Predictive modelling of Neonatal Mortality in Zimbabwe using survey data: A machine learning approach.

Absolom Mbinda\textsuperscript{1}, Rornald K. Muhumuza\textsuperscript{2}, Arsene B. Sandie\textsuperscript{2}, Richard Makurumidze\textsuperscript{1}, Agbessi Amouzou\textsuperscript{3}

\textsuperscript{1}University of Zimbabwe Family Medicine, Global and Public Health, \textsuperscript{2}APHRC, \textsuperscript{3}John Hopkins University School of Public Health
## Background, Objectives and methods

<table>
<thead>
<tr>
<th><strong>Background and Objectives</strong></th>
<th><strong>Methods</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background</strong>: Zimbabwe has a high neonatal mortality rate at 32 neonatal deaths per 1,000 live births with little decline in the past decades.</td>
<td><strong>Data source</strong>: Pooled secondary data analysis of three rounds of DHS 2005, 2010 and 2015.</td>
</tr>
<tr>
<td><strong>Objective</strong>: To identify the most important predictors of neonatal mortality in Zimbabwe using survey data.</td>
<td><strong>Data source</strong>: All births in the 5 years preceding each survey (N = 9,258) with complete records.</td>
</tr>
<tr>
<td></td>
<td><strong>Predictive modelling</strong>: Handle high dimensional data well and interaction factors without need to specify them in the model in advance.</td>
</tr>
<tr>
<td></td>
<td>Multiple machine learning models were built and compared using classification evaluation metrics.</td>
</tr>
<tr>
<td></td>
<td>Best model used to select important variables.</td>
</tr>
</tbody>
</table>
Machine learning allows us to assess a large number of variables at the same time.

68 possible predictors of neonatal mortality examined; 21 selected for model.

Assessing model of the best fit:
- Area under the curve (ROC)
- Accuracy
- Sensitivity
- Specificity
Gradient boosting (gbm) was the best performing model to identify the most important factors affecting neonatal mortality.

- Tested 12 models such as Random Forest (RF) and KNN.
- ROC and Sensitivity showed consistently that GBM had the best results.
- GBM is a better model in identifying predictors of neonatal mortality among models compared.
Birthweight and fertility-related factors emerged from the model as the most important features/predictors of neonatal mortality:

- **Individual factors (child):**
  - Birth weight
  - Birth order
  - Birth interval

- **Mother’s factors:**
  - Age at first birth
  - Education
  - Occupation
  - History of pregnancy termination

- **Household:**
  - Household size
  - Wealth index
  - Cooking fuel
  - Improved toilet

- **Geography/Community:**
  - Region (Harare)
  - Rural residence

Important features selected using best model: gbm

- Birthweight
- Household_size
- Birth_order
- Age_at_first_birth
- Education_secondary
- Wealth_index_middle
- Wealth_index_poor
- Occupation_mgmnt_services
- Preg_termination
- BirthInterval_48months+
- Improved_toilet
- Region_Harare
- Education_no_or_primary
- Wealth_index_rich
- Clean_cook_fuel
- Rural_residence
Conclusion

- The identified individual risk factors could be used to identify vulnerable newborns.

- The mother’s indicators, including place of residence are measures of deprived populations which could be used in equity measurements and targeting of programs.

- Machine learning is a feasible method to identify risk factors for neonatal mortality.

- The model’s accuracy and sensitivity were average, and this could be due to missing quality of clinical delivery and postnatal care variables – not available in DHS.
Acknowledgements