Predicting Patient Presentation Rates at Mass Gatherings using Machine Learning

Peter Serwylo
Monash University
peter.serwylo@monash.edu

Paul Arbon
Flinders University
paul.arbon@flinders.edu.au

Grace Rumantir
Monash University
grace.rumantir@infotech.monash.edu.au

ABSTRACT

Mass gatherings have been defined as events where more than 1,000 people are present for a defined period of time. Such an event presents specific challenges with respect to medical care. First aid is provisioned on-site at most events in order to prevent undue strain on the local emergency services. In order to allocate enough resources to deal with the expected injuries, it is important to be able to accurately predict patient volumes. This study used machine learning techniques to identify which variables are the most important in predicting patient volumes at mass gatherings. Data from 201 mass gatherings across Australia was analysed, finding that event type is the most predictive variable, followed by the state or territory, heat index, humidity, whether it is bounded, and the time of day. Variables with little bearing on the outcome included the presence of alcohol, whether the event was indoors or outdoors, and whether it had one point of focus. The best predictive models produced acceptable predictions of the patient presentations 80% of the time, and this could be further improved using optimization techniques.

Keywords

Mass Gathering, Patient Presentation Rate, Data Mining, Machine Learning

INTRODUCTION

Mass gatherings, such as sporting events and concerts, occur frequently and the potential for injury at such events is considered higher than the general population (Arbon, 2004). Providing on-site first-aid at these events is important to provide rapid access to triage of casualties, stabilize and transport seriously ill patients to the most appropriate hospital, and to cope with minor complaints (Hnatow and Gordon, 1991). All of these factors aim to reduce the strain on local emergency medical services (EMS) which may not be designed to cope with the larger amount of injuries present at a mass gathering.

In order to provide on-site medical care, there must be an adequate amount of first aid staff, such as paramedics and physicians. This must be balanced by the need to efficiently use available resources, especially as many Australian mass gatherings are serviced by volunteer first-aid organisations such as St. Johns Ambulance (St.John, 2010). Hence, there is a need to accurately predict the volume of patients expected during an event in order to provide the correct amount of resources. A patient is defined as an individual presenting for care to on site first aid or medical services at the event. Measures used in the literature to quantify the number of patients are the Patient Presentation Rate (PPR) and the Transport to Hospital Rate (TTHR), both of which are measured in patients per 1,000 attendees.

Often the staffing requirements for events are based solely on "local experience and anecdotal knowledge" (Zeitz, Schneider and Dannielle Jarrett, 2001). The aim of this study is to build on earlier research which investigated the relationship between event variables such as time of day, weather and humidity, and how they interact to influence the PPR and TTHR. This paper will focus on the PPR, although the techniques used could

Reviewing Statement: This full paper has been fully double-blind peer reviewed for clarity, relevance, significance, validity and originality.

also be applied to predicting TTHR and investigating which variables influence it.

BACKGROUND

The literature on injury patterns at mass gatherings can be divided into three categories: single descriptive case studies, aggregate case studies and predictive analysis. Case studies of individual events analyse the frequency and severity of injuries and often describe the types of medical support which was present. Examples range from the 2003 Toronto Rolling Stones concert with over 450,000 attendees (Feldman, Lukins, Verbeek, MacDonald, Burgess and Schwartz, 2004), to a Japanese festival with over 1 million attendees, which only allowed medical staff to attend for the first time in 2004 (Yazawa, Kamijo, Sakai, Ohashi and Owa, 2007). While these are important in understanding how medical care is organised for different events and how it can be improved, they do not focus as strongly on what may cause injuries and how to prevent them.

Improving on these single case studies are those which analyse multiple events over a longer time period. Rose, Laird, Prescott and Kuhns (1992) studied data from 35 college football games over a period of 6.5 years. They identified that the injury rate was underrepresented, perhaps because people were not aware of the availability of first aid, or because they refused medical attention in favour of returning to watch the game. Grange, Green and Downs (1999) analysed 405 different concerts over a five year period, and identified rock concerts as having a larger impact on patient presentations than any other type. These longitudinal studies provide useful information in identifying factors which influence injuries, and they provide a good stepping stone to the third type of study.

Using data from past events, predictive models have been built which can estimate the number of patients at an event. Many of these are based on research from Milsten, Maguire, Bissell and Seaman (2002) who provided a comprehensive analysis of over 30 research articles from between 1977 and 2002, looking at factors which influenced injury patterns at each event. They found that the injury rate was related to a large number of factors, including weather, environmental factors, alcohol and drug use, attendance, crowd density, crowd mood, duration, outdoor vs indoor, seated vs mobile, locale, age of attendees and event type (Milsten, Seaman, Liu, Bissell and Maguire, 2003).

Arbon, Bridgewater and Smith (2001) produced a statistical model which attempts to predict the number of patients based on the variables identified by Milsten et al. (2003). A linear regression model was derived from data collected from 201 different events across Australia. The variables measured for each event are shown in Table 1. This is also the data set used for analysis in the current study. The problem with using linear regression to build such a predictive model is that predicting patient rate is an inherently non-linear problem. This can be seen, for example, by the positive correlation between temperature and PPR until the temperature reaches a certain point, after which the PPR begins to fall, possibly due to extra precautions taken by the attendees

Attribute	Description
Event Type	Rugby, Cricket, Soccer, Tennis, etc
State	Australian state where the event occurred
Bounded	Events which occur within a boundary (often fenced)
Indoors	Either Indoors, Outdoors, or Both
Heat Index	Combination of humidity and temperature
Time	Either Day, Night, or Both
Humidity	Percentage humidity
Number attending	Total crowd in attendance
Temperature	Measured in Celsius
Focused	Focused events occur in a clearly defined venue
Alcohol	Whether alcohol is served or not
Wind Chill	Perceived temperature influenced by the wind
Mobile/Seated	Seated events include large stadium events, while mobile events include fairs

Table 1: List of variables included in the dataset from Arbon (2002) and used in this study (Arbon, 2002).

The model developed by Hartman, Williamson, Sojka, Alibertis, Sidebottom, Berry, Hamm, O'Connor and Brady (2009) uses five key variables in analysing data from 55 events which took place at a College in the united states (Table 2). Each variable has three possible values, and each value is assigned a score from 0 to 2

depending on its severity. Events are then categorised into either major, intermediate or minor based on the sores assigned to each variable. Such a model is better able to deal with non-linear problems, however it fails to prioritise any one variable over another. Also, the simplicity of the model, while allowing anybody to use it, also hinders the ability to make predictions about more specific or complicated mass gatherings. It is unlikely that the majority of events can be categorised into one of only three categories.

Heat Index	Number Attending	Alcohol	Crowd age	Crowd mood	Score
> 32.2°C	> 15,000	Significant	Older	Animated	2
< 32.2°C	1,000 - 15,000	Limited	Mixed	Intermediate	1
Climate Controlled	< 1000	None	Younger	Calm	0

Table 2: Five attributes used by Harman et al. (2009) to predict the PPR. (Adapted from Harman et al. (2009), and renamed so that the attributes match those from Arbon (2001) and the current study)

Other models for predicting the PPR have also been developed. For instance, Baird, O'Connor, Williamson, Sojka, Alibertis and Brady (2010) focus on the weather conditions and its impact on injuries. Their model builds on results from Kman, Russell, Bozeman, Ehrman and Winslow (2007) who identified a positive correlation between temperature and patient presentations. To solve the problem that predicting the PPR at an event is a non-linear problem, Baird et al. (2010) explicitly take into account the PPR and heat index from previous events. As this is the most variable aspect of a recurring event, it provides a mechanism to control for all other variables (e.g. crowd mood, attendance, presence of alcohol), most of which will not change much between different recurrences of an event. This relies on the assumption that recurring events will have strong similarities and attract similar people in the future.

Another retrospective method for predicting the PPR was devised by Zeitz, Schneider et al. (2001). The Zeitz model is based on retrospective data from one multiday event over the past seven years. Having such a specific data set to base predictions on is beneficial because it can predict quirky but important issues. For instance, half price admission days consistently exhibited a higher incidence of injuries, despite not always drawing the largest crowd.

These retrospective model types have been found to be useful for recurring events (Zeitz, Zeitz and Arbon, 2005), but are not able to make predictions about a new event or one for which there is no prior observations available. For this, more general models are required.

While each of the models reviewed here make important contributions to the field, more research is required to better understand the relationship between event variables and the PPR. The aim of this research is to identify the most important variables which influence patient rates, so that they can be further investigated at future mass gatherings. It will also identify the least influential variables that should not be included in models for predicting the PPR at mass gatherings.

METHOD

This study analysed the relationship between different variables and their effect on the PPR using machine learning. Learning algorithms are used to build predictive models that in turn will be used to estimate how many injuries can be expected at an event. During the modeling process, variable selection is performed to identify the most "useful" variables. These are the variables which, although not necessarily having a strong correlation with the PPR, provide the highest information gain when being used in the models.

The data set contains 201 different events which occurred over a period of 12 months. The data was collected by St John Ambulance Australia personnel (St.John, 2010), and has previously been used for building a linear regression model by Arbon et al. (2001). The variables which were measured are shown in Table 1. Of these variables, three contained missing values: *heat index* was missing 15/201, *wind chill* was missing 81/201 and *alcohol* was missing 1/201. All algorithms used in this study have the capability to appropriately deal with missing values, therefore no effort was undertaken to replace them.

Some models are easy to interpret and can be used without any special knowledge or software, such as decision tree learners which produce simple "if *x* then *y*" type rules. Others, such as neural networks, are treated as more of a black box where the input variables are provided, and an estimation of the PPR is returned. For a model to be useful to event organisers in predicting the PPR and planning for future gatherings, it should fall into the simple and easy to understand category. Despite this, the abstract models are still useful in identifying which variables have the most predictive power, so that these can be further investigated during subsequent data collection.

Type	Algorithm	Abbreviation	Reference
Regression	M5P Decision Tree	M5P	(Wang and Witten, 1997)
	Neural Network	MLP	(Hall et al., 2009)
	Linear Regression	LR	(Hall et al., 2009)
Classification	Naïve Bayes	NB	(John and Langley, 1995)
	Support Vector Machine	SMO	(Keerthi, Shevade, Bhattacharyya and Murthy, 2001)
	C4.5 Decision Tree	J48	(Quinlan, 1993; Wang and Witten, 1997)
	Reduced Error Pruning Decision Tree	REP	(Hall et al., 2009)

Table 3: The seven learning algorithms used. Their abbreviations will be used to reference them throughout this article.

Different learning algorithms require different input data types (categorical or numerical) and have varying biases in the way they make their predictions. This can be influenced by the properties of the data being used. To cater for the chance a particular algorithm may be unfairly biased with this particular dataset, multiple algorithms were used in predicting the PPR and the results averaged (Table 3). All machine learning tasks were performed using the Weka software (Hall, Frank, Holmes, Pfahringer, Reutemann and Witten, 2009).

The data is first split into two groups: 80% (161) of the events for training and the remaining 20% (40) for testing. The training data is used to build and test the predictive models, in addition to performing variable selection in order to rank the variables. The remaining 20% is not used in any capacity until we have successfully ranked the variables to identify which are the most useful in making predictions. The reason for retaining 20% of the events for purely testing of models is because during the model building phase, each of the 80% of training events will have a chance to contribute to a particular model at some stage. This means that when it comes to testing, it is not a completely independent data set. What the reserved 20% testing data does is to provide a means for testing models in real world situations, on data they have never seen before. The results from this type of testing are important, because they can be seen as representative of the accuracy of the model in practice.

In total, three regression and four classification algorithms were used for building the models. The difference between the two is that with regression, the predictive models will output specific values for the PPR (e.g. 0.134), whereas classification models can only output discrete classes (e.g. between 0.0 and 0.2). For the classification models, the PPR was discretized into six groups each with an equal number of members. The ranges of each class are not equal (e.g. 0.5 to 0.7, and 0.7 – 2.3), and that reflects the non-uniform distribution of different PPRs at each event. The algorithms used throughout this study are shown in Table 3.

In order to rank the variables for "usefulness", the training data is split into 10 folds. Each will be used to train a different model and select the variables which provide the most accurate results for that particular model. This will result in 10 different predictive models for each of the seven algorithms. The variables are then ranked based on how many of these 10 models they were included in. The variables which are selected the most can be considered to be the most useful. To generate the folds for use in training each model, the data is divided into 10 sections. These can be combined in different ways to generate 10 different folds of data, each using a single different section for testing, and the remaining 9 sections for training (Table 4). This provides 10 different training-test data sets from only 80% of the one original data set, still reserving the final 20% for final testing.

		Sections			
		#1	#2	…	#k
Folds	Fold 1	Test	Train	…	Train
	Fold 2	Train	Test	…	Train
	Fold …	…	…	…	…
	Fold k	Train	Train	…	Test

Table 4: Description of how to obtain k folds (and hence k unique training/testing sets) from the one data set.

It is rare that the best model will be created by using every available variable. Therefore, a subset of the available variables should be selected to be used in the final model. Given that it is not feasible to train models with every possible combination of variables, heuristic search strategies can be used to reduce the search space

of feasible variable combinations. While they are not guaranteed to find the best available combination, they will usually identify a good selection which helps produce accurate models. This research will use three different heuristic search strategies (best first, greedy stepwise and a genetic search algorithm) provided by Weka (Hall, et al., 2009).

This provides us with a technique to perform machine learning to identify useful variables. All results will be derived directly from the data, without the need for any previous observations other than the data being analysed.

RESULTS

The final variable rankings are shown in Table 5. Of the seven different algorithms, each one performed heuristic variable selection on 10 distinct folds, using three different search strategies (best first, genetic search and greedy stepwise). This produces 210 (7 x 10 x 3) possible opportunities for a variable to be selected and used in building a model. Table 5 shows that *event type* was selected most frequently (70.48%) while *wind chill* was selected the least (29.05%).

Variable	Selected	Total (%)
Event Type	148	70.48%
State	126	60.00%
Heat Index	97	46.19%
Bounded	96	45.71%
Humidity	94	44.76%
Time	92	43.81%
Mobile/ Seated	83	39.52%
Number attending	83	39.52%
Temperature	75	35.71%
Indoors	74	35.24%
Alcohol	65	30.95%
Focussed	63	30.00%
Wind Chill	61	29.05%

Table 5: The 13 variables, ranked by "usefulness" according to how many times they were selected for inclusion in a model.

Based on the rankings in Table 5, we can then decide which variables to include in the final models. With these rankings, it is possible to identify the cutoff point where useful variables are included in the models and the less useful variables are ignored. This is done by rebuilding all of the models several times, each time including one less variable from the bottom (less useful) end of the list. This resulted in the 70 models shown in Table 6 which depicts the accuracy obtained for each model during validation. The models begin with only the top four variables, and move all the way down to include each of the 13 available variables. These results indicated that the most appropriate cutoff point for useful variables is after the top seven. This was despite the fact that the seventh and eighth variables exhibit exactly the same ranking. Three out of the seven best models (LR, J48 and REP) achieved their best accuracy with these top seven variables.

The most accurate regression model was the MLP neural network with a root mean squared error (RMSE) of 2.33. This model made use of all but the two lowest ranked variables: *wind chill* and *focused*. The J48 decision tree, built using either the top six or seven variables (they both produced exactly the same decision tree) was the most accurate classification model with 52% of events classified correctly. The confusion matrix for this decision tree is shown in Table 7, and highlights all of the correctly classified events, as well as the events which were “nearly” classified correctly. Knowing that the classes which represent different PPR values are ordinal, predicting only one class either side of the correct class may still be accurate enough for event organisers to make informed decisions. From this matrix, it can be calculated that 51.55% of events had their PPR correctly predicted, and a further 24.84% were nearly classified correctly, leaving 23.60% misclassifications.

Regression – Root mean squared error (Lower is better)											
Algorithm	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Trend**
LR	2.82	2.82	2.76	**2.75**	2.79	2.79	2.81	2.85	2.87	2.86	
M5P	2.86	2.84	2.41	2.46	2.48	2.50	2.49	2.49	2.49	**2.40**	
MLP	3.42	4.25	3.19	3.73	2.48	3.17	2.85	**2.33**	3.10	2.50	
Average	3.03	3.30	2.79	2.98	2.58	2.82	2.72	**2.56**	2.82	2.59	

Classification – Correctly classified (Higher is better)											
Algorithm	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Trend**
SMO	45%	44%	44%	45%	44%	45%	45%	**47%**	45%	45%	
NB	34%	37%	37%	37%	38%	38%	36%	38%	38%	**40%**	
J48	37%	42%	**52%**	**52%**	49%	46%	47%	47%	47%	47%	
REP	38%	40%	**43%**	**43%**	40%	39%	39%	39%	39%	39%	
Average	38%	41%	44%	**44%**	43%	42%	42%	43%	42%	43%	

Table 6: The accuracy of generated models, ordered by how many variables were selected to build the model (from a minimum of 4 variables to all 13). The regression and classification algorithms were evaluated separately because they use different metrics to measure accuracy. The trend column shows whether each algorithm operates better with more or less attributes.

		Classified as					
		0.0 – 0.2	**0.2 – 0.4**	**0.4 – 0.5**	**0.5 – 0.7**	**0.7 – 2.3**	**2.3 – ∞**
Actual Class	**0.0 – 0.2**	**13**	7	3	4	2	0
	0.2 – 0.4	6	**15**	3	3	3	0
	0.4 – 0.5	2	6	**14**	4	0	0
	0.5 – 0.7	6	6	3	**8**	3	0
	0.7 – 2.3	3	5	1	3	**16**	1
	2.3 – ∞	0	0	0	0	4	**17**

Table 7: J48 decision tree, using the top six or seven variables (51.55% correct, 24.84% nearly correct, 23.60% incorrect)

The final analysis performed was to take the 20% data which was set aside for testing and to use the various models to predict the PPR for each event. This is important because the models were built with no knowledge of these events, and therefore we can get an indication of how accurately the models can predict future events. Using this test data, the best regression model was the M5P decision tree built with the top six variables. During validation this model achieved an RMSE of 2.41, only 0.01 worse than the best performing M5P model during validation. When testing the model with the test data set, the RMSE improved significantly to a value of 0.565 (from 2.41). Figure 1 shows the accuracy of the M5P regression model with the top six variables when used to estimate the PPR for each of the test events. The average error is 127.52%, for example, an event had an actual PPR of 0.284, but the predicted value was 0.744, an error of 161.97%. Figure 1 shows all 40 events from the test set and how accurately the PPR was estimated for each one. The average error is depicted by the horizontal line, and from this we are able to see how the outliers skew the average error, with most events well below the average error (127.52).

Two competing classifiers exhibited the best accuracy using the test data, the SMO Support Vector Machine and the Naïve Bayes classifier. The SMO classifier (Table 9) made use of the top 10 variables and classified the highest number of events as either correct or nearly correct (85%), with only 15% completely misclassified. The Naïve Bayes classifier using all but the least ranked variable, *wind chill* (Table 8) had the most events classified in the correct category (47.5%). However it also had a 20% misclassifications rate.

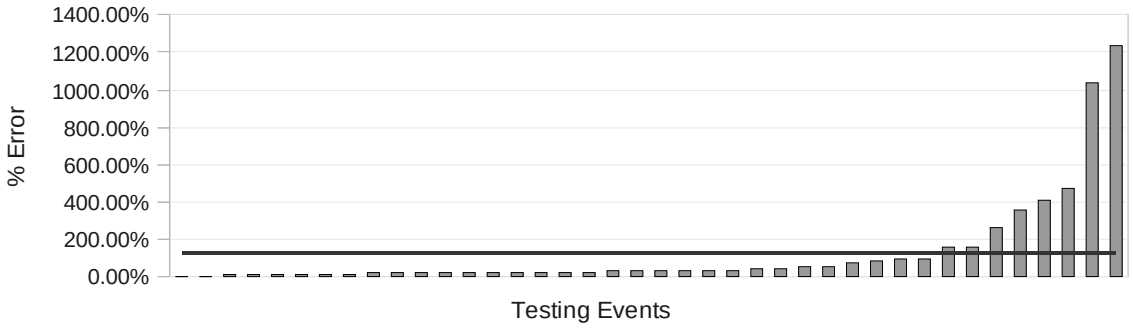

Figure 1: The accuracy of the M5P decision tree when performing regression to estimate the PPR in the 20% unseen testing data. The horizontal line is the average, although this is skewed by the outliers to the right.

		Classified as					
		0.0 – 0.2	**0.2 – 0.4**	**0.4 – 0.5**	**0.5 – 0.7**	**0.7 – 2.3**	**2.3 – ∞**
Actual class	**0.0 – 0.2**	**5**	0	1	0	0	0
	0.2 – 0.4	0	**1**	5	0	0	0
	0.4 – 0.5	0	2	**5**	0	1	0
	0.5 – 0.7	3	0	4	**2**	0	0
	0.7 – 2.3	1	0	2	0	**3**	0
	2.3 – ∞	0	0	0	0	2	**3**

Table 8: NB classifier using 12 out of 13 variables (47.5% correct, 32.5% nearly correct, 20% incorrect)

		Classified as					
		0.0 – 0.2	**0.2 – 0.4**	**0.4 – 0.5**	**0.5 – 0.7**	**0.7 – 2.3**	**2.3 – ∞**
Actual class	**0.0 – 0.2**	**5**	1	0	0	0	0
	0.2 – 0.4	1	**4**	1	0	0	0
	0.4 – 0.5	0	6	**1**	0	1	0
	0.5 – 0.7	1	3	3	**2**	0	0
	0.7 – 2.3	0	1	0	0	**3**	2
	2.3 – ∞	0	0	0	0	2	**2**

Table 9: SMO classifier using 10 out of 13 variables (45% correct, 40% nearly correct and 15% incorrect)

DISCUSSION

In this discussion, the performance of notable variables is considered. Most of the variable rankings agree with the literature, while some, such as *alcohol* do not. Some variables can be seen as masks for others, either having a similar impact (e.g. *bounded* and *focused*) or being a higher level variable with constituent parts which need further investigation, such as *state/territory* in the way it encapsulates a certain set of weather parameters based on its geography.

The most frequently selected variable is *event type*, being selected in 70.48% of the models. Each individual type of event has a unique set of circumstances which impacts on the PPR. A likely factor which can be ascertained from the type of event is the *crowd mood*. This has been mentioned in the literature as in important influence (Milsten, et al., 2003) and has been used in a predictive model (Hartman, et al., 2009). Given the qualitative nature of the *crowd mood* variable, the *event type* may provide a suitable way to predict it.

The *state/territory* variable (used 60% of the time) warrants more investigation. If other variables in this study are found to cause a higher incidence of injury, they can be easily controlled for. Attendance can be capped, more drinking fountains can be deployed for hot days, and alcohol consumption can be limited. The *state/territory* cannot be so easily changed to influence the rate of injury. However, if the event organizers were aware of the unique environmental makeup of each *state/territory*, then more useful information can be gleamed. For example, people attending an event in the northern part of Australia will often be attending hot and humid events. However, these attendees would be more acclimatized to these hot temperatures, and therefore may be better prepared.

The *heat index* is used by 46.19% of the models. This validates claims made by Arbon et al. (2001) as well as the model by Baird, et al. (2010) which use heat index as an input variable. This is in contrast to *wind chill*, which scored the worst with only 29.05% utilisation. The disparity between the two may be an artifact of Australia's relatively hot climate. It could be expected that in a hot climate, the warming effect of humidity would have more of an impact on injuries than the cooling effect of the wind. Given this, it would be interesting to see if the results from data originating in a colder country were reversed. Perhaps in such an environment, *humidity* would have very little effect and the *wind chill* would dominate.

The variable *indoors* (35.24%) is ranked below all of the temperature measurements except *wind chill* (*heat index* 46.19%, *humidity* 44.76% and *temperature* 35.71%). This is surprising, because if an event is held indoors, then the effect of weather would be greatly dampened. That is not to say that there is no effects of the weather felt inside, but that it should have a greater influence on outside events. It is worth noting here that of the 201 events analysed, only five were classified as indoor, 32 were both indoor and outdoor, and the remaining 164 solely outdoor. It may be that the *indoor* variable would rank differently if there were a greater proportion of indoor events to provide a more balanced data set.

Given that the literature mentions *alcohol* as a predictor (Milsten, et al., 2003, Moore, Williamson, Sochor and Brady, 2010), it therefore comes as a surprise that the presence of *alcohol* is not regarded as very useful. It is selected for use in only 30.95% of the models. Events which serve alcohol are usually in hotter conditions with specific crowd moods, which may be the reason research has shown an increase in PPR with the presence of alcohol. Due to these complex relationships, more research is needed before alcohol can be ruled out as an important influence.

The variable *focused* performed very poorly, selected for inclusion in only 30% of the models. The most likely reason is its strong correlation with the *bounded* variable ($r = 0.871$). When there are two or more variables closely correlated, machine learning and statistical algorithms are designed in such a way as to only utilise one or the other. Simpler models are prioritised and there is little information to be gained by adding the second variable.

Finally, it is interesting that although *mobile/seated* and *number attending* were both equally selected in 39.52% of the models, the subsequent rebuilding of models with only a subset of available variables (Table 6) prioritised *mobile/seated* over *number attending*. Given the fact that they were initially ranked equally, it would be worthwhile for future research to continue to investigate the number of attendees and how it influences the PPR.

LIMITATIONS

The two highest ranked variables, *event type* and *state/territory*, are specific to Australia. Thus, this analysis needs to be replicated for different geographies to provide a more widely applicable model. Event types will most likely differ between countries and perhaps even within a single country. The state where an event is held is obviously only applicable to Australian environments and cannot be used to generalise predictions across multiple countries.

CONCLUSION

This study used machine learning to identify which variables are the most useful to understand when trying to predict the patient presentation rate at mass gatherings. The aim was to perform a data centric analysis of the different variables, based on data collected from 201 mass gathering events held across Australia. It found that *event type*, *state/territory* and *heat index* were the most useful variables to include in the predictive models, confirming previous research in the field. The less useful variables include *wind chill* and *alcohol*. The low ranking of the *alcohol* variable is in contrast to past research, and will need further investigation to identify how it affects the PPR, while *wind chill* is ranked far below *humidity* for this Australian data set. This research will allow future data collection efforts to be refined, resulting in more accurate models for predicting the PPR in the future.

FUTURE WORK

We believe that more accurate and fine grained prediction models can be developed if PPR is measured as "PPR per hour". Such models would be able to refine their predictions based on patterns which are observed throughout a day, rather than just at the broad level of an entire event.

More robust and accurate models can be built by, for example, combining multiple models into a large

ensemble. Such predictive models have shown to be more effective, but obviously have the drawback of being large and complex. These models could then be compared to other models described in the literature to see if they exhibit more accurate predictions. Because the scope of this paper was to investigate the different variables contributing to mass gathering PPRs, its primary goal was not supreme accuracy.

In this research, the models were only validated on unseen test data from the past. The models could also be validated with data from newly documented mass gatherings, to see if they are useful for predicting future PPRs.

Also, including the results of predictive models into simulation systems would allow event organisers to experiment with different resource allocations and configurations to see how they fare. These simulations could also be used to train first aid staff and event organisers.

REFERENCES

1. P. Arbon (2002) - The development of a web-based algorithm for the prediction of patient presentation rates at mass gatherings, *Australian Journal of Emergency Management, The*, 17, 1, 60-64.
2. P. Arbon (2004) - The development of conceptual models for mass-gathering health, *Prehospital and Disaster Medicine*, 19, 3, 208-212.
3. P.A. Arbon, F.H. Bridgewater, C. Smith (2001) - Mass Gathering Medicine: A Predictive Model for Patient Presentation and Transport Rates, *Prehospital and Disaster Medicine*, 16, 3, 109-116.
4. M.B. Baird, R.E. O'Connor, A.L. Williamson, B. Sojka, K. Alibertis, W.J. Brady (2010) - The impact of warm weather on mass event medical need: a review of the literature, *The American Journal of Emergency Medicine*, 28, 2, 224-229.
5. M.J. Feldman, J.L. Lukins, R.P. Verbeek, R.D. MacDonald, R.J. Burgess, B. Schwartz (2004) - Half-a-million strong: the emergency medical services response to a single-day, mass-gathering event, *Prehospital and Disaster Medicine*, 19, 4, 287-296.
6. J.T. Grange, S.M. Green, W. Downs (1999) - Concert Medicine: Spectrum of Medical Problems Encountered at 405 Major Concerts, *Academic Emergency Medicine*, 6, 3, 202-207.
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten (2009) - The WEKA data mining software: An update, *ACM SIGKDD Explorations Newsletter*, 11, 1, 10-18.
8. N. Hartman, A. Williamson, B. Sojka, K. Alibertis, M. Sidebottom, T. Berry, J. Hamm, R.E. O'Connor, W.J. Brady (2009) - Predicting resource use at mass gatherings using a simplified stratification scoring model, *The American Journal of Emergency Medicine*, 27, 3, 337-343.
9. D.A. Hnatow, D.J. Gordon (1991) - Medical planning for mass gatherings: a retrospective review of the San Antonio Papal Mass, *Prehospital and Disaster Medicine*, 6, 4, 443-450.
10. G.H. John, P. Langley (1995) - Estimating Continuous Distributions in Bayesian Classifiers, *Eleventh Conference on Uncertainty in Artificial Intelligence*.
11. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy (2001) - Improvements to Platt's SMO Algorithm for SVM Classifier Design, *Neural Computation*, 13, 3, 637-649.
12. N.E. Kman, G.B. Russell, W.P. Bozeman, K. Ehrman, J. Winslow (2007) - Derivation of a Formula to Predict Patient Volume Based on Temperature at College Football Games, *Prehospital Emergency Care*, 11, 4, 453-457.
13. A.M. Milsten, B.J. Maguire, R.A. Bissell, K.G. Seaman (2002) - Mass-gathering medical care: a review of the literature, *Prehospital and Disaster Medicine*, 17, 3, 151-162.
14. A.M. Milsten, K.G. Seaman, P. Liu, R.A. Bissell, B.J. Maguire (2003) - Variables influencing medical usage rates, injury patterns, and levels of care for mass gatherings, *Prehospital and Disaster Medicine*, 18, 4, 334-346.
15. R. Moore, K. Williamson, M. Sochor, W.J. Brady (2010) - Large-event medicine--event characteristics impacting medical need, *The American Journal of Emergency Medicine*, In Press, Corrected Proof.
16. R. Quinlan (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers.
17. W.D. Rose, S.C. Laird, J.E. Prescott, G.B. Kuhns (1992) - Emergency medical services for collegiate football games: A six and one half year review, *Prehospital and Disaster Medicine*, 7, 157-159.

18. St. John Ambulance (http://www.stjohn.org.au/) (2010).
19. Y. Wang, I.H. Witten (1996) - Induction of model trees for predicting continuous classes, *Poster papers of the 9th European Conference on Machine Learning*.
20. K. Yazawa, Y. Kamijo, R. Sakai, M. Ohashi, M. Owa (2007) - Medical Care for a Mass Gathering: The Suwa Onbashira Festival, *Prehospital Disaster Medicines*, 22, 5.
21. K.M. Zeitz, D. Schneider, B.N. Dannielle Jarrett (2001) - Mass gathering events: Retrospective analysis of patient presentations over seven years at an agricultural and horticultural show, *Prehospital and Disaster Medicine*, 585, 1-14.
22. K.M. Zeitz, C.J. Zeitz, P. Arbon (2005) - Forecasting medical work at mass-gathering events: predictive model versus retrospective review, *Prehospital and Disaster Medicine*, 20, 164-168.