**Automated data-driven development of clinical prediction models with application to hypoxic-ischaemic encephalopathy**

Matthew S Lyon (0000-0002-2500-1013)1,2,3\*, Heather White, Tom R Gaunt1,2,3, Deborah Lawlor (0000-0002-6793-2262)1,2,3, David Odd4

1. National Institute for Health Research (NIHR) Bristol Biomedical Research Centre (BRC), University of Bristol
2. Medical Research Council (MRC) Integrative Epidemiology Unit (IEU), University of Bristol
3. Population Health Sciences, Bristol Medical School, University of Bristol
4. Division of Population Health, Cardiff University

\* To whom correspondence should be addressed. Tel: +44 (0) 117 331 4094; Email: [matt.lyon@bristol.ac.uk](mailto:matt.lyon@bristol.ac.uk)

**Abstract**

**Introduction**

*Start with something that summarises what we mean by prediction and how is it used in health*

In healthcare prediction is used to stratify individuals into different groups of risk in order to optimise prevention and treatment of disease (refs). This approach has been successfully applied for decades to identify individuals who benefit from statin medication reducing population levels of cardiovascular disease (refs). Prediction modelling involves combining several variables in a statistical equation (for example a multivariable regression model) to estimate the occurrence of a future outcome. Clinical prediction models should be accurate (i.e. be able to discriminate between those who go on to get the disease and those who remain healthy and calibrated such that the proportion predicted to get disease is similar to that subsequently observed), feasible (i.e. using variables that are possible and relatively easy to obtain) and cost-efficient. New prediction models should be compared to those already used in terms of accuracy, feasibility and cost. Traditionally, prediction models have been based on established risk factors (risk factors by definition being causally related to the outcome of interest), for example cardiovascular risk prediction tools in common use internationally, such as the pooled cohort equation (refs) and Q-risk (refs), include established cardiovascular risk factors such as cigarette smoking, systolic blood pressure and diabetes. However, it is increasingly recognised that whilst some risk factors are good predictors, disease predictors do not need to be (causal) risk factors. Hence cardiovascular risk prediction tools also include high density lipoprotein cholesterol (HDLc) but do not include, the biologically related, triglyceride levels, despite evidence from randomized trials and genetic studies supporting a causal role for triglycerides (refs) but not HDLc (refs) in cardiovascular disease risk. This is because HDLc is a stronger predictor than triglycerides, most likely because it does not vary so much within people.

*Then introduce how there might be a role for a more hypothesis-free and ML approach, whilst acknowledging this is increasingly used, particularly with ‘omics data*

With the availability of large scale multi ‘omic biomarker data in many population cohorts, the potential to take a more hypothesis-free approach to prediction has increased. In contrast to a hypothesis driven risk factor approach, this has the potential to identify many more variables that might improve prediction accuracy. Machine learning (ML) methods are often used in these studies used to reduce the potential number of variables that remain in the model, as the scale of the data (several 100s or 1000s of variables) would be beyond conventional approaches involving comparing different multivariable regression models with a smaller number of variables selected on the basis of ‘independence’ based on a p-value and/or coefficient threshold. For example, ML methods have been used to combine multiple risk factors and >150 nuclear magnetic resonance metabolite traits to predict pregnancy complications (refs), multiple risk factors and >700 mass spectrometry metabolites to predict pregnancy outcomes, (refs) and to compare a polygenic risk score based on ~7million single nucleotide polymorphisms with established clinical risk scores for predicting cardiovascular disease (refs), with the two metabolite studies suggesting some improvement in prediction for some pregnancy outcomes when metabolites were added to the risk factor predictors, whereas the polygenic risk score did not improve cardiovascular risk in comparison to the established clinical score.

*Then something that introduces the novel idea of time-updated prediction*

With the increased use of electronic records in healthcare practice and the place of PCs, laptops and/or tablets in primary and secondary care it should be feasible to develop predication tools with the potential to do ‘real-time-updated’ risk stratification– i.e. having validated prediction tools for different outcomes embedded within healthcare systems and be able to add new patient/person data to those at any clinic appointment so that their risk is (re)estimated on the basis of new data and their treatment tailored to that new risk. In parallel with this, all new data from all patients could be used to regularly update the underlying prediction models for different outcomes. Whilst such ideas would require major changes to the understanding of prediction/ risk stratification by the public and healthcare providers, together with major operational changes to healthcare systems and provision, determining the extent to which a ML approach applied to large scale routine linked health record data could improve prediction of health outcomes over and above any established risk factors or prediction models would be useful for knowing whether the aim of ‘real-time updated prediction’ for ‘personalised’ healthcare is feasible, and if so for which conditions. A small number of studies have done attempted to do this… This approach could be particularly beneficial in pregnancy, given antenatal and intrapartum (i.e. around the time or labour and birth) care consist of detailed and specified monitoring over a defined period of time.

<https://www.nature.com/articles/s41746-021-00426-3.pdf>

<https://www.nature.com/articles/s41746-018-0029-1>

<https://www.nature.com/articles/s41591-018-0300-7>

<https://jamanetwork.com/journals/jama/fullarticle/2588761>

<https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0387>

<https://ojs.aaai.org//index.php/aimagazine/article/view/2438>

<https://www.bmj.com/content/361/bmj.k1479>

<https://www.nejm.org/doi/10.1056/NEJMra1814259?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed>

<https://obgyn.onlinelibrary.wiley.com/doi/10.1111/1471-0528.16487>

<https://pubmed.ncbi.nlm.nih.gov/29989977/>

<https://www.nature.com/articles/s41746-018-0048-y>

*Then on to the para about HIE*

A key aim of antenatal care is accurate prediction of adverse perinatal outcomes. Ideally pregnancy prediction tools should enable updated prediction across pregnancy so that monitoring can be increased or decreased as appropriate and key decisions related to when to deliver an infant that optimises the balance between delivering too soon (and risking adverse offspring outcomes related to preterm birth) or too late (risking potential brain injury or neonatal mortality). A significant cause of perinatal brain injury is perinatal asphyxia, leading to hypoxic-ischaemic encephalopathy (HIE). HIE is often devastating, with life-long impacts for the infant [1] and their family, as well as costing society millions of pounds in medical compensation, lost earnings and welfare support [2]. While interventions such as induction of labour or operative delivery can be employed if the risks of continuing the pregnancy are higher than those of early delivery [3] there is a lack of clear data on when to intervene. Current UK clinical guidelines recognise this and suggest that a research priority is to accurately “identify babies at particularly high risk of morbidity and mortality who will benefit from induction and therefore avoid induction for babies who do not need it” [4].

Badawi et al identified 35 potential risk factors for HIE based (antenatal n=20, intrapartum n=14 and growth n=1) on them being independently (of each other) associated with HIE [5], [6]. Whilst these independent associations make it possible these could predict future HIE, to the best of our knowledge these 35 variables have not been researched as predictors, most likely because they were derived in case control studies with ‘risk factor’ status determined retrospectively. As HIE is rare, very large prospective cohort studies would be necessary to explore prediction. In a search of publications in the last ten years we identified just xxxx. Here we investigate the potential of automated features selection approaches and ML classification to simplify the development of a clinical prediction modelling for HIE. We use this approach to explore prediction accuracy of the 35 Badawi risk factors and to compare that with models derived from all available clinical recorded data from the large US Collaborative Perinatal Project [7] (CPP). Collection of data was from 14 units across the United States and showed little evidence of selection bias [8]. The dataset includes data on approximately 60,000 pregnancies, and 58,000 live born infants born between 1959 and 1965. Data was collected throughout the prenatal period, labour and delivery, postpartum and as the child grew.

The aim of this study is to explore the feasibility of undertaking real-time updated risk stratification in pregnancy using a large linked antenatal and intrapartum dataset applied to prediction of HIE as an exemplar. To do this we used the xxxx data, which includes data on xxxxx pregnancies with deliveries occurring between xxx and yyy. We used automated feature selection and engineering approaches to automate model development using all available variables and compared this approach with prediction models using 35 potentially casual risk factors manually identified by Badawi et al.

**Methods**

Outcome

Hypoxic-ischaemic encephalopathy (HIE) was defined as having definite seizures (? Defined how), hypertonia, jitteriness, hypotonia, abnormal reflexes, or abnormal cry; after having a low 5 minute Apgar score (<7) [9].

Defining training and testing datasets

Pregnancies were ordered chronologically and split into two equal subsets for training (infants born 1959-1962) and testing purposes (infants born 1963 to 1965). All variables were identified as either antenatal (measurable before 37 weeks’ gestation), growth (birth measures of growth), and intrapartum (measures only available at or after 37 weeks, up to the point to delivery), and classified as either unordered categorical, ordinal or numerical.

Feature (variable) selection

Two broad variable selection approaches were used. The first took the 35 Badawhi potential risk factors for HIE (**Supplementary Table 1**)[5], [6]. The second approach used automated data-driven methods to rank variables by their predictability of HIE using the training dataset.

Model assessment

Models were trained using the training data and each of the feature sets and applied to predict HIE status in the testing cohort. Receiver-operator curves (ROC) were used to derive the area under the curve measure with 95% confidence interval using the pROC R-package (v1.16.2). These were used to compare discrimination between different models (i.e. the Badawi risk factors model and the models defined by different ML variable selection methods. We also compared discriminaton between models where variable selection had been stratified by antenatal only, antenatal and growth, and antenatal and intrapartum variables..

Analysis using established risk factors

Training data were cleaned and harmonised where possible with the measures previously proposed [5]. A logistic regression model was developed using the training data and evaluated using the testing data. The primary model included only antenatal variables (**Supplementary Table 1**), but this was extended to additionally include fetal growth or intrapartum variables. This analysis was performed using Stata v16.

Automated feature selection and feature engineering

Data-driven feature selection approaches were applied to the three sets of variables (antenatal, antenatal and growth, antenatal and intrapartum) and ranked using the training data. First, variables were excluded if they contained >5% of missing data values (28 out of 518) leaving a potential 490 exposure data fields for the prediction models. Second, unordered categorical variables were recoded as dummy variables. Third, the training data was used for feature selection using a range of algorithms from the scikit-learn Python package (v0.23.2) as follows with default parameters except where specified. These methods were chosen to cover a range of commonly used methods (tree, regularisation and recursive elimination) as outlined in the scikit-learn documentation [10]. Prediction models were subsequently developed using 20, 40 or 60 of the best predictors.

*Reverse feature elimination (RFE) with cross-validation*

During the first iteration all input variables were included as predictors in a logistic regression model trained using five-fold cross validation. On each subsequent iteration the five weakest predictors (determined by the smallest absolute coefficient) were eliminated, with iterations continuing until only one predictor remained. This approach uses the effect of each predictor on the cross-validation mean AUC to rank features (lowest rank is most predictive).

*Elastic-Net and LASSO regression*

Logistic regression with L1 (LASSO) or L1 & L2 (Elastic-Net) regularisation penalties was trained using five-fold cross-validation to determine the optimal value of alpha. The Elastic-Net mixing parameter (l1\_ratio, representing the ratio of L1 to L2 penalty) was set to 0.5. The penalty term shrinks weak predictors to zero which were subsequently eliminated from downstream analyses. The remaining features were ranked by their absolute regression coefficient (smallest is least predictive).

*Extra-trees classifier*

The feature importance metric obtained from an extra-trees classifier with five-fold cross-validation was used to rank features (smallest is least predictive).

*Linear support vector classifier (SVC)*

The linear SVC was trained with five-fold cross-validation using the AUC metric. The default L1 penalty term (C=1.0) was applied to shrink weak coefficients to zero. The coefficients were taken as a measure of feature importance (absolute value; smallest least predictive).

Binary classification

The following models were trained and evaluated using default hyperparameters to facilitate an automated approach and the top n=20, n=40 or n=60 predictors of each feature selection method: logistic regression, random forest, naïve Bayes and neural network (one hidden layer with number of nodes equal to number of predictors using the rectified linear activation function and Adam optimiser implemented in Tensorflow v1.15).

**Results**

Participants

The dataset was based on the full CPP variable file dataset, containing data on 58,760 infants (**Figure 1**). A total of 12,005 infants were born preterm (<37 weeks of completed gestation), 5476 were born after 42 weeks, and 964 were born to a mother of less than 16 years age; leaving a total of 40,315 for the analyses. 19,487 infants were born between 1959 and 1962 (and were placed in the first cohort), while 20,828 were born between 1963 and 1966 (and were placed in the second).

Table 1 shows the demographics of the population, split by HIE status. Overall, 209 (0.5%) had evidence of HIE, 549 (1.4%) died in perinatal period, 1228 (3.1%) had a low Apgar score at 5 minutes and 2013 (5.1%) required resuscitation after birth. Regarding antenatal factors, infants with HIE were more likely to have older (P < 0.001) but primiparous mothers, without private health insurance (P=0.02). Mothers were also more likely to have placenta previa (P=0.02) and infants more likely to be male (P<0.001) and from multiple births (P=0.006); but otherwise, antenatal risk factors did not appear to differ substantially. For growth measures, infants with HIE were more likely to be poorly grown (P<0.001). Infants with, and without HIE, differed for most of the intrapartum factors except the recording of a nuchal cord (P=0.64).

Feature engineering and selection

Of the 518 variables provided in the CPP dataset (**Figure 2**), 28 (5%) were removed due to high missingness (>5%). The variables were split by type into ordinal (n=26), continuous (n=27) and unordered categorical (n=437). Unordered categorical variables were one-hot encoded separating the fields into multiple binary (dummy) variables (n=2126). The combined set contained n=2179 features which were split by collection point during pregnancy into antenatal (n=1729), antenatal and intrapartum (n=2230), antenatal and growth measures (n=2115).

Broadly, two approaches were used for feature selection at each pregnancy timepoint (**Figure 2**). First, clinically defined variables from Badawi *et al* were selected while the second approach was purely data-driven using a range of algorithms. Feature importance metric rankings of elastic-Net, LASSO, RFE and extra-trees were highly correlated (Spearman’s correlation coefficient 0.73 to 1.00), SVC was more weakly correlated with all other methods (Rho = 0.42 or 0.43 for all except the extra-trees feature importance method which was 0.53; **Supplementary Figures 1-3**).

Classification

The discriminatory ability of each feature set was measured using the AUC on the second (later) half of pregnancies with logistic regression.

The established predictors of HIE gave equally good discrimination (**Figure 3**) using all three feature set time point collections: antenatal period gave an AUC of 0.71 (95% CI 0.64-0.77; n=20 predictors), antenatal and infant birth weight was 0.73 AUC (95% CI 0.67-0.79; n=21 predictors) and antenatal and intrapartum measures had an AUC 0.70 (95% CI 0.64-0.77; n=35 predictors).

The data-driven approach using antenatal, growth and intrapartum measures we applied a range of automated approaches for feature ranking and applied logistic regression using the best 20, 40, or 60 features for comparison with established predictors (**Figure 3**). Broadly, there was no strong difference in discrimination between the clinically defined and automated feature selection approaches although SVC and RFE performed worse than the clinically defined features at the antenatal timepoint. There was a trend towards better discrimination with larger number of features included in the model, while this could be a symptom of overfitting, we estimated the AUC in a final holdout set.

We found no strong difference in discrimination when comparing logistic regression to a range of other classifiers (**Supplementary Figure 4**).

**Discussion**

Through this work we have developed prediction models for HIE using a range of automated feature selection approaches and compared these with models developed from clinically defined feature sets. Secondly, we evaluated a representative selection of ML classifiers including logistic regression.

Using freely available historical data, we demonstrated that automated data-driven feature selection and engineering produced classification models with similar or slightly improved HIE discrimination over expert-developed clinical prediction models despite minimum data preparation of the cohort. Secondly, we found that logistic regression generally performed as well as or slightly better than other machine learning classifiers. Finally, we identified a small improvement in prediction performance using growth measures in combination with established risk factors when using logistic regression.

These findings indicate that low-cost automated approaches to developing clinical prediction models may be comparable with expensive human-driven feature selection. The main advantage of automated model development is the requirement for minimal human input, ease of automation and application to a range of clinical outcomes. Other studies have demonstrated similar findings. For example, AutoPrognosis [11] is an end-to-end workflow for complete ML automation and has shown improvements in prediction of cardiovascular disease [12] and survival of cystic fibrosis patients [13] over existing models and clinical guidelines. Simpler strategies similar to those described here have also shown benefit; elastic-net regression applied to prediction of pregnancy outcomes using high-dimensional metabolomic data improved discrimination over clinical features alone [14].

Logistic regression was among the best performing classifiers in this study, this finding is consistent with a systematic review of ML classifiers [15] which showed contemporary ML methods offer no benefit over logistic regression. Since these models do not make strong assumptions, they may provide benefits over logistic regression when there are non-linear and interaction effects. However, the data used for modelling through this work were categorical which may explain the lack of improvement with relaxed assumptions.

A key strength of our analysis was the comprehensive evaluation of a wide range of feature selection techniques and classification algorithms. However, there were also limitations. First, there were some pregnancies discarded due to missing data (1947/13146 [14.8%]) and selection of complete records could induce collider bias if the values were not missing at random. Second, although the data was derived from a large cohort study considered to be of high methodological quality, it is clearly limited by the age of the data. Nevertheless, the risk factors proposed in Badawi et al, derived from an Australian population some decades later, were still strong predictors of outcome suggesting that the main underlying causes of HIE remain significant over this time frame. In addition, for this to provide a valid use to clinicians it need to be tested on recent, but also routinely collected data with likely less cleaning and more missing data points than is present in this research dataset.

This work is consistent with other prediction work, suggesting that poor birth condition can be predicted, although in this study we have attempted to test if ‘raw’ data, mostly unprepared by the research team, could be used by the ML model to aid clinicians. The advantages of this enables application across the multiple data standards and systems in place within healthcare services, and the development of adaptive risk scores to aid decisions around and before birth. Women in the highest decile of risk had risks of between 20% and 40% of their infant developing brain injury around birth (compared to the background risk of only 0.5%), with this risk predictable from antenatal measures. If replicable using live data, this would give the mother and clinicians valuable data to help guide management as we know interventions exists to reduce or modify the risk. We also wanted to test if growth measure, at the time measures after the birth of the infant but now measurable in-utero using ultrasound would add prediction to the model. Using logistic regression and established risk factors addition of growth measures added additional value for HIE (P=0.026). This may reflect other measures of growth or correlates of it (e.g. number of outpatient appointments) stored in the antenatal record and further work using antenatal measures should interrogate this.

**Conclusion**

In this work, on a historical cohort, ML models with minimum data preparation were able to match and, in some examples, exceed the prediction of expert clinical prediction model analysis in identifying which infants would develop HIE after birth. Some predictions improved when measures of growth were included, supporting the role for routine antenatal measures of growth during pregnancies using modern imaging techniques. Routine growth measures, and automated ML models on other routinely collected health data may provide an additional tool to obstetric services to help identify infants at high risk of brain injury around birth and help target additional observation or interventions.

**Data and open-source code availability**

Stata, R and Python3 code are available from <https://ieugit-scmv-d0.epi.bris.ac.uk/ml18692/hie-ml>.

The CPP data files and documentation are available for download from the National Archives Catalog (https://www.archives.gov/research/electronic-records/nih.html)

**Acknowledgments**

This study was funded by the NIHR Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work was also funded by the UK Medical Research Council as part of the MRC Integrative Epidemiology Unit (MC\_UU\_00011/4 and MC\_UU\_00011/6).

**Competing interest**

T.R.G receives funding from GlaxoSmithKline and Biogen for unrelated research.

**References**

[1] D. V. Azzopardi *et al.*, “Moderate Hypothermia to Treat Perinatal Asphyxial Encephalopathy,” *N. Engl. J. Med.*, vol. 361, no. 14, pp. 1349–1358, Oct. 2009, doi: 10.1056/nejmoa0900854.

[2] D. E. Odd, D. Gunnell, G. Lewis, and F. Rasmussen, “Long-term impact of poor birth condition on social and economic outcomes in early adulthood,” *Pediatrics*, vol. 127, no. 6, Jun. 2011, doi: 10.1542/peds.2010-3604.

[3] G. Molina *et al.*, “Relationship between cesarean delivery rate and maternal and neonatal mortality,” *JAMA - J. Am. Med. Assoc.*, vol. 314, no. 21, pp. 2263–2270, Dec. 2015, doi: 10.1001/jama.2015.15553.

[4] “Inducing labour Clinical guideline,” 2008. Accessed: Mar. 17, 2021. [Online]. Available: www.nice.org.uk/guidance/cg70.

[5] N. Badawi *et al.*, “Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study,” *Bmj*, vol. 317, no. 7172, pp. 1554–1558, 1998.

[6] N. Badawi *et al.*, “Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study,” *Bmj*, vol. 317, no. 7172, pp. 1549–1553, 1998.

[7] R. E. Cooke, “The Johns Hopkins Collaborative Perinatal Project. A symposium. Introduction.,” *Johns Hopkins Med. J.*, vol. 128, no. 5, p. 237, May 1971.

[8] “The Women and Their Pregnancies: The Collaborative Perinatal Study of the ... - Kenneth R. Niswander, Myron Gordon - Google Books.” https://books.google.co.uk/books?id=dttsAAAAMAAJ&printsec=frontcover&source=gbs\_ge\_summary\_r&cad=0#v=onepage&q&f=false (accessed Mar. 17, 2021).

[9] D. E. Odd, G. Lewis, A. Whitelaw, and D. Gunnell, “Resuscitation at birth and cognition at 8 years of age: a cohort study,” *Lancet*, vol. 373, no. 9675, pp. 1615–1622, 2009, doi: 10.1016/S0140-6736(09)60244-0.

[10] “scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation.” https://scikit-learn.org/stable/index.html (accessed Sep. 10, 2019).

[11] A. M. Alaa and M. Van Der Schaar, “AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning,” 2018.

[12] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, “Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants,” *PLoS One*, vol. 14, no. 5, p. e0213653, May 2019, doi: 10.1371/journal.pone.0213653.

[13] A. M. Alaa and M. van der Schaar, “Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning,” *Sci. Rep.*, vol. 8, no. 1, p. 11242, Dec. 2018, doi: 10.1038/s41598-018-29523-2.

[14] N. McBride *et al.*, “Do nuclear magnetic resonance (NMR)-based metabolomics improve the prediction of pregnancy-related disorders? Findings from a UK birth cohort with independent validation,” *BMC Med.*, vol. 18, no. 1, p. 366, Dec. 2020, doi: 10.1186/s12916-020-01819-z.

[15] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 110. Elsevier USA, pp. 12–22, Jun. 01, 2019, doi: 10.1016/j.jclinepi.2019.02.004.

Figure 1. Participant inclusion flowchart



Table 1. Demographics of study population split by HIE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Characteristic** | | **Non-HIE infants** | **HIE infants** | **P** |
| **Ante-natal Measures** | | | | |
| Late Booking\*\*\* |  | 11,405 (29.1%) | 60 (28.7%) | 0.905 |
| Thyroid Disease |  | 1,028 (2.6%) | 5 (2.4%) | 0.837 |
| Maternal Age | < 20 years | 11,057 (28.2%) | 57 (27.3%) | <0.001 |
|  | 20-24 | 11,690 (29.8%) | 49 (23.4%) |  |
|  | 25-29 | 8809 (22.5%) | 36 (17.2%) |  |
|  | 30-34 | 4,644 (11.8%) | 39 (18.7%) |  |
|  | 35 or more | 3,022 (7.7%) | 28 (13.4%) |  |
| Parity | 0 | 10,434 (26.7%) | 78 (37.3%) | 0.001 |
|  | 1 | 8579 (22.0%) | 29 (13.9%) |  |
|  | 2 or more | 20,049 (51.3%) | 102 (48.8%) |  |
| Employed |  | 5,989 (15.5%) | 23 (11.2%) | 0.084 |
| Private Insurance |  | 2546 (7.0%) | 5 (2.7%) | 0.022 |
| Race | White | 19,560 (49.9%) | 63 (30.1%) | <0.001 |
|  | Black | 16,898 (43.1%) | 123 (58.9%) |  |
|  | Other | 2,764 (7.1%) | 23 (11.0%) |  |
| Family History of Seizures |  | 2560 (6.7%) | 16 (7.9%) | 0.500 |
| Family History of Neurology\* |  | 1479 (3.9%) | 12 (5.9%) | 0.133 |
| Fertility Investigations |  | 1008 (2.6%) | 6 (2.9%) | 0.797 |
| Hypertension |  | 167 (0.4%) | 1 (0.5%) | 0.911 |
| Preeclampsia |  | 1284 (3.3%) | 19 (9.1%) | <0.001 |
| Maternal Height | <160cm | 13,221 (36.4%) | 81 (41.3%) | 0.354 |
|  | 160-164cm | 10,961 (30.2%) | 54 (27.6%) |  |
|  | >164cm | 12,172 (33.5%) | 61 (31.1%) |  |
| Pre-labour bleeding |  | 10,792 (28.1%) | 69 (33.8%) | 0.071 |
| Antenatal Viral Illness |  | 2,688 (6.9%) | 15 (7.2%) | 0.846 |
| Alcoholism |  | 44 (0.11%) | 0 (0.0%) | 0.628 |
| Fever |  | 5,068 (13.0%) | 26 (12.4%) | 0.817 |
| Male |  | 19,842 (50.6%) | 134 (62.1%) | <0.001 |
| Placental Previa |  | 160 (0.41%) | 3 (1.5%) | 0.020 |
| Multiple Birth |  | 290 (0.74) | 5 (2.4%) | 0.006 |
| **Intra-partum measures** | | | | |
| OP presentation |  | 2512 (6.6%) | 34 (16.8%) | <0.001 |
| Breech Presentation |  | 1023 (2.7%) | 31 (15.3%) | <0.001 |
| ROM>12 hours |  | 5706 (16.5%) | 50 (30.5%) | <0.001 |
| Caesarean Section |  | 2076 (5.3%) | 38 (18.2%) | <0.001 |
| MIE\*\* |  | 3000 (7.7%) | 42 (20.1%) | <0.001 |
| Nuchal cord |  | 10,225 (26.3%) | 52 (24.9%) | 0.636 |
| Prolapsed cord |  | 311 (0.8%) | 12 (5.7%) | <0.001 |
| Onset | No Labour | 1,146 (3.0%) | 12 (5.8%) | 0.019 |
|  | Spontaneous | 35,124 (90.3%) | 177 (85.1%) |  |
|  | Induced | 2636 (6.8%) | 19 (9.1%) | 0.019 |
| Shoulder Dystocia |  | 230 (0.6%) | 9 (4.3%) | <0.001 |
| Epidural |  | 617 (1.6%) | 10 (4.9%) | <0.001 |
| **Growth Measures** | | | | |
| Birth weight centile | Less than 3rd | 1208 (3.1%) | 29 (14.1%) | <0.001 |
|  | 3rd to 10th | 2898 (7.4%) | 28 (13.6%) |  |
|  | 10th to 90th | 31,265 (79.8%) | 125 (60.7%) |  |
|  | Above 90th | 3824 (9.8%) | 23 (11.7%) |  |

HIE, hypoxic-ischaemic encephalopathy. FHx, family history. OP, occiput posterior fetal position. CS, caesarean section. ROM, rupture of membranes.

\* Motor, sensory or developmental disorder in siblings

\*\* APH, eclampsia, uterine rupture or ruptured cord

\*\*\* >26 weeks of gestational age

Figure 2. Feature engineering workflow



Figure 3. HIE prediction using logistic regression with a range of feature selection approaches



ROC for prediction of hypoxic-ischaemic encephalopathy using logistic regression and a range of feature selection approaches. The model was trained using the first 50% infants (born 1959-1962) and evaluated using the latter 50% of infants (born 1963 to 1965). AUROC, area under the receiver operator curve. CI, confidence interval. RFE, reverse feature elimination. SVC, support vector classification.

Supplementary Table 1. Established risk factors for HIE

|  |  |  |
| --- | --- | --- |
| Antenatal Factors (n=20) | Intrapartum Factors (n=14) | Growth Measures (n=1) |
| * Maternal age (<20, 20-24, 25-29, 30-34, >35) * Parity 0, 1,>1 * Maternal Employment * Health Insurance * Maternal race * FHx of seizures (recurrent non-febrile seizures) * FHx of neurological disorder (excludes seizures) * Infertility Treatment * Maternal Hypertension * Maternal height (<160, 160-164, >164) * Maternal Thyroid Disease * Pre-eclampsia * Antenatal bleeding (mod or severe) * Viral Illness * Alcohol (some, none, unknown) * Birthweight centile (>90th, 10-90th, 3rd-9th, <3rd) * Sex * Abnormal placenta * Late or no antenatal care * Multiple births | * Gestation (37-42) * OP presentation * Maternal Pyrexia * Maternal Intrapartum Event (Haemorrhage, convulsions, uterine rupture, snapped cord, out of hospital birth) * Membrane rupture >12 hours * Blood Pressure abnormalities – Captured above * Nuchal cord * Cord prolapse * Onset of labour (spontaneous, induced, none) * Mode of delivery (Spontaneous, induced vaginal, elective CS, emergency CS, breech manoeuvre) * Shoulder dystocia * Epidural Anaesthetic * Breech Presentation * ROM>12 hours | * Birth weight centile (>90th, 10-90th, 3rd-9th, <3rd) |

Previously reported predictors of hypoxic-ischaemic encephalopathy obtained from Badawi et al (1998). FHx, family history. OP, occiput posterior fetal position. CS, caesarean section. ROM, rupture of membranes.

Supplementary Figure 1. Distribution and correlation of automated feature selection scores using all antenatal, intrapartum and growth feature sets

Supplementary Figure 2. Distribution and correlation of automated feature selection scores using all antenatal, intrapartum and growth feature sets

Supplementary Figure 3. Distribution and correlation of automated feature selection scores using all antenatal, intrapartum and growth feature sets

L1 penalised regression (ElasticNet & Lasso), absolute regression coefficient. Random forest classifier (Tree), feature importance. L1 penalised linear support vector classifier (SVC), absolute coefficient. Reverse feature elimination (RFE) using logistic regression, descending rank of predictors. Corr, Spearman’s rho rank correlation coefficient.

Supplementary Figure 4. Discrimination of HIE with a range of classifiers



Features were selected using elastic net from the antenatal and fetal growth dataset. AUROC, area under the receiver-operator curve. LR, logistic regression. NB, naïve bayes. NN, neural network. RF, random forest.