# Vision over Incidents and Claims Research Proposal

Simardeep Kaur, Merve Sahin, Brayden Tang, Xugang (Kirk) Zhong

## Summary

To investigate the potential factors related to TransLink's increasing insurance cost, the research is proposing a four-phase research plan, deploying data science skills such as exploratory data analysis, regression analysis and machine learning models to identify the incidents/claims patterns and root causes among predictors such as driver characteristics, vehicle characteristics and other third-party predictors. A fully reproducible and interactive report generated via a data pipeline is expected at the end.

## Introduction

With the largest public transit service area in Canada, TransLink is operating more than 245 bus routes, 79 kilometers of rapid transit, etc. to meet the transportation need of 2.5 million people in Metro Vancouver as of the end of 2018 (TransLink 2018). Legislatively, TransLink is required to carry $1 million per occurrence liability policy on each of its revenue vehicles and a $200,000 per occurrence liability policy on each of its non-revenue vehicles. Since 2014/2015, the premium paid to ICBC has increased by over 200% to cover onboard passenger injuries, cyclist injuries, pedestrian injuries and losses from collisions with third party vehicles. For at-fault physical damage losses to its vehicles, the premium paid to its own captive insurance company has been increased by 33%. In responding to the soaring insurance cost and road safety concern, this research will investigate the factors that influence the frequency and severity of transit incidents and claims, identifying the incidents/claims pattern, and root causes. The research objectives are to:

- Characterize the current incidents in terms of driver-end predictors, vehicle-end predictors, and other predictors based on the incidents/claims, and operator data
- With the help of external data such as weather conditions, and geographic data, examine the main predicators of incidents and potential incidents/claim pattern
- Recommend actions for TransLink to reduce the insurance cost strategically
- Identify research needs for future study

## Data Product

A fully reproducible and interactive report will be emerging at the end of the project. The report will give a visual representation of relationships between the frequency or severity of claims and specific variables interactively. The project will be done using a fully reproducible data pipeline so that the user can run the entire analysis using simple and understandable make commands. A docker container for reproducibility on any operating system will also be included.

## Data Description

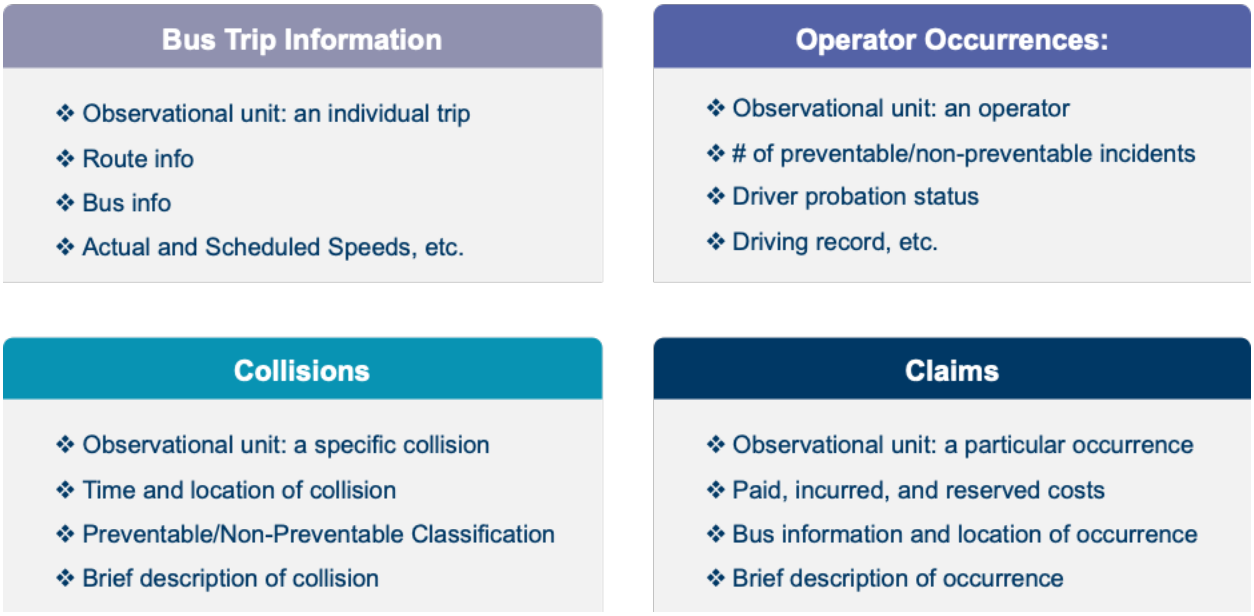
**Bus Trip Information**: This dataset consists of individual trips taken by busses over a span of five days in March 2020 with routes, speeds, vehicle information, and metadata included.

**Operator Occurrences**: This dataset provides the number of incidents (preventable and non-preventable) for all operators with at least one incident in the past three years. Information regarding each operator's

characteristics is also included.

**Collisions**: This data set provides a detailed description of both preventable and not preventable collisions that took place. Columns describing the location and time of each collision are provided, along with a brief description.

**Claims**: This dataset describes all occurrences as well as their associated costs. It includes the location, time and description of the occurrence, along with vehicle information.

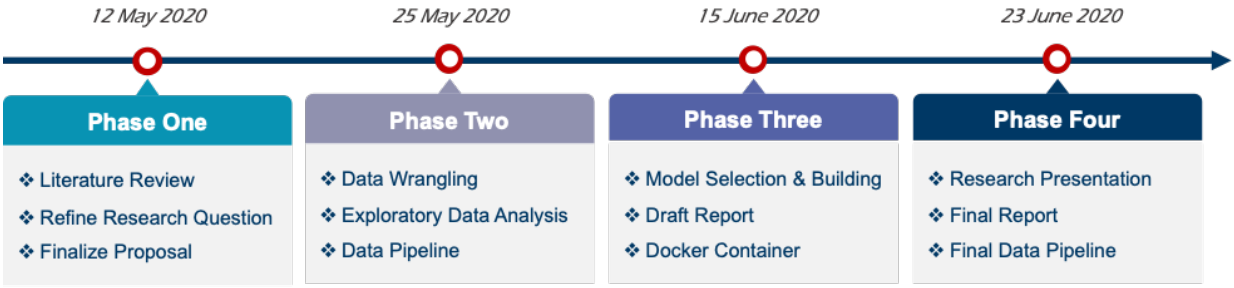



## Methodology

The research will mainly focus on preventable claims since the non-preventable ones are less likely to lend themselves to easy interventions used to reduce insurance-related costs. To model the occurrences which do not yield any cost, we propose a hurdle model, combing a binary classification model, predicting "equal to zero" and "greater than zero" over all occurrences with a regression model that is fit only to non-zero cost occurrences. For simplicity, the research will start with linear models. Alternatively, a different framework like Tweedie regression (which allows for exact zeros) will also be investigated. The model that predicts a held-out test set the best will be used. In terms of bias in the operator dataset, the research will examine methods such as zero-truncated models or Bayesian regression models. To find out whether certain descriptions and/or codes exhibit similar loss experience, methods such as standard cluster analyses will be deployed to identify potential groupings. Furthermore, topic modelling for the manually written accident descriptions will be used to further determine potential groupings or patterns in claim behaviour.

## Timeline

The project will be completed with a four-phase research plan. After refining the research objectives, the proposal is expected to be composed during the first phrase. The second phase consists of three tasks: data preprocessing and getting some preliminary results via exploratory data analysis. Findings will be communicated directly to TransLink to further explore and implement any possible feedback to improve the analysis. Meanwhile, a Makefile is expected throughout the project to establish a data pipeline. In the third phase, the final predictive model will be developed, and the final interactive report will be generated. With a docker container, the data pipeline will be completed to make the project fully reproducible. The fourth phase is dedicated to final reviews and completion of the project.




# Reference

TransLink. 2018. *2018 Accountability Report*. Vancouver, Canada: TransLink. https://view.publitas.com/translink/2018-accountability-report/page/1.