counts, we modelled mortality counts as a smooth function (a cubic spline) of day of the year(degrees of freedom =5) while adjusting for day of week and year over the time period of our study (1998–2006)."

Response: We thank the reviewer for this important comment. Controlling for seasonal and long-term trend is indeed a common approach in environmental studies. However, our goal in this study was to adopt a purely data-driven approach, i.e., one where none or very few preconceptions of underlying mechanisms (causal or otherwise) are assumed a priori. Instead, we feed the data as-is into the ML algorithms to see what they can learn from it. Therefore, picking up seasonal signals is essentially part of the overall design of the study.

While the Zhang 2014 paper does correct for such effects by subtracting a baseline for expected mortality, its goal was not to do actual forward projections of the mortality, but instead to identify what predictors are most likely to lead to excess mortality. If forward projections of future mortality were to be carried out on that basis, the subtracted baseline would have to be added again to arrive at the final predictions. The same holds true for our study design. In our case we therefore decided that letting the algorithms decide on how to deal with seasonality, rather than constraining their potential by presupposing any trend/seasonality statistics. Another aspect of this is that our study aims at providing a first step towards projecting MI occurrence under climate change. At such timescales (at least 30 years) seasonal trends may change gradually which would not be reflected by any fixed trend derived from current or historical data. Therefore, extracting the trend from the data alone is a key aspect of our study.

We suggest to address this issue in the revised paper by more clearly outlining this limitation of the study in the introductory section. This includes a more clear and concise description of our reasoning behind choosing data-driven approach and application to the timeseries data.

Comment 3:

The methods description might well be clear to readers well familiar with ML approaches, but if it is designed to be accessible to others some “unpacking” would I think be useful. In particular, the meaning of the “importance” measures of each variable. Zhang 2014 explained this a bit, but even there I was not quite sure. I think it is a measure of reduction in prediction accuracy if the variable in question is dropped from the algorithms’ consideration. But if that is right won’t the measure be highly dependent on how much remaining variables are associated with the omitted variable? For example apparent and dry-bulb temperature are typically very highly correlated, and absolute humidity pretty highly correlated with both, so dropping one while leaving others in will give a misleading picture of any one variable’s importance.

Response: Thank you for the comment. We do agree that a brief but clear description of variable importance is missing from the draft, which we will revise. The notion of variable importance differs depending on context. For some models, such as the linear regression and its variants, these simply relate the magnitude of the trained weights (coefficients) of the model to their associated predictors. In such cases care must be taken to consider the relative magnitudes of the predictors, but this has been addressed in our study by scaling the input data. In other cases, such as the decision tree and variants, the variable importance is based on the mean impurity decrease with respect to splits based on a given predictor, i.e., by how much the squared sum of deviations from the mean decreases for the observations within the two resulting nodes created by the split. Additionally, there are other measures that work by systematically dropping features and observing the decrease in prediction accuracy. In this study however, we did not use these, but rather relied on the built-in importances of the models as described above.

We suggest to address this in the paper by outlining the variable importance used for each of the regression methods applied.

Comment 4:

No descriptive data is given for the outcome data. It would be usual and is I believe useful to give measures like daily mean, SD, min and max counts overall and for each region.

Response: We agree that such a table would give the reader the opportunity to more easily get an overview of the results at first glance. We will add the proposed information in a table in the appendix of the paper.

Comment 5:

Readers would be further informed if the comparisons of prediction accuracy could include a conventional (a priori selected predictors) time series regression model. For example that used by Chen (Eur H J 2019, with some of the same authors as the current paper). Also, for the comparisons of annual predictions, it would be very useful to state how much prediction accuracy did ALL the environmental variables add to that with just the demographic variables.

Response: We thank the reviewer for bringing up these two important issues. We believe that conducting a thorough time series analysis based on distributed lag non-linear models is out of the scope of the current paper. While we believe that a direct comparison of the methods on the same data would be very insightful, we see it as a potential step for further research. Instead we would like to address this first issue by adding a comparison of the methods to the paper, outlining differences as well as similarities of both approaches.

For the second issue, we propose to run an additional experiment with all demographic predictors turned off and to compare in the revised paper the prediction accuracy to the original results. In the interest of speed and brevity we would propose to do this only for the overall population.

Comment 6:

I found the paper much longer than I thought necessary to make it's main points, and believe its length will put off many readers.

Response: We also agree that the draft could be made more concise. We propose to shorten it substantially, without taking away the main points of the paper.

Comment 7:

I am aware of a couple of reviews of the temporal association of MI with weather and pollution, which could I think be usefully cited. Sun Env Poll 2018 and Mustafie JAMA 2012.

Response: We thank the reviewer for making these two good suggestions. We propose to reference these in the appropriate sections of the paper.