

# Bringing Predictive Modeling to the DSP: An Investigation of Claim Severity

BY EMILIE DUBOIS, ALISON MILFORD, AND P. DIVYA PARIKH

**In studying the drivers of medical professional liability (MPL) claims, there is an ongoing need to find reliable ways to identify and mitigate risk and to control costs. In a data-intensive world, MPL needs to be analytic, conclusive to the extent possible, and predictive. For these reasons, predictive models are more important than ever for making sense of the information we have and for estimating and planning for what might happen in the future.**

The Medical Professional Liability Association (MPL Association) has been collecting data on closed claims, via its Data Sharing Project (DSP), since 1985. Today, this data includes details on the type of events that gave rise to the claim, among many other data elements. Analyses of the DSP data have provided a foundation for risk management and patient safety programs for MPL Association insurers for decades, and this effort continues today. The DSP database contains more than 300,000 claims closed between 1985 and 2017 (approximately 29% of the claims closed with an indemnity payment), representing \$22 billion in indemnity payments and \$9 billion in defense costs.

Information is voluntarily submitted to the DSP semiannually from a subgroup of MPL Association member companies.

This article presents a summary of one predictive model that was constructed with the DSP data. (Greater detail is presented in a forthcoming white paper on this project, to be published later this year.) In this context, predictive modeling is understood to be the use of statistical techniques to extract knowledge from large volumes of data. It is the use of data and models that are empirically derived and statistically valid to make decisions and inform actions.

## Methodology

Predictive modeling, one type of *multivariate analysis*, makes possible simultaneous consideration of many variables, as well as the assessment of their overall effect. When a large set of claims is analyzed, patterns begin to emerge, revealing a number of characteristics to consider.

More specifically, the analysis discussed here relies on generalized linear models, a predictive modeling approach that is commonly used in the insurance industry. Such models have many potential uses throughout the insurance business, including pricing and underwriting, marketing, claims, and risk management.

The goal in this analysis is to understand the drivers of claim severity, on the assumption that if we understand the particular circumstances that are likely to give rise to large claims, the severity can be mitigated, avoided altogether, or at least better managed once reported to the insurance carrier. Note that the DSP data cannot be used to analyze claim frequency because criteria were established at the outset of the

**Table 1. Model Variables by Importance**

### Model Variables by Importance

Outcome (e.g., infection, disease, cancer)  
State

#### Important

Chief medical factor (e.g., communication, diagnostic error)  
Patient age  
Procedure (e.g., physical exam, surgery)

#### Less Important

Facility (e.g., hospital, physician office)  
Full time / part time  
Insured gender  
Patient gender  
Patient gender / patient age interaction  
Specialty

#### Footnote:

Within each importance group, variables are listed alphabetically, not ranked. In addition, we have excluded the control variables of coverage type, policy limit, and report year.

DSP to intentionally exclude any information that might be used for competitive rate-making purposes, including insured counts.

For this analysis, we focus on DSP claims settled between 2006 and 2015. Two models were built, one for each of two target variables: (1) paid indemnity and (2) paid expense. The final data set for the indemnity model, which will be the focus of our discussion here, includes approximately 24,000 claims. (The complete analysis is reviewed in the white paper.)

The data set includes 24 potential explanatory variables. Various interactions are studied by analyzing the combined impact of multiple explanatory variables. Because the objective of the modeling exercise is to provide information for claims or risk management professionals, the variables reviewed are limited to those that could reasonably be known from an initial investigation of the claim. For this reason, claim outcome variables are excluded from the modeling exercise.

The indemnity model includes nine variables and one interaction variable that combines multiple factors, in addition to the control variables of state, policy limit, coverage type, and report year. In Table 1,

the full listing of variables included in the model is shown with a broad ranking of their relative importance, as defined by the magnitude of their impact on claim severity. The two most important variables are found to be medical outcome and state. Chief medical factor, patient age, and med-



*Emilie Dubois, FCAS, MAAA, and Alison Milford, ACAS, MAAA, are with Willis Towers Watson; and Divya Parikh, MPH, is with the Medical Professional Liability Association.*

ical procedure also ranked as important.

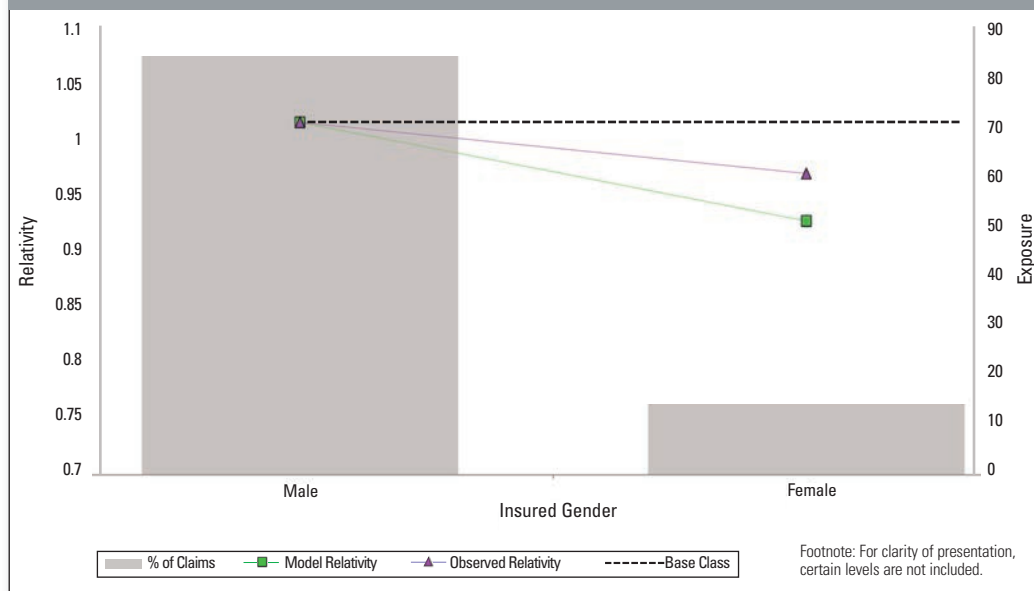
The results for the key model variables are summarized as a relative for each variable level. For each variable, a base level was selected as the reference point. The relativity for this base level is 1.00, and the results for all other levels are expressed relative to this base level. For example, a variable level with an indicated relativity of 2.00 implies that, in the context of this multivariate analysis, the level has an average severity that is twice that of the base level, *holding all other factors constant*. Thus, each relativity indicates the impact of that particular variable alone.

Also included in the results is the observed average relativity. This is the typical result of a univariate analysis, which will thereby reflect the impact of any correlation across variable levels, since all of the other factors are not held constant. The results of a few variables are highlighted below; a more complete discussion of the model results can be found in the white paper.

In Figure 1, a univariate review of the data indicates that female physicians have lower-severity claims, on average. However, female physicians in this data set are more likely to work part-time, are disproportionately concentrated in certain specialties, and are more likely to perform certain types of procedures, thereby potentially confounding the impact of gender on claim severity. Our model was used to isolate the impact of the single factor, physician gender, and in fact indicates a greater impact from physician gender on indemnity severity—an approximately 8.5% lower severity for female physicians with otherwise similar claims.

In Figure 2, medical procedures are ordered from less invasive (including the base class of general physical exam) on the left to more invasive (surgical) procedures on the right. Procedures are grouped by bodily system. For several less invasive procedures, the relativity as obtained from the model is broadly in line with the relativity derived from a univariate analysis. However, the observed average severities based on the data for claims related to most surgical procedures are lower than the average severity for claims associated with general physical exams, an unexpected result. The results for the model-derived relativities for surgical procedures are more in line with what one might anticipate. Once other factors are controlled for, the average severity for surgical procedures is expected to be higher than that associated with general physical exams, as indicated by the fact that most of the relativities are greater than 1.00, and only the relativities for the generally less invasive surgeries (skin, eyes, nose, mouth, throat) are less than 1.00.

Figure 1. Comparison of Model and Observed Relativities - Insured Gender Variable



### Potential model applications

Risk management/patient safety and claims departments are the two areas most likely to benefit from applications of these results. One potential application is based on a claim-scoring exercise. Scoring a claim involves deriving the expected indemnity for the claim, based on its specific characteristics and the associated model-derived relativities. The expected severity for each claim can then be used in sorting a batch of claims.

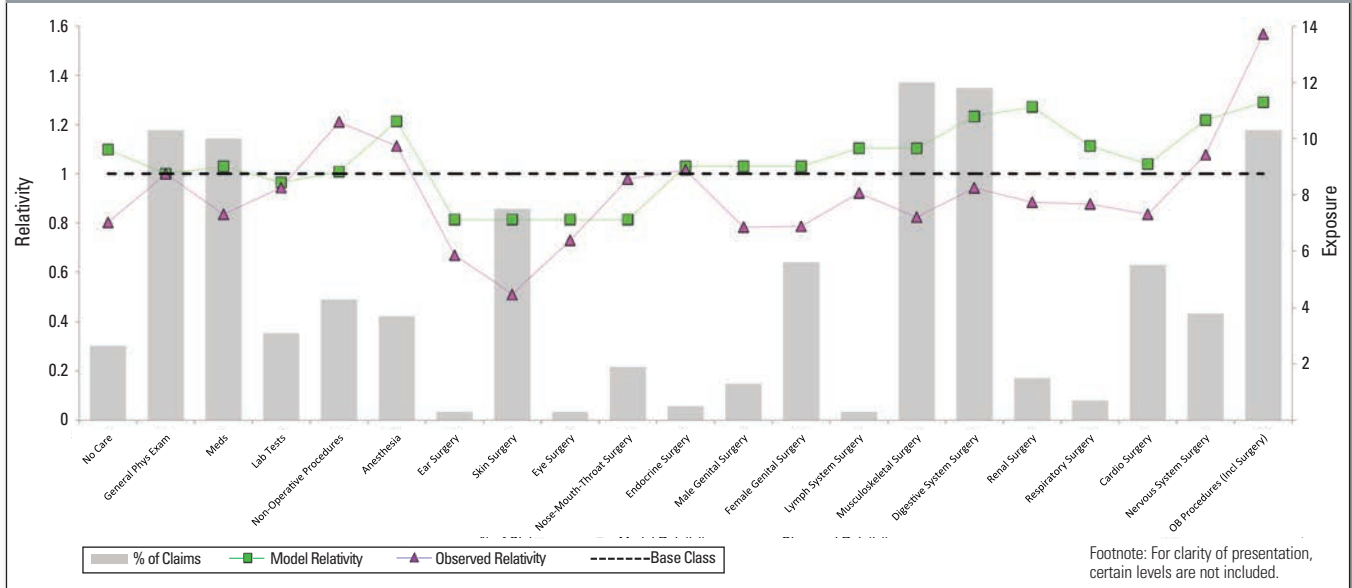
Once the claims are sorted, claims identified as high or low severity can be analyzed to determine which groups of factors tend to be associated with particular severity outcomes. The results of such an exercise can be used to answer any number of questions, including:

- What risk management programs can be designed or improved to reduce payouts?
- Are the limited resources of claims departments being deployed in an optimal way?

The results of this sort of exercise may serve to confirm current knowledge of claim outcomes. For instance, it is well known in the MPL community that claims related to adverse birth outcomes have a higher than average severity. It is likely, however, that such an exercise, if based on a sufficiently rich data source, will identify pairs, or groups, of characteristics associated with high-severity outcomes not previously known that may be difficult to identify based on univariate analysis alone. The multivariate analysis would of necessity combine medical and analytical expertise.

Thus, the analysis described here is only a starting point for the many possibilities in future research and investigation. As expected, the multivariate analysis of thousands of claims showed that variables related to the medical features of the case, including the outcome of the medical interaction at issue, the medical procedure involved, and the chief medical factor in the claim, have a significant relationship to the eventual size of the claim payment. Multivariate analysis of claim

**Figure 2. Comparison of Model and Observed Relativities - Procedure Variable**



severity has significant potential benefit for risk management and claims departments: it allows combinations of characteristics contributing to claim severity to be identified and, possibly, mitigated.

Future research opportunities beyond this review of the DSP are abundant, with the potential for additional detailed modeling of claim severity, similar modeling of claim frequency, and deeper analysis of

the applications of the results presented here. The key is leveraging all of the available information to produce estimates and forecasts that are relevant to the medical community and insurers, thereby improving quality, lowering costs, and optimizing decision making. **MPL**

For related information, see [www.mplassociation.org](http://www.mplassociation.org).

