**METHODOLOGY OUTLINE**

1. **Data cleaning and preparation**
   1. Study period
      1. In models utilizing counts of violations or counts of significant and substantial violations as predictor variables, the sample is restricted to 2000 Q1 - 2016 Q1. This restriction is due to data availability and quality.
      2. In models utilizing penalty points as predictor variables, the sample is further restricted to 2007 Q1 - 2016 Q1. This restriction if a consequence of the fact that the meaning of penalty point assignments changes over time. The last change to the manner in which penalty points are assessed occurred in 2007 Q1.
      3. , which was when the manner in which violations are assessed was changed for the last time.
   2. Other mine inclusion criteria
      1. Minetype = Underground
      2. Subunit = Underground
      3. Coalormetal = Coal
   3. Mine status inclusion criteria
      1. Only mine-quarters for which a mine was not sealed, abandoned and sealed, nonproducing, or temporarily idled are eligible for inclusion in the analysis; however, mine status information is available only at the mine level, not at the mine-quarter level. Therefore, we dropped observations for mines with any of the above mine statuses (e.g., abandoned) when their mine status date comes before the current quarter (at the mine quarter level).
      2. Mine-quarters with no hours worked are excluded, because only quarters with active operation are presumed to generate violations and injuries by the same processes.
   4. Files taken in from <http://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp>
      1. The following
         1. [Accident Injuries Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/Accidents.zip)
         2. [Assessed Violations](http://arlweb.msha.gov/OpenGovernmentData/DataSets/AssessedViolations.zip)
         3. [Contractor Employment/Production](http://arlweb.msha.gov/OpenGovernmentData/DataSets/ContractorProdQuarterly.zip)(Quarterly)
         4. [Contractor Employment/Production](http://arlweb.msha.gov/OpenGovernmentData/DataSets/ContractorProdYearly.zip)(Yearly)
         5. [Controller/Operator History](http://arlweb.msha.gov/OpenGovernmentData/DataSets/ControllerOperatorHistory.zip)(Yearly)
         6. [Employment/Production Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/MinesProdYearly.zip) (Quarterly)
         7. [Employment/Production Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/MinesProdYearly.zip) (Yearly)
         8. [Inspections Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/Inspections.zip)
         9. [Mines Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/Mines.zip)
         10. [Violations Data Set](http://arlweb.msha.gov/OpenGovernmentData/DataSets/Violations.zip)
   5. Other files taken in:
      1. Pinning and Striking training set (sent by NIOSH on January 29, 2016)
      2. Maintenance and Repair training set (sent by NIOSH on August 28, 2015 )
      3. Additional maintenance and repair fatalities manually collected by Sarah Levine from John Heberger’s research: Leanna M. Reardon, John R. Heberger, Patrick G. Dempsey “Analysis of Fatalities During Maintenance and Repair Operations in the U.S. Mining Sector,” *IIE Transactions on Occupational Ergonomics and Human Factors* (2014), 2: 37–38.
         1. “Fatalgrams” corresponding with the document numbers on pages 37-38 were extracted from <http://arlweb.msha.gov/fatals/fabc.htm>.
      4. Dataset assembled by Morantz team itemizing parts and subparts of the CFR code that were marked as “relevant” or “maybe relevant” to each subtype in a meeting between Linda McWilliams, Miguel Reyes, John Heberger, Ellen Rubinstein, Alison Morantz, Nikhil Saifullah, and Sarah Levine at NIOSH in Pittsburgh on February 12th, 2016.
   6. Variable conversion and renaming
   7. Merging the datasets
      1. Merge employment and production data onto mines by mine ID
      2. Merge mine data onto accidents by mine ID
      3. Classify accidents as maintenance and repair (Y/N) and pinning and striking (Y/N)
      4. Merge assessments data onto violations data by violation number, and merge violations and assessments onto inspections data, by event number
      5. Merge “CFR key” onto inspections-violations-assessments data by subpart
   8. Recoding the training sets
      1. In email correspondence with Linda McWilliams and Miguel Reyes, details of the subtype definitions were refined, and individual accident observations we recoded (for example, recoded from a “yes” for pinning and striking to a “no”). These recodings were implemented before training the classification algorithms for both injury subtypes.
   9. Preparation of specific variables
      1. Missing value imputation for variables with null values (this is necessary only before running specific classification algorithms such as a random forest).
         1. We experimented with the following methods, and tested the performance of each:
            1. Method 1: Replacement with the mean of a given continuous variable and the mode of a given categorical variable
            2. Method 2: Replacement with the median of a given continuous variable and the mode of a given categorical variable
            3. Replacement of all missing values by randomly sampling from the distribution of a given continuous or categorical variables. This method ultimately performed best.
         2. Missing values in the following variables were imputed using the above methods:
            1. Categorical variables: totalexperience, mineexperience, jobexperience, hourspershift, numberofemployees.
            2. Continuous variables: accidenttypecode, classificationcode, sourceofinjury, natureofinjury, bodypart, mineractivity, occupation.
      2. Cleaning up operator names
      3. Cleaning up inspection hours data
         1. In order to utilize the total inspection hours data at our desired unit of analysis (the mine quarter) we needed to ensure that inspections hours were properly distributed across quarters. For example, if a single inspection had a duration of one year (according to the inspection start and end date), then we want to divide the total number of inspection hours by four (for each quarter), for the purposes of not double-counting hours across quarters.
      4. Keyword indicators were extracted from narrative analysis (these variables were extracted from the narrative portions of the injury data and are subtype specific – see Description of Injury Classification Algorithms for details).
      5. Formulations of the outcome variable (injuries)
         1. Continuous variable
         2. Binary variable
      6. Formulations of violation variables
         1. Types
            1. Counts of violations per mine-unit time
            2. Counts of significant and substantial violations per mine-unit time
            3. Number of penalty points assessed per mine-unit time
         2. Lags
            1. One lag: number of violations in the preceding unit of time
            2. Four cumulative lags: cumulative number of violations over the preceding four units of time
            3. Complete cumulative lags: cumulative number of violations since the beginning of operation at that mine
   10. Collapsing the data
       1. Finally, we collapsed the data to the mine-quarter level. For each mine-quarter, we sum the each variable of interest (e.g. number of violations, number of significant and substantial violations.)
       2. We complete the same collapsing process to prepare data at the mine-year level.

**DESCRIPTION OF INJURY SUBTYPES**

1. **Description of Maintenance and Repair Subtype**
   1. Subtype definition
      1. Cases of workers being injured while conducting maintenance and repair activities. Existing miner activity classifications include:
         1. Electrical maintenance/repair
         2. Machine maintenance/repair
         3. Ventilation maintenance/install
   2. Included cases
      1. All cases in which the miner occupation is recorded as “maintenance or repairman” should be included in this category, except in cases in which the accident was unrelated to the maintenance/repair activity (e.g. a roof fall).
      2. This subtype can include cases with any recorded occupation, injury source, accident class, or miner activity.
      3. Shoveling coal at the belt is regarded as maintenance, and therefore injuries sustained during this activity are included in the subtype.
   3. Excluded cases
      1. Accidents that occur during maintenance/repair work but are not immediately related to that activity, such as falling rock accidents, are excluded from the subtype.
      2. Accidents that occur during installation work are excluded from the subtype.
2. **Description of Pinning and Striking Subtype**
   1. Subtype definition
      1. All cases of pinning or striking accidents involving machinery or powered haulage and either a:
         1. Vehicle-to-vehicle collision, or a
         2. Vehicle-to-person collision
   2. Included cases
      1. Non-powered vehicle accidents: Cases in which a vehicle struck another vehicle or a person, but the initial vehicle was not powered at the time of the collision, will be considered pinning and striking accidents.
      2. Certain roof-bolting accidents: Cases of roof-bolting accidents that will be included in the pinning and striking subtype include injuries in which an employee was pinned, struck or caught by/between the roof bolting machine as a result of the powered motion of that machine.
      3. Intermediate-object collisions: Cases in which a vehicle strikes an object (neither another vehicle nor a person), and that object in turn strikes a person or vehicle, will be included as pinning and striking accidents.
      4. Vehicle accidents with injured remote operator: If a vehicle can be operated remotely, it is possible that the operator was outside of the vehicle at the time of the injury. As a practical matter, we often cannot determine from the narrative field whether the operator was inside or outside of the vehicle at the time of the accident. (This is the case for all vehicles in our training dataset except shuttle cars and roof bolting machines.) In such ambiguous situations, we included the case as a pinning/striking incident as long as it meets all of the other inclusion criteria.
      5. Brake failure: Cases of vehicle brake failure cannot be prevented using proximity detection technology, because the failure is purely mechanical. However, it is impossible to ascertain from the narrative fields alone whether the brakes of a particular vehicle in fact failed, or whether they were simply not engaged properly by the operator. Therefore, we include such cases in our subtype.
      6. Vehicle to vehicle accidents with no injuries.
   3. Excluded cases
      1. Certain roof-bolting accidents: Injuries caused by the rotational movement of the drill or bolts will not be included in the subtype. Cases of entanglement (in which clothing or body parts are pulled by the rotational movement of the drill) will also not be included in the subtype. Moreover, cases of the drill steel bending or breaking will not be categorized as pinning and striking accidents.
      2. Vehicles colliding with walls/other unmotorized machinery.
      3. Vehicle accidents caused by bumps in the road.  
         Vehicle accidents with injured operator inside the vehicle: If a vehicle is involved in a pinning/striking incident, but it is apparent that the injured miner was inside the vehicle when the accident occurred, then the case will be omitted from the pinning/striking subtype. This is because, in most such cases, the operator was simply jarred or jostled while driving.
      4. Vehicular accidents with no collision: Cases of derailing, a vehicle slipping off the track, or sliding on an oil patch will not be included.

**DESCRIPTION OF INJURY CLASSIFICATION ALGORITHMS**

1. **Maintenance and Repair Algorithm**
   1. Narrative field analysis
   2. Variable selection techniques
      1. PCA
      2. LASSO
      3. Random forest
   3. Attempted algorithms
      1. Logit
      2. CaRT
         1. rpart
      3. Random forests
      4. Boosting
      5. Composite algorithms
   4. Other attempted improvements
      1. Oversampling
      2. Downsampling
      3. Various parameterization techniques
2. **Pinning and Striking Algorithm**
   1. Narrative field analysis
   2. Variable selection techniques
      1. PCA
      2. LASSO
      3. Random forest
   3. Attempted algorithms
      1. Logit
      2. CaRT
      3. Random forests
      4. Boosting
      5. Composite algorithms
   4. Other attempted improvements
      1. Oversampling
      2. Downsampling
      3. Various parameterization techniques

**GAME PLAN**

1. **Re-run preliminary models on sample of big mines and sample of bad mines (TIMELINE: ONE WEEK)**
   1. How to define big mines?
      1. Hours – average, min, max, median (across all quarters)
   2. How to define bad mines?
      1. Relevant violations – average, min, max, median (across all quarters)
      2. Total violations – average, min, max, median (across all quarters)
      3. Proportion of relevant/total violations – average, min, max, median (across all quarters)
2. **Select best-fitting models and re-run them (on all data) (TIMELINE: ONE WEEK)**
   1. Separate data into training and test sets
      1. By quarters
         1. Try several combinations of test set sizes, e.g.,
            1. Training: 14 years, Test: 2 years
            2. Training: 12 years, Test: 4 years
      2. By mines
         1. Randomly select mines to include in training and test sets
         2. Conduct many iterations (with new sets of mines in training and test set)
         3. Also test size of training and test set, e.g.,
            1. Training: 50% mines, Test: 50% mines
            2. Training: 60% mines, Test: 40% mines
3. **Select best-fitting models and re-run them (on big/bad mines) (TIMELINE: ONE WEEK)**
   * 1. By quarters
        1. Try several combinations of test set sizes, e.g.
           1. Training: 14 years, Test: 2 years
           2. Training: 12 years, Test: 4 years
     2. By mines
        1. Randomly select mines to include in training and test sets
        2. Conduct many iterations (with new sets of mines in training and test set)
        3. Also test size of training and test set, e.g.
           1. Training: 50% mines, Test: 50% mines
           2. Training: 60% mines, Test: 40% mines

**DESCRIPTIVE STATISTICS OUTLINE**

**Goal:** Assess whether and to what extent relative prevalence and/or distribution of maintenance and repair and pinning and striking injuries have fluctuated over time

**Potential plots:**

1. All injuries vs. Time
2. MR injuries vs. Time
3. PS injuries vs. Time
4. MR/all injuries vs. Time
5. PS/all injuries vs. Time
6. All injuries vs. All violations
7. MR injuries vs. All violations
8. PS injuries vs. All violations
9. MR injuries vs. MR violations
10. PS injuries vs. PS violations
11. MR/all injuries vs. All violations
12. MR/all injuries vs. MR violations
13. MR/all injuries vs. MR/all violations
14. PS/all injuries vs. All violations
15. PS/all injuries vs. PS violations
16. PS/all injuries vs. PS/all violations
17. Distribution of miner experience at time of MR/PS injuries
18. Recorded occupations for MR/PS injuries
19. Recorded injury classifications for MR/PS injuries
20. Recorded body parts involved in MR/PS injuries

**Variations:**

1. Lose mine dimension or not

2. Time – quarters or years (how to collapse to years)

3. Control for hours worked

4. Stratify by mine size

5. Stratify by mine badness

6. Rates based on individual mines/times or overall

**Also consider:** For each mine, could we create a measure of trends over time (e.g., correlation, regression lines)?

**Potential maps:**

1. Total number of MR/PS injuries per mine divided by total hours worked (dot size would correspond to proportion)
2. Time-lapse maps: showing MR/PS injuries per mine quarter (could do minimaps, or animate a GIF)
3. Display MR/PS injuries according to total miner experience (in number of years)
4. Display MR/PS injuries according to miner occupation