**Possible points:**

**EDA:**

* More samples missing an associated MMR estimate post 2011 and especially post 2015 for high and upper-middle income countries
* For example, the dataset for the lowest income level countries was missing 58% of data regarding ‘women participating in own health care decisions (% of women age 15-49)’ while the dataset for the highest income countries was missing 99.9% (see Table 4). This feature was sourced from a Demographic Health Survey (DHS), which is normally collected in lower income countries, explaining the data disparity. This is a good trend to be aware of when analysing feature importance for features measured more frequently using DHS data.
  + Increasing standard deviation could show less uniform quality of care and health outcomes
* According to Table 10, the national MMR estimates were subject to large outliers, as the mean values were larger than the median values for all income levels. Additionally, the standard deviation for the MMR estimates was large. The difference between mean and median, as well as the magnitude of the standard deviation, decreased as income level increased. This indicates that outliers will be more common for lower income countries, as they have a wider variety of possible MMR values than higher income countries.
* The merged input data was projected onto its first two principal components for better visualisation of patterns and clusters in the data (See Figure 11). The dense cluster of data in the bottom left of Figures 11a and 11b correspond to the countries with the lowest MMRs and highest income levels. A country’s income level tended to decrease and its MMR tended to increase travelling up and to the right of Figures 11a and 11b, indicating an relationship between the two trends. There was no clustering by year (see Figure 6c), potentially due to heterogeneity in countries’ MMR ratios at time points (see Figure 3). The greater influence of income level on MMR compared to year was shown in Figure 3.

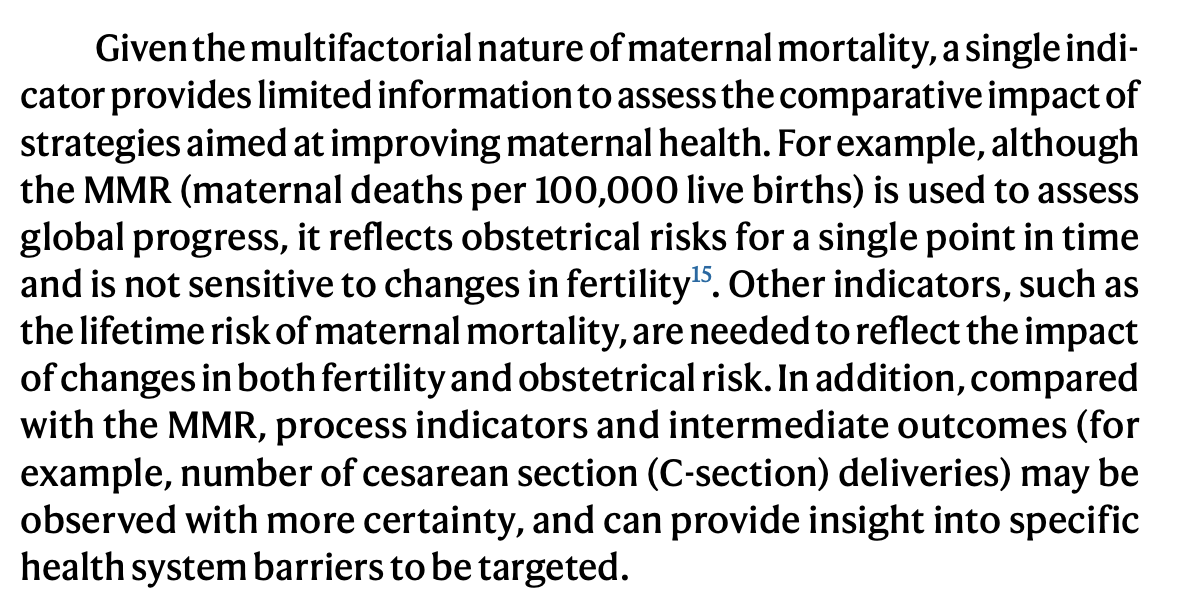
Base estimator performance

* Lower performance on the 0.8 correlation feature subset
* Not hugely different performance on the other feature subsets

Different ensembles

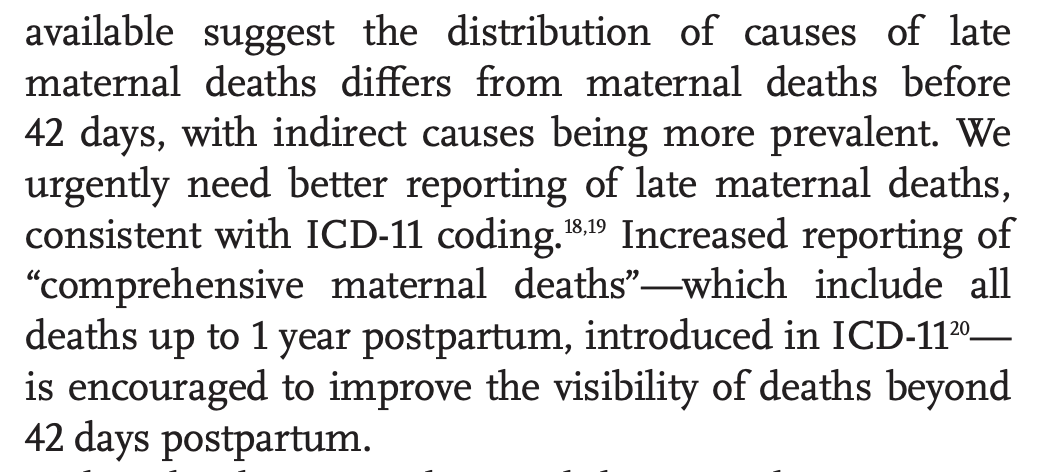
* Difference between the mse and mre metrics between which pa and mda best model (elastic net versus rf)
* Voting doing worse for mse mda than mse pa
  + Potentially could work less well for predictive if outliers for predictive majorly affected a highly weighted model

Sensitivity analysis:

* Mse sensitivities greater than original
* Potentially due to datapoints that would be considered outliers for that income-level, but not for the dataset as a whole.
* Explore semi-supervised learning to instead be able to use datapoints with missing MMR data
* Selective use of features in ensembles
* Uncertainty in ensembles
  + Especially important given known wide uncertainties in literature
* Use of high quality base data?
* Overfitting to low income data due to small sample size
  + Consideration of smote
* Why no imputation
  + Possible future extension
* Did not consider under-reporting in model
* Possibly extend framework to more specific causes and/or lifetime risk due to limitations of the mmr measurement itself
  + 
    - <https://www.nature.com/articles/s41591-023-02310-x>
  + Especially with respect to late maternal deaths

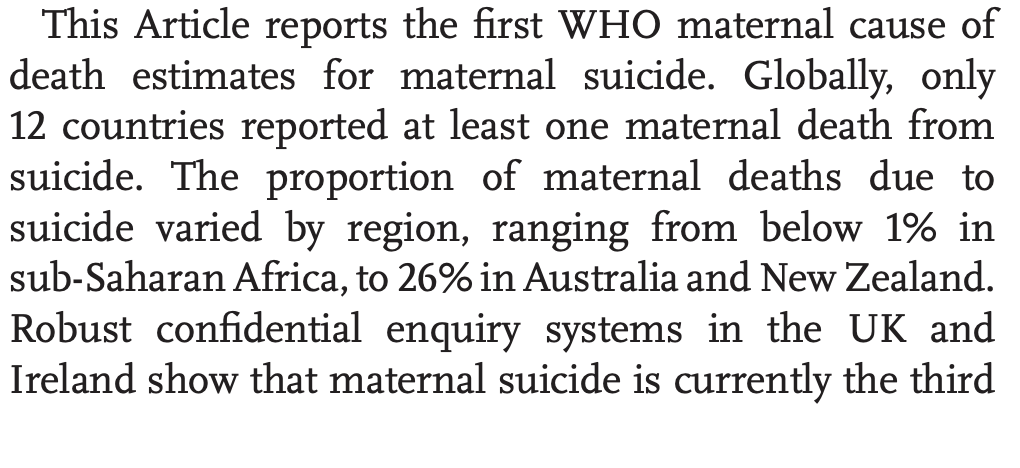
A close up of text

AI-generated content may be incorrect.



<https://www.clinicalkey.com.au/service/content/pdf/watermarked/1-s2.0-S2214109X24005606.pdf?locale=en_AU&searchIndex=>

* + Could also use it to expand cause of death for maternal deaths due to suicide



A close-up of a text

AI-generated content may be incorrect.

* + - same article as above point
* Primarily, this study relied exclusively on data from the World Bank database, which, though extensively used and recognized for its reliability, may not wholly encapsulate the intricacies and nuances of MMRs across the diverse African continent. The reliability and completeness of the World Bank’s MMR data could potentially vary across countries and regions within Africa due to the disparities in data collection, reporting standards, and healthcare infrastructure. Consequently, our analysis might be subject to some degree of imprecision.
  + Also use as a limitation for my study
  + ARIMA Africa study’
* Interesting extention: what if we used completely unsupervised machine learning to do cluster analysis
  + Lose fine-grained, but will potentially gain in not needing data subject to misreporting?
* Other studies finding that missing data imputation would be better than just splitting
  + <https://www.sciencedirect.com/science/article/pii/S0020025515001838#s0085>

reason for using Bayesian methods was to incorporate uncertainty, which I do not do

* <https://pmc.ncbi.nlm.nih.gov/articles/PMC4196216/#S5>

<file:///Users/rosalitarosenberg/Downloads/16-AOAS1014%20(2).pdf>

A table with numbers and text

AI-generated content may be incorrect.

<file:///Users/rosalitarosenberg/Downloads/16-AOAS1014%20(2).pdf> (2017 un mmeig Alkema paper)

* The second limitation of our study is due to the limited information on AIDS maternal mortality (Wilmoth et al. 2012), which complicates the reconstruction of trends in maternal mortality in countries with generalized HIV/AIDS epidemics. This limitation is of lesser concern for more recent years in which the contribution of AIDS maternal deaths to the overall number of maternal deaths has decreased. Lastly, because of the dependency of the maternal mortality estimation on the estimation of all-cause deaths to women of reproductive ages as well as the number of births, the challenges and limitations that apply to the estimation of these demographic indicators are also applicable to maternal mortality estimation. We did not include the uncertainty surrounding these demographic indicators into the uncertainty assessment for maternal mortality because such uncertainty assessments are generally not available. Future research work regarding the assessment of uncertainty in these demographic indicators may result in the reporting of uncertainty intervals for a wider range of demographic indicators and allow for a more complete uncertainty assessment for maternal mortality
  + This could contribute to explaining why my results are not within the bounds

1. [**https://www.sciencedirect.com/science/article/pii/S2589537024002323**](https://www.sciencedirect.com/science/article/pii/S2589537024002323)
   1. Using our individual-level structural model of maternal health which allows for flexible aggregation of model outcomes, we leverage information on observed subgroup differences in intermediate factors to estimate differences in maternal health outcomes. Although we do not have subgroup-specific estimates of these outcomes, we do have subgroup-specific parameters that yield model predictions consistent with empirical data for women overall.3 This approach could be generalized to other topics where only marginal (overall) outcomes are observed, but subgroup-specific information on intermediate factors are available. Although empirical data are often lacking, and there is typically wide uncertainty around estimates that are available, our modelling approach can be used to examine trends in maternal health indicators by subgroup and the potential impact of policies to improve health equity
      1. Motivating my approach

<https://pmc.ncbi.nlm.nih.gov/articles/PMC5674830/pdf/40748_2017_Article_59.pdf>

* Maternal mortality is much higher in developing countries compared to developed nations owing to lack of adequate medical care; high prevalence of infectious diseases,higher total fertility rate and due to health care system difference. Countries with high maternal mortality ratio have less reliable vital statistics registry system; as a result level of maternal mortality is usually underestimated and little information is available regarding locally specific risk factors for maternal death [2, 3].
* or developing countries, neither antenatal care nor the presence of a skilled birth attendant was related to maternal mortality until 40 to 60% coverage was achieved; only then did mortality decrease. In contrast, there was a clear decrease in maternal mortality as cesarean section rates increased from 0 to about 10%, with little change thereafter.

A table of statistics with numbers and text

AI-generated content may be incorrect.

* difference between test set and training set (especially for split by year)
* priors used by literature (gbd different definition of maternal mortality)
  + but may not have a difference due to low number of late maternal deaths (contrast this with lancet paper)
* feature correlations can make ‘feature importance’ scores from my models less clear, as one feature may be easily substituted for another
* overall purpose of the research is to predict, allowing causal mechanisms to be further developed by others? (versus trying to do everything)
* following gather framework for presenting method as a strength?
  + <https://pmc.ncbi.nlm.nih.gov/articles/PMC4924581/>
* Lack of samples about low income makes it difficult to extrapolate performance quality

<https://www.sciencedirect.com/science/article/pii/S2589537025004389?via%3Dihub>

* In contrast, estimates of maternal death by cause vary substantially across the GMatH and GBD models, except for deaths from abortive causes, hypertensive disorders, and hemorrhage. Globally, the GMatH model estimates higher maternal deaths from sepsis, obstructed labour, other direct, indirect causes, and late maternal deaths. Women who experience these complications may be more likely to face delays in obstetric care, with resulting maternal deaths less likely to be accurately captured in traditional data sources, highlighting the importance of adjusting for underreporting by location of death (e.g., home vs. facility) in the GMatH model.
* Future study?
* Other types of data (e.g., MMR, pregnancy-related mortality) must first be converted (via a series of assumptions and calculations) to the PM, which introduces additional assumptions (and potential errors) outside of the modelling framework, and which may explain differences in model estimates for countries with only pregnancy-related mortality estimates. In contrast, structural models such as the GMatH model can directly incorporate various types of indicators (without pre-processing assumptions/transformations) when calibrating to empirical data
  + Strength of mine too
* Note that there are differences in the definitions for these indicators across models with respect to age range and the inclusion of late maternal deaths. GMatH (v2023) considers maternal deaths at all ages, and includes late maternal deaths in estimates of total maternal deaths but not in the MMR (consistent with the WHO definition of “maternal death”).3 UN (v2023) generally considers women of reproductive age as 15–49, and does not include late maternal deaths in any estimates.2 In contrast, GBD (v2021) includes late maternal deaths in all maternal mortality estimates and considers ages 10–54 years.
* Although these models can provide insight into relative trends, the maternal mortality estimates (both global and countrylevel) from these models often differ (sometimes substantially), which can result in confusion, mistrust in reported results, and uncertainty about the effects of policies
  + Make sure to put this in aim/introduction
* Because we model the entire reproductive life course of individual women, we can integrate various data from multiple sources into the model. We also account for uncertainty around all model parameters across the reproductive life course, including accounting for uncertainty around underreporting of maternal deaths, which are reflected in our uncertainty intervals. This approach provides the potential for more robust estimates of maternal health outcomes as we can incorporate data for multiple indicators along the reproductive life course that are generally more accurately observed than maternal mortality outcomes.
  + Justification for my use of multiple datasets
  + Despite these similarities, Ward et al. (2025) found large variation across estimates from the 3 models for certain countries, including Nigeria and Afghanistan [49]. The inter-model variation is likely a result of their different methodologies and input datasets. For example, Ward et al. (2025) noticed that the differences between the models’ estimates was often greatest when the country only had survey-based data about pregnancy-related mortality available [49]. In contrast to maternal mortality, the cause of pregnancy-related death does not need to be related to pregnancy, childbirth, or termination [49]. To be used in the BMat and CODEm models, pregnancy-related mortality must be converted into a model-recognisable metric using a series of calculations and assumptions [49]. In contrast, the pregnancy-related mortality data can be used directly with the GMat model without pre-processing [49]. Ward et al. (2025) attributes this methodological difference with the inter-model variation in MMR estimates [49].
* The variation in MMR estimates can produce confusion and uncertainty about the type of policy that should be implemented [49]. Consequently, the authors of the GMatH model describe their hope that their intrinsically different modelling approach could provide insight into the reason why the models results have diverged [9].

A limitation of all three models is their high computational complexity, as they each require significant calibration on large quantities of data [25, 26]. This computational effort is compounded by the need to not only develop the sophisticated statistical models, but also to consider the various sources of uncertainty and transform low quality data into a usable form. As described above, the MMEIG developed the BMis model entirely for this purpose and the GBD Study estimates were produced after using complex algorithms to re-label unspecific or incorrect causes of death within the input data [2, 4, 8]. The GMatH model introduced separate parameters solely to model site-specific underreporting of maternal deaths [9]. These processes are computationally intensive.

* Say that mine is just as bad

Therefore, pandas correlation calculation method may behave strangely for columns with a lot of missing data, where the lack of data may make it look like there is either a false strong or weak correlation between the feature and MMR measurement.

Groupkfold method ensures same group member in only one validation set, creating larger difference between the folds and thus more variation int eh base estimators

* Could improve ensemble performance

Look at correlation between the metrics, and can use later as a discussion point about comparative model performance

**SMOTE Discussion**

|  |  |  |  |
| --- | --- | --- | --- |
| **Income Level** | **Number of Samples** | | **Proportion of Samples Remaining (%)** |
| **Before Removing Samples with MMR Missing** | **After Removing Samples with MMR Missing** |
| **Low** | 884 | 78 | 8.8 |
| **Lower-Middle** | 1734 | 310 | 17.88 |
| **Upper-Middle** | 1802 | 996 | 55 |
| **High** | 2176 | 1405 | 65 |

*Problems caused by small proportion of low-income samples*

* Lack of generalisation to low-income countries, who would benefit most from this technology
* Performance measures may not reflect poor performance on low-income samples due to their relative scarcity
  + Although this may be somewhat balanced by the larger magnitude of the errors, as the low-income samples tend to have much larger MMR estimates.

*Oversampling by randomly replicating low occurrence datapoints*

* Can cause overfitting given the small number of cases covered by the low occurrence datapoints, with replication potentially amplifying noise

*Synthetic minority oversampling technique (SMOTE)*

* Synthetic samples are generated using interpolation on the K nearest neighbours of low occurrence instances.
* Limitations:
  + Can generate overlapping and noisy samples.
    - This is particularly important in my dataset, as the PCA shows that the MMR estimates for the upper-middle, lower-middle, and low-income classes overlap. This means that the neighbours of low-income samples could actually be from another income level, meaning that the generated sample would not be representative of the low-income level samples.
  + Prior work has found that applying SMOTE on high-dimensional data produces no real benefit in classification performance.
    - Lower performance of SMOTE for high-dimensional data is partially attributed to hubness, where the same small subset of datapoints are more frequently chosen as the neighbours, and thus are used to generate a higher proportion of the synthetic samples. This can bias the synthetic samples, especially if the datapoints in the hub are unrepresentative.
    - As a result, many methods propose use of feature selection or dimensionality reduction techniques before applying SMOTE.
    - Particularly limitation for my data, given that I have 721 features.
  + A review discusses how a variety of techniques have been developed that perform data cleaning and filtering before applying SMOTE to increase data quality and thus have more representative points.
  + In a review of SMOTE that covered 15 years of progress, there was no explicit mention of modifications to SMOTE using missing data.
* My thoughts (not from the paper):
  + Imputation could add additional bias, causing the generated samples to be unreflective of the true data distribution.
  + These limitations could result in biased and/or inaccurate generated samples. Thus, SMOTE was not used in this project, but would be an interesting avenue to explore in the future, as potential modifications could be made to the primary SMOTE method to allow it to work with sparse and overlapping data.

*Undersampling*

* Removing samples from the more common income levels to reduce imbalance in the data.
* However, this can reduce generalisation if important samples are removed
  + Important given the dataset is a little bit small
* <https://jair.org/index.php/jair/article/view/11192/26406>
* [SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary](https://jair.org/index.php/jair/article/view/11192)

Use of shap

* Bias for high cardinality features in scikit learn feature importance

High correlation among features could:

* Could have noise/overfitting if similar info
* Could have unstable trees and feature importance