Systematic Reviews and Meta- and Pooled Analyses

Risk Factor Models for Neurodevelopmental Outcomes in Children Born Very Preterm or With Very Low Birth Weight: A Systematic Review of Methodology and Reporting

Louise Linsell®, Reem Malouf, Joan Morris, Jennifer J. Kurinczuk, and Neil Marlow

* Correspondence to Louise Linsell, National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF, United Kingdom (e-mail: louise.linsell@npeu.ox.ac.uk).

Initially submitted January 7, 2016; accepted for publication March 3, 2016.

The prediction of long-term outcomes in surviving infants born very preterm (VPT) or with very low birth weight (VLBW) is necessary to guide clinical management, provide information to parents, and help target and evaluate interventions. There is a large body of literature describing risk factor models for neurodevelopmental outcomes in VPT/VLBW children, yet few, if any, have been developed for use in routine clinical practice or adopted for use in research studies or policy evaluation. We sought to systematically review the methods and reporting of studies that have developed a multivariable risk factor model for neurodevelopment in surviving VPT/VLBW children. We searched the MEDLINE, Embase, and PsycINFO databases from January 1, 1990, to June 1, 2014, and identified 78 studies reporting 222 risk factor models. Most studies presented risk factor analyses that were not intended to be used for prediction, confirming that there is a dearth of specifically designed prognostic modeling studies for long-term outcomes in surviving VPT/VLBW children. We highlight the strengths and weaknesses of the research methodology and reporting to date, and provide recommendations for the design and analysis of future studies seeking to analyze risk prediction or develop prognostic models for VPT/VLBW children.

data reporting; development; preterm infants; prognosis; research methodology; risk factors; systematic reviews; very low birth weight

Abbreviations: VLBW, very low birth weight; VPT, very preterm.

Prematurity and its associated neonatal morbidity have a pervasive effect on neurodevelopment, leading to a number of conditions, including cerebral palsy, visual and auditory deficits, impairments in motor and cognitive function, and behavioral problems (1). Early identification of factors that mediate long-term outcomes is necessary to guide the clinical management of children born preterm, provide information to parents, and help develop, target, and evaluate interventions. A large body of literature reports on risk factor models for neurodevelopmental outcomes in very preterm (VPT) children (≤32 weeks’ gestation), yet few, if any, have been developed for use in routine clinical practice or adopted for use in research studies or policy evaluation. Our aim in this article is to describe and review the conduct and reporting of these studies over the last 2 decades and provide methodological recommendations for future research in this area.

We reviewed and summarized the methods and reporting in 78 studies that were recently included in a systematic review of risk predictors for neurodevelopmental outcomes in surviving children born at or before 32 weeks’ gestation (VPT) or with a birth weight less than or equal to 1,250 g (very low birth weight (VLBW)) (2–4). Studies were included if the aim (or one of the aims) was to perform a multivariable (>2 variables) risk factor analysis for neurodevelopmental outcomes in this population, in one or more of the following domains: motor skills, cognition, hearing, vision, or behavior. In this article, we report the main elements of study design, model development, reporting, and validation (if performed) of the 78 studies included in the review. The findings are then discussed within the framework of recommended approaches for risk factor analysis and reporting advocated by experts advising on prediction modeling in the medical and statistical literature.
Despite the recent publication of reporting guidelines for studies developing and validating risk prediction models (5), there is no equivalent central resource that provides guidance for the design and analysis of studies seeking to perform a risk factor analysis or develop a prognostic model.

METHODS

Methods for the systematic review of risk predictors for poor neurodevelopment in VPT and VLBW children have previously been published in a protocol (http://www.crd.york.ac.uk/PROSPERO; registration number CRD42014006943) and in 3 published review articles on motor, cognitive, and behavioral outcomes (2–4). They are described in brief below.

Search strategy

Three electronic search strategies were devised in the MEDLINE, Embase, and PsycINFO databases (available with the protocol) using the National Institutes of Health Medical Subject Headings (see Web Appendix 1, available at http://aje.oxfordjournals.org/). The searches identified journal articles published between January 1, 1990, and June 1, 2014, presenting a multivariable risk prediction model for a neurodevelopmental outcome assessed at or after the age of 18 months in VPT/VLBW children. No language restrictions were made. The bibliographies of all articles included for data extraction were hand-searched for further eligible articles.

Eligibility criteria

Articles were included in the review if they satisfied the following eligibility criteria: 1) they contained original data; 2) the study population was born after January 1, 1990; 3) the study population had a gestational age less than or equal to 32 weeks (VPT) or a birth weight less than or equal to 1,250 g (VLBW) and was not a highly select group (based on other clinical criteria); and 4) one of the authors’ objectives was to perform a multivariable risk factor analysis (>2 variables) of a neurodevelopmental outcome in survivors assessed at or after 18 months of age.

A birth weight cutoff of ≤1,250 g was used, instead of the official definition of VLBW (≤1,500 g), to exclude the subset of more mature but extremely growth-restricted children included in the typical ≤1,500-g VLBW cohort, which can cause heterogeneity and lead to confounding bias when examining the relationship between risk factors and outcomes (6). Explanatory prognostic factor studies which investigated the causal pathway between a single prognostic factor and an outcome (ideally adjusted for confounders) and estimated effect size were not included in the review. In these types of studies, other risk factors are included based on the change in the regression coefficient of the prognostic factor under study, whereas in multivariable outcome prediction models, risk factors are included in the model based on their ability to predict the outcome. Current guidelines recommend not combining these 2 distinct types of studies, as their objectives and model-building strategies differ and when synthesized could lead to biased results (7, 8).

Data extraction and reporting

All articles identified by the search strategies were screened on title and abstract for definite exclusions and duplicates (screen 1). For the remaining articles, the full text was retrieved and the eligibility criteria were applied (screen 2). The 2 screens were initially performed by the first author (L.L.), but if there was uncertainty about the eligibility of an article, it was screened independently by the second author (R.M.). If a decision could not be reached, it was referred to the rest of the author review team (J.J.K., N.M., and J.M.). Non-English articles included in the review were fully translated. Multiple articles based on the same cohort of children underwent a panel review (L.L., R.M., and N.M.). Studies reporting on the same outcome domain (cognitive, motor, behavior, hearing, or vision) at the same age of assessment (<5 years or ≥5 years) were assessed on relevance to the review, and only 1 article was selected for data extraction. For all articles eligible for inclusion, both reviewers (L.L. and R.M.) independently completed a full data extraction form in a customized Microsoft Access 2010 database (Microsoft Corporation, Redmond, Washington). These were cross-validated for discrepancies and referred to the rest of the author review team if agreement could not be reached. If critical information was missing or unclear, the corresponding author was contacted once by e-mail for clarification.

Summary statistics were provided for each item of data extracted (listed in Tables 1–3), and the results from this review are presented in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Web Appendix 2) (9). If more than 1 model was presented in an article, we selected the first model reported to summarize model characteristics to avoid the overrepresentation of studies presenting multiple models.

RESULTS

The search strategy retrieved 44,500 articles (Figure 1), and 2,284 articles were screened on full text, applying the full set of eligibility criteria. Ninety-one articles (from 48 cohort populations) containing multivariable risk factor analyses were eligible for inclusion. Following panel review (L.L., R.M., and N.M.), a further 13 articles were excluded because they reported on the same outcome domain at the same age of assessment in the same cohort as another article with a more relevant objective. The remaining 78 articles, reporting 222 risk prediction models for a neurodevelopmental outcome in surviving VPT/VLBW children, were included in the data extraction (10–87).

Study design

The design characteristics of the 78 studies that were included for data extraction are shown in Table 1. The main study design was cohort (91%), and there were 5 randomized controlled trial populations (51, 52, 55, 60, 69), 1 case-control study (53), and 1 cross-sectional study (31). Of the 71 cohorts followed up prospectively, half (n = 36) were ascertained from all live births in a geographically defined region and half (n = 35) were recruited from neonatal intensive care
unit admissions. Forty-five percent ($n = 35$) of the studies recruited subjects from a single research center, and 79% ($n = 62$) of studies extracted risk factor and outcome data prospectively. Overall, 32% ($n = 25$) of studies were based on a cohort study of all live births in a geographically defined region and followed up the subjects prospectively.

Fifty-five percent ($n = 43$) of studies defined the study cohort using gestational age only, and 21% ($n = 16$) used birth weight only; the remaining studies used some combination of the two. The most common age of assessment was 18–24 months (45%; $n = 34$), and no studies followed up participants beyond the age of 12 years. The median sample size was 219 (interquartile range, 141–461), and 11 studies had more than 1,000 participants (11, 12, 14, 16–18, 24, 25, 54, 69, 82). Only 1 published article mentioned the issue of statistical power, and it referred to the number of events per variable in the modeling process (39).

### Model development

The median number of risk factor models presented per study was 2, and 25% of studies presented 4 or more models. A summary of the model-building techniques used is presented in Table 2 (for the first model presented in each article). The median number of candidate risk factors considered at the outset was 17 (range, 3–51). Seventy-two percent ($n = 56$) of studies provided a rationale for their choice of candidate factors or used a comprehensive list with wide coverage (≥20 factors), but 28% ($n = 22$) of studies with fewer than 20 candidate factors gave no rationale at all. The derivation and coding of outcomes and risk factors were described clearly in 85% ($n = 66$) of studies, and outcomes were generally measured using comprehensive, well-validated tests or assessed using standardized diagnostic criteria, though information on blinding to previous medical history at assessment was frequently not reported. Sixty-nine percent ($n = 54$) of studies categorized some or all of the risk factors that were measured on a continuous scale. In many cases there was a clinical rationale, but $28%$ ($n = 22$) of studies with fewer than 20 candidate factors gave no rationale at all.

### Table 1

**Designs of Studies Presenting a Multivariable Risk Factor Model for a Neurodevelopmental Outcome in Very Preterm and Very Low Birth Weight Children Included in a Systematic Review, 1990–2014**

<table>
<thead>
<tr>
<th>Study Design Characteristic</th>
<th>No. of Studies</th>
<th>% of Studies ($n = 78$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>71</td>
<td>91</td>
</tr>
<tr>
<td>RCT</td>
<td>5</td>
<td>6.4</td>
</tr>
<tr>
<td>Case-control</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>Participant selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live births</td>
<td>36</td>
<td>46</td>
</tr>
<tr>
<td>Admissions to NICU</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>RCT participants</td>
<td>5</td>
<td>6.4</td>
</tr>
<tr>
<td>Selected from a database</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>Not clear</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>Research center selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single center</td>
<td>35</td>
<td>45</td>
</tr>
</tbody>
</table>
| All centers in a geographi
cally defined region          | 30             | 38                     |
| Multicenter (nonrandom)    | 13             | 17                     |
| Data extraction            |                |                        |
| Prospective                | 62             | 79                     |
| Retrospective              | 13             | 17                     |
| Not clear                  | 3              | 4                      |
| Data collected as part of routine follow-up | |                  |
| No                         | 57             | 73                     |
| Yes                        | 21             | 27                     |
| Continent of study         |                |                        |
| Europe                     | 46             | 59                     |
| North America              | 17             | 22                     |
| Australia and New Zealand  | 12             | 15                     |
| Asia                       | 2              | 2.6                    |
| International              | 1              | 1.3                    |
| Starting years of recruitm
tent                      |                |                        |
| 1990–1995                  | 25             | 32                     |
| 1996–2000                  | 33             | 42                     |
| 2001–2010                  | 20             | 26                     |
| GA and BW inclusion criteria |            |                        |
| VPT (≤32 weeks)            | 24             | 31                     |
| EPT (≤29 weeks)            | 19             | 24                     |
| ELBW (≤1,000 g)            | 16             | 21                     |
| Other GA/BW combination    | 19             | 24                     |
| Exclusion criteria         |                |                        |
| Congenital anomalies       | 34             | 44                     |
| Severity of disability     | 22             | 28                     |
| Multiple births            | 7              | 9                      |

### Table 1. Continued

<table>
<thead>
<tr>
<th>Study Design Characteristic</th>
<th>No. of Studies</th>
<th>% of Studies ($n = 78$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of assessment, years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5–2</td>
<td>34</td>
<td>45</td>
</tr>
<tr>
<td>3–5</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>6–12</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Sample size*</td>
<td>219 (141–461)</td>
<td>40–7,632</td>
</tr>
</tbody>
</table>

*Values are presented as median (interquartile range) and range.

**Abbreviations:** BW, birth weight; ELBW, extremely low birth weight; EPT, extremely preterm; GA, gestational age; IQR, interquartile range; NICU, neonatal intensive care unit; RCT, randomized controlled trial; VPT, very preterm.
but the use of arbitrary thresholds for no reason was also a widespread practice.

Forty-three percent \( (n = 33) \) of studies entered all of the candidate variables into the multivariable model; 28\% \( (n = 22) \) only included candidate factors with a \( P \) value below a set threshold in a univariable test of association with the outcome, and 8\% \( (n = 6) \) screened variables using multivariable analysis. The most popular method of model-building

Table 2. Development of the First Model Presented in Each Study Included in a Systematic Review of Risk Factor Models for Neurodevelopmental Outcomes in Very Preterm and Very Low Birth Weight Children, 1990–2014

<table>
<thead>
<tr>
<th>Model Development</th>
<th>No. of Studies</th>
<th>% of Studies ( (n = 78) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of candidate risk factors at outset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear(^a)</td>
<td>17 (11–24)</td>
<td>3–51</td>
</tr>
<tr>
<td>Unclear</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Rationale given for candidate risk factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>29</td>
<td>37</td>
</tr>
<tr>
<td>No, but wide coverage ( (\geq 20) )</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>No</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>Treatment of continuous risk factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All left as continuous</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Some categorized</td>
<td>33</td>
<td>42</td>
</tr>
<tr>
<td>All categorized</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>No continuous risk factors</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Unclear</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Coding of risk factors and outcome clearly described</td>
<td>66</td>
<td>85</td>
</tr>
<tr>
<td>Statistical model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear/logistic regression</td>
<td>74</td>
<td>95</td>
</tr>
<tr>
<td>Multinomial/ordinal logistic regression</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Multilevel modeling</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Method of initial screening of candidate risk factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No screening—all included</td>
<td>33</td>
<td>43</td>
</tr>
<tr>
<td>( P ) value from univariable model</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>( P ) value from multivariable model</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Unclear</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Model-building strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All candidates/screened candidate risk factors included</td>
<td>30</td>
<td>39</td>
</tr>
<tr>
<td>Stepwise selection</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>Backward selection</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Forward selection</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Unclear</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Any interaction terms fitted</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Any nonlinear terms fitted</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Any factors forced into the final multivariable model</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>Multicollinearity of risk factors mentioned/discussed</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Model fit/assumptions checked</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>No. of risk factors in final multivariable model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear(^a)</td>
<td>5 (4–9)</td>
<td>0–25</td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) Values are presented as median (interquartile range) and range.
after initial screening (or no screening) was to simply include all factors (39%; n = 30) regardless of statistical significance, and the second-most popular method was stepwise selection (35%; n = 27). The median number of risk factors included in the final prediction model was 5 (interquartile range, 4–9).

Reporting and model validation

A summary of reporting and model validation (if performed) is presented in Table 3. Generally the reporting of study attrition and missing data was quite poor, making it difficult to assess how representative of the original recruited study sample the analyzed children were. A comparison of baseline characteristics between children lost to follow-up and those assessed was not performed by 30% (n = 21) of studies. Furthermore, only 25% (n = 19) of studies presented the amount of missing data for each variable included in the final model, and only 31% (n = 24) reported the number of infants included in the final model. Although 52 (67%) studies reported point estimates, confidence intervals or standard errors, and the P value or significance level of all model coefficients included in the final multivariable model, the remaining studies only did so for selected variables or failed to report them at all.

Four studies (12, 32, 42, 72) assessed the performance of the final model using statistical validation techniques. All 4 reported the area under the receiver operating characteristic curve, which measures the discriminatory ability of the model to differentiate between individuals by severity of outcome. One study (42) produced a decision tree/algorithm based on their model, and another (12) produced a Web-based tool with which to predict either death or neurodevelopmental impairment for...
In this study, the model was developed by randomly splitting
the data set into a development set (70%) and a validation set
(30%), and the model was tested in the validation set.

DISCUSSION

This review has highlighted some strengths and weakness
in the research methodology and reporting of multivariable
risk factor models for neurodevelopmental outcomes in the
VPT/VLBW population in the published literature. Recom-
mendations for future studies, based on the latest literature
in the field, are discussed below in relation to the findings of
the review. These guidelines are intended as an aid for re-
searchers planning risk factor studies and are not intended
as a prescribed checklist.

Study design

Selection of participants. The optimal study design for
prognostic research in the VPT/VLBW population is a
prospective cohort study of all live births in all research cen-
ters in a geographically defined region. The strength of this
design is that the sample represents the source population
and exposure to risk factors can be measured prior to the oc-
currence of an outcome (88). Studies based in a single re-
search center, while convenient, have less generalizability
due to between-center differences in service provision, re-
ferral, and management practices and the socioeconomic/ethnic characteristics of the local population. In addition,
neonatal intensive care unit populations are less representa-
tive because they tend to have a different case-mix than the
general VPT/VLBW population; they contain a higher pro-
portion of infants with poorer outcomes due to referrals of
higher-risk infants from lower-level centers. Data from ran-
donized controlled trials can be used to study prognosis,
provided that the treatment allocation is included as a pre-
dictor variable, though such studies may have reduced gen-
eralizability because of restricted trial eligibility criteria.

There was a lack of consistency in the gestational age and
birth weight criteria used nationally and internationally to
define cohorts, with some studies using prematurity or birth

Figure 1. Search and selection process used to locate published studies that developed a multivariable risk factor model for neurodevelopmental outcomes in very preterm and very low birth weight children, 1990–2014.
weight alone and others using a combination of both. Before the routine use of ultrasound, cohorts were generally defined by birth weight due to the unreliable measurement of gestational age. Although there has been a shift away from using birth weight–defined cohorts in the last 2 decades, even studies conducted more recently have used varying criteria. The typical VLBW cohort can be a fairly heterogeneous group due to the inclusion of more mature but extremely growth-restricted infants, and it is recommended that epidemiologic studies of immature newborns be based on gestational age rather than the birth weight criterion (6).

Sample size and number of events per variable. It is generally recommended that there be a minimum of 10 outcome events per predictor variable when using logistic regression models for risk factor analysis (89), though the evidence supporting this rule is weak, and some advocate the view that a minimum of 20 events per variable generally eliminates bias when many low-prevalence predictors are included (90, 91). Simulation studies have shown that as the number of events per variable declines below 10, there is an increase in bias and variability, unreliable confidence interval coverage, and problems with model convergence (92). Overfitting is a particular problem when small samples are used for risk prediction modeling, which can lead to the model’s performance being overoptimistic in the data set from which it was developed (93). The median sample size of studies included in the review was 219, and the median number of risk factors in the final models was 5, which means that, on average, studies lacked the power to detect associations for any outcome with an event rate less than 20%. It is difficult to calculate a suitable sample size for risk prediction studies, and risk factor analysis is often a secondary aim when planning a cohort study; however, it is advisable to not stray too far below the recommended number of events per variable.

Model development

Definitions of outcomes and risk factors. The measurement and assessment of outcomes were generally clearly defined and robust in the studies included in this review, though it would be helpful to have more international agreement and standardization—for example, in the diagnostic criteria used by studies to define cerebral palsy, which are available (94). There was more consistency across studies in the tests used to measure motor and cognitive outcomes than there was for tests used for the other outcome domains; however, studies did not always use the same cutpoints for impairment, and some used continuous scores, which made it difficult to compare findings. Few studies reported whether assessments were blinded to previous medical history. For risk factor variables, some conditions such as bronchopulmonary dysplasia and intraventricular hemorrhage were defined quite consistently across studies, whereas other factors such as sepsis and necrotizing enterocolitis varied in definition or were not clearly defined at all. It is recommended that outcomes be assessed prospectively using comprehensive, well-validated tests or using standard diagnostic criteria with a strict protocol, blinded to previous medical history. The definitions of outcomes and risk factors should be described clearly in sufficient detail if the model is to be correctly interpreted or applied by other researchers.

Coding and modeling of continuous variables. The factors retained in the final model and the value of their coefficients are strongly influenced by the coding and methods used to model continuous and categorical variables (95). Many studies included in the review categorized some or all of the continuous risk predictors used in the modeling, often without a clear rationale. The arbitrary categorization of continuous predictors should be avoided, because this results in a loss of information and statistical power. It also results in the classification of individuals who are close to but on either side of the chosen cutpoint as having very different levels of risk (96). Using cutpoints that are data-driven, such as the sample median, are problematic if a model is transported to a different study population (97). Assessing whether the relationship between a continuous predictor and an outcome is linear or nonlinear is important, yet few studies reported doing this. If a nonlinear relationship exists, then categorization or fitting of a nonlinear term—for example, a quadratic term—may be a reasonable option. The use of splines and fractional polynomials could be also be considered in larger samples (95).

Selection of risk predictors. There is no overall consensus on the best strategy with which to select variables for inclusion in a risk factor model; however, some approaches are not recommended. This includes screening candidate factors using univariable tests of association with the outcome, as the correlation with other risk factors is not controlled for. This can result in the rejection of important risk factors that only become predictive after adjustment for other factors (92, 98). Nearly half of the studies included all candidate variables, but many started with fewer than 20 variables at the outset, so it is likely that some preselection process was applied but not reported. It is important to report any procedure used to reduce the number of candidate variables in sufficient detail so that the degree of coverage (of factors considered) can be assessed.

We caution against the use of automated variable selection processes, such as forward selection, backward elimination, or stepwise approaches, as they are data-driven and ignore clinical plausibility. This can lead to poorly performing models with biased regression coefficients (93, 99, 100). Some important risk factors that are well-established in the literature may not always be statistically significant in a particular data set, but it is advisable to include these in the development process. The preferred approach is to start with the full model and eliminate factors one by one (taking both statistical significance and clinical relevance into consideration), because this avoids overfitting and selection bias and provides correct estimates of standard errors (101). The level of significance has a major effect on the number of variables retained; a 1% level will almost always result in a more parsimonious model than a 5% level. However, starting with the full model can be problematic when there are a large number of candidate factors, as is usually the case when predicting neurodevelopmental outcomes in the VPT/VLBW population, due to the vast amount of data usually collected during the neonatal period and following

discharge. Six studies (12, 30, 45, 47, 52, 85) in the review adopted a sequential multivariable approach to prediction, fitting candidate factors in stages according to the time frame in which they occurred or according to themes, such as clinical and sociodemographic factors. This approach seems reasonable, given that prognosis in these children is a complex and dynamic process, with environmental factors potentially superseding the influence of early biological factors as the child grows up.

Reporting and model validation

Attrition and missing data. One area that could easily be improved is the reporting of study ascertainment and attrition. While this was excellent in some studies, it was lacking in many others, which made it difficult to determine how representative the study populations were. In prospective cohort studies where children are followed up at multiple time points, the numbers and reasons for exclusions and dropouts at each stage should be clearly reported in a flow diagram. In retrospective studies, the number of infants assessed for inclusion should be reported, in addition to the number selected for inclusion, which is important for assessing the risk of selection bias. There is evidence of selective dropout in studies of preterm children, with those lost to follow-up being more likely to be severely impaired or to have mothers with a lower level of education (102–104); therefore, it is important to report and assess the potential impact of dropout on the results. While the majority of studies reported that a comparison of children lost versus not lost to follow-up had been conducted, the data relating to this were not always fully reported in the article (or not provided as supplemental material). If further participants are excluded from the final model due to missing data, this number should be reported and the representativeness of the analysis population should be assessed.

Model reporting and validation. Some studies only reported the results for selected variables in the final model or only provided $P$ values or regression coefficients with no confidence intervals. The absolute minimum data that should be reported are regression coefficients with confidence intervals for all factors retained in the final model, but it is also helpful to report the results of important intermediate stages of model development (as supplemental material if space is not permitted). If authors wish to go further and develop a prognostic model, then its performance can be assessed for calibration (comparing observed and predicted outcomes for groups), accuracy (comparing the observed and predicted outcomes for individuals), and discrimination (ability to distinguish between individuals at low and high risk of developing an outcome). These tests can be carried out on the data used to derive the model but should ideally be carried out on new data, either from the same source (internal validation) or from an independent source (external validation) to evaluate the transportability of the model. Methods of internal validation include split-sample (splitting the sample randomly into a “training” set for model development and a “test” set for model validation), cross-validation (developing a model in each set and testing it on the other set), and bootstrapping (resampling with replacement). The latter 2 resampling methods are more effective than the split-sample approach, which is inefficient and results in a loss of power, and therefore are the preferred approaches for internal validation (88, 105).

Strengths and limitations of the review

The search filter used in this review was intentionally broad at the expense of precision in order to capture all studies reporting risk factor analyses, which resulted in a large number of articles retrieved. This approach is recommended for reviews in fields in which clinical prediction models are largely underdeveloped, rather than a more specific search filter, which would have had a high false-negative rate, potentially leading to many articles being missed (106). No language restrictions were imposed and no further articles were identified in the hand search of bibliographies of all studies included, so it is unlikely that there were any major omissions. Some studies included in the review were published before the proliferation of comprehensive guidelines on the conduct and reporting of research studies (http://www.equator-network.org/) and may not reflect the standards of more current work. Furthermore, many of them preceded the publication of the large body of literature that now exists on risk prediction modeling. Specific reporting guidelines for this field have only just emerged (5).

Conclusion

This systematic review of 78 published articles reporting on multivariable risk factor models for neurodevelopmental outcomes in surviving VPT/VLBW children has revealed some shortcomings in methodology and reporting that could be improved in future studies, and has confirmed that there is a dearth of properly designed and well-conducted prognostic modeling studies in this field. Modeling long-term outcomes in such a heterogeneous population is challenging, often with the existence of multiple impairments within the same individual and with multiple risk factors acting sequentially over time. In the 3 published reviews of risk factors for cognitive and motor impairment and behavioral problems that were based on these studies, the evidence for most risk factors was mixed or unclear (2–4). This may be due to the difficulty of modeling prognosis in this population, but it may also be due to differences in study design, study population, methodological quality, and lack of standardization of measures. The findings and recommendations of this critical review should be used as a basis for the design and analysis of future studies seeking to develop multivariate risk factor or prognostic models in this population.

ACKNOWLEDGMENTS

Author affiliations: National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom (Louise Linsell, Reem Malouf, Jennifer J. Kurinczuk); Queen Mary University of London, Centre for Environmental and Preventive Medicine, Barts and The London School of Medicine and Dentistry, London, United Kingdom (Joan Morris); and

Institute of Women’s Health, University College London, London, United Kingdom (Neil Marlow).

This work was supported by National Institute for Health Research doctoral research fellowship NIHR-DRF-2012-05-206.

We thank Nia Wyn Roberts, Outreach Librarian at the Bodleian Health Care Libraries (Oxford, United Kingdom), for her input and expertise during the search phase of the review.

The funder had no role in the study design, data collection, data analysis, manuscript preparation, or publication decisions. This article presents independent research funded by the National Institute for Health Research (NIHR), United Kingdom. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the United Kingdom Department of Health.

Conflict of interest: none declared.

REFERENCES


28. Emsley HC, Wardle SP, Sims DG, et al. Increased survival and deteriorating developmental outcome...


