STATE OF WASHINGTON
DEPARTMENT OF LABOR & INDUSTRIES

Report Supporting the Review of Early Case Reserves Modeling System as of June 30, 2023

Deloitte Consulting LLP
December 6, 2023

Deloitte Consulting LLP
555 West 5th Street, Suite 2700
Los Angeles, CA 90013
USA
Tel: (213) 600-0100
www.deloitte.com

December 6, 2023

Mr. Steve Wendling
Audit Manager
Washington State Auditor's Office
3200 Capitol Boulevard
P.O. Box 40031
Olympia, WA 98504

RE: Department of Labor & Industries' Early Case Reserves Models Review

Dear Mr. Wendling,

Deloitte Consulting LLP ("Deloitte Consulting") previously completed its review of the Washington State Department of Labor & Industries' ("the Department") Early Case Reserves ("ECR") models, including data preparation, modeling methodology, quality control procedures, performance monitoring, and other relevant aspects of the models in December 2020, with a subsequent ECR model review and report update in December 2021. This report serves as an update to our original report issued on December 4, 2020 and the updated report issued on December 17, 2021, and includes comments based on the Department's efforts to enhance the ECR modeling system since our last review, in addition to the previous findings and conclusions.

Please note that starting on page 1 of the report below:

- Original text from the December 4, 2020 report as well as text from December 17, 2021 report update is in black font and not italicized.
- *New text from this report is in blue font and italicized.*

The Deloitte Consulting team appreciates the time and effort dedicated by the Department's modeling team to help us understand their modeling process, as well as the resources devoted to providing us with the appropriate information needed to perform our review.

Please contact us at the phone numbers below if you would like to discuss any aspect of this report or have any questions or comments.

Matthew Carrier is an Associate of the Casualty Actuarial Society. Vera Sakalova is a Fellow of the Casualty Actuarial Society. Matthew and Vera are Members of the American Academy of Actuaries and meet their qualification standards for rendering the opinions in this letter.

Sincerely,

Matthew Carrier, ACAS, MAAA
Principal
Deloitte Consulting LLP
(603) 234-4742
macarrier@deloitte.com

Vera Sakalova, FCAS, MAAA
Manager
Deloitte Consulting LLP
(617) 437-3723
vsakalova@deloitte.com

Table of Contents

Page

I. INTRODUCTION AND PURPOSE 1

II. RESULTS 2

III. COVID-19 IMPACT 7

IV. DISTRIBUTION AND LIMITATIONS 8

Distribution and Use 8

Data Reliance 10

V. INFORMATION RELIED ON 11

VI. APPENDIX 12

ECR Modeling System Overview 12

Deloitte Consulting's Detailed Recommendations 19

I. Introduction and Purpose

Deloitte Consulting LLP ("Deloitte Consulting", "us", "we" or "our") was retained by the Washington State Auditor's Office to provide professional services in support of the actuarial review of the Washington State Department of Labor & Industries' ("the Department") Early Case Reserves ("ECR") predictive modeling system. The system was developed by the Department to automate estimated case reserves for open claims which had been a challenge in the past due to resource constraints.

The case reserving process at the Department differs from a typical process used by workers' compensation insurance carriers. Not all open claims get case reserves assigned to them, and for those that do, case reserve amounts are not determined by claims adjustors. The Department's Case Reserve Unit ("CRU") assigns case reserves to open claims that have injury dates between 8 months and 7 years old and have been determined to have an indemnity component or are medical-only but with a high amount of paid medical losses. For claims that have a pension amount awarded, the pension managers estimate the reserve for the pension annuity portion of the claim. It should be noted that, prior to the ECR modeling system implementation, the fact that not all claims were case reserved caused settled and case reserved claims to be unfairly affected by loss development factors used in retrospective rating procedures when compared to open claims which were not case reserved.

Case reserving claims to obtain incurred loss amounts serves three main purposes for the Department:

- Premium rating, where it helps with more accurate estimation of the historical costs by class and employer (this is particularly important for retro-rated policies);
- Providing guidance to employers on which claims may require intervention and which claims are more likely to resolve quickly;
- Helping claim managers identify claims that are likely to become costly and complicated.

The ECR modeling system was developed to enable the Department to case reserve all open claims in order to improve the Department's premium rating, provide insights into potential claim severity and help with early identification of severe claims.

II. RESULTS

The ECR predictive modeling system was developed to enable the Department to case reserve all open claims in order to improve the Department's premium rating, provide insights into potential claim severity and help with early identification of severe claims.

Our work in reviewing the design and effectiveness of the ECR modeling system consisted of reviewing over 90 pages of materials provided by the Department based on the information requested by us, supplemented with over six hours of working sessions with William Vasek, the Department's Chief Actuary, Joshua Ligosky and Henry Cheng of the Department's modeling team.

Conclusion

> Deloitte Consulting was engaged to perform a review of the ECR predictive modeling system. Based on our review we conclude that the ECR models are fit for their intended purposes and provide value for the business case defined by the Department.

We also identify in this report certain improvements that can be made to variable analysis, modeling methodology, quality control, model maintenance and performance monitoring to improve the overall ECR solution and increase business value.

2021 Conclusion Update

> The Department's modeling team has begun developing enhancements to the ECR modeling system based on the recommendation made in our December 4, 2020 report. Based on our review, we affirm our previous conclusion that the ECR models are fit for their intended purposes and provide value for the business case defined by the Department.

2023 Conclusion Update

> *The Department's modeling team has not been able to make additional enhancements to the ECR modeling system since our 2021 review due to significant resource constraints. Based on our review as of this report date, we find that the ECR models are still fit for their intended purposes and provide value for the business case defined by the Department, although model performance is showing slight deterioration, particularly for Stage 2 models. Please see more detailed comments on this conclusion under Recommendations # 6 and #11 below. We recommend additional resources should be added to the modeling team to support model enhancements outlined in the Recommendations section below.*

Contained in this report are 13 recommended enhancements (see Deloitte Consulting's Detailed Recommendations section in the Appendix below) to the ECR models covering the areas of data analysis and variable selection, modeling methodology, as well as quality control, model maintenance and performance monitoring. These recommendations are organized by subject area and prioritized as High / Medium / Low using the following categorization:

- High Priority (4 total): these recommendations address potential quality and model design concerns and are expected to provide high value to the Department when implemented;
- Medium Priority (6 total): these items include additional recommendations to consider;
- Low Priority (3 total): suggestions to explore when time and resources permit.

In the remainder of this section, we discuss the four High Priority recommendations in order of importance.

Single Modeling Resource (Recommendation #10 in Appendix)

In our discussions, we noted that the Department currently relies on the knowledge and expertise of the ECR developer, Henry Cheng, for extracting data, developing, testing, and maintaining the models, as well as being responsible for various aspects of production use of the ECR. This is not ideal considering the number of models that currently need maintenance in the ECR and other modeling projects the Department has in production. The capacity of the Department's current modeling team does not allow for making on-going model improvements or implementing more rigorous quality control and performance monitoring procedures. Moreover, without the appropriate knowledge transfer to other members of the team, the Department is open to a significant risk of production failures if the ECR developer is unable to run the models for a period of time. We therefore recommend the Department consider options to increase modeling capacity to enable knowledge transfer, contingency planning, and model enhancements.

2021 Update: As of the time of this report, the Department has not been able to hire and train additional modeling resources due to budgetary constraints. Increasing the modeling team capacity to allow knowledge transfer, model enhancements and contingency planning remains one of our top recommendations.

2023 Update: Same comments apply as mentioned in 2021 update above.

Modeling System Complexity and Predictive Power (Recommendations #6 and #7 in Appendix)

The current ECR system design includes over 90 variables and over 400 models. Compared to industry standards, this is a very complex modeling system. In our discussions the Department mentioned the difficulty in maintaining this large number of models. A small portion of models did not perform well during validation exercises, specifically, months 1 and 2 and the models for Miscellaneous Accident Fund claims. We note that month 1 and 2 show the lowest coefficient of determination values, which is expected given data sparsity at these early evaluation points. Since the models use claim age from the date of accident and not from reporting date, a portion of claims do not get entered into the system until month 3 or later, and not all relevant claim data is coded electronically to be available for the models. Furthermore, as noted by the Department, Miscellaneous Accident Fund claims, which have some of the lowest coefficient of determination values, are significantly less frequent than other types of claims, and this can lead to lower observed model performance. Also, the Department has indicated that Miscellaneous Accident Fund benefits make up about 1% of claim costs. The main statistic used to assess the predictive power of stage 2 models (coefficient of determination, or R2) indicates that this portion of models did not perform as strongly as is typically expected by industry standards. The remainder of the models performed well for the most part.

We recommend enhancing variable analysis to identify weak predictors that can be eliminated and researching new strong variables that can be added to enhance the models. We also recommend researching ways to reduce the number of models that comprise the ECR, which could help simplify model maintenance. We believe the model design can be improved and simplified without compromising (and potentially even enhancing) the overall predictive power of the models. Longitudinal models, which are currently being researched by the Department's modeling team, could help improve the model design.

<u>2021 Update</u>: The Department modeling team has begun the process of developing longitudinal models based on our recommendation above. This process has not been completed as of the date of this report due to technology constraints. Longitudinal modeling allows the Department to reduce the number of models in the ECR modeling system from over 400 to around 60, while also increasing the number of observations used by each model – in some cases, to over 1 million rows. Initial statistical results from the first of ~ 60 models indicates that the longitudinal approach should provide a superior ECR solution from the version currently in production. However, the current technology utilized by the Department's modeling team is several years old *(the technology was 6 years old in 2021)* and has limited processing power. This technology limitation is causing significantly extended development timelines for each

longitudinal model due to the large sample size and number of predictive variables. The net effect of the technology limitation is that appropriate development, testing and implementation of a longitudinal ECR solution may not be possible. We recommend the Department consider performing a technology assessment to determine what alternative technology solutions are viable to support the Department's modeling and production solution requirements.

Additionally, the department is currently contemplating testing variables from an additional data source, electronically submitted Activity Prescription Form ("APF"), to include in the solution as a separate model. We note that as of the date of this report the Department has not conducted the analysis to identify weak predictors among the existing variables to eliminate from the models.

2023 Update: The Department modeling team has not been able to make progress on longitudinal model development or other model enhancements to address the above recommendation since our last review in December 2021 due to significant resource constraints.

The Department has provided additional model validation results for our current review per our request. These included observed and predicted correlations for medical aid, time loss and PPD benefits in Stage 2 models, calculated on model validation data. Based on our review of this data and comparison to model statistics shared in 2021, as well as discussions with the Department, we noted that models show much better performance on train data than on validation data. Models performing poorly on validation data could be indicative of overfitting or underfitting. Although this is typically not as much of a concern for random forest models as it is for GLMs, model performance between train, test and validation data could be made more consistent by refining the variable list. One suggestion is to evaluate variable significance across 10 cross-validation samples to ensure that there are no significant shifts in variable rankings. Standard deviation of variable rank between samples should be small to indicate that the variable performance is stable across different data samples. The variables showing significant shifts in performance across samples should be removed from models.

The Department also shared documentation showing a comparison of observed and predicted correlations between 2017, 2018, 2019, 2020 and 2021 ECR models for claims closed at 180 days from date of prediction. We noted a slight deterioration of Stage 2 model performance based on these additional model validation results. As mentioned in our conclusion, dedicating additional resources to make model enhancements to increase predictive power has become even more important. Please see additional comments under Model Validation / Performance Tracking recommendation #11 below.

Adjusting Historical Losses (Recommendation #1 in Appendix)

The Department shared during our working sessions that no adjustments were made to historical claim data to reflect trends in losses over time due to aspects like inflation, increasing medical costs, settlement practices, etc. Since the models were built on a historical dataset going back many years, leading practice would be to bring historical claims to present level by adjusting for trends. For example, the same 3 months old medical only claim occurring in 2007 vs 2015 will incur different paid benefits due to medical inflation between 2007 and 2015. Without trending, the same exact type of injury with the same corresponding medical procedures would have different medical costs in 2007 vs 2015, which would degrade the predictive power of the model. Trending all claims at each age to a common current level aligns with leading practices and could improve predictive power of the models.

2021 Update: The Department has incorporated medical aid, time loss, partial permanent disability and miscellaneous fund cost adjustment level factors used by the Actuarial staff into the ECR models based on the recommendation above. These factors account for historical benefit changes, including changes due to inflation and cost of living adjustment. All historical claim payments are adjusted using the cost adjustment level factors based on the date of injury.

2023 Update: This recommendation was addressed prior to our December 2021 model review and no further efforts are needed at this time.

III. COVID-19 IMPACT

The current ECR modeling methodology does not consider any potential direct or indirect effects of the COVID-19 pandemic. Currently, the course of the pandemic remains unclear and represents a significant source of uncertainty with respect to estimating workers' compensation case reserves and ultimate claim costs for individual claims. We note that it is likely that the distribution of claims between Medical Only, Time-Loss and Permanent Partial Disability claims, as well as claim benefits and durations will change due to COVID-19 and the effect of the pandemic may reach into 2021.

Although there may be changes to claim case reserves due to COVID-19 claims, we believe that attempting to estimate these differences would be difficult with immature data and may not be necessary if the effects of the pandemic subside prior to 2021. We note that the Department has made the decision to exclude all COVID-19 experience from experience rating, retrospective rating programs, and future retrospective rating adjustments. Therefore, we conclude that it is reasonable to exclude an estimate of the effect of COVID-19 claims from the current ECR models.

2021 Update: The Department has confirmed that the ECR modeling system does not consider the effects of the COVID-19 pandemic. COVID-19 claims are excluded from the models.

2023 Update: Same comment applies as mentioned in 2021 update above.

IV. Distribution and Limitations

Distribution and Use

This report has been prepared for the internal use of the SAO and the Department solely for the purpose of evaluating the appropriateness of the ECR models developed and implemented by the Department's modeling team. It is neither intended nor necessarily suitable for any other purpose.

The report may be provided to other parties ("Recipient"), for the purpose of evaluating the appropriateness of the ECR models if the following conditions are met:

- Deloitte Consulting is provided a list of Recipients to whom this report is provided.
- This report is being provided to the Recipient solely for its information and cannot and shall not be relied upon by the Recipient. The Recipient agrees that access to the report is not a substitute for the Recipient undertaking appropriate inquires and procedures in relation to its assignment.
- The Recipient agrees not to reference or distribute the report to any other party.
- The Department is solely responsible for providing accurate and complete information requested by Deloitte Consulting, and Deloitte Consulting has no responsibility for the accuracy or completeness of the information provided by, or on behalf of, the Department, even if Deloitte Consulting had reason to know of or should have known of such incompleteness.
- Deloitte Consulting has no responsibility to advise the Recipient of other services or procedures that might be performed and makes no representation as to the sufficiency or appropriateness of this report for the purposes of the Recipient.
- The Recipient acknowledges that the Department and the SAO have participated in the preparation of this report and the information, including, without limitation, by reviewing and commenting on prior drafts of this report and the information, and such participation may have resulted in the addition, modification or deletion of information which might be considered material by the Recipient.

- The Recipient acknowledges that Deloitte Consulting is currently providing and may in the future provide professional services to the Department, and the Recipient agrees that Deloitte Consulting and its personnel shall have no responsibility to the Recipient relating to such services nor any responsibility to use or disclose information that Deloitte Consulting possesses by reason of such services or otherwise, whether or not such information might be considered material by the Recipient.

- The Recipient acknowledges and agrees that the Recipient does not acquire any rights as a result of access to this report and Deloitte Consulting does not assume any duties or obligations as a result of access to this report.

- In the event that the Recipient is required by order of a court of competent jurisdiction, administrative agency or governmental body, or by any law, rule, regulation, subpoena, or any other administrative or legal process to disclose this report, the Recipient may disclose this report without liability hereunder, provided that the Recipient gives Deloitte Consulting prompt notice of any such requirement and, at our discretion, either (1) cooperates with us, at our expense, to prohibit such disclosure, or (2) uses all reasonable efforts to get confidential treatment of this report under a protective order or other appropriate mechanism. Furthermore, the Recipient may reference or disclose this report without liability hereunder in the event that such reference or distribution is required by professional standards bodies.

- By retaining a copy of this report the Recipient understands that such Recipient is deemed to have accepted these terms and conditions.

Deloitte Consulting shall have no liability, regardless of form, to any third parties (an entity other than the State Auditor's Office and the Department) for any action taken or omitted to be taken by such parties in respect of this except for matters that are finally judicially determined to be caused by Deloitte Consulting's own bad faith or willful misconduct. Third parties should recognize that the furnishing of this report is not a substitute for their own diligence and should place no reliance on this report or data contained herein that would result in the creation of any duty or liability by Deloitte Consulting to the third party. Any release or distribution of this report to any third party must include the report in its entirety (with discussions and exhibits).

This report has been prepared for use by individuals who have a degree of technical competence in insurance and predictive modeling matters. This report should be studied in its entirety before any judgments are made about the conclusions in the report. It is our intention that this report be used in its entirety, as a whole, and not segmented for other purposes. Deloitte Consulting personnel are available to discuss any questions or concerns regarding this report.

The services we performed throughout this engagement did not constitute an audit, review, examination, or other form of attestation as those terms are defined by the American Institute of Certified Public Accountants ("AICPA"). Any use of the word "review" within this report should be interpreted in the common use of that term, and not in the definition of "review" promulgated by the AICPA.

DATA RELIANCE

Deloitte Consulting has relied upon data provided by the Department for this review. A specific audit to verify the accuracy or completeness of the data is beyond the scope of this engagement. While we have reviewed the data in regard to its reasonableness and consistency for our review, we have relied on such data without audit or verification and our conclusions are based on the assumption that it is accurate and complete. In addition, Deloitte Consulting facilitated six total hours of working sessions with William Vasek, the Department's Chief Actuary, Joshua Ligosky and Henry Cheng of the Department's modeling team to gather further information on the ECR modeling process and methodology. If the underlying information provided is inaccurate or incomplete, the results of our analysis may likewise be inaccurate or incomplete.

V. INFORMATION RELIED ON

Over the course of our ECR models review, we have relied on the following information provided to us by the Department's modeling team:

- ECR modeling system documentation contained in "ECR_8_18_20.docx";
- Comments on 2021 ECR modeling system improvements contained in "improvement2021.docx";
- *Model validation documentation provided in "ECR2023Cheng.docx";*
- Answers provided to our initial 2020 data request and follow-up questions;
- Answers provided to our initial 2021 data request and follow-up questions;
- *Answers provided to our initial 2023 data request and follow-up questions;*
- Additional data shared by the Department, including histograms of error terms, R^2 values, variable analysis examples, actual vs expected medical aid benefit log-plots and observed vs predicted matrices for stage 1 models;
- Information gathered by us during six plus hours of working sessions conducted with William Vasek, the Department's Chief Actuary, Joshua Ligosky and Henry Cheng of the Department's modeling team for the December 4, 2020 report;
- The additional 2 hours of working sessions conducted for the 2021 report update. Note that one of the 2021 working sessions also included Ali Ishaq, an Actuary hired by the department in 2021.
- *The additional 2 hours of working sessions conducted for the 2023 report update, which included William Vasek, Henry Cheng, Joshua Ligoski and Ali Ishaq.*

We note that we did not review the actual model code or model output as part of this analysis.

VI. Appendix

ECR Modeling System Overview

2023 Update: The ECR Modeling System Overview section below still applies to the ECR models currently in production, with a couple of updates. The Department has informed us that the response variables in Stage 2 models have changed as follows:

- *Incurred Time-Loss benefits at 24 months has been updated to Incurred Time-Loss benefits at 36 months*
- *Incurred Medical benefits at 24 months has been updated to Incurred Medical benefits at 36 months.*

Furthermore, the duration Stage 2 models are not currently used in production.

2021 Update: The following ECR Modeling System Overview section still applies to the ECR models in production with one exception; as noted above, the current ECR production solution is developed with historical losses that have on-leveled to current costs while the ECR production solution in place as of December 4, 2020 was developed using nominal historical losses.

Data and Variable Overview

The ECR models were built using the data available in the Department's data warehouse. The Department pointed out that while some of the data warehouse tables used in modeling were updated daily, others were only updated weekly, monthly or quarterly, causing some difficulties in synchronizing data. The resulting modeling data consisted of a sample set of 727,796 data points based on open claims between 1 and 18 months old (for example, a claim that is open for 18 months could have 18 different observations in the sample set), with age of claim calculated from the accident month. Since most of the data recorded in the Department's data systems is intended to pay bills and process claims and not for predictive modeling purposes, the data needed to be manipulated to handle missing information and restructured to make it appropriate for modeling. The following approaches were used to handle missing information:

- For categorical variables, an "unknown" or "missing" level was created;
- For continuous variables, missing information was replaced with either the mean value (calculated on the training dataset) or zero, whichever was more appropriate.

Some of the continuous variables were converted to categorical to make them more suitable for the modeling process. All categorical variables were limited to 25 levels (where levels are separate values of the variable, for example a variable that can have a value of either "true" or "false" has 2 levels), except for one variable that has 28 levels. Variable levels with frequencies below the 23 highest were grouped into a 24th level, with the 25th level reserved for records with missing information.

The modeling process excluded any rejected and self-insured claims and claims that had negative paid-to-date benefits. The latter were reported to the Department's quality control units as incorrect data. The Department confirmed during our working sessions that they are confident that over 99% of the claims that go through the ECR system today are valid claims, although there is a small chance of rejection in rare cases, e.g. when an employer decides to appeal the claim.

Another data question that needed to be addressed by the Department was the treatment of "Kept-on-Salary" claims. These are cases when employers choose to keep the injured workers on salary, which means these workers become ineligible for Time Loss ("TL") benefits. Employers typically choose this approach to increase their retrospective rating refunds. For the purposes of the ECR models, the decision was made to include "Kept-on-Salary" claims with Medical Only ("MO") claims. However, in some cases these claims become long duration and workers eventually start receiving TL benefits or Permanent Partial Disability ("PPD") awards, which means the claim type changes from MO to TL or PPD.

Several hundred claim-related fields were initially pulled from the Department's database. The current version of the ECR modeling system uses approximately 90 variables. About half of these variables are extracted directly from the data, with the other half being synthesized. The predictor variable selection process relied heavily on expert opinion. The initial variables were selected based on claim staff interviews. Consideration was given to variable availability and relative stability, and presence of potential data issues in the variables. The Department's model developer shared that he did not have much time for testing individual variables during the variable selection process, although several summary statistics were reviewed and variables that were deemed unlikely to provide predictive information were rejected.

One of the advantages of the ECR modeling system utilizing tree-based type models (please refer to the section on modeling methodology below) is that they are able to handle interactions between predictor variables, which would be problematic for general linear models. The ECR system uses the same variable set for all models and relies on model fitting to determine the most important variables to use for each model at each age, based on maximum predicted accuracies or largest coefficients of determination. We note that while this might simplify model maintenance, it also slows down the model training process and adds a degree of risk of overfitting some of the models, where the models fit too closely to the training dataset and may fail to predict future observations reliably.

Modeling Methodology Overview

The ECR modeling system consists of hundreds of cross-sectional models projecting individual claim case reserves for claims 1 to 18 months old. The models were built on a sample set of 727,796 open claims at varying ages. Currently, separate sets of models are in production for claim ages 1, 3, 6, 9, 12, 15, and 18 months. Month 3 models are built on data for claims 2 to 3 months old, month 6 models – on claims 4 to 6 months old, etc. This approach results in inclusion of all relevant claim data for model training purposes and allows for smoothing some of the differences in predictions between the models at different claim ages.

The ECR is a two-stage modeling system, with stage 1 models predicting the type of claim at 72 months, and stage 2 models projecting the associated benefit costs and durations. The Department notes that the main reason for choosing claim type at 72 months as the response variable for stage 1 models is the high duration of open claims in the Department's systems. For instance, analysis showed the median claim age when total permanent disability pensions are granted to be seven years, with claims appearing as TL or PPD prior to pensions being granted. In addition, PPD awards are typically only granted after a TL claim closes.

One of the important reasons the Department chose the 2-stage modeling approach is the importance of predicted claim type in the downstream actuarial calculations – for example, different loss development factors apply to different claim types, and some employers receive indemnity claim-free discounts as part of experience rating.

Stage 1 Models

The goal of stage 1 models is to predict how each claim will be categorized at age 72 months: as PPD, TL or MO claims. The ECR system uses Classification Tree models to make these predictions, which is done with the *rpart()* package in R, an open source programming language and statistical software. Several other software packages and model types were tested by the Department for this purpose, including *tree()*, *cforest()*, *randomforest()* and *GLM()* packages in R, support-vector machines, neural networks, the C4.5 algorithm and Salford Systems' CART. The selected *rpart()* package was found to perform better based on the overall comparison of prediction accuracy, run time and memory usage.

The size of the decision trees built by the *rpart()* package is controlled with the complexity parameter. The complexity that resulted in the lowest ten-fold cross validation error was selected for each tree.

Please note that for the purposes of the ECR models, fatal claims are grouped with PPD claims. The ECR documentation shared by the Department also states that the system is not designed to predict pension award amounts or structured settlement claims.

The ECR documentation provides a table summarizing the number of claims with accurately predicted claim type (PPD, Time-loss or Medical Only) at 72 months, along with the percentage of predictions that were accurate, from the 10-fold cross validation procedure performed for stage 1 models. We were also provided with actual vs predicted claim types for each of the claim age months included in current production models. Based on this information, the models performed reasonably well in predicting the type of claim, with overall prediction accuracies averaging 91%. PPD prediction accuracies start at 61% for month 1 models and gradually increase to 96% for month 18 models, which can be explained by the proportion of open PPD claims gradually increasing with the claim age (only 10% of open claims in the modeling sample at age 1 month are PPD, whereas at 18 months PPD claims comprise 73% of all open claims due to the majority of TL and MO claims being closed by that time). TL claims prediction accuracies range between 78% and 89%, while MO prediction accuracies range from 87% to 98%.

The Department also performed a model validation test on closed claims to assess the performance of models being used in production. For stage 1 models in production, overall prediction accuracies were estimated on closed claims from August 8, 2016 to the date of the report. Prediction accuracies were estimated at 24.04% for PPD claims, 65.65% for TL claims and 97.42% for MO claims. Although validating the models on closed claims less than four years old can create a bias due to PPD claims typically taking longer to close, significantly lower prediction accuracies for TL and especially PPD claims in this model validation exercise compared with the model training/testing data could indicate potential overfitting of the models. It could also be an indication of the model not being responsive to the changes in claims over time.

Table 1 below summarizes predicted accuracies from the 10-fold cross validation procedure performed for stage 1 models along with the corresponding accuracies from the model validation test performed on closed claims.

Table 1: Comparison of stage 1 model accuracies on model data and validation data

	Predicted Accuracies			
Data set	**Overall**	**PPD**	**TL**	**MO**
10-fold Cross Validation	91.04%	90.65%	84.45%	96.37%
Validation Closed Claim Data (≥08/08/2016)	76.39%	24.04%	65.65%	97.42%

Stage 2 Models

Stage 2 models predict benefit amounts and durations for each claim applicable for the claim type using Random Forest Models. This is done via a *randomforest()* package in R. Several other packages and model types were tested for stage 2 models, including *party()*, *gbm()* and *GLM()* R packages and Salford Systems' Random Forest. The *randomforest()* R package was selected after comparing the coefficients of determination, total sum of squares, computation time, and plots of observed and predicted values for both small and large datasets.

Stage 2 models currently include the following response variables:

- Incurred PPD benefits at 72 months;
- Incurred Time-Loss benefits at 24 months;
- Incurred Miscellaneous Accident Fund ("Misc") benefits at 24 months;
- Incurred Medical benefits at 24 months;
- Incurred Time-Loss benefits duration (in days) at 24 months;
- Medical benefits duration (in days) at 24 months (was 48 months prior to 2/21/2020).

For certain benefit types, stage 2 modeling is performed in multiple parts. This includes Miscellaneous Accident Fund Benefit models and PPD Award models. For Miscellaneous Accident Fund benefits, this is done due to a predominant number of claims having a zero benefit. The first part of the models predict which claims will have non-zero payments using a Classification Tree model, while the second part utilizes a Regression Tree model to predict the benefit amounts for these non-zero claims.

PPD benefit amounts are also modeled in two parts, where the award is first classified into one of seven categories based on the type of injury, then a separate model is run for each of the predicted injury types, resulting in a total of 8 models.

The table below summarizes the number of different models run at each claim age month.

Table 2: Number of stage 1 and 2 models used at each claim age month

Claim Type*	Claim Subtype	PPD Amount	TL Amount	TL Duration	Misc Amount	Medical Amount	Medical Duration
PPD	Hearing Loss	8	—	—	2	1	1
	Carpal Tunnel		1	1	2	1	1
	Kept on Salary		1	1	2	1	1
	All Other		1	1	2	1	1
Time-Loss	Carpal Tunnel	—	1	1	2	1	1
	Kept on Salary	—	1	1	2	1	1
	All Other	—	1	1	2	1	1
Medical Only	Hearing Loss	—	—	—	2	1	1
	Carpal Tunnel	—	—	—	2	1	1
	Kept on Salary	—	—	—	2	1	1
	All Other	—	—	—	2	1	1

**Determined in stage 1 models*

We note that, in order to avoid negative outstanding case reserves, the ECR models predict case outstanding benefits and not total case incurred benefits.

For stage 2 Random Forest models, the ECR documentation uses the coefficient of determination (R^2) to assess model performance. The R^2 values range from 0.13 to 0.75, with only 54% of models having coefficients of determination above 0.5. The R^2 is an evaluation metric that calculates the proportion of variance in the response variable that can be explained by the predictor variables, and low R^2 values can indicate that the model does not fit the data well.

We note that for stage 2 model validation, we were provided with plots of observed vs predicted values for medical aid benefits on all closed claims 2010 through mid-2016, separately for PPD, TL and MO claims. These plots indicate that the models predicted medical aid benefits for validation data reasonably well. We have not been provided with more recent stage 2 model validation results or validation results for response variables other than medical aid benefits.

Quality Control Procedures

The following statistics are currently being gathered from model input and output data at every model run for the purpose of maintaining quality control and detecting changes or potential problems that occur within the ECR modeling system:

- Missing information;
- The levels (or distribution of values) of each categorical variable;
- Presence of negative values in variables that should be real and positive

In addition, the ECR system recorded all historical estimates for models run between April 2015 and May 2020 to track reasonability of changes over time. This process outputs several charts, including the number of open claims and corresponding total case reserves, which allowed for identification of data problems, such as an issue with unreliable closed claims indicator that was detected in May and June 2019. Another chart showed the proportion of open claims at various ages by claim type and mean case incurred values by claim type.

Performance Monitoring and Model Maintenance

The ECR predictive system is currently run and maintained by the developer, Henry Cheng, who also performs data extraction and reviews model results and summary statistics. The models are being run on a weekly basis. The output data from the models is kept by Henry Cheng on the servers and transferred to a mainframe database where it can be used in downstream processes.

The ECR models are being updated every three to six months. The update process involves training new models using updated historical data and testing the proposed new models against two historical datasets to ensure reasonability of results.

L&I's Plan for Future Model Enhancements

As stated in ECR documentation, one of the weaknesses of tree-based regression is that it could take more variables to pick up on and describe trends in the data compared to using a mixture of tree based, linear and nonlinear regression models. One of the contemplated improvements to the ECR modeling system is using a hierarchical modeling technique, so that the final model is a combination of non-parametric models such as tree-based regression, and parametric models such as generalized linear models. Certain R packages are being considered for this hybrid models, such as mob() and cubist().

Additionally, the Department is considering using longitudinal models to replace cross-sectional models, which could be built using R REEMTREE() package. Longitudinal models provide several advantages, such as increasing the number of observations available for each model, enhanced insights into the effects of each predictor variable over time and reducing the number of predictive models. However, developing these models is a very time-consuming process that will require a significant amount of time and resources.

The ECR documentation also discusses using specific models for certain claim types in the future based on deeper understanding of these types of claims, such as hearing loss. The Department notes that the majority of hearing loss claims tend to only have hearing loss benefits, although there are some complicated cases involving additional benefit types. Using parametric or semi-parametric models for modeling specific claim types such as hearing loss would be more feasible due to smaller datasets available for these claims. The Department is also considering the option of using a model for hearing loss claims which would be consistent with the actuarial reserving process. This would involve using a hearing loss annuity table as claimants are typically eligible for these benefits for life. The actuarial reserves for hearing loss claims would then be derived from the sum of all predicted case incurred amounts.

Other potential enhancements considered by the ECR include extracting information from unstructured data such as medical reports for use in models, using the new claim complexity code developed for the Auto Adjudication predictive modeling system, and extending the use of ECR modeling system to claims older than 18 months.

DELOITTE CONSULTING'S DETAILED RECOMMENDATIONS

The section below outlines our recommendations and suggestions for potential adjustments that can be made to the ECR modeling system to improve performance of the models and increase business value. These recommendations are organized by subject area and prioritized as High / Medium / Low using the following categorization:

- High Priority: these recommendations address potential quality and model design concerns and are expected to provide high value to the Department when implemented;
- Medium Priority: these items include additional recommendations to consider;
- Low Priority: suggestions to explore when time and resources permit.

Data Analysis and Variable Selection

1. Adjusting Historical Losses (High): The Department shared during our working sessions that no adjustments were made to historical claim data to reflect trends in losses over time due to aspects like inflation, increasing medical costs, settlement practices, etc. Since the models were built on a historical dataset going back many years, leading practice would be to bring historical claims to present level by adjusting for trends. For example, the same 3 months old MO claim occurring in 2007 vs 2015 will incur different paid benefits due to medical inflation between 2007 and 2015. Without trending, the same exact type of injury with the same corresponding medical procedures would have different medical costs in 2007 vs 2015, which would degrade the predictive power of the model. Trending all claims at each age to a common current level aligns with leading practices and could improve predictive power of the models.

   2021 Update: The Department has incorporated medical aid, time loss, partial permanent disability and miscellaneous fund cost adjustment level factors used by the Actuarial staff into the ECR models based on the recommendation above. These factors account for historical benefit changes, including changes due to inflation and cost of living adjustment. All historical claim payments are adjusted using the cost adjustment level factors based on the date of injury.

   2023 Update: This recommendation was addressed prior to our December 2021 model review and no further efforts are needed at this time.

2. Variable Analysis (Medium): Although expert opinion and domain knowledge can provide valuable information on which variables should be included in the models, it is often beneficial to conduct variable selection using a more analytical approach in addition to relying on expert opinion. Not only can univariate analysis be helpful in detecting correlations between potential predictor variables, it can provide helpful information about the predictive strength of each potential variable and help eliminate unpromising variables early in the process. This can in turn significantly speed up the model training process.

   2021 Update: The Department's modeling team has not conducted univariate analysis to enhance the ECR modeling system variable selection process as of the date of this report, due to competing priorities and resource constraints. As the Department continues to develop longitudinal models, new techniques will be added to help address this recommendation.

   2023 Update: Same comments apply as mentioned in 2021 update above.

3. Third Party Data/Information (Medium): The ECR models could benefit from supplementing its internal data with the additional insights provided by external data sources. Some of these data sources could be used to add predictive variables on the insured businesses or individual claimants (e.g. Experian, Dun & Bradstreet); others can add helpful geographical information (e.g. US Census data). Example variables from these various sources include:

   a. Financial condition of the insured company

   b. Distance between business address and injured employee home address

   2021 Update: The Department has not included third party data in the ECR models as of the date of this report due to time and resource constraints.

   2023 Update: Same comments apply as mentioned in 2021 update above.

4. True MO vs Medical on "Kept-on-Salary" (Medium): The Department has shared that case reserves are used in individual account pricing, which is most important for retro-rated policies. In cases when an employer chooses to keep the injured workers on salary, these workers are not eligible for TL benefits, which benefits employers if the claims are short duration. These claims are currently combined with MO claims, however, it would be beneficial to separate these claims into a different category from true MO, since they are typically associated with higher medical costs and longer duration.

2021 Update: The Department clarified that "kept-on-salary" claims are modeled separately from the true MO claims.

2023 Update: Same comments apply as mentioned in 2021 update above.

5. Seasonality (Low): For modeling seasonality, the Department might consider using the season of the year, or something similar, in place of the quarter, or a variable that ties more closely to the intended use. These could provide better alignment with seasonality of certain types of work. For example, summer work might be from May through September, peaking in July and August. In this case you could have a variable that is zero for January through April, Perhaps 0.5 and 0.75 for April and May, a 1 for July and August, 0.75 for September, and maybe zero for October through December.

   2021 Update: The Department has not included seasonality in ECR models as of the date of this report due to the low priority of this recommendation. However, this recommendation is being evaluated as part of the longitudinal models; initial testing indicates that this variable does not have a strong predictive power.

   2023 Update: Same comments apply as mentioned in 2021 update above.

Modeling Methodology

6. Model Complexity (High): The current ECR model design includes over 90 variables and over 400 models. Compared to industry standards, this is a very complex modeling system. In our discussions the Department mentioned the difficulty in maintaining this large number of models. We believe the model design can be improved and simplified using longitudinal models without compromising (and potentially enhancing) the overall predictive power of the models.

   2021 Update: The Department modeling team has begun the process of developing longitudinal models based on our recommendation above. This process has not been completed as of the date of this report due to technology constraints. Longitudinal modeling allows the Department to reduce the number of models in the ECR modeling system from over 400 to around 60, while also increasing the number of observations used by each model – in some cases, to over 1 million rows. Initial statistical results from the first of ~ 60 models indicates that the longitudinal approach should provide a superior ECR solution from the version currently in production. However, the current technology utilized by the Department's modeling team is several years old and has limited processing power. This technology limitation is causing significantly extended development timelines for each longitudinal model due to the large sample size and number of predictive variables. The net effect of the technology limitation is that appropriate development, testing and implementation of a longitudinal ECR solution may not be possible. We recommend the Department consider performing a technology assessment to determine what alternative technology solutions are viable to support the Department's modeling and production solution requirements.

Additionally, the department is currently contemplating testing variables from an additional data source, electronically submitted Activity Prescription Form ("APF"), to include in the models. We note that as of the date of this report the Department has not conducted the analysis to identify weak predictors among the existing variables to eliminate from the models.

2023 Update: The Department modeling team has not been able to make progress on longitudinal model development or other model enhancements to address the above recommendation since our last review in December 2021 due to significant resource constraints.

The Department has provided additional model validation results for our current review per our request. These included observed and predicted correlations for medical aid, time loss and PPD benefits in Stage 2 models, calculated on model validation data. Based on our review of this data and comparison to model statistics shared in 2021, as well as discussions with the Department, we noted that models show much better performance on train data than on validation data. Models performing poorly on validation data could be indicative of overfitting or underfitting. Although this is typically not as much of a concern for random forest models as it is for GLMs, model performance between train, test and validation data could be made more consistent by refining the variable list. One suggestion is to evaluate variable significance across 10 cross-validation samples to ensure that there are no significant shifts in variable rankings. Standard deviation of variable rank between samples should be small to indicate that the variable performance is stable across different data samples. The variables showing significant shifts in performance across samples should be removed from models.

The Department also shared documentation showing a comparison of observed and predicted correlations between 2017, 2018, 2019, 2020 and 2021 ECR models for claims closed at 180 days from date of prediction. We noted a slight deterioration of Stage 2 model performance based on these additional model validation results (some of the deterioration could be driven by the impact *of COVID-19 on claims). As mentioned in our conclusion, dedicating additional resources to make model enhancements to increase predictive power has become even more important. Please see additional comments under Model Validation / Performance Tracking recommendation #11 below.*

7. Predictive Power (High): It appears from the range of coefficients of determination (R2) that there are some stage 2 models with good predictive power, while for some of them the predictive power is considerably lower (for almost 20% of the models, the R2 values are below 0.4). The concern is that while the system might have good accuracy in aggregate across all claims, it may have very poor accuracy for individuals or employers. Two common causes of this are missing important variables or overfitting to the training data. A third cause could be the overall complexity of the system of models. Reducing the number of models and variables used in the ECR system could help enhance predictive power of the models.

   2021 Update: Please refer to the update for Recommendation #6 above.

   2023 Update: Please refer to the update for Recommendation #6 above.

8. Probability vs Class (Medium): It is our understanding that the classification tree models used in stage 1 and some stage 2 models choose one class or the other, even if two classes were nearly equally likely. What could provide additional value is designing the classification models to not output a class, but a probability that a claim is of each class. Then when applying the second stage models, a weighted average of their predictions is used. It provides a smoothing or blending of the answers from different branches of possibilities. This would also be particularly effective for the Miscellaneous Accident Fund Benefit model. Although this adds some computation time to the prediction phase, the benefits of this approach typically outweigh this drawback.

   2021 Update: The Department has not included this recommendation in their 2021 ECR modeling system update due to competing priorities and modeling resource constraints. However, the Department plans to investigate this as part of the longitudinal modeling effort.

   2023 Update: Same comments apply as mentioned in 2021 update above.

9. Incremental Enhancements (Medium): If the Department determines that a simpler solution architecture is insufficient, we would recommend focusing on individual aspects of the solution one at a time. For instance, the Department could develop a method for measuring how well

each part of the model is performing individually – not at a system level, but at an individual model level. Identify a piece that has the most room to improve, using the appropriate metrics, and focus your effort on that single piece. We have found this approach particularly effective on complex multi-model systems such as the ECR. Focusing on a specific model to find areas of improvement can also provide insights that can be applied to enhance other models.

2021 Update: The Department has not included this recommendation in their 2021 ECR modeling system update due to competing priorities and modeling resource constraints.

2023 Update: Same comments apply as mentioned in 2021 update above.

Quality Control, Performance Monitoring and Model Maintenance

10. Single Modeling Resource (High): We note that the Department currently relies on the knowledge and expertise of one ECR developer for extracting data, running, testing, and maintaining the models. This is not ideal, especially considering the large number of models involved in this process. There is also significant risk to production use of the model if the developer becomes unavailable for an externed period. We recommend the Department look for additional modeling capacity to perform knowledge transfer and contingency planning.

    2021 Update: As of the time of this report, the Department has not been able to hire and train additional modeling resources due to budgetary constraints. Increasing the modeling team capacity to allow knowledge transfer, model enhancements and contingency planning remains one of our top recommendations.

    2023 Update: Same comments apply as mentioned in 2021 update above.

11. Model Validation / Performance Tracking (Medium): We recommend tracking model performance of each individual model as well as performance of the overall system. If the model input and output data is being archived at every model run, it could be used to perform ongoing comparisons against closed claims to assess model performance. This could be done across different versions of the model. In addition, since stage 2 models project all but PPD benefits and durations at 24 months, model validation can be performed on an ongoing basis as new data becomes available and model predictions can be compared against actual claim data at 24 months (using claims that have been reviewed by case reserving staff, for example). For PPD benefits, limited validation can be performed on closed claims.

2021 Update: The Department has not enhanced the model validation process as described above as of the date of this report due to competing priorities and modeling resource constraints.

2023 Update: The Department has provided additional model validation results for our current review per our request, including a comparison of observed and predicted correlations between 2017, 2018, 2019, 2020 and 2021 ECR models for claims closed at 180 days from date of prediction. Based on this data, we noted a slight deterioration in Stage 2 model performance in recent years. We recommend implementing additional model performance tracking mechanisms to increase visibility into model performance changes over time. The Department should keep track model performance statistics (see the paragraph below for some examples) to understand how they are changing over time.

Final assessment of each updated model version should depend on multiple criteria – e.g., RMSE, observed and predicted correlations, Lift Curves – to determine if model performance is improving. Deterioration in model performance metrics should indicate that a model refresh is needed.

12. Data checks (Low): Another helpful metric could be tracking changes in data over time and comparing them against the changes in model predictions. For example, track proportion of PPD, TL and MO claims over time and contrast them to the proportion of predicted claims. Same could be applied to predicted benefits and durations, as well as distributions of independent variables over time. This could allow for early identification of trends and potential issues in the data which could affect model performance.

    2021 Update: The Department has not included this recommendation in their 2021 ECR models update due to its low priority and modeling resource constraints.

    2023 Update: Same comments apply as mentioned in 2021 update above.

13. Model Updates (Low): As models evolve, it's important to compare performance of individual models and the overall system to previous versions of the model, ideally maintained in a source-code control system, such as Git. To do this effectively, the Department should balance two conflicting goals. First, you want to test models on recent hold-out data because that is most likely to represent what the model will experience in production. Second, you also want to test against a stable, consistent hold-out data set so that you can do a fair comparison between different versions of each model. While there is no perfect solution, one approach is to maintain two different hold-out test sets, one for each purpose above. One test set is maintained for,

perhaps a period of 3-6 months and is used to compare versions of models. Periodically it is replaced with more recent data set, and tests are run on available models for both the old and new versions of this stable dataset to provide some level of continuity. The other data set might be updated monthly to represent "current" claims. This data set should typically be based on information that is more recent than the data used to train the models. Perhaps obvious, but it might be worth mentioning that it is important not to have a version of the same claim in the training and test or validations sets. This can occur when the model has claim X at age 2 months, and test set has claim X at age 3 months. Both claims would be associated with the 3-month model.

2021 Update: The Department has not included this recommendation in their 2021 ECR models update due to its low priority and modeling resource constraints.

2023 Update: Same comments apply as mentioned in 2021 update above.