Comment on nhess-2021-389
Anonymous Referee #1

Referee 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-RC1, 2022

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 interpreted.

- 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)."")

- 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.
- 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.
- I thought that particular strengths of the work were the comparison of several different ML methods, and the careful use of data splitting to get unbiased measures of prediction performance in test data.
- 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.
- 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.
- 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.