Validation of the Children's Hospital of Philadelphia Retinopathy of Prematurity (CHOP ROP) Model

Gil Binenbaum, MD, MSCE; Gui-shuang Ying, PhD; Lauren A. Tomlinson, BS; for the Postnatal Growth and Retinopathy of Prematurity (G-ROP) Study Group

IMPORTANCE The Children's Hospital of Philadelphia Retinopathy of Prematurity (CHOP ROP) model uses birth weight (BW), gestational age at birth (GA), and weight gain rate to predict the risk of severe retinopathy of prematurity (ROP). In a model development study, it predicted all infants requiring treatment, while greatly reducing the number of examinations compared with current screening guidelines.

OBJECTIVE To validate the CHOP ROP model in a multicenter cohort that is large enough to obtain a precise estimate of the model's sensitivity for treatment-requiring ROP.

DESIGN, SETTING, AND PARTICIPANTS This investigation was a secondary analysis of data from the Postnatal Growth and Retinopathy of Prematurity (G-ROP) Study. The setting was 30 hospitals in the United States and Canada between January 1, 2006, and June 30, 2012. The dates of analysis were September 28 to October 5, 2015. Participants were premature infants at risk for ROP with a known ROP outcome.

MAIN OUTCOMES AND MEASURES Sensitivity for Early Treatment of Retinopathy of Prematurity type 1 ROP and potential reduction in the number of infants requiring examinations. In the primary analysis, the CHOP ROP model was applied weekly to predict the risk of ROP. If the risk was above a cut-point level (high risk), examinations were indicated, while low-risk infants received no examinations. In a secondary analysis, low-risk infants received fewer examinations rather than no examinations.

RESULTS Participants included 7483 premature infants at risk for ROP with a known ROP outcome. Their median BW was 1070 g (range, 310-3000 g), and their median GA was 28 weeks (range, 22-35 weeks). Among them, 3575 (47.8%) were female, and their race/ethnicity was 3615 white (48.3%), 2310 black (30.9%), 233 Asian (3.1%), 93 Pacific Islander (1.2%), and 40 American Indian/Alaskan native (0.5%). The original CHOP ROP model correctly predicted 452 of 459 infants who developed type 1 ROP (sensitivity, 98.5%; 95% CI, 96.9%-99.3%), reducing the number of infants requiring examinations by 34.3% if only high-risk infants received examinations. Lowering the cut point to capture all type 1 ROP cases (sensitivity, 100%; 95% CI, 99.2%-100%) resulted in only 6.8% of infants not requiring examinations. However, if low-risk infants were examined at 37 weeks' postmenstrual age and followed up only if ROP was present at that examination, all type 1 ROP cases would be captured, and the number of examinations performed among infants with GA exceeding 27 weeks would be reduced by 28.4%.

CONCLUSION AND RELEVANCE The CHOP ROP model demonstrated high but not 100% sensitivity and may be better used to reduce examination frequency. The model might be used reliably to guide a modified ROP screening schedule and decrease the number of examinations performed.

Published online July 13, 2017.

© 2017 American Medical Association. All rights reserved.
Retinopathy of prematurity (ROP) is a disease of the developing retinal vasculature that can lead to retinal detachment and vision loss. Treatment with laser retinal photocoagulation or intravitreous anti-vascular endothelial growth factor agent injection can reduce the risk of progression to retinal detachment, making timely diagnosis by ophthalmologists of treatment-worthy disease important. Current ROP screening guidelines are based on birth weight (BW) and gestational age at birth (GA). These criteria have high specificity but low sensitivity for identification of premature infants at risk for severe ROP, with less than 10% of examined infants requiring treatment. Serum insulin-like growth factor 1 (IGF-1) is a permissive factor in vascular endothelial growth factor-induced retinal vessel growth, which is inhibited by low serum IGF-1 levels. Both low serum IGF-1 levels and its surrogate measure slow postnatal weight gain are predictive of the development of severe ROP, and multiple statistical approaches have been applied to incorporate slow postnatal weight gain into the prediction of ROP to improve the specificity of screening. The resultant models have included WINROP (Weight, IGF, Neonatal ROP), PINT ROP (Premature Infants in Need of Transfusion ROP), ROPScore, CHOP ROP (Children’s Hospital of Philadelphia ROP), and CO-ROP (Colorado ROP).

The PINT ROP model was developed using data from a randomized trial of blood transfusion in 369 infants with BW less than 1000 g. It consists of a logistic regression equation, including BW, GA, and weight gain rate, calculated using weekly weight measurements. Numerous other risk factors, such as sepsis and necrotizing enterocolitis, were considered but were no longer statistically significant when weight gain was included in the model. The model equation was calculated on a weekly basis to predict the risk of Early Treatment of Retinopathy of Prematurity (ETROP) type 1 or 2 ROP, and if the risk was greater than a predetermined alarm level, examinations were indicated. The PINT ROP model correctly predicted all 33 infants requiring laser treatment in a high-risk cohort of 369 infants, while reducing the number of infants requiring examinations by 30%. The CHOP ROP model was developed by applying the same approach to a broader risk group of 524 infants meeting current US ROP screening guidelines. The CHOP ROP model has a structure identical to that of the PINT ROP model, but the model coefficients and alarm risk level were changed. After updating the coefficients and alarm level to fit the new cohort, the CHOP ROP model correctly predicted all infants developing type 1 ROP (sensitivity, 100%; 95% CI, 84%-100%), while reducing the number of infants requiring examinations by 49%. Although promising, these studies were limited by sample sizes too small to provide precise estimates of sensitivity. Sufficient confidence to use the CHOP ROP model in clinical practice would require much narrower 95% CIs (eg, width <1%), with the width being driven by the number of type 1 ROP cases. In addition, it is important to validate a predictive model in a new cohort to determine its generalizability before clinical use. Recently, the CHOP ROP model correctly predicted all 44 infants with type 1 ROP in an Italian cohort of 445 infants. That study provided validation in a new cohort, but the sample size was again too small to provide a sufficiently precise estimate of sensitivity (sensitivity, 100%; 95% CI, 92.0%-100%).

We sought to validate the CHOP ROP model in a diverse, multicenter cohort large enough to obtain a precise estimate of the model’s sensitivity for treatment-worthy retinopathy of prematurity. Secondarily, we evaluated whether updating of the model improved its performance, and we considered alternative screening schedules in which examination timing and frequency are guided by the degree of risk predicted by the model.

### Methods

**Study Design**

We performed a secondary analysis of data from the Postnatal Growth and Retinopathy of Prematurity (G-ROP) Study, the design of which has been reported previously in detail. Briefly, the G-ROP Study was a multicenter, retrospective cohort study of infants who underwent ROP screening at 30 hospitals in the United States and Canada between January 1, 2006, and June 30, 2012. The dates of analysis were September 28 to October 5, 2015. Institutional review board approval for the study was obtained, and waiver of informed consent was granted at the study headquarters (Children’s Hospital of Philadelphia), the study data coordinating center (University of Pennsylvania), and at all study hospitals (listed in the Acknowledgments at the end of the article).

The study enrolled infants born between January 1, 2006, and December 31, 2011, who underwent ROP examinations and had a known ROP outcome. Infants received ROP examinations by meeting the screening guidelines being used during that time at each institution, typically BW less than 1501 g, GA less than 30 weeks, or larger-BW and older-GA infants with an unstable clinical course as determined by the neonatologist. A known ROP outcome included (1) type 1 ROP, type 2 ROP, or ROP treatment in either eye or (2) retinal vasculature maturity, immature vasculature extending into zone III without prior disease in zone I or II, or regression of ROP not reaching criteria for type 1 or 2 ROP in both eyes. Detailed demographic, ophthalmologic, and medical data, including BW, GA, and weight gain rate measurements, were collected from the medical record by certified data abstractors and entered into a web-based database. Data quality was...
Validation of the CHOP ROP Model for Retinopathy of Prematurity

Figure. Sample Nomogram to Predict the Risk of Severe Retinopathy of Prematurity Based on the CHOP ROP Model

A straight line is drawn between the values for birth weight and daily weight gain rate. The intersection of this line with the auxiliary axis is then connected to the value for gestational age. The intersection of this second line with the probability line provides the predicted probability of severe retinopathy of prematurity. If the risk is greater than 0.0140, eye examinations are indicated. CHOP, Children’s Hospital of Philadelphia. Reprinted with permission from Binenbaum et al.14

Statistical Analysis

The CHOP ROP model consists of a logistic regression–based equation, which has terms for BW, GA, and weight gain rate. Rate of weight gain is calculated by taking the difference between the mean of the immediately preceding week’s daily weights and the mean of the penultimate week’s daily weights.14 Alternatively, weekly weight measurement values can be used in lieu of weekly averages. For this validation analysis, daily weights with weekly averages were used because in the model development study there was a small (3%) advantage to using daily vs weekly weight measurements with regard to the number of infants that would not need examinations.14 Weights during the first week after birth are excluded from the analysis owing to the common weight loss that occurs in very low-BW infants during this period.20

The probability of severe ROP is calculated using the CHOP ROP model equation14 on a weekly basis to predict the risk of developing ETROP type 1 or 2 ROP. Diagnostic examinations are indicated if the predicted risk for any single weekly calculation is greater than a threshold level, and no further weekly calculations are necessary. The model can also be represented as a nomogram (Figure).14

In the primary analysis of this CHOP ROP model validation study, the same model terms (BW, GA, and weight gain rate), coefficients, and risk threshold level of 0.0140 as the original description of the model were used to make all-or-none ROP screening decisions (ie, infants with a predicted risk above the threshold level received examinations; the remaining infants did not receive examinations). The study outcomes were sensitivity for type 1 or 2 ROP (the proportion of infants who developed type 1 or 2 ROP for whom examinations would be indicated by the model), sensitivity for type 1 ROP, and the reduction in the number of infants receiving examinations, which is a more intuitive measure of the specificity of the model. The 95% CIs for the measures of sensitivity were calculated using the Clopper-Pearson exact method.27

In a secondary analysis of this validation study, model updating and a modified screening schedule were considered. In the validation of a predictive model, the initial step is to simply evaluate the model using the original structure (terms, coefficients, etc); however, if the performance is not as high as in the model development study, the model should be updated.23,28,29 Model updating may involve adjusting the model coefficients or even adding new variables if necessary.29 Therefore, an a priori plan was made to update the CHOP ROP model if the sensitivity of the original model was not 100% by lowering the risk threshold level to capture all type 1 ROP cases and by keeping the same model structure but refitting the coefficients using multivariable logistic regression. In addition, a modified screening schedule was considered in which high-risk infants (infants with a predicted risk above the threshold level using the original model and threshold level) received examinations per routine clinical care, and the remaining low-risk infants (infants whose predicted risk never surpassed the threshold level) received a single examination and were followed up further only if ROP of any stage was present at that examination. The timing of the examination was based on postmenstrual age (PMA) and was determined by identifying the latest PMA at which an infant with type 1 ROP who had been identified by the model incorrectly as low risk could be examined before the development of type 1 ROP (ie, the latest possible developmental age at which implementation of the modified screening schedule still would result in the diagnosis of ROP on clinical examination before the need for treatment). The reduction in the number of examinations was calculated instead of the reduction in the number of infants receiving examinations.
any examinations because all infants would receive at least one examination using the modified screening schedule.

All analyses were performed using statistical software. SAS (version 9.3; SAS Institute Inc) was used.

Results

The G-ROP Study cohort included 7483 infants, all of whom were included in this validation study (Table 1). The median BW was 1070 g (range, 310-3000 g). The median GA was 28 weeks (range, 22-35 weeks). Retinopathy of prematurity developed in 3224 infants (43.1%); 459 infants (6.1%) developed type 1 ROP, 472 infants (6.3%) developed type 2 ROP, and 524 infants (7.0%) were treated.

The original CHOP ROP model correctly predicted 917 of 931 infants who developed type 1 or 2 ROP (sensitivity, 98.5%; 95% CI, 97.5%-99.1%) and 452 of 459 infants who developed type 1 ROP (sensitivity, 98.5%; 95% CI, 96.9%-99.3%), while 2563 infants (34.3%) would not receive examinations. These results are summarized in Table 2.

Model updating was undertaken because the sensitivity for type 1 ROP was not 100%. When the model coefficients were kept the same but the threshold level was lowered to capture all type 1 ROP cases, the sensitivity for type 1 ROP was 100% (95% CI, 99.2%-100%), but the reduction in the number of infants receiving examinations was only 6.8%. When the model terms (BW, GA, and weight gain rate) were kept the same but the model was refitted to the G-ROP Study cohort data, resulting in adjusted coefficients for these terms, and a risk threshold was chosen to capture all type 1 ROP cases (sensitivity, 100%), the reduction in the number of infants receiving examinations (10.5%) was not increased.

A modified screening schedule was considered. The CHOP ROP model was applied in its original form, including original terms, coefficients, and risk threshold level. Under this hypothetical modified screening schedule, high-risk infants (predicted risk surpassed the threshold level) received examinations per standard clinical care, and low-risk infants (predicted risk did not surpass the threshold level) received an examination at 37 weeks’ PMA and were followed up with additional examinations only if stage 1 or worse ROP was present at that examination. Applied in this fashion, all 459 infants who developed type 1 ROP would have been diagnosed at the appropriate time to receive treatment (sensitivity, 100%; 95% CI, 99.2%-100%). The number of examinations among all 7483 infants would have been decreased from 32,296 to 28,664, a reduction of 3632 examinations (11.2%). The number of examinations among 4310 infants born at GA exceeding 27 weeks would have been decreased from 12,046 to 8624, a reduction of 3422 examinations (28.4%).

Discussion

We validated the CHOP ROP model in a large, diverse cohort of North American infants. The size of the cohort, with 931 infants who developed severe ROP, allowed us to estimate the
sensitivity of the model with a high degree of precision. The model can be applied clinically to reduce the number of infants requiring examinations by one-third, but up to 2.5% of infants with type 1 or 2 ROP may be missed (the lower boundary of the 95% CI of the sensitivity for type 1 or 2 ROP was 97.5%). Alternatively, the model could be used more confidently, ensuring that all infants with type 1 ROP are identified, by using the model to guide a modified screening schedule and reduce by 28.4% the number of examinations for lower-risk, older-GA infants.

There are multiple reasons why a predictive model may not perform well when applied to new individuals, including overfitting of the model to the development cohort and differences in relevant demographic and medical characteristics between the development and validation cohorts. Therefore, model validation is an important step before clinical implementation. Similar to CHOP ROP, both the WINROP predictive model and the CO-ROP predictive model initially had 100% sensitivity for severe ROP in development studies but demonstrated decreased sensitivity in multicenter validation studies of US infants. Therefore, neither model has been proposed as a replacement for current screening criteria. The WINROP model has been used to guide examination scheduling, as opposed to all-or-none screening decisions.

A trade-off is apparent between reducing the number of infants receiving examinations and the need to identify all infants requiring treatment to prevent disease progression to retinal detachment. Outlier, higher-BW, older-GA infants who develop severe ROP may always exist and not become apparent until a large enough study is undertaken. While slow weight gain, a presumed surrogate measure for low serum IGF-1 levels, helps to distinguish many of these outliers among a large pool of low-risk infants, none of the predictive models developed to date have captured all such outliers. Further research, building on the statistical techniques used to develop the existing models, might result in an even better-performing “hybrid” approach. Clearer definition of conditions causing nonphysiologic weight gain, which may cause weight gain despite low IGF-1 levels (eg, fluid retention), and conditions associated with severe ROP that do not lower IGF-1 levels (eg, excessive early oxygen exposure), which would not be recognized by measuring postnatal weight gain, may help to flag outliers. Such conditions might be incorporated into the predictive model structure, as was done with ROPScore, which includes terms for oxygen use in mechanical ventilation and the need for blood transfusions, and with WINROP, which requires that infants with hydrocephalus receive ROP examinations.

Consensus among ophthalmologists and neonatologists has yet to be reached about which trade-offs are acceptable to save many low-risk infants from unnecessary examinations. Arguably, missing some outlier type 2 ROP cases, which do not require treatment, may be acceptable during the acute phase of ROP if all infants are examined when older to identify visual comorbidities of prematurity, such as strabismus and high refractive errors. In contrast, missing treatment-requiring type 1 ROP cases that would have been identified using current screening guidelines likely would not be acceptable to most physicians. Notably, the appropriate context for such evaluation of a predictive model is indeed comparison with current ROP screening guidelines, the BW and GA thresholds of which also do not capture all outliers and necessitate the use of the poorly defined third criterion of “a complicated postnatal course” for larger-BW and older-GA infants. In this regard, a model that correctly predicts treated infants who surpass current BW and GA screening levels might have a higher sensitivity for type 1 ROP than do current BW and GA thresholds alone.

### Strengths and Limitations

There are important strengths and limitations of this study to consider. A primary strength was the large, diverse, multicenter cohort, representative of infants being examined under current ROP screening guidelines in the United States and Canada. This cohort provides external validity of model performance and precise measures of the model’s sensitivity for predicting severe ROP, represented by narrow 95% CIs around the point estimates of sensitivity. Although retrospective data collection can be a limitation if bias is introduced into the study design, steps were taken to minimize bias. A feasibility study was carried out to identify data consistently documented in the medical record (including BW, GA, and weight gain rate) and that are possible to collect retrospectively without inference; and data quality measures were taken, including a rigorous data collection procedure.

#### Table 2. Prediction of Type 1 and 2 ROP by the CHOP ROP Model Based on Birth Weight, Gestational Age, and Daily Weight Gain Rate

<table>
<thead>
<tr>
<th>Prediction Model and Alarm Cut Point*</th>
<th>ROP</th>
<th>No. (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Reduction in the No. of Infants Requiring Examinations, No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOP ROP equation with cut point of 0.0140</td>
<td>Type 1 (n = 459)</td>
<td>452 (98.5)</td>
<td>2556 (36.4)</td>
<td>2563 (34.3)</td>
<td></td>
</tr>
<tr>
<td>CHOP ROP equation with cut point of 0.0026</td>
<td>Type 1 (n = 459)</td>
<td>459 (100)</td>
<td>509 (7.8)</td>
<td>509 (6.8)</td>
<td></td>
</tr>
<tr>
<td>CHOP ROP equation with cut point of 0.0034</td>
<td>Type 1 (n = 459)</td>
<td>459 (100)</td>
<td>784 (11.2)</td>
<td>784 (10.5)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: ROP, retinopathy of prematurity.

*Infants without sufficient weight data (n = 68) were considered to have received an alarm for ROP examinations.

© 2017 American Medical Association. All rights reserved.
collector certification process and extensive procedures to identify and correct data collection errors. While retinal examinations were not performed in an a priori fashion for the study, they were completed by ophthalmologists with expertise in ROP using standardized International Classification of Retinopathy of Prematurity terms and reflect the variety and, most importantly, the reality of clinical practice, which is the setting in which it is most critical to evaluate the performance of the model. Similarly, the weight measurements reflect the regular variations seen in practice. Using real-world data provides the most valid test of the model’s generalizability to clinical practice. Another limitation is the inability to generalize our findings to countries with less highly developed neonatal care systems. The performance of the model depends greatly on the medical setting in which it is applied. A weight-gain-based model, such as the CHOP ROP, may not perform well in countries where higher-BW and older-GA infants routinely develop severe ROP. Excessive oxygen use is more likely to have a central pathogenic role because endogenous IGF-1 production is already higher in older-GA infants. Therefore, postnatal weight gain would not reliably predict severe ROP, and separate predictive model development and validation studies are necessary in those settings.

Conclusion

More accurate prediction of severe ROP through incorporation of weight gain measurements into ROP screening deciles has numerous potential benefits. Updated screening guidelines could reduce the number of children receiving stressful eye examinations, as well as the frequency of examinations for low-risk infants. Professional and infrastructure resources can then be better allocated to high-risk infants. With this study, the CHOP ROP model has been validated for clinical use but would be more safely used to guide a modified screening schedule than replacement of current BW and GA examination criteria. To replace these criteria with lower-BW and younger-GA thresholds, additional study will be necessary to develop criteria for slow postnatal weight gain that more fully capture the higher-BW, older-GA outlier infants who develop severe ROP.
duPont Hospital for Children: Christopher M. Facorotta, MD (PI), Dorothy Hendricks, MD (PI), Alicia Olivant Fisher, MS, and Mark Paulin, MS.

Cost-effectiveness Component: Beth Israel Deaconess Medical Center: John Zupanick, MD, MS, ScD (PI).

Editorial Committee: Edward F. Bell, MD, Gil Binenbaum, MD, MSCE, Pamela Donohue, ScD, Graham E. Quinn, MD, MSCE, Laurea A. Tomlinson, BS, and Gui-shuang Ying, PhD.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Meeting Presentation: This study was presented in part at the 2016 Annual Meeting of the American Association for Pediatric Ophthalmology and Strabismus, April 7, 2016, Vancouver, British Columbia, Canada.

REFERENCES


