Department of Transportation
Federal Motor Carrier Safety Administration

SUPPORTING STATEMENT
"CRASH RISKS BY COMMERCIAL MOTOR VEHICLE (CMV) DRIVER SCHEDULES" ICR
OMB Control No. 2126-TBD

Part B. Collections of Information Employing Statistical Methods

1. DESCRIBE POTENTIAL RESPONDENT UNIVERSE AND ANY SAMPLING SELECTION METHOD TO BE USED.

The potential respondent universe consists of CMV carriers that hold and maintain a DOT number that include drivers that operate under the ELD mandate (excluding short-haul operations that operate with 150 air-mile radius). Drivers at eligible carriers must hold a valid CDL (Commercial Driver's License) and be at least 21 years of age. According to the FMCSA Pocket Guide, as of December 2022, there were 813,844 FMCSA-regulated carriers (interstate motor carriers and intrastate HM motor carriers) with recent activity operating in the United States, and these carriers had a total of 5,800,289 drivers, 4,172,559 of which were CDL drivers. Among CDL drivers, there are estimated to be 3.5 million drivers operating under the ELD mandate. Eligible drivers will not be excluded based on the type of schedule, driving pattern, or operational condition.

CMV operators may be segmented based on operational classification – they may be a for-hire motor carrier (where transporting goods for other people is the core business) or a private motor carrier (typically transporting their own goods or persons as part of their internal logistics chain). The table below sets out the segmentation of CMV operators by operation type.

Table 11: Number of carriers by operation classification (FMCSA Pocket Guide 2023).

Operation Classification	Number of Carriers
For-Hire	540,093
Private	200,904
Both	70,129
Neither (e.g., Government)	2,718

Carriers vary by size, ranging from large motor carriers (i.e., having several thousand power units) to small carriers (i.e., having only a single power unit, such as independent owner-operators).

Table 12: Number of carriers by size (FMCSA Pocket Guide 2023).

	Number of Power Units	Number of Carriers
Small	1-10	739,727
Medium	11-100	56,361
Large	101 or more	4,908
Other	No Power Units/Unreported	12,848

Motor carriers can also be segmented by interstate versus intrastate carriers, by geographic extent (i.e., long-haul, regional, or local trucking), or by type of transport (i.e., freight, passengers, or hazardous materials).

Table 13: Number of carriers by operation (FMCSA Pocket Guide 2023).

	Number of Carriers
Interstate Freight	763,867
Interstate Passenger	10,019
Intrastate Hazardous Materials	39,958

Among motor carriers, there is a wide range of vehicle configurations that are represented, including buses, single-unit trucks, and tractor/trailer (combination trucks). The configuration can vary as well, including no trailer (bobtail), single trailer, double or triple trailers, flat-bed, intermodal, and refrigerated.

We will aim to recruit study participants from a variety of carriers in order to capture a diverse sample of driver types and driving schedules. For this study we will target the carrier types as illustrated in Table 14 below.

Table 14: Targeted study enrollment plan.

Year	Carrier Type	Number of Carriers	Average Number of Drivers per Carrier	Number of Years Participation per Carrier	Expected Number of Driver Years
1	Small	5	5	4	100
1	Medium	5	50	4	1,000
1	Large	5	1000	4	20,000
2	Small	5	5	3	75
2	Medium	5	50	3	750
2	Large	5	1000	3	15,000
3	Small	5	5	2	50
3	Medium	5	50	2	500
3	Large	5	1000	2	10,000
4	Small	5	5	1	25
4	Medium	5	50	1	250
4	Large	5	1000	1	5,000
1-4	Total	60	Avg: 351	-	52,750

To meet our participant recruitment targets, we will use randomized and targeted outreach to generate a non-probability sample. A non-probability sample is necessary due to the voluntary nature of participation and the absence of any mechanism to mandate involvement among CMV carriers. We will use a model-based analytical approach that includes relevant covariates to account for potential sources of selection bias introduced by a non-probability sample.

The sampling frame will consist of all carriers in the FMCSA census that maintain an active DOT number and operate under the ELD mandate (with no other exclusion criteria for carriers). Only drivers at least of the age of 21 with a valid CDL will be included from each carrier. In the first study period (year 1) we will take the sampling frame, randomize the order, and begin by contacting the first 1500 carriers to solicit participation in the study. We will support recruitment through coordinated efforts with FMCSA, by advertising in trade publications, and by promoting the study at industry conferences. Advertising in leading trade publications is expected to reach over 30,000 fleet executives. The study team regularly attends industry conferences (TCA Truckload Convention hosts over 1,300 trucking executives). To further encourage carrier participation, the study team will offer carriers safety reports with an analysis of fatigue risk in their operation. For each subsequent study period (years 2-4), we will repeat the above steps.

The enrollment plan attempts (to the extent practicable) to achieve a data set representative of the trucking industry by recruiting broadly across a wide range of segments and driver types. However, bias is anticipated in part due to the promotional efforts that will be made by the study team (e.g. targeted outreach with limited audience) as well as self-selection bias. Self-selection bias is anticipated due to the fact that participation is voluntary, and it has been demonstrated that companies who participate in these health and safety studies generally already have mature safety culture and have adopted better safety measures than most other companies. Also, because this study relies on the cooperation and assistance of companies to gather data, larger companies that are well resourced may be overrepresented in the study data. These biases will be accounted for by the modeling approach, but have the anticipated effect of underestimating the true risk of fatigue in the industry. We will document our recruitment and data collection methods in publications to guide interpretation relative to potential bias.

DESCRIBE PROCEDURES FOR COLLECTING INFORMATION, INCLUDING STATISTICAL METHODOLOGY FOR STRATIFICATION AND SAMPLE SELECTION, ESTIMATION PROCEDURES, DEGREE OF ACCURACY NEEDED, AND LESS THAN ANNUAL PERIODIC DATA CYCLES.

1.1. Data collection plan

To participate in the data collection effort, carriers must hold and maintain a valid DOT number. Drivers at eligible carriers must hold a valid CDL (Commercial Driver's License) and be at least 21 years of age. They also must be currently employed as a CMV driver at an eligible carrier. As mentioned above, recruitment will be done using the FMCSA census, through coordinated efforts with FMCSA, in trade publications, and at industry conferences. The sampling will aim to recruit broadly to represent the industry.

1.2. Description of Analysis Data Set

Once carriers are enrolled in the study, the driver HOS data will be collected through an integration with the ELD provider used by the carrier. The driver's hours-of-service (HOS) records are required for regulatory compliance and are recorded in electronic form by in-cab ELD compliant devices. Throughout the duty day, drivers update their HOS data to reflect their duty status and that data is transmitted automatically to the ELD servers. Through Pulsar's Fatigue Meter integration with the ELD vendors, the carrier's ELD data is transmitted via API (Application Programming Interface) to the research team in an automated manner in near real-time. For any driver involved in DOT-reportable crash, the crash report is submitted to FMCSA as part of standard procedures and is stored in the federally maintained MCMIS (Motor Carrier Management Information System) database. The MCMIS database, which includes both DOT-reportable crashes and inspection violations, will be transmitted by DOT to the research team through regular (monthly) exports of the database during the entire study period. Carriers will also be able to separately provide the research team with additional SCE events or crashes (including non-DOT reportable crashes).

In early discussions about potential participation in this study, CMV carriers identified data retention as a concern. Current regulations require carriers to retain HOS data for a period of six months. To minimize litigation risk, many carriers delete HOS data after this minimum retention period. The feedback received from such carriers is that any primary HOS data transmitted to the study team would need to be deleted in a similar timeframe. The data processing plan devised by the research team to accommodate these data retention concerns will encapsulate and transform the data into de-identified feature tables. The data in the feature tables will be stored in a form that supports present (and future) data analysis while remaining suitably deidentified. In the discussion that follows, we describe the scheme for producing the analysis data set.

The study aims to understand the relationship between a dependent variable (e.g., the occurrence of a crash) and one or more independent variables (factors). The feature table approach supports this type of analysis. Each row of a feature table represents a single driver's duty day, identified based on the HOS logs (i.e., a new duty day starts when a driver resumes duty after a minimum of ten consecutive hours off-duty). Then, for each duty day, a set of features (or variables) will be extracted based on the duty history (ELD data) and inspection violation, crash, and incident records associated with the duty. An example structure of a Duty Day feature table is presented in Table 15 below, and an illustration of a duty day is presented in Figure 1.

Table 15: Example Duty Day feature table structure. Each row represents a driver duty day and contains features related to the day, such as HOS data, crash data, and inspection data.

Duty Day ID	Set of Duty Date Features	Set of HOS Features	Set of Crash Features	Set of Inspection Violation Features	Set of Driver Features	Set of Carrier Features
1						
2						

The complete set of Duty Day features is too exhaustive to show in tabular form, but a simplified version is shown below.

Table 16: Example Duty Day feature table structure, with subset of HOS features (hours driving in last 7 days, hours driving in duty, hours on-duty in duty), and subset of crash features (number of crashes) shown.

Duty Day ID	Hours Driving in Last 7 Days Prior to Duty Day	Hours Driving in Duty Day	Hours On-Duty in Duty Day	Number of Crashes
1	60	10	3	0
2	52	8.5	1	1
3	52	8.5	2	0

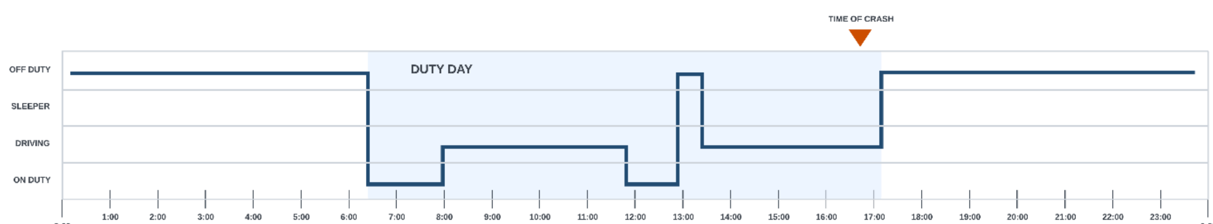

Figure 1: HOS logs are analyzed at the unit of a duty day. Here, the duty day begins at 6:30am and ends at 5:10pm. The occurrence of a crash at 4:45pm becomes one of the features in the analysis table.

To support an analysis that examines the relationship between crash rates and cumulative hours driven, we have created a second feature table that divides the duty day into separate driving hours. In this Driving Hour feature table, each row represents one hour driven during a duty period. The occurrence of a crash during the period is indicated with a Boolean flag (i.e., 0 = no, 1 = yes). The Duty Day table and the Driving Hour feature table are linked by the Duty Day ID field. An example Driving Hour feature table is presented in Table 17, below. The Driving Hour table shown below is simplified, but it can be joined with the Duty Day table based on the common Duty Day ID to create a table that includes a full set of features calculated for the Duty Day.

Table 17: Example Driving Hour feature table. The column labeled "Crash" contains a Boolean value.

Duty Day ID	Driving Hour	Hours Driven (0 – 1)	Crash
1	1	1	0
1	2	0.5	1
1	3	0	0

Together, the Duty Day and Driving Hour feature tables contain a full set of features that will support the analyses outlined above. They are envisioned to also serve the needs of future unplanned analyses. A Driver ID field is intentionally not included in either feature table to support the deidentification objective (a de-identified driver ID is used temporarily during data processing but is removed in the final version of the feature tables). While the removal of the driver ID is necessary for creating a deidentified data set, we recognize that its omission will limit certain analyses (i.e., longitudinal designs and within-subjects designs).

<u>Features</u>
The following variables are included in each row of the feature tables (* denotes primary independent variables of interest).

Duty Date Variables:
- Calendar year
- Calendar month
- Calendar day
- Day of the year (e.g., 173rd day of the year)
- Duty start/end time of day (binned into 1-hour bins)
- Hour of driving start/end time of day (binned into 1-hour bins) (*)

Work Variables:
- Cumulative hours on-duty during duty
- Cumulative hours driven during duty (*)
- Total number of breaks during duty (*)
- Cumulative duration of breaks during duty
- Cumulative duration of time in sleeper berth during duty
- Total duration of rest prior to current duty period
- Total duration in sleeper berth during rest prior to current duty period (5)
- Consecutive duration in sleeper berth during rest prior to current duty period (5)
- Total duration in personal conveyance during rest prior to current duty period (6)
- Total duration in personal conveyance during rest after current duty period (6)
- Cumulative hours of personal conveyance in prior *P* days (*P* = up to 180 days)
- Total duration of last restart
- Number of nights of last restart (*)
- Cumulative hours worked/driven since last restart at start of duty

- Cumulative hours worked/driven since last restart, up to a period of *Q* days (*Q* = 1 to 30) at start of duty
- Cumulative duty days since last restart at start of duty
- Cumulative hours in sleeper berth in recent past period of *Q* days since the restart (*Q* = 1 to 30)
- Cumulative hours worked/driven over the past *Q* days (*Q* = 1 to 180) (*)
- Model-based fatigue level estimate for the duty period at duty start, end, and peak
- Average model-based fatigue level estimate over the past *Q* days (*Q* = 1 to 180)
- Model-based sleep estimate prior to duty
- Average model-based sleep estimate over the past *Q* days (*Q* = 1 to 180)

Crash and Inspection Violations Outcome Variables:

- Number of crashes/violations in the current duty day
- Binned time of any crash/violation (binned into 1-hour bins)
- Time to event (hours of driving from start of duty) of crash/violation in current duty day
- Cumulative number of Crashes/violations over the past *Q* days

Driver Variables:

- Age
- Sex
- Total number of years of driving experience

Carrier Variables:

- Carrier size (Small/Medium/Large)
- Carrier type (Interstate/Intrastate/Freight/Passenger/Hazardous)
- Carrier operation classification (For-Hire, Private, Both, Neither)

Procedure for Generating Analysis Data

The feature tables described above would normally be produced once all the data had been collected, validated, and linked between the disparate data sources to produce the desired data set for analysis. However, to accommodate the desired rolling six-month data retention limit for raw HOS data, the analysis data set must be generated in a continuous fashion during the data collection study period. A second consideration is the asynchronous manner that data is received (see Table 18, below). For example, while HOS data is received in near-real time, crashes and inspection violations in the MCMIS database are reported monthly.

Table 18: Data sources and timing of when data from each source are obtained.

Data source	Received	Notes
ELD HOS	Real-time	HOS corrections (~1 week)
MCMIS Crashes	Monthly	Reporting lag of up to 4 months for crashes to be entered in MCMIS
MCMIS Inspection Violations	Monthly	Reporting lag of up to 4 months
Crashes directly reported by carrier	As provided	Reporting lag of up to 4 months

The approach we have devised involves ingesting and storing HOS data in a database that will fully represent the driver logs. An automated process then deletes any HOS entry that is older than six months. Any carrier-provided data (incidents and crashes) will also be ingested and deleted in a similar manner. The MCMIS data will be ingested into the database as received. Since MCMIS data is not subject to a six-month retention limit, such data will be kept in its entirety for the full duration of the study. See Figure 2 below.

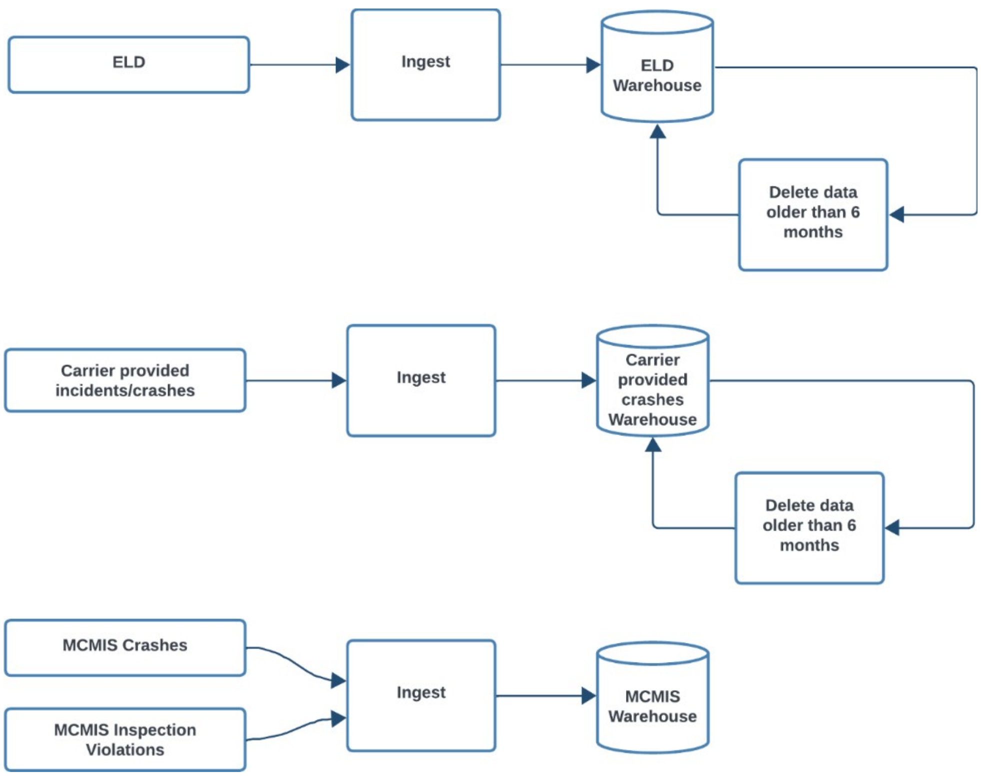

Figure 2: Data ingestion and deletion process flows to accommodate retention requirement applicable to HOS data.

To create the deidentified data set while adhering to the retention limit applicable to HOS data, a new row of the Duty Day feature table (which represents one duty day) will be generated initially when that day’s HOS data is first received. Corresponding to the same duty day, the new rows will be generated in the Driving Hour feature table (with one row for each duty hour in the duty day). At the time of creation, this entry will be marked with the current date and time. Any new data associated with that entry (e.g., MCMIS data) that is received within the six-month period following the date/time stamp will be stored in the corresponding feature variable records. See Figure 3 for an illustration of this process.

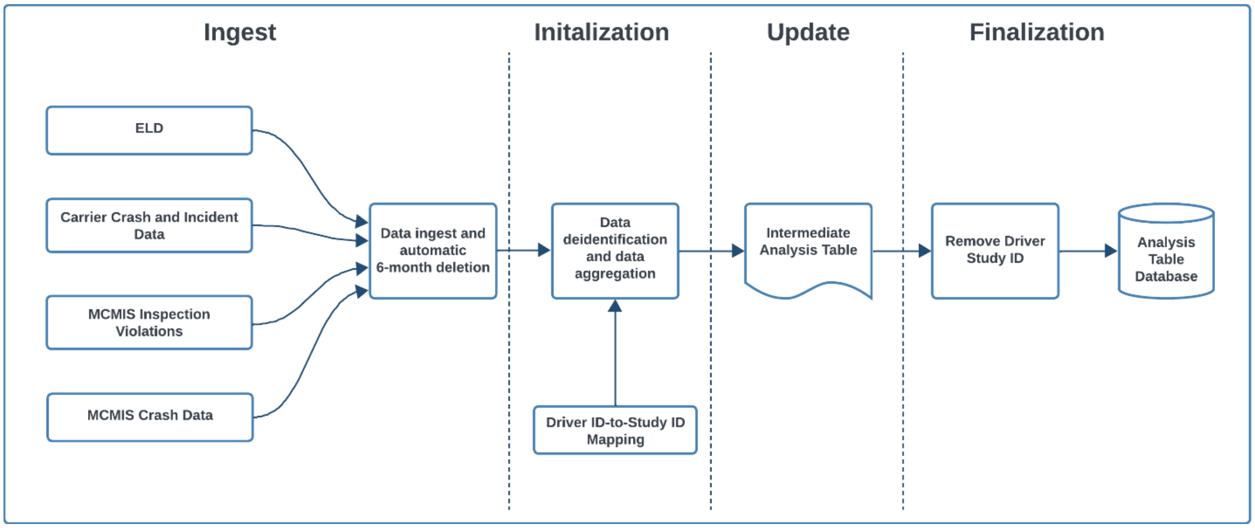

Figure 3: Data processing workflow to generate final analysis data set. Source data is ingested and deleted based on the six-month retention policy. The primary HOS and crash data are deidentified and aggregated into an intermediate analysis table that retains the deidentified subject ID for six months. Data in the intermediate data table is mutable for six months while relevant data (e.g., crash data) is received and updated. The table entries in the intermediate analysis table that reach the six-month mark are stripped of the subject ID, and the row entry is copied into the final analysis table, then deleted from the intermediate analysis table.

<u>Initialization (Day 1)</u>

On a daily basis as HOS records are received and new duty days based on the HOS logs are identified, a new row is created in the intermediate table. This table is identical to the analysis table with the important addition of driver *study* ID information. The driver study ID is anonymized, assigned, and used only for generating and updating the intermediate analysis tables (e.g., Joe Smith is assigned a driver study ID of CR0001). Upon creation, all of the feature variables of the table are set based on the available data. The HOS data for the particular driver would extend back 180 days, and any features that can be extracted based on the work patterns in the prior six months are added.

<u>Update (Day 2 to Day 179)</u>

As new data is received (e.g., MCMIS and carrier reports), any crashes and inspection violations that are associated with the duty day entry in the intermediate table will have the relevant features updated. The crash that is reported may not have occurred on the specific duty day but may be still relevant to that entry because of a metric that counts the number of the crashes in the prior 30 days. Updates of data may also affect the HOS logs due to corrections that are made from time to time. When HOS corrections are made that affect the table entry, those entries are recalculated based on the latest available data.

<u>Finalization (Day 180)</u>

In the finalization step, the driver study ID is removed from an intermediate Table A, and the resulting entry is copied over into the Deidentified Data Table. Once the entry has been copied to Table B, then no further edits are possible. (If, for instance, a MCMIS data export is received that shows a crash relevant to that duty day, it cannot be added.)

At the time of finalization, elements of the data record may be known to be missing. For example, driver demographic information may be missing from the carrier, and no updated information was available prior to the end of the six-month cutoff period. Features that are known to be missing will be marked as missing, and rows with missing values may be excluded from analysis or have the missing values imputed. To minimize missing data, ongoing data validation will be performed to identify and correct errors and gaps throughout the process.

During this six-month period, HOS data will be linked with any incident data from the MCMIS database. However, the reporting and updating of incidents in the MCMIS database have a lag that may be several months removed from the time of the incident. While most of the incidents are expected to be reported and available within the six-month data retention period, there may be incidents that are not reported, or updates to incidents that are not reported, within the six-month period. Missing incidents may bias the results if there are a significant number of events that are not reported, or if specific types of events are not reported within a timely manner. As part of the study, we will log and report the number of MCMIS events that are received beyond the six-month period following a duty day associated with an incident.

The figure below illustrates the processing and intermediate tables produced at various points over a representative six-month period.

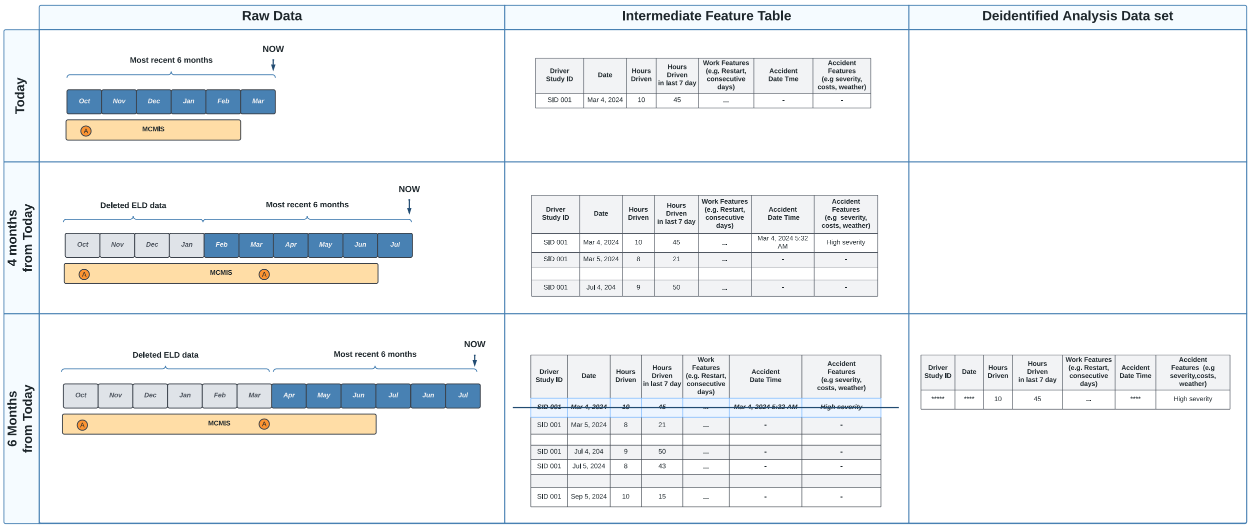

Figure 4: Illustration of relationship between raw data, intermediate feature tables, and final feature tables.

Description of Analysis Plan

HOS data will be processed to identify duty cycles and duty days for a driver based on the HOS regulations. Typically, duty days are demarcated by periods of at least ten consecutive hours off-duty. Duty cycles, or “work sets,” are the series of duty days demarcated by restart periods of at least 34 consecutive hours off-duty. For each duty day, relevant features are calculated:

- Hours of driving in the 7 days (144 hours) prior to start of duty day
- Number of 30-minute (or more) breaks within the duty day
- Number of nights (1-5AM periods) in restart period at the beginning of a work set
- Hours in sleeper berth during rest periods between duties
- Number of crashes in duty period

The duty day can be subdivided into the hour of driving (up to 11 hours of driving for compliant logs, and up to 13 hours for compliant logs during adverse driving condition exception). Within each hour of driving for a duty day, we calculate:

- Hours driven within driving hour (i.e., from 0 to 1h)
- Number of crashes within the driving hour

An example data table illustrating this analysis is presented in Table 19 below.

Table 19: Hour of driving data table example.

Driver	Duty	Hour of Driving	Hours Driven in Driving Hour	Number of Crashes
Driver A	1	1	1	0
Driver A	1	2	1	0
Driver A	1	3	1	0
Driver A	1	4	1	0
Driver A	1	5	1	0
Driver A	1	6	1	0
Driver A	1	7	1	0
Driver A	1	8	1	0
Driver A	1	9	1	0
Driver A	1	10	1	0
Driver A	1	11	0.5	1
Driver A	2	1	1	0

<u>Unit of Analysis: Hour of Driving</u>

To assess relative risk, we will use generalized multiple linear regression (GLM) models to estimate the effects of individual factors on study measures (e.g., crash rate, etc.), while adjusting for covariates. Through the appropriate choice of link function in the GLM, e.g.log-link, the parameters of the GLM model when exponentiated can be interpreted as the relative risk of crash associated with each factor. The generalized multiple linear regression models follow the general structure of the example below. The covariates in the analysis can include a broad range of factors, including driver (e.g., age, sex) and carrier (e.g., size, type) factors and may also include other factors including model-based estimates (e.g., model-based estimated prior sleep and model-based prior fatigue).

$$f(Y_{ij})=\beta_0+\beta_1 X_{1ij}+\beta_2 X_{2_{ij}}+\beta_3 X_{3_{ij}}\ldots+\beta_{11} X_{11_{ij}}+\beta_{12} X_{12_{ij}}+\ldots+\beta_n X_{n_{ij}}+\epsilon_{ij}$$

where:

- Y_{ij} is the number of crashes observed at i^{th} hour of j^{th} duty day
- $f(Y_{ij})$ is the generalized link function
- $X_{1ij}\ldots X_{11ij}$ is an indicator variable for driving at 1^{st} through 11^{th} hour, of j^{th} duty day
- $X_{12ij}\ldots X_{nij}$ are covariate factors
- $\beta_1\ldots\beta_{11}$ are the corresponding parameters for the driving hour factors
- $\beta_{12}\ldots\beta_n$ are the corresponding parameters for the covariate factors
- ϵ_{ij} is random error

More compactly, and including covariates, we can write:

$$f(Y_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_{cov} X_{covij} + \epsilon_{ij}$$

where:

- Y_{ij} is the number of crashes observed at *i*th hour of *j*th duty day
- $f(Y_{ij})$ is the generalized link function
- X_{1ij} is the vector of indicator (0 or 1) variables to denote the drive hour (e.g., one indicator variable for each hour, hour 1 through hour 11 of driving)
- β_1 is the corresponding (vector) parameter
- X_{covij}, β_{cov} are the covariate features and parameters
- ϵ_{ij} is random error

The effect of time of day on crash risk is a key factor to the risk of crash, based on both traffic patterns and circadian effects on fatigue. The time-of-day factor will be included as a covariate in all analyses by encoding as either a dummy variable or a continuous variable using a sine-cosine expansion.

For each driving hour, the time of day (TOD) can be encoded using dummy indicator variables based on a binned approach:

$$f(Y_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_{tod} X_{¿d_{ij}} + \beta_{cov} X_{covij} + \epsilon_{ij}$$

Alternatively, for each driving hour, the start (end or middle) time of day of each driving hour can be encoded using a sine-cosine expansion:

$$f(Y_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_{¿d_{sin}} \sin\left(\frac{\pi}{12} x_{¿d_{ij}}\right) + \beta_{¿d_{cos}} \cos\left(\frac{\pi}{12} x_{¿d_{ij}}\right) + \beta_{cov} X_{covij} + \epsilon_{ij}$$

Both time-of-day representations will be considered, with the final selection determined by model fit.

<u>Unit of Analysis: Duty Day</u>

The analytical approach based on the count of crashes/violations for the Duty Day unit of analysis is similar:

$$f(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_{cov} X_{covi} + \epsilon_i$$

where:

- Y_i is the number of a crashes observed on *i*th duty day
- $f(Yi)$ is the generalized link function.
- $X_{1i}$is the independent variable of interest (e.g., cumulative hours of driving in prior 7 days)[a]
- β_1 is the corresponding (vector) regression parameter

[a] The variable of interest can be treated as continuous variable or a vector of indicator variables, when the variable is binned into discrete. For example, the cumulative hours of driving in the prior 7 days may be treated as a continuous variable (e.g., 0 – 80 hours) or as binned into discrete groups (e.g., 0 – 20 hours, 21 – 40 hours, etc.), with a vector of indicator variables to encode which bin the duty belongs.

- X_{covi}, β_{cov} are the covariate features and covariate parameters
- ϵ_i is random error

The effect of time of day on crash risk is a key factor to the risk of crash, based on both traffic patterns and circadian effects on fatigue. The time-of-day factor will be included as a covariate in all analyses by encoded as either a dummy variable or a continuous variable using a sine-cosine expansion.

For each duty day, the duty start (duty end or middle) time of day can be encoded using dummy variables based on binned approach:

$$f(Y_i) = \beta_0 + \beta_1 X_{1ij} + \beta_{tod} X_{\text{¿}d_i} + \beta_{cov} X_{covij} + \epsilon_i$$

Alternatively, for each duty day, the duty start (end or middle) time of day can be encoded as a continuous variable using a sine-cosine expansion:

$$f(Y_i) = \beta_0 + \beta_1 X_{1ij} + \beta_{\text{¿}d_{\sin}} \sin\left(\frac{\pi}{12} x_{\text{¿}d_i}\right) + \beta_{\text{¿}d_{\cos}} \cos\left(\frac{\pi}{12} x_{\text{¿}d_i}\right) + \beta_{cov} X_{covij} + \epsilon_{ij}$$

An analytical approach based on the timing of crashes in the duty day extends the duty day analysis and allows the event rate to be time varying $h_0(t)$ instead of a fixed rate β_0 above, can be expressed:

$$h(t) = h_0(t) * \exp\left(\beta_1 X_1 + \beta_{cov} X_{cov}\right)$$

where:

- $h(t)$ is the hazard rate during the duty period for driving hour t
- $h_0(t)$ is the baseline hazard rate during the duty period for driving hour t
- $X_1$is the independent variable of interest (e.g., cumulative hours of driving in prior 7 days)[b]
- β_1 is the corresponding (vector) regression parameter
- X_{cov}, β_{cov} are the covariate features and covariate parameters

In the generalized linear models, the crash counts will be modeled with Poisson (or Negative Binomial or zero-inflated variants) distribution. The outcome is the number of crashes within each sampling unit (e.g., hours of driving or duty day), modeled as a function of fixed factors and covariates, and follows a similar approach in a prior study of CMV driving[1]. The generalized linear model uses a log link function and an offset equal to the natural log of exposure time (hours driven within a duty day) to account for differences in driving duration across observations.

With the use of a log-link function, the relative risk to be estimated by exponentiating the model parameters. For example, comparing the crash rate associated with specific schedule factors to the baseline crash rate assuming all other factors fixed yields the following relative risks:

$$Relative\ Risk = \frac{E[Y \vee X = exposed\ factor]}{E[Y \vee X = baseline]}$$
$$\text{¿}\ \frac{\exp\left(\beta_0 + \beta_1 (X_1 = 1) + \beta_{cov} X_{cov}\right)}{\exp\left(\beta_0 + \beta_1 (X_1 = 0) + \beta_{cov} X_{cov}\right)}\ \text{¿}\ e^{\beta_1}$$

The duty day and by driving hour analysis may be extended to a multiple regression model where the multiple independent factors (e.g., number of breaks, number of nights in restart, cumulative hours prior to duty; in the case of driving hour analysis, this includes both the driving hour factor and other independent factors, such as number of night restarts, cumulative hours prior to duty, etc.) are all included in a single regression model to better account for correlation among factors.

The analysis on the time-to-event data will be based on a proportional hazards model (Cox regression) that models the hazard rate as a function of both a time-varying baseline hazard function and factors (hazards). Similar to the Poisson regression models, the coefficients of the Cox model can be interpreted as the relative risk (or, more precisely, hazard ratios) associated with each factor (e.g., 1-night vs 2-night restarts). The use of the proportional hazards model may depend on the nature of the events (crashes or inspection violations), as well as the data quality of the time provided of specific events (e.g., crash times provided in MCMIS).

Model assumptions will be validated to ensure suitable model specification through a variety of diagnostic methods. We will use residual plots to assess the model form – including appropriateness of the link function, linearity of predictor effects, potential omitted variable bias, and homoscedasticity. This includes checking that the mean of the residuals does not vary systematically with the predictors, and that the variance of the residuals remains constant across the range for each predictor.

In addition to residual plots, we will examine overdispersion statistics to evaluate the variance structure. The Poisson model regression assumes that the data variance equals the mean, and the observed crash and inspection data may exhibit overdispersion where the variance is greater than the mean. The dispersion can be calculated from the ratio of residual deviance to the residual degrees of freedom, and dispersion greater than 1 (> 1.5-2.0) can be a strong indicator of overdispersion in the Poisson model, in which case the Negative Binomial model may be more appropriate.

Overdispersion can be indicative of a mismatch in the between the observed and expected count distribution due to a zero inflated data set. Zero inflation is the condition where the number of observations without crashes greatly exceeds what is expected in a Poisson (or Negative Binomial) distributed process. Given the infrequent nature of crashes, zero inflation is a likely issue with using the standard count models, in which case zero-inflated variants (zero-inflated Poisson/zero-inflated Negative Binomial) are considered.

Model selection among Cox, Poisson, Negative Binomial, and zero-inflated variants will be conducted by fitting all candidate models and comparing their goodness-of-fit using tests such as likelihood ratio test, as well as information criterion including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

We will perform additional model diagnostics checks for highly influential observations (e.g. outliers), and multicollinearity. Influence diagnostics, such as Cook's distance and leverage values, will be used to identify observations with disproportionate influence on model estimates. Multicollinearity among predictors will be assessed through Variance Inflation Factors (VIF) and correlation matrices.

Independence among observations is a primary assumption in regression models. Individual drivers will contribute multiple duty days (observations) in the study, and we expect to see some dependence between observations because of this. Due to the infrequency of crashes, we anticipate that crash occurrences within the same driver will be largely independent across days. A mixed-effects model can account for interdependencies due to repeated observations from drivers and carriers. One limitation of the proposed approach is that the study design excludes driver information from the analysis table and does not enable a driver to be associated across duty days. As such, a mixed-effects model cannot be used, within-subject correlations cannot be explored, and the set of variables available for modeling is constrained. Using a fixed-effects model, we will include covariates to capture driver and carrier effects in a deidentified manner.

The study data collection design deidentifies data after six months, but it maintains subject identifiers to enable data to be updated within a six-month rolling window. We will validate the approach by performing a sensitivity analysis comparing mixed-effects versus fixed-effects study designs using a six-month snapshot before the data is deidentified and mixed-effects design is still possible. This validation may be limited, however, because of: 1) the low rate of occurrence of crashes; and 2) the lag in reporting incidents through MCMIS, which may further limit the usable time period to less than six months.

2.5. Power Analysis and Sample Size Calculations

DESCRIBE PROCEDURES FOR COLLECTING INFORMATION, INCLUDING STATISTICAL METHODOLOGY FOR STRATIFICATION AND SAMPLE SELECTION, ESTIMATION PROCEDURES, DEGREE OF ACCURACY NEEDED, AND LESS THAN ANNUAL PERIODIC DATA CYCLES.

Our overall target sample size, over the 4-year study, of 52,750 driver years was based on our power analysis. The power analysis was conducted to determine the sample size needed to answer the research questions. Based on a literature review, we found that the relative risk of the various schedule factors ranged from 1.15x to 1.5x. These effect sizes are operationally meaningful and translate into an additional number of crashes.

For example, in 2021 there were 119,462 crashes involving combination trucks[7]. Assuming 98% of hours driven are in the first 10 hours, and 2% of hours driven are in the 11th hour. If driving in the 11th hour presents no increased risk, we expect there to be 2,389 crashes (i.e., 2% of crashes) in the 11th hour. If there was 1.5x increased risk of a crash (i.e., 50% more or 3% of crashes), then we would expect 3,584 crashes in the 11th hour. This effect size translates to 1,195 additional crashes (i.e., 3,584-2,389=1,195) per year due to driving in the 11^{th} hour.

The study aims to detect differences at least as great as the prior published estimates. The sample size estimates were based on assumptions of the prevalence of different conditions (e.g., frequency of 1-night, 2-night restarts), and based on a power of 80% ($\beta = 0.80$) and a significance level of 5% ($\alpha = 0.05$), the study requires between 10,000 and 55,000 driver years of data (based on an average of 100,000 miles driven per driver year) to detect that difference for all the different research aims.

Once we empanel a carrier the expectation is that drivers from that carrier will continue to contribute to the study sample for the remaining number of years left in the study. Our enrollment plan of 60 carriers over the study period (Table 14) will provide 52,750 driver years and is sufficiently powered to meet the study objectives. The study is not powered to detect true effect sizes below the minimum detectable difference (Table 20) but does establish an upper bound for these estimates. For example, focusing on the cumulative driving hours on duty, based on the literature we anticipate the effect of 1.5x. If the observed effect size is below a relative risk 1.39, the minimum detectable effect size, the sample size will not have sufficient power to detect the effect but will establish the effect is below 1.39x.

We will periodically assess observed incident frequency to confirm a priori assumptions about prevalence and effect sizes and update the sample size requirements as appropriate.

Table 20: Minimum Detectable Rate for Target Sample Size

Study Aim	**Minimum Detectable Relative Rate Given 52,750 Driver Years**	**Target Relative Rate (Effect Size)**	**Minimum Sample size Given Target Effect Size**
Cumulative driving hours on duty	1.39	1.5	35,000
Cumulative hours in prior 7 days	1.18	1.2	35,000
Number of nights restart period	1.17	1.2	40,000
Use of sleeper berth	1.14	1.5	10,000
Consecutive hours of sleeper berth	1.26	1.25	55,000
Number of breaks within duty	1.14	1.15	45,000
Use of personal conveyance	1.25	1.5	15,000

The details of the calculations follow and provide a more detailed view of how a different effect size may affect the true sample size needed. The power analysis was performed based on an α = 0.05 and power ($1 - \beta$) = 0.8, and assumptions are made on the rate of crashes, drive durations, and other schedule factors. The power analysis excludes covariates in the generalized linear model.
Given the low rate of crashes, knowing the expected number of crashes that would be observed is important for understanding if the study is sufficiently powered. Tables 21 and 22 provide estimates for crash rates for large trucks/buses and combination trucks (tractor/trailer), respectively.

Table 21: Crashes, Vehicle Miles Traveled, and Crash Rate for Large Trucks and Buses 2019-2021 (Crash data from MCMIS; Miles Traveled from Federal Highway Administration (FHWA), Highway Statistics 2021, Table VM-1) [7, 8].

Year	Large Truck/Bus Fatal Crashes	Large Truck/Bus Total Crashes	Vehicle Miles Traveled (in Million Miles)	Large Truck/Bus Fatal Crashes per 100 Million Miles	Large Truck/Bus Crashes per 100 Million Miles
2019	4,793	182,088	318,030	1.5	57.3
2020	4,652	156,913	312,686	1.5	50.2
2021	5,378	184,508	343,771	1.6	53.7

Table 22: Crashes, Vehicle Miles Traveled, and Crash Rate for Combination Trucks 2019-2021 (Crash data from MCMIS; Miles Traveled from Federal Highway Administration (FHWA), Highway Statistics 2021, Table VM-1) [7, 8].

Year	Number of Combination Trucks Involved with Fatal Crash	Number of Combination Trucks Involved with Crash	Vehicle Miles Traveled (in Million Miles)	Fatal Crash Vehicles per 100 Million Miles	All Crash Vehicles per 100 Million Miles
2019	3,442	114,165	175,305	2.0	65.1
2020	3,271	101,679	179,817	1.8	56.7
2021	3,859	119,462	195,389	2.0	61.1

Based on an analysis of this historical data (Tables 21-22), we use conservative estimates of 50 crashes per 100 million miles and 1.5 fatal crashes per 100 million miles as the basis of our power analysis. These estimates come from the lowest crash rates observed in the data and allow us to estimate a realistic minimum power for the study. Table 23 shows the number of expected crashes for the entire study population over the duration of the study, assuming the estimated crash rates, given different scenarios of average carrier size and driving miles.

Table 23: Number of expected crashes for scenarios with different carrier sizes and driving miles per driver. Expected crashes are based on assuming a crash rate of 50 crashes per 100 million driving miles, and 1.5 fatal crashes per 100 million miles.

Average Number of Drivers per Carrier	Average Number of Driving Miles per Driver	Carrier Years	Driver Years	Total Number of Driving Miles (Million)	Expected Number of Fatal Crashes	Expected Number of Crashes
100	50,000	150	15,000	750	11	375
100	100,000	150	15,000	1,500	22	750
500	50,000	150	75,000	3,750	56	1,875
500	100,000	150	75,000	7,500	112	3,750

A power analysis for the effect of cumulative driving hours in the duty period was performed based on the WebPower package in R to estimate the sample size needed for comparing the rate of crashes in the start of the duty (1st hour of driving) versus the rate of crashes in the 11th hour of driving. Blanco et al. (2011)[2] estimated the relative risk of crash on driving hour; the reported parameter estimate (u = 0.0363) translates to 1.49x relative risk ($1.49 = e^{0.0363\times11}$) at the 11th hour. Assuming that 20% of duty periods would extend into the 11th hour of driving, and an increase of 1.5x in the rate of crashes at hour 11, the study would require 3,500 million driving miles.

Table 24: The number of duty periods needed and expected crashes based on a comparison of the first and 11th hour of driving assuming a 1.5x relative risk of crashes and 20% percent of shifts driving into the 11th hour.

Duty Periods with at Least 1 Hour	Duty Periods with 11th Hour of Driving	Expected Crashes in Hour 1	Expected Crashes in 11th Hour	Total Number of Duty Periods	Total Number of Miles Driven
8,275,262	1,655,052	207	62	8,275,262	3,310,104,800

The power analysis for the effect of cumulative driving hours was based on the 1.5x relative risk observed in crash data. The relative risk of crashes for driving hours may be different, and the sample size needed to observe statistically significant difference depends on the true underlying effect size. To determine the sample size needed when the relative risk is greater or less than our base assumption of 1.5x relative risk, we calculate the minimum detectable difference in crash risk for a range of sample sizes. The table below shows the detectable differences in crash rates between hour 1 and hour 11 of driving for different study sizes.

Table 25: Detectable relative risk between crash rate during 1st hour vs 11th hour of driving for different study sample size, based on assumed prevalence of duties with at least the 1st hour of driving (100% of duties) and 11th hour of driving (20% of duties).

Total Million Miles Driven	Number of Driver Years	Number of Duty Days	Number of Duties with 1st Hour of Driving	Number of Duties with 11th Hour of Driving	Minimum Detectable Relative Risk
1000	10,000	2,500,000	2,000,000	500,000	1.95
1500	15,000	3,750,000	3,000,000	750,000	1.76
2000	20,000	5,000,000	4,000,000	1,000,000	1.65
2500	25,000	6,250,000	5,000,000	1,250,000	1.58
3000	30,000	7,500,000	6,000,000	1,500,000	1.53
3500	35,000	8,750,000	7,000,000	1,750,000	1.49
4000	40,000	10,000,000	8,000,000	2,000,000	1.45
4500	45,000	11,250,000	9,000,000	2,250,000	1.43
5000	50,000	12,500,000	10,000,000	2,500,000	1.40

The sample size estimate for other research questions is performed in a similar manner, and the results are summarized in Table 26. For the sample analysis for the effect of cumulative duty prior to work start, we compare duties with less than 60 hours of driving in the last 7 days versus duties with more than 60 hours of driving in the last 7 days. Based on data from prior studies, the prevalence of having more than 60 hours of driving in the prior 7 days was observed in 10% of the duty days. The analysis is based on assuming 8 hours of driving per duty, and a base crash rate of 25 crashes per 100 million driving hours for duties with less than 60 hours of cumulative driving in prior 7 days. In studies on workplace, working greater than 60 hours was associated with a 23% increase in SCEs[3].

For the sample analysis relative to the effect of restart type, we assume that the more recent 34h restart was a 1-night restart for 15% of duty days, 45% for a 2-night restart, 40% for a 3-or-more-night restart. Then, we look at the relative risk between duties preceded by 1-night restarts versus duties preceded by 2-night restarts. The analysis is based on assuming 8 hours of driving per duty, and a base crash rate of 25 crashes per 100 million driving hours for duties preceded by a 2-night restart. Given that prior research observed limited impact of restart on safety events, we consider an effect of 20% between restart conditions.

For the sample analysis relative to the effect of sleeper berth use, the use of sleeper berth can vary depending on the carrier operations. FMCSA[1] reported 91% of long-haul drivers use sleeper berths, but other operations with more local routes may not include any sleeper berth use. For the power analysis, given the large differences in operation types, we assume that only 20% of duty days will have sleeper berth use. The analysis is based on assuming 8 hours of driving per duty, and a base crash rate of 25 crashes per 100 million driving hours for no sleeper berth use. In previous studies of sleeper berth use, sleep time was comparable[4] but fatality risk was 3.05 odds ratio for consecutive shifts with sleeper berth[5].

For drivers using the sleeper berth provision, a minimum of 7 consecutive hours in the sleeper berth is required during the 10-hour rest period. For drivers using the sleeper berth provision, we look at the relative risk of incidents for drivers that spend only the minimum 7 consecutive hours in the sleeper berth compared against drivers that spend more than 7 consecutive hours in the sleeper berth. For the power analysis, we assume that 20% of duty days will have sleeper berth days, and, of those duty days, we assume an even split of days with the minimum 7 consecutive hours (50%) and more than 7 consecutive hours (50%).

For the sample analysis relative to the effect of personal conveyance (i.e., the movement of a CMV for personal use while the driver is off-duty), we assume that drivers use personal conveyance on 5% of duty days. The analysis is based on assuming 8 hours of driving per duty, and a base crash rate of 25 crashes per 100 million driving hours for duties with no personal conveyance use. Our literature review did not return any studies on the effect of personal conveyance on the driving risk, but if personal conveyance is used immediately after a driving duty, then the risk would be similar to driving into the 11th hour (50% increase, as previously stated).

For the sample analysis relative to the effect of the number of 30-minute driving breaks, we assume for the power analysis that drivers take 0, 1, or 2 breaks with equal probability. Jovanis et al.[6] reported a 51% decrease in crash odds for drivers taking 2 breaks instead of no breaks, and a 15% decrease in crash odds when taking 1 break instead of no breaks (breaks were between 15 and 90 minutes in duration within the duty day).

Table 26: Detectable relative risk of crashes for different study sample sizes for each research aim. Cumulative Driving Hours compares 1st vs 11th hour of Driving. Cumulative Hours in Prior 7 days compares duties with less than vs more than 60 hours driving in prior 7 days. Number of nights restart period compares duties with 1-night vs 2-nights in restart. Use of sleeper berth compares duties with and without sleeper berth use. Consecutive hours in sleeper berth compares duties with 7 consecutive hours vs more than 7 consecutive hours of sleeper berth use. Number of breaks within duty compares duties with zero 30-minute breaks vs two 30-minute breaks within duty. Use of personal conveyance compares duties with and without personal conveyance.

Total Million Miles	Num. of Driver Years	Detectable Relative Rate						
		Cum. Driving hours	Cum. Hours in prior 7 days	Nights in restart period	Use of sleeper berth	Hours in sleeper berth	Num. of breaks	Use of PC
1,000	10,000	1.95	1.43	1.39	1.32	1.65	1.33	1.59
1,500	15,000	1.76	1.35	1.32	1.26	1.52	1.27	1.48
2,000	20,000	1.65	1.30	1.27	1.23	1.44	1.23	1.41
2,500	25,000	1.58	1.27	1.24	1.20	1.39	1.20	1.37
3,000	30,000	1.53	1.24	1.22	1.18	1.35	1.19	1.34
3,500	35,000	1.49	1.23	1.21	1.17	1.32	1.17	1.31
4,000	40,000	1.45	1.21	1.19	1.16	1.30	1.16	1.29
4,500	45,000	1.43	1.20	1.18	1.15	1.28	1.15	1.27
5,000	50,000	1.40	1.19	1.17	1.14	1.27	1.14	1.26
7,500	75,000	1.33	1.15	1.14	1.12	1.22	1.12	1.21
10,000	100,000	1.28	1.13	1.12	1.10	1.19	1.10	1.18
20,000	200,000	1.20	1.09	1.08	1.07	1.13	1.07	1.13

2. DESCRIBE METHODS TO MAXIMIZE RESPONSE RATE AND TO DEAL WITH THE ISSUES OF NON-RESPONSE.

To maximize carrier participation, the research team will implement a promotion plan for advertising the website and recruiting property-carrying and passenger-carrying CMV carriers to participate. The research team intends to raise awareness of the study in trade publications and targeting recruitment based on existing industry contacts. To further encourage participation from carriers who are less likely to enroll, the research team will consider offering carriers proprietary reports with an analysis of fatigue risk in their operations based on an analysis of their HOS data. We recognize that certain carriers (e.g., those with smaller fleets) may be underrepresented. As discussed in Section B1, we will address potential selection bias using model-based analytic techniques that include covariates reflecting systematic differences that may influence participation. After approval of the ICR, researchers will commence recruiting carriers to volunteer for this study.

Pulsar’s research team is well positioned within the CMV industry to support efforts to recruit carriers during the project to provide ELD and crash data on an ongoing basis. Pulsar has:

- Existing active integrations with ELD and telematics providers covering most of the CMV market
- Relationships with CMV and motorcoach industry associations (American Trucking Associations, Truckload Carriers Association, National Private Truck Council, National Safety Council, American Bus Association, Bus Industry Safety Council, etc.)
- Large customer base of CMV carriers that use Trucking Fatigue Meter, described on page 4
- Strategic partnerships with insurers that underwrite CMV carriers

After a carrier completes the sign-up process, Pulsar can utilize its established Fatigue Meter technology and existing integrations with ELD and telematics service providers to support direct, automatic transfer of the carrier's ELD and crash data without the need for any manual steps to be completed by the CMV carrier. It is not a requirement for the CMV carrier to be or become a Fatigue Meter customer. This approach offers more scalability, enhanced data security, and less opportunities for errors to be introduced into the data set. Providing CMV carriers an approach that is automatic and less burdensome has two strong advantages: (1) it will lower the barrier for CMV carriers to sign up to provide their data; (2) it will make it easier for carriers to continue to provide data on an ongoing basis.

Because the research team has demonstrated experience successfully using the Fatigue Meter technology to collect ELD and crash data from CMV carriers automatically, the expected risk of non-response is minimal. Further, the integration of Fatigue Meter with carriers' TSPs reduces the risk of receiving unreliable data.

3. DESCRIBE TESTS OF PROCEDURES OR METHODS TO BE UNDERTAKEN.

The research team will leverage its established Fatigue Meter technology and existing integrations to streamline the collection of ELD, SCE, and crash data from carriers. Given the historical success of this data collection method, a test of this collection procedure is not necessary.

While the research team has experience with collecting and processing demographic data from carriers, a test to trial this process with up to three carriers is needed to ensure all requested data is readily transmissible and a streamlined means of receiving and parsing the data is established prior to initiating information collection. Carriers selected for this test will only be asked to provide a single transfer of demographic data. The goal of this test is to reduce the burden on carriers during the study.

4. PROVIDE NAME AND TELEPHONE NUMBER OF INDIVIDUALS WHO WERE CONSULTED ON STATISTICAL ASPECTS OF THE INFORMATION COLLECTION AND WHO WILL ACTUALLY COLLECT AND/OR ANALYZE THE INFORMATION.

Project Leads for this information collection request:

Theresa Hallquist
Project Manager, Research Division
Federal Motor Carrier Safety Administration
1200 New Jersey Ave SE
Washington, DC 20590
+1 202.366.1064
theresa.hallquist@dot.gov

Daniel Mollicone, PhD
CEO & Chief Scientist
Pulsar Informatics

505 Broadway E #136
Seattle, WA 98102
+1 877.791.0989
daniel@pulsarinformatics.com

REFERENCES:

1. Federal Motor Carrier Safety Administration. (1999). Commercial Truck Driver Fatigue, Alertness, and Countermeasures Survey.
2. Blanco, M., Hanowski, R., Olson, R., Morgan, J., Soccolich, S., Wu, S., Guo, F. (2011). The Impact of Driving, Non-Driving Work, and Rest Breaks on Driving Performance in Commercial Motor Vehicle Operations.
3. Dembe, A., Erickson, J., Delbos, R., Banks, S. (2005). The impact of overtime and long work hours on occupational injuries and illnesses: new evidence from the United States. Occupational and Environmental Medicine. 62(9):588-97.
4. Darwent, D., Roach, G., Dawson, D. (2012). How well do truck drivers sleep in cabin sleeper berths? Applied Ergonomics. 43(2):442-6.
5. Hertz, R. (1988). Tractor-trailer driver fatality: The role of nonconsecutive rest in a sleeper berth. Accident Analysis & Prevention. 20(6):431-9.
6. Jovanis, P., Wu, K-F., Chen, C. (2012). Effects of Hours of Service and Driving Patterns on Motor Carrier Crashes. Transportation Research Record. 2281:119-27.
7. MCMIS A&I Crash Statistics (http://ai.fmcsa.dot.gov)
8. Federal Motor Carrier Safety Administration (2021). Crashes, Vehicle Miles Traveled, and Crash Rate for Combination Trucks 2019-2021.