On the development of predictive models for use in Emergency Medical Dispatch centers

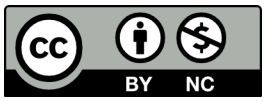
Douglas Spangler
Uppsala University Hospital, Uppsala Center for Prehospital Research
douglas.spangler@akademiska.se

October 2, 2018

Abstract

Triage tools based on Machine Learning offer a promising approach to enhancing the decisional capacity of staff working at Emergency Medical Dispatch centers. There are however many challenges to implementing such systems. In a 2 year project funded by the Swedish Agency for Innovation, the Uppsala Ambulance Service aims to develop a machine-learning based tool to augment the current rule-based decision support system. In this analysis, we outline our approach to addressing these challenges, provide a thorough descriptive analysis of the data available for this purpose, and provide a brief report regarding the overall performance of a set of preliminary models based on retrospective data. Overall, our current non-optimized models have a fair predictive value across a broad range of measures covering ambulance interventions, patient vital signs, and hospital outcomes. We also provide an online tool for exploring our free-text data.

Contents



External resources:

Summary document - (source code)

Online Freetext Explorer - (source code)

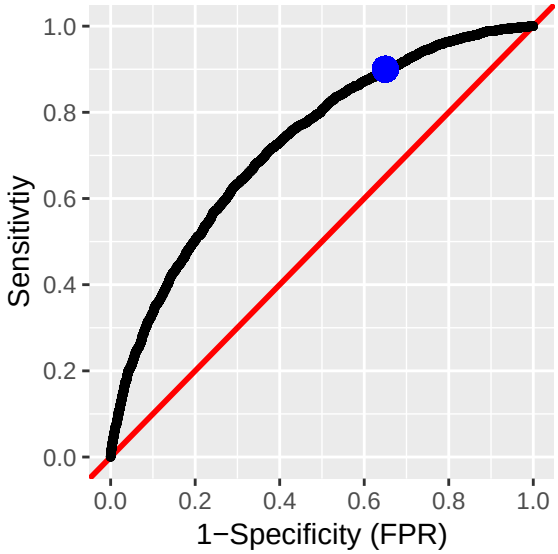
Introduction

It is the goal of the Uppsala Ambulance Service (UAS) to provide the right resource, to the right patient, at the right time. Given an increasingly frail elderly population often with multiple comorbidities and significant underlying care needs, this can be a difficult goal to achieve without sacrificing patient safety. A high degree of sensitivity is demanded of the ambulance service for emergent conditions necessitating an ambulance response, but given a resource-constrained healthcare system, the need to differentiate between patients with various levels of need is a fact of life.

The ability of an Emergency Medical Dispatch (EMD) system to differentiate between acute conditions requiring an ambulance, and conditions that are more appropriately treated using other healthcare resources is dependent on a large number of factors. These include a well educated and trained staff, the provision of feedback and robust clinical guidelines, a healthy workplace culture, well-established links with other healthcare providers, and an effective Clinical Decision Support System (CDSS). While this document with deal exclusively with this final point, it should be recognized that the development of CDSS in the absence of efforts to improve other preconditions for delivering high-quality medical decision making in the practice of EMD are likely to be fruitless.

CDSS have traditionally been based on explicit rules determined based on clinical expertise. Such "expert systems", including the widely utilized Medical Priority Dispatch System and various Criteria-Based Dispatch systems, have the advantage of being relatively simple, and have been found to be reasonably effective in identifying specific high-acuity conditions in the context of EMD. This methodology has been extended to to the identification of low-acuity patients in EMD, and thus far these efforts have had some success. Shah et al. (2005) for instance found that a set of "low priority" dispatch codes could be identified which contained at least 90% calls which were not provided with ALS-level care, while Hinchey et al. (2007) found a comparable rate of less than 1%. Studnek et al. (2012) found that among callers triaged to a low acuity dispatch code, 70-75% of patients could be discharged from the emergency department.

An increasingly popular approach to clinical decision support is the use of predictive models in various forms. Within the domain of emergency medicine, a number of research groups have begun developing machine learning based approaches to triaging patients at the Emergency Department (ED). To describe the overall predictive value of predictive models, the machine learning literature tends to refer to the *Area Under the Reciever Operating Characteristics curve* or AUROC. This value constitutes a summary of all potential pairs of sensitivity and specificity afforded by a given model. As implied by it's name, the AUROC can be visualized by plotting each possible sensitivity against specificity:



In the plot of randomly generated data above, the black line represents the predictive value of the model - in this case, an AUROC of 0.74. Random guessing would result in an AUROC of 0.5 and a curve laying flat along the red line, while a model making perfect predictions would have an AUROC of 1. From this plot, any given sensitivity can be selected and the corresponding specificity read off - Selecting a sensitivity of 0.9 here results in a specificity of roughly 0.35 (Equaivalent to a False Positive Rate (FPR) of 0.65, indiated by the blue dot in the above plot).

The AUROC value provides a useful overall summary, but is also problematic in a number of ways. AUROC values will tend to be higher in cases where the outcome is rare, and predictors with fewer levels will have artificially low AUROC values. Being a summary, the AUROC includes combinations of sensitivity and specificity which may be unimportant clinically. For instance, having anything less than an extremely high degree of sensitivity for cardiac arrests would be unacceptable. While better measures of overall predictive value do exist (the area under a precision-recall curve for instance), for better or worse, AUROC is the measure which has become the "industry standard", and which we will report in this analysis.

Levin et al. (2017) used Random Forest models to achieve predictive values of 0.73 - 0.92 AUROC for a number of outcomes using data gathered at the time of ED triage. Hong et al. (2018) meanwhile achieved an AUROC of 0.92 in the prediction of hospital admission by integrating data from the patient's historical medical records using a number of methods. Horng et al. (2017) used support vector machines to analyze free-text and clinical data, achieving an AUROC of 0.86 in detecting diagnosed infections. Tadahiro et al. (2018) were able to achieve AUROC values of 0.80 and 0.83 respectively in predicting critical care and admission among Asthma and COPD patients in the ED.

We are aware of one other effort to apply machine learning methods to the problem of triage at the EMD center, with Blomberg et al., (2017) achieving a high degree of accuracy in identifying cardiac arrests using an unspecified methodology. We are also aware of efforts to improve operational aspects of EMD, including the City of Cininatti, and the Västerbotten EMS depertment. We are not aware of any peer-reviewed results in this field thus far however.

The problem of developing predictive models for use in EMD is complex and multifaceted. Many type of patients with a wide variety of conditions are handled in the prehospital care system, and given the context in which it is collected, documentation of the care provided can be poor. Numerous decisions must be made during the analytical process in terms of selecting appropriate inclusion and exclusion criteria, developing a robust set of outcome measures, training models to predict these outcomes, and delivering their outputs to clinical staff in a manner which augments rather than replaces their decisional capacity.

This paper aims to provide insight into how these problems are being addressed at the UAS in a two-year project funded by the Swedish Agency for Innovation (Vinnova). Based on the findings of the project thus

far, and the analysis of two years of data from the Uppsala region (2016-2017), this document will:

1) Describe the assumptions and general approach taken in developing a CDSS based on predictive modelling for use by qualified clinical staff at the EMD center.

2) Provide a descriptive analysis of the EMD data available for use in predictive modelling at out agency, as well as the outcome data available from ambulance records and the regional Electronic Medical Record (EMR) system, and note some interesting findings from exploratory data analysis.

3) Assess the predictive value of preliminary models for the outcome measures developed in this project, and compare these to existing triage methods.

4) Discuss the findings and how these models may be used as the basis for a clinically useful tool by nurses operating at the EMD center.

While we hope that these results and the source code used to generate them will be useful to other researchers and clinicians, we believe that to be considered of research quality, the results of our predictive modelling efforts in particular must be validated against an unobserved dataset. While we report model performance based on cross-validated test data in this report, in a future publication we plan to validate our final models against unexamined data collected in 2018.

1) Approach

Our approach to developing models for prehospital care is based on the consideration of a number of factors which must be weighted against each other. What data do we have access to, and is it reliably documented? For what patient cohorts are the models based on these data valid? What specific decisions can we help our clinicians make when triaging a patient? Over the course of the project thus far we have found that there is a preference among EMD nurses towards simplicity in how the CDSS is presented, and towards a focus on short-term patient outcomes over longer term outcomes. Simultaneously, simple models based on single outcome measures may be criticized for being insufficiently comprehensive (e.g., a prediction of the likelihood of transport to a hospital might mischaracterize patients requiring treatment in the home but not transport).

As such, we've chosen to pursue the development of 3 outcome "families" which represent respectively the prehospital interventions provided to a patient, a patients prehospital vital signs (as scored on the National Early Warning System (NEWS) scale), and the patients hospital outcomes. Each of these families has strengths and weaknesses:

Prehospital interventions may include for instance transport with lights and sirens, the administration of medications, spinal immobilization, insertion of IV catheters, oxygen administration, etc. It is the aim of this family to include as comprehensive a set of interventions as possible, allowing us to differentiate between patients in need of ambulance care from those who could reasonably be transported via alternate means to an ED or other healthcare facility. This set of measures includes those outcomes which are of most interest to dispatchers based on a questionnaire distributed to EMD nurses in Uppsala and Västmanland, but also come with weaknesses. These measures can be said be somewhat subjective - The criteria for transporting a patient with lights and sirens for instance are fairly vague, and different crews are likely to make different decisions in the same situation. There is also an overhanging risk that biases on the part of the dispatcher may influence outcomes, particularly if patients who do not receive an ambulance are included and considered to be negative for these measures. These measures could be thought of in the context of the Donebedian quality framework as measures of care processes.

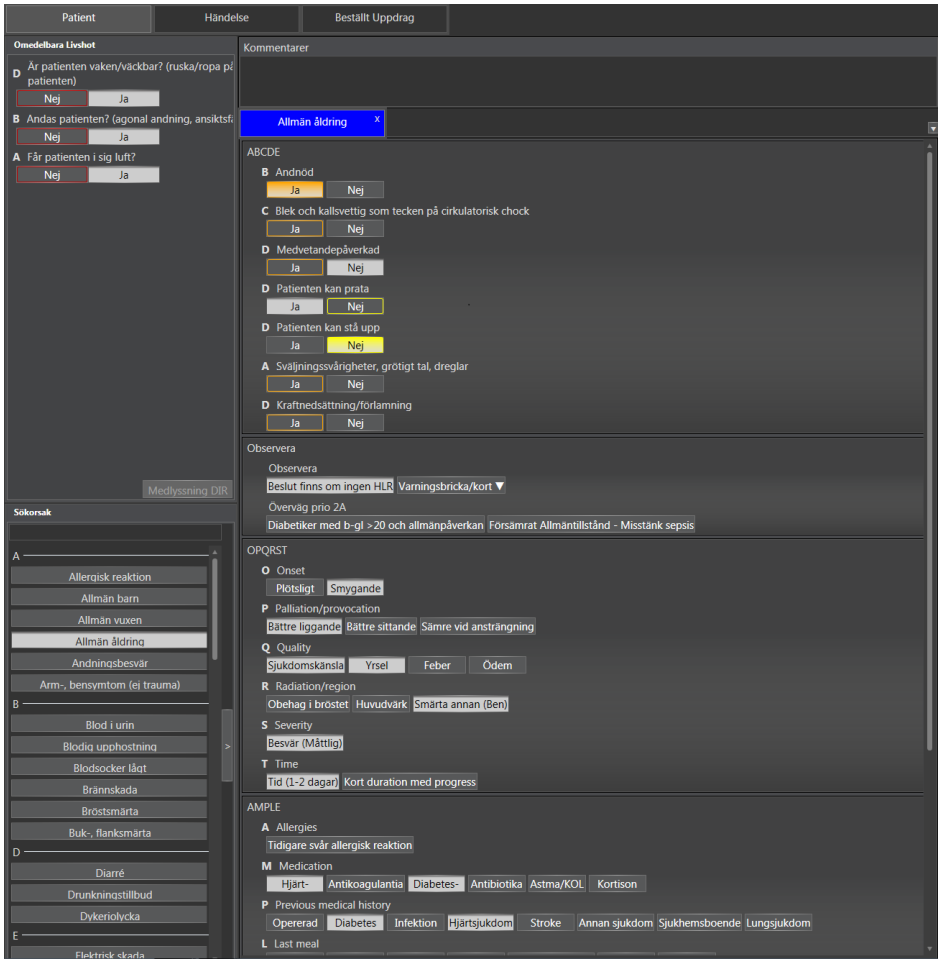
Patient vital signs per NEWS have the benefit of being a clinically validated and objective measure of patient acuity, for instance by Silcock et al. (2015) and Hoikka et al. (2018). However, the interpretation of what a high or low NEWS score means in the context of an individual patient is difficult to postulate beforehand in the context of EMD. Does a high NEWS value indicate the need to an ambulance response, or does it indicate a response by a mobile geriatric care doctor? Furthermore, these data are only captured for patients receiving

an ambulance response, and therefore models based on these measures risk mischaracterizing patients who commonly do not receive an ambulance response.

Hospital outcomes include for instance admission to an in-patient care ward, treatment in an intensive care unit, hospital length of stay, provision of surgical interventions or radiological diagnostic procedures, etc. While these measures are likely to be less subject to dispatcher bias, hospital data are only available for patients for whom a valid Personal Identification Number (PIN) is captured, and as such these measures risk misrepresenting patient populations for whom a valid PIN is not commonly documented. Furthermore, given the complexity of hospital medical care and the relatively limited structured data available, this measure family is not likely to represent a comprehensive picture of the care provided to a patient at the hospital, and can at best provide a limited set of indicators of care intensity. There are also risks in terms of loss to followup as we only capture data from the regional EMR system (though we can exclude patients calling from municipalities close to neighboring hospitals), and we only capture healthcare contacts up to 72 hours following the call to the EMD center.

The goal then is to present three risk models representing each of these families to clinicians at the EMD center. Many of the issues presented (e.g., which patients each model can reliably represent) cannot be resolved statistically. Only by training staff in the strengths and weaknesses of each model and their appropriate use, presenting information which allows clinicians to interpret the model predictions, and continuously monitoring system performance can their safe implementation be ensured.

Just as outcomes for our patients are multifaceted, so too are the data available to predict them from. The CDSS used at the Uppsala, Västmanland, and Sörmland EMD centers in central Sweden is of the traditional rule-based type. Known in Swedish as *Medicinska Beslutsstödet* (MBS), it consists of 60 distinct call types, Within which clinicians are able to document answers to a structured set of questions distinct to each call type (though some call types share questions, resulting in 36 unique sets of structured data). The user interface for this system looks like this:



The MBS is structured based on the AMLS triage concept, and consists of several question groups. Three *Initial* questions at the top left of the screen are designed to identify and detect cardiac arrest, and result in an immediate, top-priority ambulance dispatch. Upon completing the *Initial* questions, dispatchers select a call type, each of which is tied to a base priority; The "General Elderly" call type shown above is for instance has a base priority of Referral. The *ABCDE* questions associated with each call type are designed to rule out acute conditions based on the patient signs and symptoms. This set of questions always follows a yes/no format and often tied to the rule-based triage system, resulting in an adjustment to the priority determination as indicated by the color of the answer buttons. In his case, sine the dispatcher has indicated that the patient has difficulty breathing (Andnöd), the call is upgraded to a lights and sirens response. *Observera* questions are intended to capture additional items important to the triage determination, but which are not tied explicitly to the rule-based determination. *OPQRST* questions are designed to capture details of the patient's condition, while *AMPLE* questions are designed to capture information related to the patient's medical background. Only some of these questions have answers which lead to a change in call priority, and the *OPQRST* and *AMPLE* sections unfortunately have a poor rate of documentation. Additionally, dispatchers lack the ability to document negative findings for these questions, making these data unsuitable for use in predictive models (Nor does inclusion of data from these sections in our models seem to yield improvements in their predictive value). We plan on implementing improvements which will (hopefully) address these issues, but for now, only data from the *ABCDE* group will be used in the context of developing predictive models.

Based on the call type, and modified by the MBS answers, a recommended priority is generated:

1A - Lights and sirens, high priority
1B - Lights and sirens, low priority
2A - No Lights and sirens, high priority

2A - No Lights and sirens, low priority
Referral - No ambulance sent

Calls at the Uppsala EMD center are answered and triaged exclusively by nurses, and a high degree of latitude is afforded as to whether to adhere to the recommended priority from the MBS. Indeed, while highly encouraged, it is not mandatory to utilize the system, and nurses may also choose to assign a priority based on their discretion without using the MBS.

In addition to the structured data documented in the MBS, information is available regarding the demographics of the patient (Age, Gender, Prior EMD contacts, call location and time), as well as unstructured text data documented by the dispatcher in a comment field located above the MBS proper. These comments have a median length of ~130 characters, and follow no particular structure. These varying types of data allow for varying approaches to predictive modelling. It is quite feasible to model demographic and structured MBS data using traditional regression-based approaches given the comparatively small number of features. The analysis of text data in this context can be trickier, though certainly possible through the application of regularization and/or dimensionality reduction techniques.

While a number of methods have been tested on these data, the primary estimation method presented here in section 3 will be based on gradient boosting as implemented in the xgboost R package. In our experience, this method has performed as well or better than comparable regularized regression, Random Forest, and neural network models in terms of overall predictive value (We'll be presenting a formal comparison of these approaches as validated in 2018 data in a future publication). While similar in performance, we've found gradient boosting to be more robust in the context of unbalanced outcome data, to handle certain types of missing data gracefully, account appropriately for interactions between predictors, and to be relatively straightforward in terms of implementation. We by no means however claim to have identified the optimal methods for modelling these data. As such, we limit our interpretation to their relative performance in comparison with other nested sets of predictors, and with benchmarks consisting of the priority recommended by the MBS and of the actual dispatched priority of the call. The analyses presented here were performed using R 3.5.1, and the source code associated with this document is available on github.

2) Descriptive analysis

Data quality

Data quality problems in prehospital care are common - The situations encountered by EMD nurses are extremely varied, often chaotic, and commonly involve a strong time pressure. Perhaps understandably, missing PINs and incomplete documentation are not uncommon. In the table below, we present some measures of "Missingness" in our data. We have here excluded planned transports, and consider only primary missions. We only began tracking detailed disposition data in August, 2016, and as such are unfortunately missing this valuable information for calls prior to this time, denoted here as *No data*:

	No data	Ambulance need	No Ambulance care need	Non-medical
Total calls (%)	22035 (27.8)	37137 (46.8)	9479 (12)	10634 (13.4)
PIN captured at dispatch (%)	12341 (56)	30335 (81.7)	6988 (73.7)	928 (8.7)
PIN captured ever (%)	14149 (64.2)	35550 (95.7)	7036 (74.2)	937 (8.8)
MBS Used (%)	14192 (64.4)	34527 (93)	6266 (66.1)	1890 (17.8)
Ambulance journals (%)	10254 (46.5)	34718 (93.5)	78 (0.8)	44 (0.4)
PINs with hospital record (%)	9993 (70.6)	29183 (82.1)	4508 (64.1)	372 (39.7)
Transported with hospital record (%)	7561 (84.1)	27240 (90.6)	27 (69.2)	5 (31.2)

We see that missing PINS are highly concentrated among the Non-medical calls. The rate of missingness among patients with a medical complaint but not in need of ambulance care is more satisfactory, though around a quarter of these calls have no associated PIN, and call type documentation rates are still somewhat low! We also see that there are a handful of non-ambulance calls with associated ambulance journals - This typically occurs when dispatchers mistakenly document the ambulance crew's findings rather than their own determination as the call disposition.

In order to mitigate dispatcher bias in our models, we also capture cases where patients re-contact with the EMD center or other healthcare providers following the call. We chose a follow-up period of 72 hours for both EMD center re-contacts and healthcare system contacts. If during this time period a patient for whom a valid PIN is captured re-contacts the EMD center, we associate any ambulance interventions from the subsequent contact to the prior contact as well. For health system contacts, the first subsequent contact with the healthcare system within 72 hours, and any associated subsequent contacts (e.g., transfers between wards in a hospital and ED contacts within 6 hours of visiting a primary care facility), are associated to the call.

Since we only have this detailed disposition data from April 2016 onwards we will limit our descriptive analysis to calls from that time to the end of 2017, and exclude non-medical calls from our analysis. We can also further divide our referrals into two broad categories: Referrals to alternate transport, and referrals to alternate forms of definitive care. Upon implementing these modifications, the distribution of missingness among the dispatched priorities reveals some interesting patterns:

	1A	1B	2A	2B	Alt. txp	Alt. care
Total calls (%)	1655 (4)	13812 (33.1)	14644 (35.1)	3742 (9)	4541 (10.9)	3386 (8.1)
Valid PIN captured at dispatch (%)	768 (46.4)	10373 (75.1)	13158 (89.9)	3544 (94.7)	3921 (86.3)	2159 (63.8)
PIN captured ever (%)	1482 (89.5)	13038 (94.4)	14310 (97.7)	3661 (97.8)	3937 (86.7)	2182 (64.4)
MBS Used (%)	1428 (86.3)	12987 (94)	13844 (94.5)	3470 (92.7)	3448 (75.9)	1956 (57.8)
Ambulance Journals (%)	1496 (90.4)	12907 (93.4)	13933 (95.1)	3557 (95.1)	285 (6.3)	218 (6.4)
PINs with hospital record (%)	1127 (76)	10788 (82.7)	11779 (82.3)	2969 (81.1)	3250 (82.6)	726 (33.3)
Transported to ED with hospital record (%)	961 (90.7)	9108 (94.5)	9747 (93.6)	2407 (91.2)	47 (97.9)	2 (100)

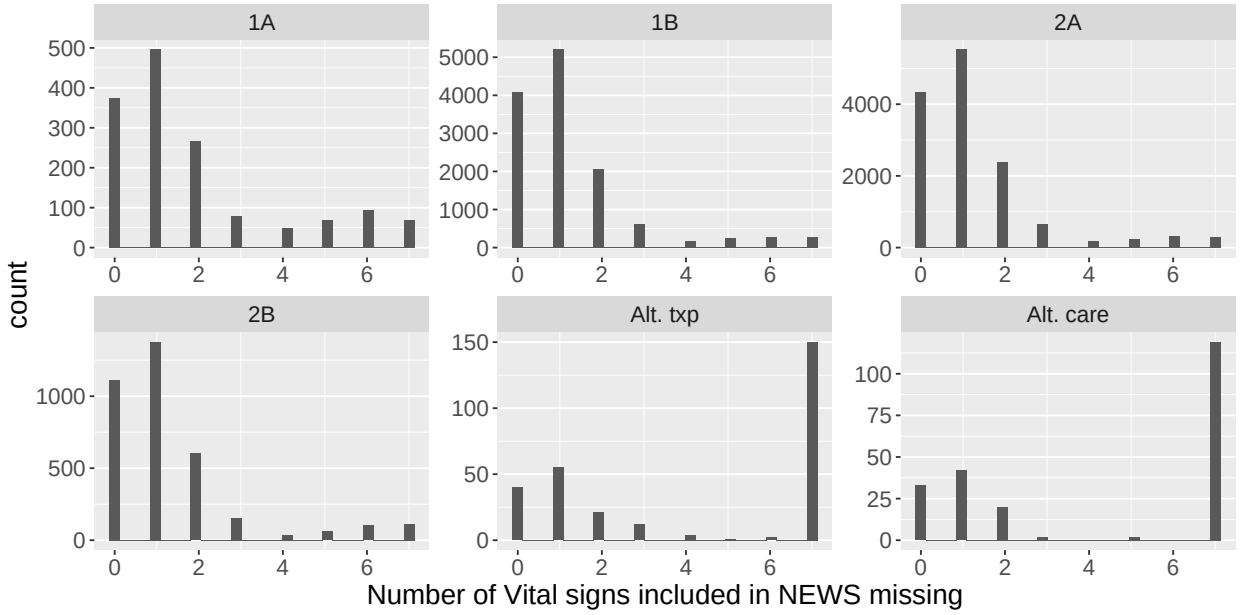
We see two effects in terms of documentation completeness: Higher acuity calls appear to have lower rates of documentation completeness, as do referred calls, and especially calls referred to alternate care. Upon including re-contacts within 72 hours, we see that around 6% of referred calls among both referrals to alternate transport and alternate care receive an ambulance within this time frame. We see that 75-80% of calls with a captured PIN have an associated health system contact within 72 hours, while patients transported by ambulance to one of the two regional emergency departments (Uppsala University hospital or Enköping hospital) with a valid PIN have a match rate of 90-95%. The extent to which this is problematic in terms of predictive modelling is largely determined by the extent to which this loss is randomly distributed across the population.

Beyond simply the ability to link records, we must also consider the completeness of the documentation from our dispatch center, our ambulance journals, and our hospital data. For ambulance data, the correctness of these data can only be checked via manual quality assurance of our data - A process which is ongoing, but has thus far shown satisfactory results, with both ambulance interventions and hospital outcomes ranging from 88% to 100% accuracy compared to a human reviewer (n = 100 so far), with one outlier at 70% due to hospitals not including flat x-rays as a radiological intervention. For our MBS questions and vital signs however, we can investigate data missingness directly by examining the percentage of ABCDE MBS Questions and vital signs that have documented values:

Call Type	Number of calls	Initial missing (%)	ABCDE missing (%)	ABCDE with engagement missing (%)
All call types	46188	13.5	28.4	11.0
Trauma	5172	6.0	27.1	10.0
Chest pain	4401	5.7	NA	NA
Difficulty Breathing	3851	7.3	36.7	23.9
Stroke	3434	4.4	55.4	17.3
MBS Unused	3329	51.2	NA	NA
Abdominal/flank pain	3200	6.3	31.1	4.5
General Elderly	2313	6.0	30.1	16.3
Dizziness	2223	3.7	60.8	10.0
General Adult	1712	6.8	32.1	15.7
Reduced Consciousness	1642	11.1	NA	NA
Intoxication	1209	11.9	36.8	16.1
Convulsions	1101	18.3	46.0	34.3

We see that among callers with a documented PIN, answers for the three initial questions are missing in 14% of calls (though the values are lower among specific call types), while answers are missing for the questions included in the MBS in 28% of cases. There are a number of causes of this missingness. There is a tendency for some nurses to simply use the MBS as a guide, without documenting negative findings which do not affect the dispatched priority. The MBS also includes functionality to automatically open additional call types depending on the answers given in the primary call type (e.g., if difficulty breathing is documented for an allergic reaction call, the "difficulty breathing" call type will automatically be added), and these "secondary" call types are not typically well documented. We see that among calls where the dispatcher engaged with the MBS by filling in at least one question, the documentation rate is higher, at about 90%. Note that some call types lack MBS questions (chest pain for instance), and immediately result in a priority 1 ambulance.

Among calls with an ambulance record, missingness of vital signs is distributed as such:



We see a similar pattern regarding documentation, with the highest and lowest priority calls having more calls

with missing vital signs, and calls referred to alternate transport/care having a large number of ambulance journals with no documented vital signs. As with questions, we can also consider how the documentation of various vitals varies across call types with more than 1000 associated calls:

Call type	n	Avg miss	> 2 miss (%)	AVPU miss (%)	SBP miss (%)	Breaths miss (%)	GCS miss (%)	Pulse miss (%)	SpO2 miss (%)	Temp miss (%)
All call types	39326	1.3	12.8	46.3	11.8	11.9	23.9	8.8	8.0	23.4
Trauma	4687	1.5	16.4	45.4	12.2	13.9	22.7	10.1	9.2	37.7
Chest pain	4143	1.0	6.2	46.0	3.1	6.5	22.7	3.1	2.9	16.2
Difficulty Breathing	3500	1.2	10.0	44.8	13.7	8.3	24.8	7.1	5.8	13.5
Stroke	3157	1.0	6.3	46.3	4.9	8.2	19.4	4.9	4.1	11.1
Abdominal/flank pain	2675	1.1	7.7	45.9	6.4	9.7	26.2	5.9	5.2	10.8
MBS Unused	2058	1.9	22.5	50.0	19.5	19.5	28.4	15.4	14.5	39.8
General Elderly	1867	1.2	9.2	45.9	7.2	9.7	27.4	7.0	6.1	11.8
Dizziness	1866	1.1	7.9	45.9	5.9	8.9	24.9	6.3	5.6	11.8
Reduced Consciousness	1649	1.2	11.8	44.6	14.4	11.7	17.7	8.7	7.9	19.7
General Adult	1270	1.4	12.4	47.1	10.2	12.1	28.3	9.8	9.1	19.0
Convulsions	1068	1.4	14.9	46.8	22.8	12.6	19.9	8.7	7.9	18.6
Intoxication	1065	1.7	19.6	46.8	16.3	16.2	21.3	14.6	12.7	40.3

We see that the missingness of vital signs varies across call types. Calls where the MBS was unused have the highest rates of missingness - It seems likely that this is due to a higher prevalence in this group of cases in which the caller is found upon the arrival of an ambulance at the scene to not have a medical complaint. Note also the high rate of missingness of AVPU determinations due to many of our clinicians preferring to document patient alertness in terms of the Glasgow Coma Scale. We also see that some specific vitals are missing from certain call types, e.g., high rates of missing temperatures from trauma calls.

The distribution of these missing vitals are important to note as predictive models based on these data are implemented - Among Trauma and Intoxication call types for instance, the high rate of missingness could entail a bias in our estimates for this measure!

Outcomes

Each of the outcome families necessitates analysis in the context of a distinct patient cohort with specific inclusion and exclusion criteria. We defined patient cohorts suitable for use with our 3 measure families as follows:

Ambulance interventions - In terms of predicting ambulance interventions, two cohorts are conceivable: Only calls receiving an ambulance, and calls with either a valid documented PIN or an ambulance journal. While descriptive statistics are most useful when described in terms of only ambulance calls, excluding patients referred to other forms of care would limit the validity of the predictive models we develop among the low-acuity patients we are most interested in assessing. While this approach risks inducing bias due to the dispatchers' decisions directly affecting outcomes, we ameliorate this by incorporating data from patient re-contacts.

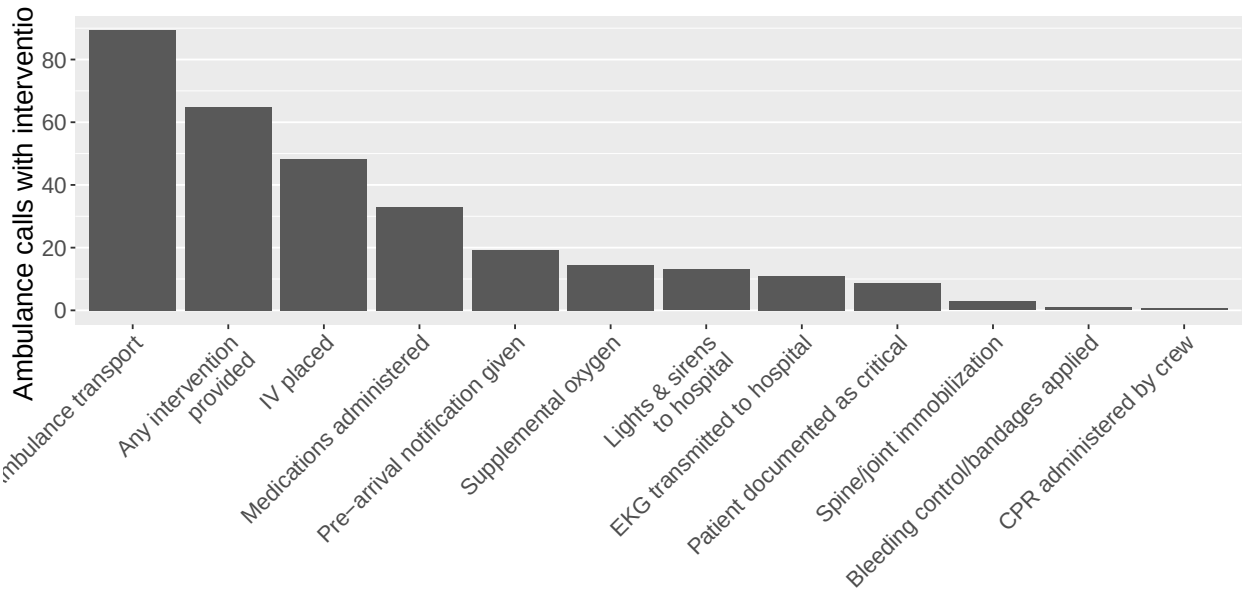
Vital signs - In this cohort, we include patients with an ambulance journal (including re-contacts) missing not more than 2 of the vital signs included in the NEWS score. Missing journals are likely to be *not missing at random*, and we do not believe that a NEWS score can be reliably imputed for patients which an ambulance does not assess. Records missing 2 or fewer vital signs are likely to be *missing at random*, and were imputed using predictive mean matching as implemented in the mice R package. NEWS is also an instrument validated in adults (there is a separate Pediatric version of NEWS), and as such, we exclude patients under 18 from this measure.

Hospital outcomes - For these outcomes, we include all patients with a documented PIN for whom we can perform a search in the regional EMR system. We exclude calls documented as transports but for whom no hospital record exists (likely false negatives due to non-matched records), and patients calling from municipalities nearby hospitals in neighboring regions not included in the regional EMR system.

In addition to these cohorts, we investigate one additional group: Patients with a NEWS score who are transported to the ED and have a matching hospital record. Within this cohort, we can develop valid predictive models based on not only data from the EMD center, but also from the Ambulance intervention and Vital Signs measures themselves in order to predict in-hospital outcomes. These models allow us to investigate how much additional performance might be gained by including ambulance data in our predictions, and potentially act as a useful tool in and of itself, for instance in care planning at the Emergency Department.

We begin by examining the specific measures included in each outcome family and describing their qualitative content and overall distribution:

Ambulance Interventions



We see that the bulk of patients receiving an ambulance are transported (90%). The majority of our patients (65%) also receive one of the interventions enumerated in the above graph. The most common intervention included in this set of measures is the insertion of an IV catheter (48%). Interestingly, this is more common than the documentation of any prehospital medication (34%). The medication list excludes oxygen which is documented separately, and is provided to 15% of ambulance patients. In 18% of cases, the crew provides a pre-arrival notification to the receiving facility.

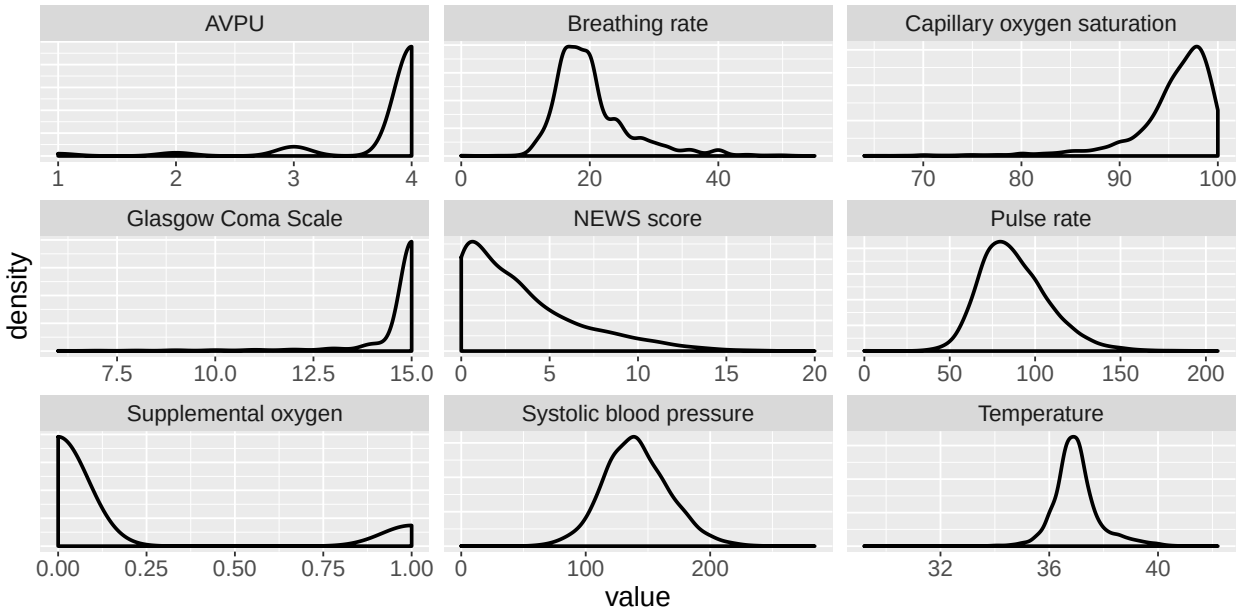
13% of our patients are transported emergently using lights and sirens, based on the documentation of a priority 1 trip to the hospital in the dispatch or ambulance records (roughly 3% of cases report differing priorities). In our experimentation thus far, we've found that the transmission of an EKG to the CICU is a better predictor of patient outcomes than the simple capture of a pre-hospital EKG, and the prior occurs in 11% of calls. We also included a marker based on the crew's subjective documentation of a "critical" patient (There is no explicit instruction provided to crews as to when a patient may be considered critical), which status was documented in 7% of calls.

Spinal and/or long bone immobilization is documented as occurring in 3% of calls, and we find documentation noting bleeding control of bandaging of wounds in 1% of calls. We suspect that that bleeding control is

under-documented, and are investigating additional markers to capture interventions performed on calls involving minor trauma. Finally, CPR is documented as being performed by an ambulance crew in 0.6% of calls.

Note that all these measures are aggregated such that on calls with multiple units responding, a patient is considered to have received the treatment if any ambulance journal contains documentation of the intervention.

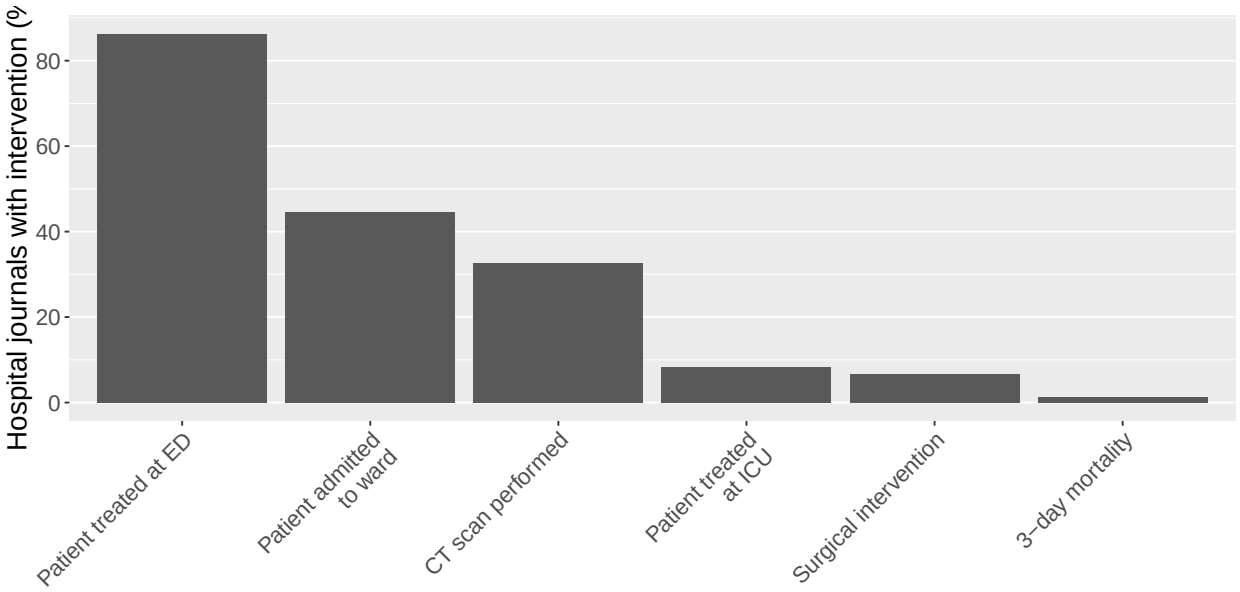
Vital Signs



We find above density plots depicting the distribution of the call NEWS values (middle), and its constituent vital signs. These values represent the first set of vitals taken by the ambulance crews. The NEWS scale includes an item for the provision of supplemental oxygen administration, which we interpreted as corresponding to the oxygen administration intervention above. Note that the AVPU scale is here coded with Alert as a 4, Verbal as a 3, and so forth. Some of our ambulance nurses prefer documenting a Glasgow Coma Scale, and where an AVPU value was missing, we considered a GCS of 15 as scoring a 0 on the NEWS scale, and a GCS of 13 or below as scoring a 2. (There seems to be some controversy, both in our data and the literature, as to whether to consider GCS 14 as Alert or Verbal). NEWS scores are typically analyzed in the literature using cutoff values of 4 and 7 to denote medium risk and high risk patients respectively. In our population, Patients with a NEWS score of > 4 constitute 29% of the population, while 13% of patients have a NEWS score of over 7.

Note that outlier values over 6 standard deviations from the median value have been omitted for visualization purposes, excluding dead patients with zero valued vital signs.

Hospital outcomes

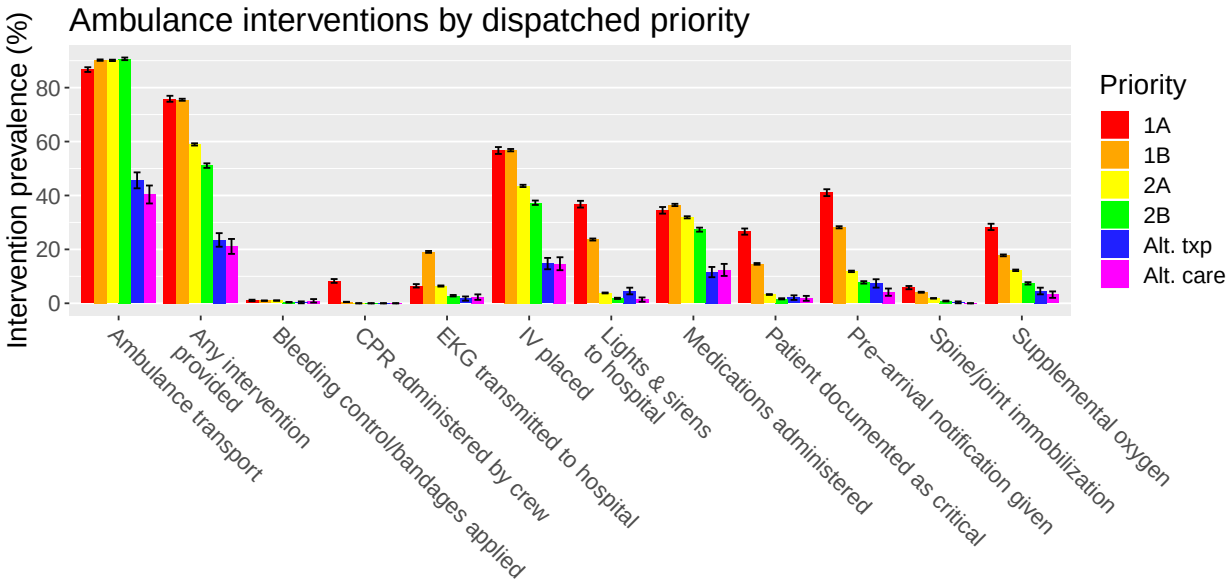


Among our 31 829 recorded health system contacts, 86% have a record indicating a visit at an ED, while 44% were admitted to an in-patient ward, and 9% were admitted to an intensive or intermediate care unit. Radiological examinations were performed in 32% of calls, including CT scans and coronary angiograms. The surgical interventions documented in 7% of contacts include a broad variety of procedures, the most common of which are the closed repositioning of femur fractures (503), the insertion of venous (316) and arterial (249) catheters (typically in concert with other more invasive procedures), and osteosynthesis of femur fractures (240). Finally, 1.3% of patients with a subsequent healthcare system contact die within 3 days.

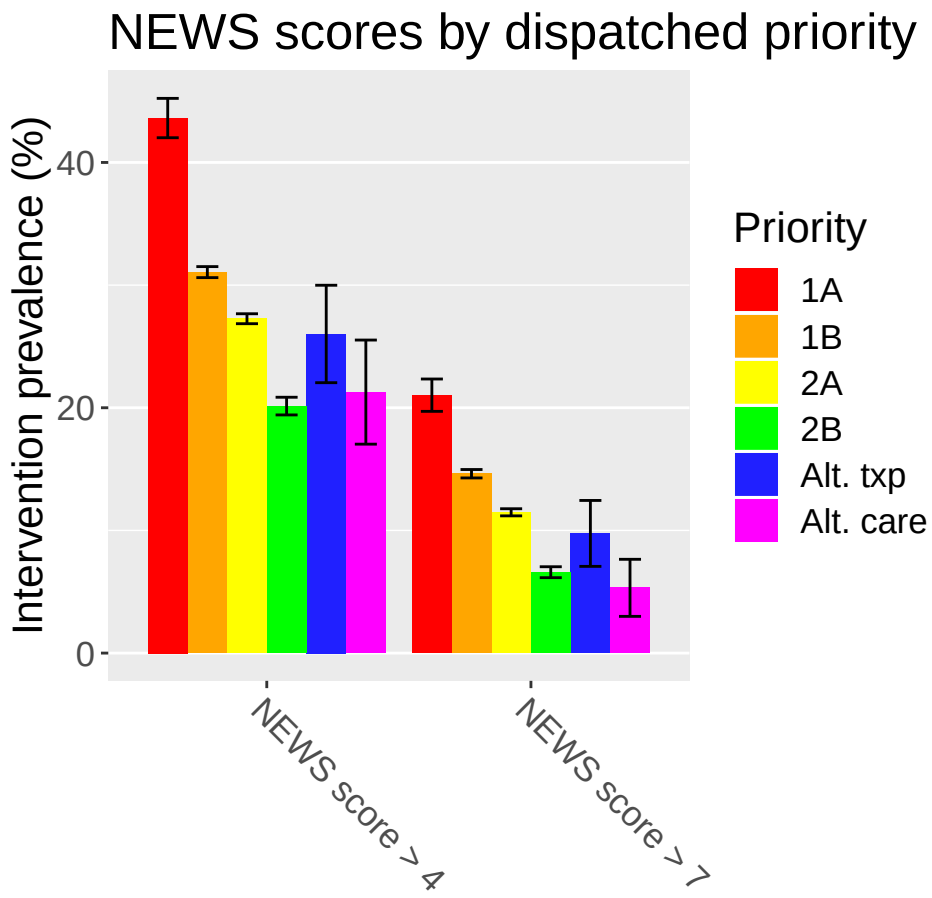
Unfortunately, a large portion of ED interventions are documented in free text at the hospitals, making automatic extraction and use in this project difficult. As such, while these measures correspond well with common markers of patient acuity found in the literature, they can only represent indicators of patient outcomes - Certainly not a comprehensive picture of the services provided by the secondary and tertiary care system.

Prioritization

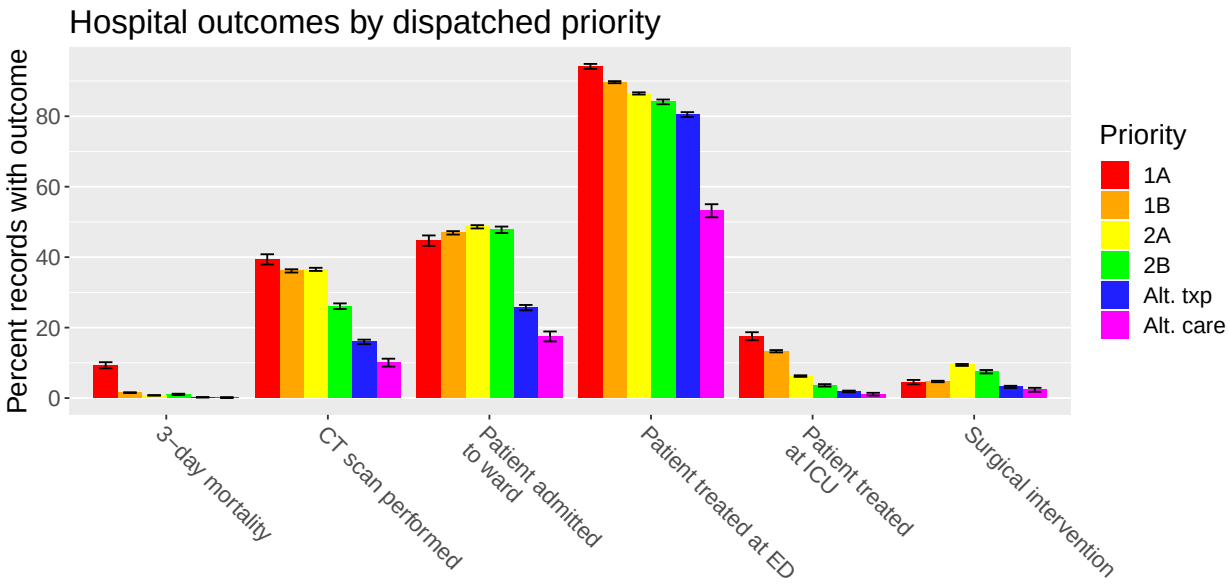
With a set of outcome measures to gauge our decisions by, a good starting point is to investigate the distribution of these markers among our calls as they are currently dispatched, i.e., do we see that these interventions and outcomes are differentially distributed among our dispatch priorities? For this investigation, we will examine patient cohorts consisting of only those calls for which a linked ambulance or hospital journal exists. In our predictive models, we include patients for whom a valid PIN was captured, and for whom we can be reasonably sure that a subsequent contact has not occurred. For the purposes of descriptive analysis however, assessing prioritizations with regards to only subsequent contacts provides a fairer comparison with regards to the true underlying care needs found in each cohort.



Taking a broad view of these outcomes, we see some interesting effects - While we by and large see decreasing prevalences of these outcomes along the prioritization scale, for some measures, Priority 1A calls have lower intervention rates than their lower priority counterparts! This could be an artifact of the lower documentation rates among high-acuity calls noted previously. Another factor may be that many Priority 1A calls are in regards to cardiac arrests, an unfortunate percentage of which are found by ambulance crews to be cases of obvious death. Interestingly, we see almost no differentiation between priorities resulting in an ambulance response with regards to transportation rates, though less than half of patients who are initially referred and subsequently receive an ambulance are transported. We also see that for many markers, Referred patients who later receive an ambulance seem to have lower scores than those receiving an immediate ambulance response.



As one might expect, we see that there seems to be a "dose response" with regards to NEWS scores among our dispatched priorities, though generally it appears that differentiation in terms of NEWS scores appears greater among high priority calls than among low priority calls. Note that the standard errors for referred calls are quite wide - as noted in the data quality section, it seems that vitals are often not documented for these patients!



Among patients with a healthcare system contact within 72 hours, dispatch priority appears to differentiate for outcomes indicative of an emergent condition such as contact with an ED and ICU care. However, the priority with which an ambulance is dispatched to the scene does not appear to be indicative of other hospital outcomes such as admission to in-patient care or surgical/radiological interventions. Indeed, the need for such longer-term hospital interventions are often not what is being considered when determining ambulance response priority! It is encouraging that for each of these measures, referred patients have the lowest intervention rates, but even a small number of patients for instance receiving ICU care following a referral to alternate care is problematic and must be addressed.

Predictors

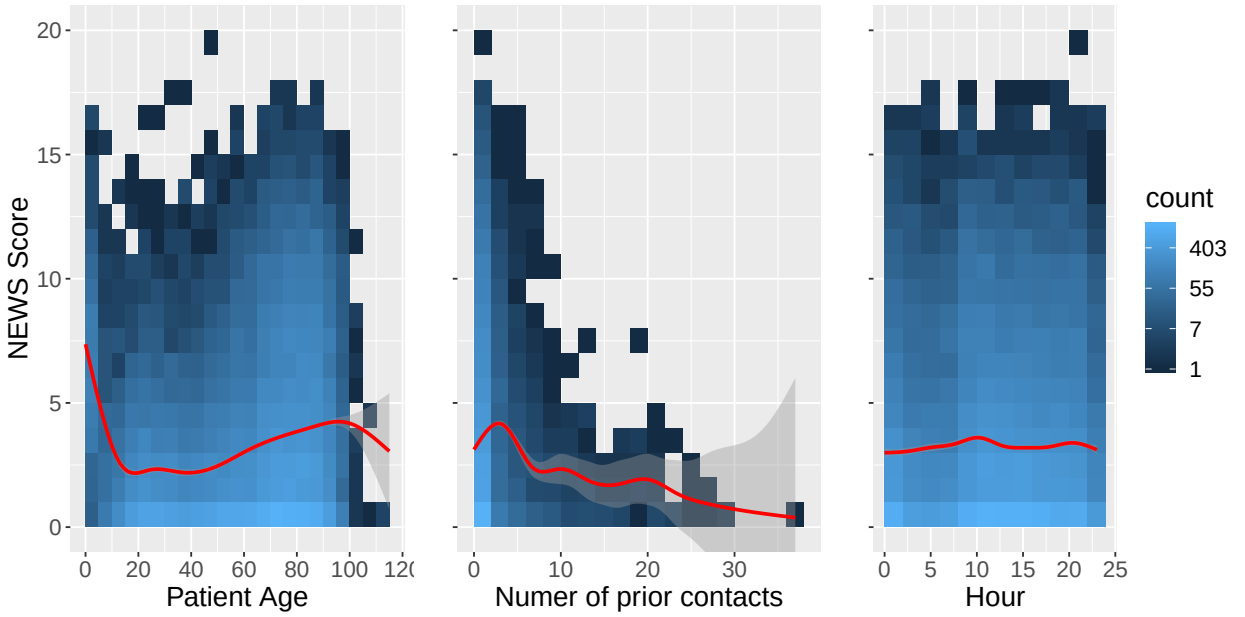
With a grasp of how these outcomes are distributed among our calls, we turn to the data available to predict these outcomes from. Conceptually, we've divided these into 4 groups:

Patient / call Characteristics

This group of predictors includes the operational characteristics of the call and demographic characteristics of the patient including:

Patient age
Patient gender
Number of prior contacts with the EMD center within 3 months Time since last contact (within 3 months)
Time of call

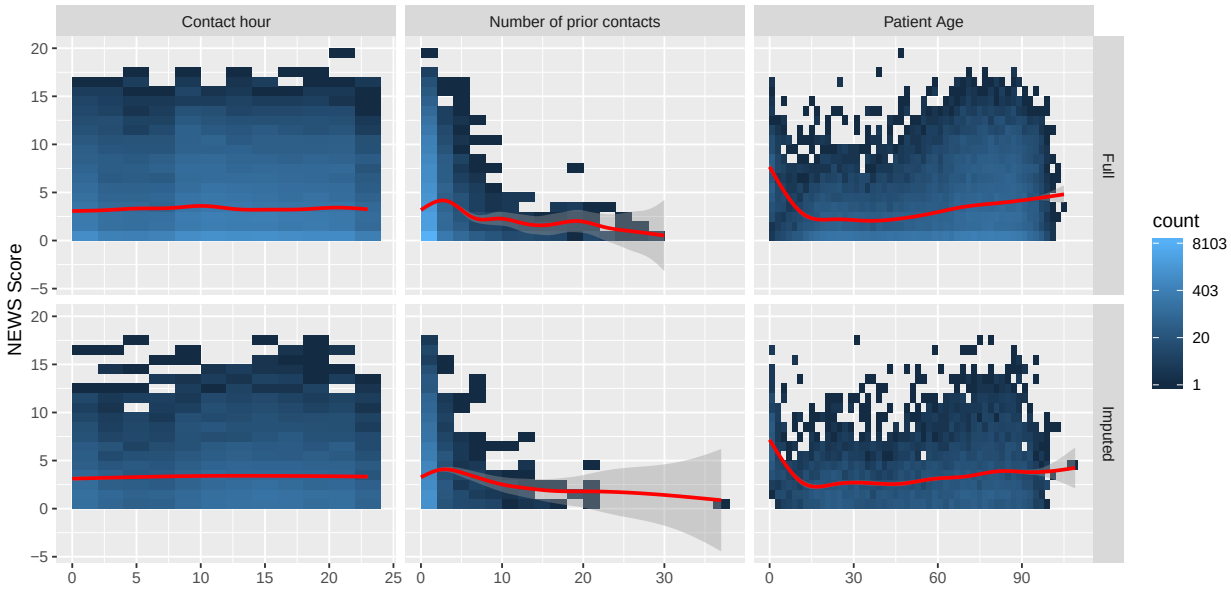
Note that a 3 month time frame for prior contacts was selected primarily due to the fact that our EMD production database retains records for this amount of time. We can begin by exploring these predictors; here we present 2d bin plots for that patient's age, number of prior contacts, and contact hour:



Excluding patients under 18 from the NEWS score cohorts does indeed seem to be a good idea. Interestingly, there appears to be quite a sharp delineation around 60 years of age where NEWS scores appear to jump. We also see that there are only a few outliers above around age 105 - Given their low news scores, it seems likely that these are documentation errors. As such, we'll set age observations over 109 to missing. Given that the methods we will use to perform our modelling do not work well for extrapolation and that we have only a few observations of the very old, it seems wise to set criteria for using these models to exclude patients over 100.

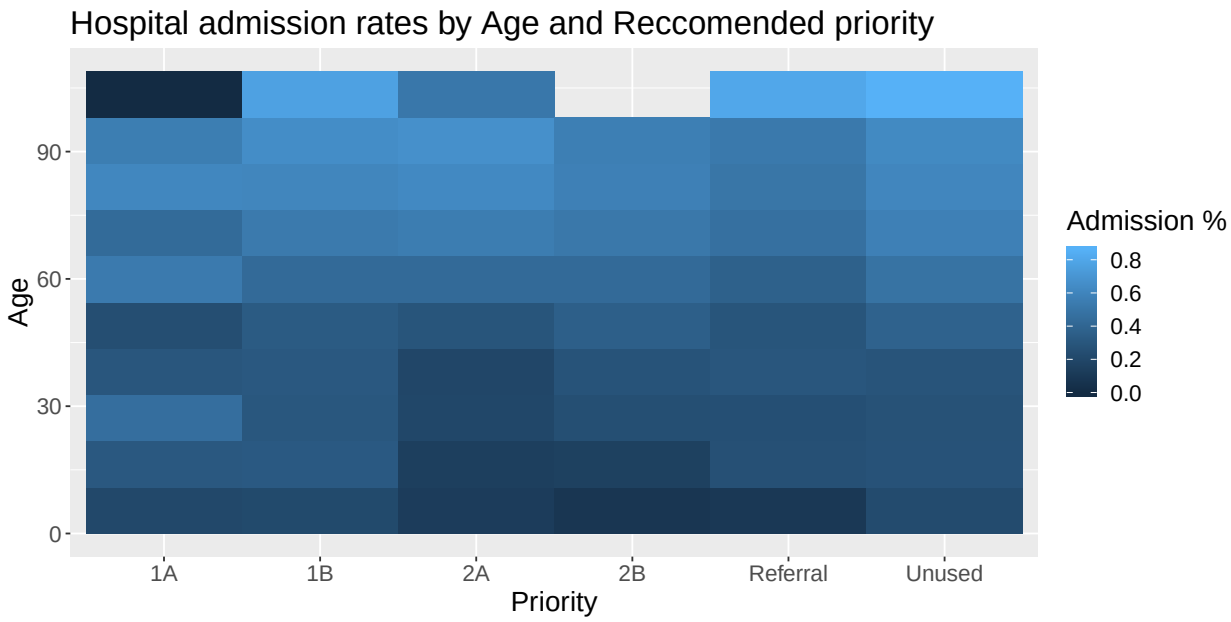
The characteristics and impact of frequent utilizers of EMS services has been widely studied in the literature, including in a study performed previously at our agency (Spangler, 2017). In line with our previous findings, we see that the relationship of contact frequency with patient acuity is non-linear and non-monotonic. That is to say, acuity appears to increase initially and peak among patients with a few prior contacts, and then return to-, or perhaps slightly below the acuity levels associated with patients with no prior contacts. With regards to contact time, one can perhaps see an increase in average NEWS scores along with an increase of the overall call volume around 10:00, and a second, weaker peak at around 21:00.

We can use the somewhat idiosyncratic distributions to verify that our multiply imputed NEWS values are distributed appropriately:



Looks pretty good! We see that many of the features of the original NEWS scores with regards to age are preserved.

It's indeed quite surprising that age in particular is not included more often in decision support algorithms. Age can serve in and of itself to substantially improve the differentiation of the current rule based decision system - Dividing each priority into age groups and shading each cell based on the percentage of patients admitted to in-patient care results in strong gradients for all but the highest-priority calls:



Call type

We've found that the call type selected by the dispatcher is a powerful predictor of patient outcomes in and of itself, often accounting for the lion's share of the predictive value in the preliminary models presented in

section 3. We present below a summary of some of the key attributes discussed previously as grouped by call types with more than 1000 observations for all patients with a PIN captured:

calltype	n	Median age	Mean prior contacts	Female (%)	Amb. intervention (%)	Median NEWS score	Admitted (%)	ICU treatment (%)
Trauma	5172	71	0.29	51.4	63.1	2	40.9	3.2
Chest pain	4401	70	0.64	49.1	78.9	2	39.7	15.9
Difficulty Breathing	3851	73	0.68	54.4	75.1	6	58.0	9.6
Stroke	3434	74	0.42	51.3	62.8	2	58.9	8.6
MBS Unused	3329	61	0.87	49.8	59.0	3	44.4	10.4
Abdominal/flank pain	3200	53	0.86	56.9	68.9	2	36.3	2.8
General Elderly	2313	82	0.58	50.2	59.2	3	64.6	7.6
Dizziness	2223	73	0.49	53.0	56.3	2	47.8	6.3
General Adult	1712	60	0.79	49.1	51.8	2	43.0	7.3
Reduced Consciousness	1642	68	0.40	48.8	70.4	3	50.3	12.5
Intoxication	1209	30	0.67	52.8	58.4	2	48.7	27.3
Convulsions	1101	31	0.88	45.1	68.4	3	36.9	7.7

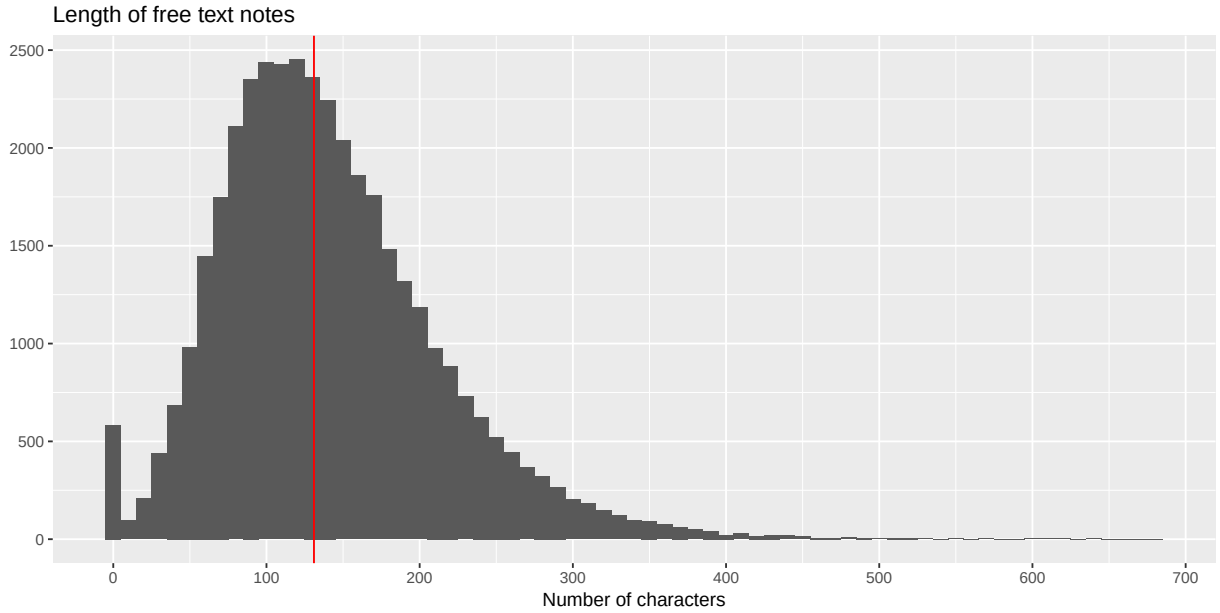
Most values seem to be what one might expect. Interesting that The "Intoxication" call type has such a high rate of intensive care - This could be due to the inclusion of intermediate care units in this outcome.

MBS data

In addition to the initial categorization, dispatch nurses are to complete a battery of questions, the answers to which can result in an upgraded dispatch priority. As noted previously, only the relatively well-documented *ABCDE* questions have been found to be of use for the purposes of predictive modelling. As the structured triage system and the questions it contains are protected by copyright, we unfortunately cannot provide a detailed breakdown of the distribution of answers to these questions. It can be noted however that these questions are typically designed to rule-out relatively rare acute conditions, and are generally unbalanced, with many more negative answers than positive answers.

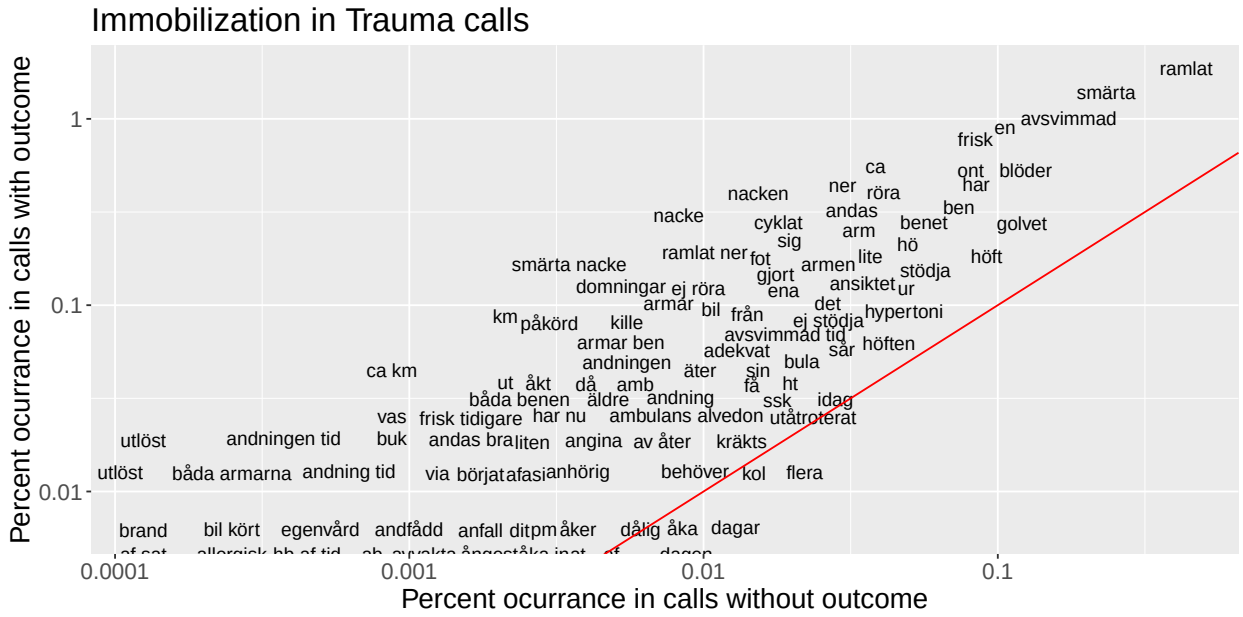
Free-text

In addition to the structured MBS, dispatchers are able to document free text notes in regards to the patients condition which is transmitted to ambulance crews. There is thus some potential to use this unstructured data for the purposes of predictive modelling. We can begin by investigating the number of characters documented in the notes associated with each call:



We see that we have a distribution of lengths with a median of 129 - Essentially a Twitter post for each call! Given this relatively short length, a "Bag-of-words" model is a good choice for analyzing these data. This model essentially treats each term in the text as an independent variable, discarding information regarding the relationship between words, and the position of the term in the text. To capture some of this information, the model can be expanded to include n-grams, i.e., terms consisting of multiple adjacent words. Based on an examination of our data to identify appropriate term lengths and cutoffs, we chose to include single words with 200 or more occurrences, two-word combinations with 100 or more occurrences, and three-word combinations with 20 or more occurrences. These cutoffs were set so as to roughly include common clinical presentations in the ED, but to exclude idiosyncratic terms, spelling errors, patient-specific information, etc. This resulted in a dataset containing 2279 distinct terms found across the 62854 calls with a documented free-text note. Note that while we have performed some basic pre-processing of these data, including the removal of non-text characters and common stop words (though retaining negatory terms in n-grams), we have not yet found a satisfactory method to perform stemming / lemmatization for Swedish medical terminology, which process may improve our ability to utilize these data.

A number of techniques may be used to explore these data to extract meaning from this quite complex dataset. These vary from the simple (word frequencies and term co-occurrences) to the sophisticated (k-means clustering and topic modelling). Our primary interest is to investigate the properties of these terms in relationship to specific binary outcomes. To perform this task visually, we developed an exploratory analysis tool based on the word-ratio approach described by Julia Silge in her excellent book Text Mining with R. In this approach, we generate a plot for each outcome of interest, whereby the x axis represents a term's prevalence among calls in which a given outcome does not occur, while the y axis represents the prevalence of the term among calls with the outcome. In this regime, terms falling falling above the red line at $x = y$ occur more frequently in "high acuity" calls as defined by a given measure, and vice versa. Finally, we plot these terms on a log-log scale. For example, we could examine our data to see which terms are associated with the documentation of immobilization in trauma calls:



As might be expected, words like “neck” “neck pain”, “bicycle”,“km” (suggestive of the involvement of a vehicle), etc., are associated with immobilization. Terms like “hip”, “floor”, “several”, “dementia”, are at the lower end of the word cloud, suggesting that falls in the home are less likely to receive immobilization. Interestingly, the majority of terms lay above the red line, denoting a higher prevalence among immobilized patients. Perhaps more documentation is produced with regards to these more “serious” trauma patients?

We implemented this methodology in an exploratory analysis tool for each of the outcomes reported in the previous sections within each of our call types with more than 1000 occurrences (for a total of 22 call types). Unfortunately, we have not translated the terms to English, but we encourage readers with some knowledge of Swedish to explore our data! An on-line version of the tool is hosted at shinyapps.io.

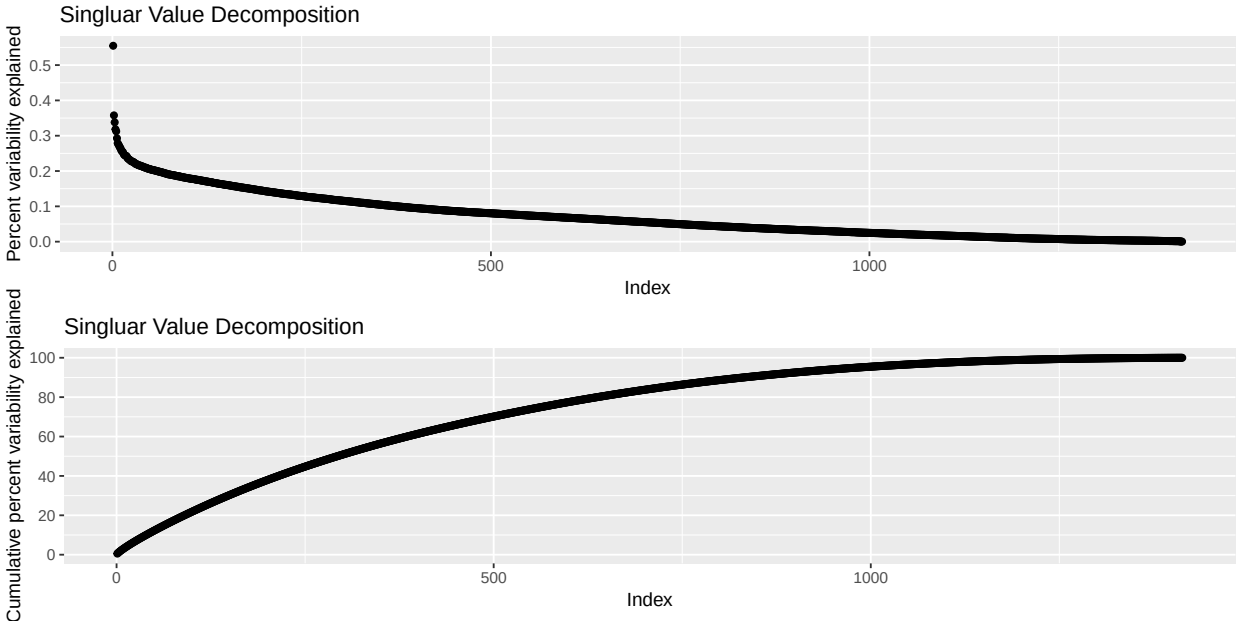
The source code (and the aggregated data the visualization is based on) is available on Github.

Dimension reduction

Given a dataset containing over 2200 distinct terms, dimensionality reduction techniques may be useful in identifying underlying commonalities between terms. Firstly, a reduced dataset results in lower training times and more responsive predictions when developing predictive models. Dimensionality reduction can also serve to remove “noise” from the dataset, and result in lower levels of over-fitting and can address co-linearity issues. Finally, the resulting reduced dataset representing underlying commonalities between terms could potentially improve model interpretability if the resulting vectors can be given meaning via human review.

Here, we implement Singular Value Decomposition for these data. This process involves performing a factorization resulting in a set of three matrices containing a reversible decomposition of the original data. This decomposition results in a set of singular values ordered by the percentage of variance they capture, some subset of which may be retained to capture a portion of the variance found in the original data while reducing the number of variables. A full explanation of Singular Value Decomposition is beyond the scope of this paper, but these are some resources we found useful in regards to the Intuition behind the method, the Theory, and its Implementation in R.

We first calculated the term frequency / inverse document frequency value for each term occurrence, and transformed the data to conform to produce bag of words model. We then applied SVD, and present below two visualizations of the ordered singular values of d, corresponding to the percentage of variability in the data explained by each value, and the equivalent cumulative sum och these percentages:



We see that while there is some gradient which we can exploit, SVD does not result in a small number of variables which can capture a large portion of the overall variability, with individual vectors capturing only fractions of a percent of the variability in the data. Nonetheless, there is some gradient which could be exploited - it seems that including the 600 or so most predictive singular values for each observation would be a good choice for "denoising" purposes and improving model training times. At this threshold, we retain just over 80% of the variability in the data, while cutting the number of variables we need to include in our models by over half.

3) Predictive modelling

Regression

With a solid grasp of the distributional characteristics of our data, we can begin the task of developing models to characterize them in a manner which can be implemented in tools to improve clinical practice. It is typically a good idea to begin with simple models to get a sense of what kind of predictive value we can expect to see in later attempts with more complex approaches.

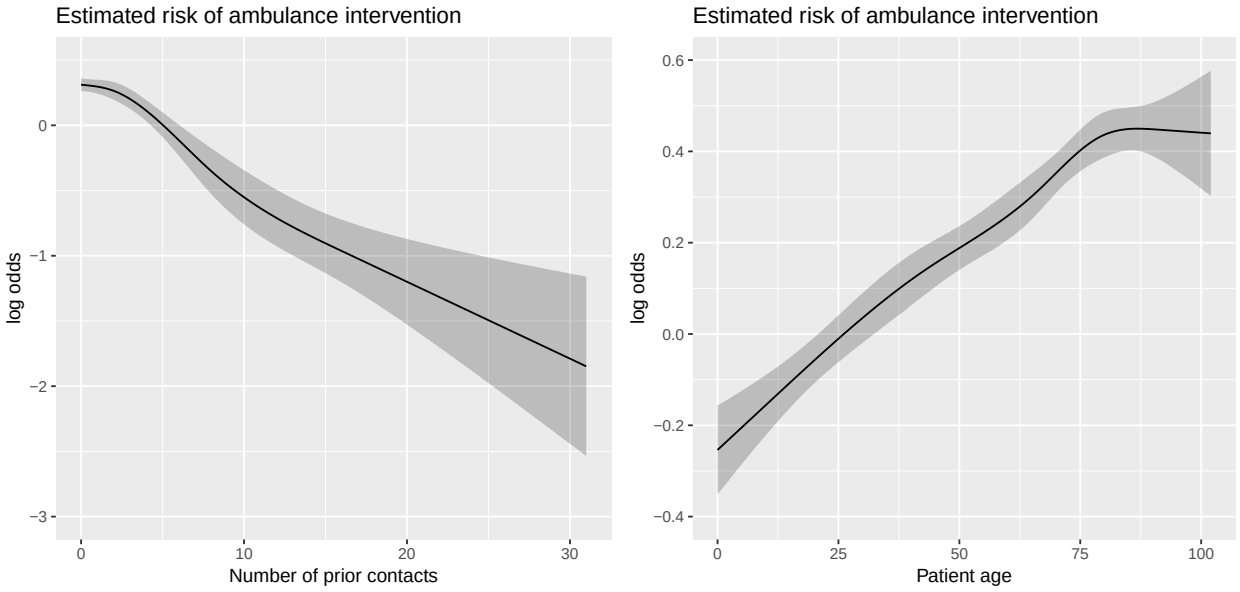
We can begin by examining the predictive value of only the patient and call characteristics which we capture. For the sake of simplicity, we will here limit ourselves to characterizing a single outcome - The documentation of any of the prehospital interventions noted above. As this outcome is binary, a reasonable choice of initial approach is logistic regression using a "sandwich" robust co-variance matrix estimate to account for clustering. We'll use the tools developed for this purpose in Frank Harrell's excellent Regression Modelling Strategies (rms) R package and described in the book of the same name. Note that while Harrell prefer "C-index" over AUROC, the terms are interchangeable.

As we saw that we have a number of predictors which are not likely to have linear effects, a simple linear model is not appropriate. A solid choice for handling non-linear predictors is the use of restricted cubic splines; This method essentially generates fixed point estimates at k intervals along the trimmed range of the variable, and fits a cubic polynomial between these points, while restricting estimates outside of this range to be linear. For contact hour, we generate sine and cosine transformations of this variable, and include the interaction between them in our model to account for the cyclical nature of the 24 hour period (See this page for some more of the thought behind this). Furthermore, in exploratory analysis, we found that age and

contact time appeared to interact - An important effect which makes sense based on theory and which should be considered in the analysis. It's also reasonable to believe that time will have different effects on weekdays and weekends. Adding these considerations to be model produces the following fit:

```
## Logistic Regression Model
##
##  lrm(formula = amb_any ~ rcs(null_pcontacts, c(0, 2, 8, 16)) +
##      null_female + rcs(null_age) * (null_time_sin * null_time_cos) +
##      null_wkday + null_wkday:(null_time_sin * null_time_cos) +
##      (null_month_sin * null_month_cos), data = ambmoddata, x = T,
##      y = T)
##
##                         Model Likelihood     Discrimination    Rank Discrim.
##                            Ratio Test           Indexes           Indexes
##  Obs          38103    LR chi2     773.71    R2       0.027    C       0.577
##   0           17539    d.f.            30    g        0.309    Dxy     0.153
##   1           20564    Pr(> chi2) <0.0001    gr       1.362    gamma   0.153
##  max |deriv| 4e-05                          gp       0.076    tau-a   0.076
##                                             Brier    0.243
```

We find that these patient characteristics account for 2.7% of the variation in the data (as indicated by the R2 value), while the AUROC (C-index) is 0.577 This model can also tell us something about the risks associated with these predictors:



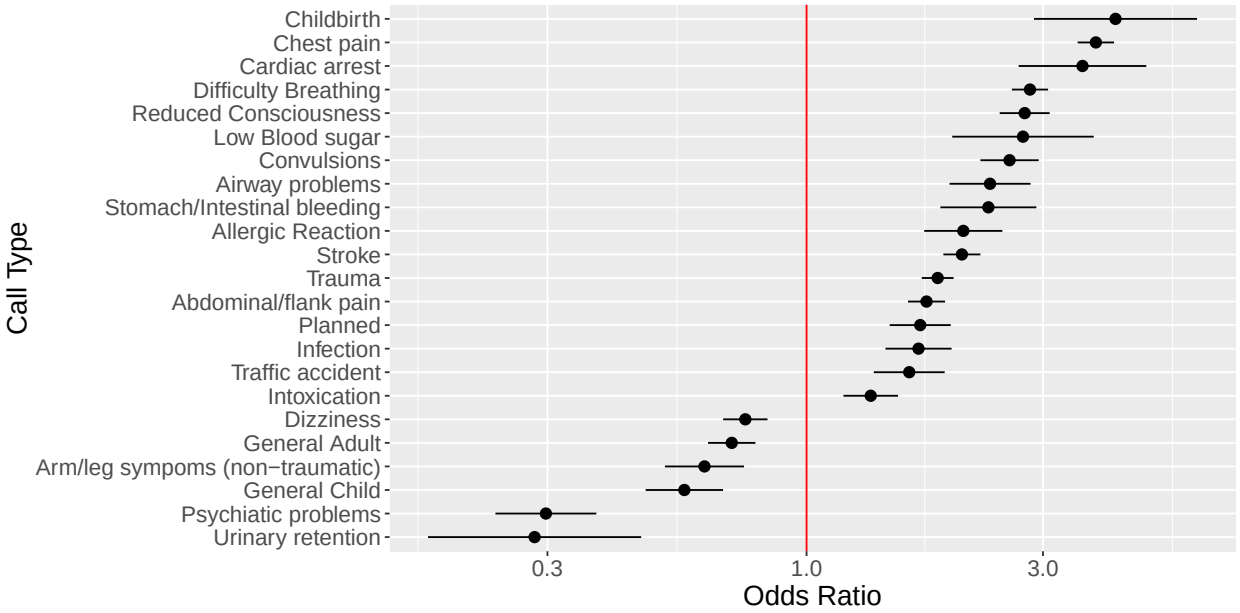
We see a similar effect with regards to ambulance interventions as we did with NEWS values, though we see no evidence of a non-monotonic effect for prior contact frequency as we did previously. Interestingly, intervention rates by age appears to plateau after age 80 or so.

We can also include call types coded as dummy variables in this type of model. Calls without a category serve here as the comparison group for the other categories, and we'll recode call types with fewer than 10 occurrences as not having used the MBS.

We find that we achieve an improved C-index of 0.679 and corresponding increases in measures of model fit. At this point, we should perhaps begin worrying about overfitting - Are we being too confident in our predictions given the large number of predictors? In terms of overall model predictions, we can investigate

this using 10-fold cross validation to estimate the out-of-sample performance, finding a C-index in the test folds of 0.675, suggesting that overfitting is not yet a huge issue.

Using the call type model, we can also investigate the ambulance intervention probability in each call type. Given the relatively large number of call types, we should be careful about interpreting individual coefficients however - To limit the number of estimates we present, we can apply a Bonferroni correction, then make it 100 times more extreme ($p < 0.000006$), and take a look at the odds ratios of interventions being performed in each call type, using cases where the MBS was not used as a comparison group:



We find that childbirth, chest pain and cardiac arrest calls have the highest intervention rates, with odds ratios of around 4. That childbirth should top this list is somewhat surprising. At the low end, we see that calls related to urine retention and psychiatric problems have odds ratios of around 0.3.

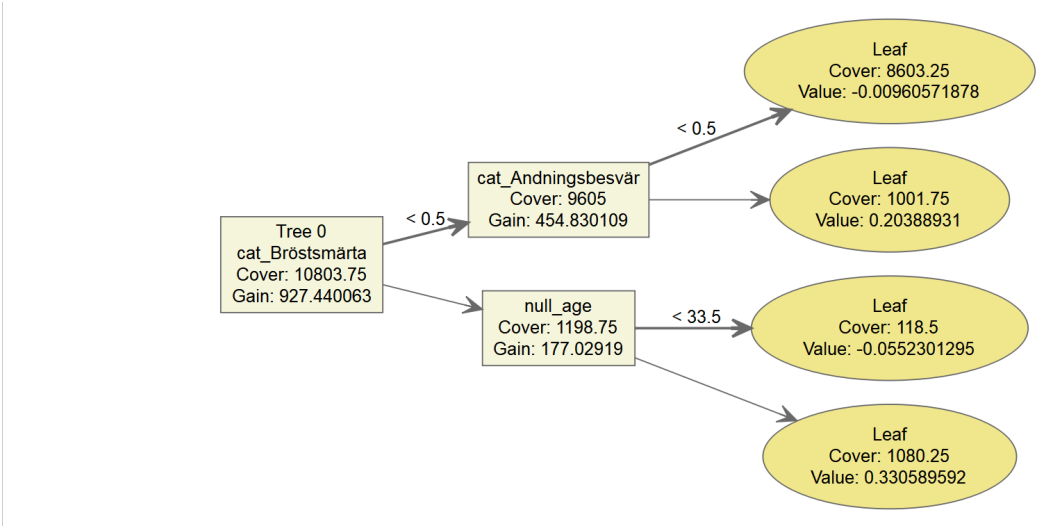
To improve this model, we need to begin considering how to account for the fact that patient characteristics are likely to have differential effects for different call types! Performing this task manually based on theory is a daunting task, and becomes increasingly difficult as additional predictors are proposed for the model. Modern variable selection techniques are available which could automate this task, and along with regularization, regression does offer a viable method to developing models with predictive values on par with machine learning based techniques. Nonetheless, at this point we will switch from the framework of regression analysis to the machine learning technique which we will apply in the remainder of this section.

Gradient Boosting

To account for interaction effects, co-linearity, and other issues which arise as we add increasingly complex forms of data to our set of predictors, we experimented with a number of techniques including random forest models, neural networks, and gradient boosting. While we will provide a comparison of the performance of these techniques in a future publication, it is not the goal of this analysis to determine an optimal modelling method for use in our data. Rather, we are interested in exploring the general and relative value of our predictors in terms of estimating the likelihood of our outcomes of interest. In this way, we can use predictive models as an exploratory tool to understand the relationships present in our data.

As such, we will limit this analysis to considering the technique we have found to work best: Extreme gradient boosting as implemented in the *xgboost* package. The method as we use it here involves the application of

decision trees which identify splits in the data which are most predictive for the outcome. For instance, a simple tree predicting ambulance interventions based only age and call type may look like this:

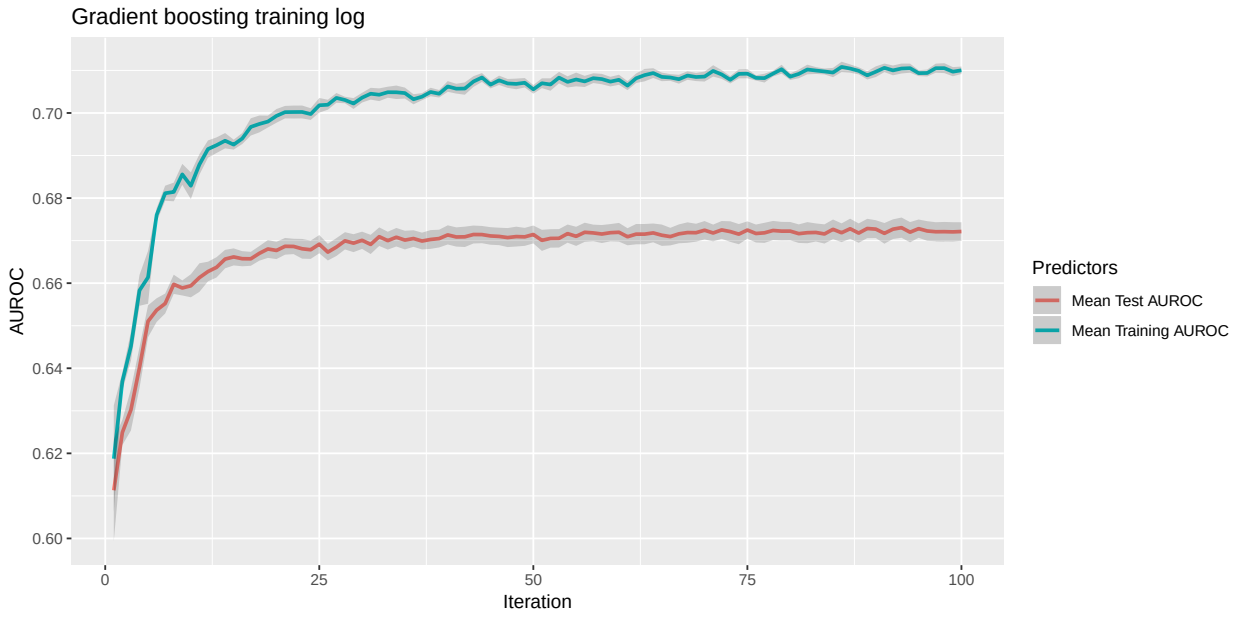


Following along this tree, we see that it first determines whether this is a chest pain call - If so, it moves downwards and checks the patients age. If the patient's age is 33 or less, the model assigns a value of -0.05, otherwise it assigns a value of 0.33. Otherwise, this tree checks if the call type is difficulty breathing, assigning a value of 0.2, or 0.01 elsewise. A number of similar trees are generated, and the values assigned by each tree are summed to result in a final predicted value. In our case of classification, and additional logistic transformation is applied to the summed predicted values, resulting in a probability between 0 and 1.

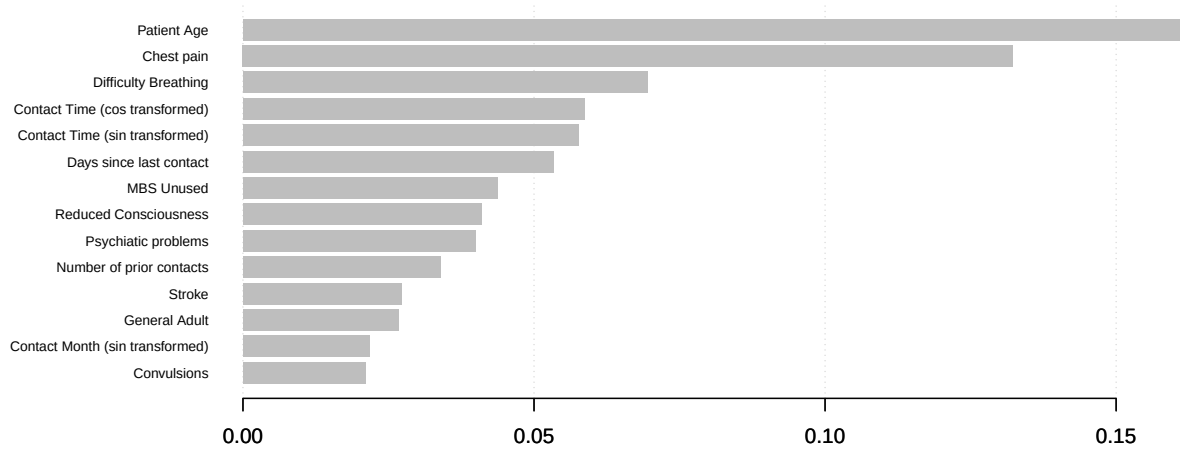
Gradient boosting involves training such trees in an iterative process whereby successive trees attempt to further subdivide groups of observations, focusing on cases where residual values are large (i.e., where the previous trees perform poorly). Gradient boosting models are trained through a number of successive iterations to improve the fit that these trees produce. The number of such generations a model is to be run, along with a number of other *hyperparameters* such as the maximum depth of trees (2 in the example above), the size of the "steps" taken in each iteration, etc. can be varied to optimize the performance of a specific model. The settings for these are typically determined by examining how changes in these values affect the performance of the model in test data - Data not shown to the model during the training procedure. As before, We'll use 10-fold cross-validation to assess this in our gradient boosting models.

We will for the purposes of this analysis however not be performing final hyperparameter tuning - Rather, we will report models based on the default settings of xgboost, with the exception of adding a drop-out procedure to control over-fitting in lieu of a more sophisticated approach. For illustrative purposes here, we will train each model for 100 iterations.

Let's begin by re-fitting our last model using xgboost. We generate a plot of how the predictive value of the model changes as additional iterations improve model fit, indicated by the mean AUROC found in the test and training folds of the data, and their standard errors:



We see that the gradient boosting model achieves a predictive value of 0.673, similar or slightly less than that produced by logistic regression. Note that these models tend to over-fit, and as such summary statistics regarding model performance in training data should generally be disregarded other than to note the presence of overfitting. While the nature of this model precludes the direct extraction of model coefficients, there are methods available to examine variable importance based on the information that is gained by the model in splits based on that variable:

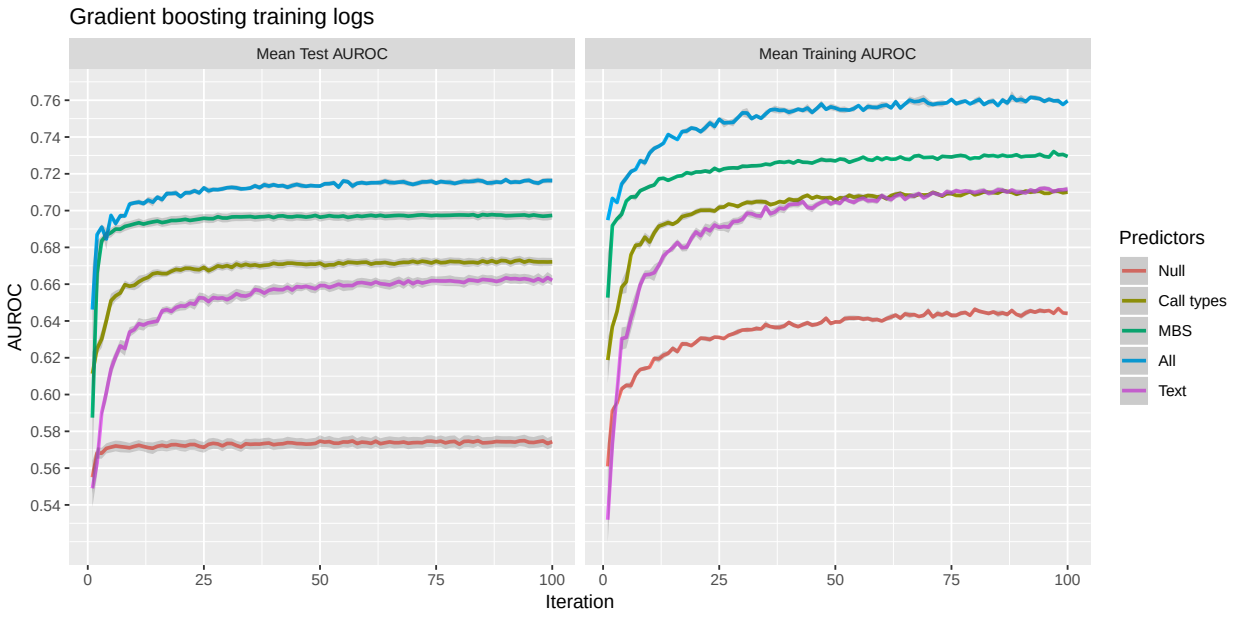


We can see that patient age was the most important predictor, followed by call types of chest pain and breathing, followed by the contact time. This does not however tell us the direction of the effect!

Adding data

Using gradient boosting, we can begin to explore the predictive value of our other datasets. We can plot similar training logs for models containing each of the sets of predictors noted in section 2. In addition to the call and patient characteristics (What we call our "Null" predictors) and call types, we also add the answers to the MBS questions, coded using -1 for a negative answer, 1 for a positive answer, and 0 for a missing answer.

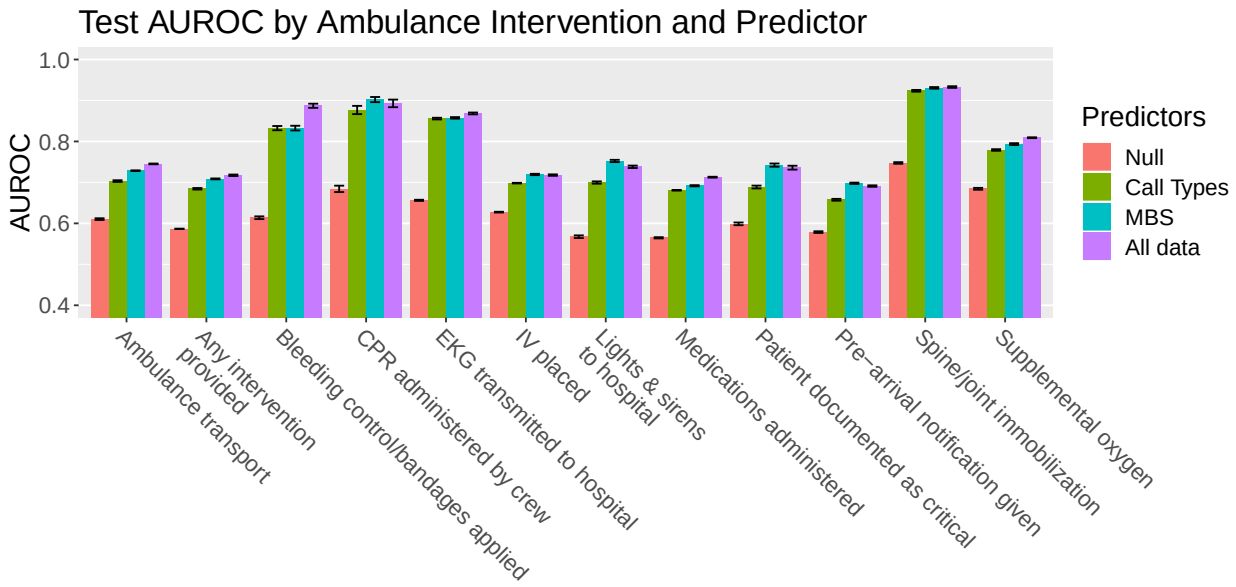
In analyzing text data, we've found that documentation practices vary from nurse to nurse at the EMD center, and taking consideration for which particular nurse is documenting the call appears to improve the predictive value of the free text data. We include this effect by generating dummy variables for each operator with more than 50 calls taken, and include it alongside the bag-of-words model described earlier. Finally, we include all of these predictors in a full model:



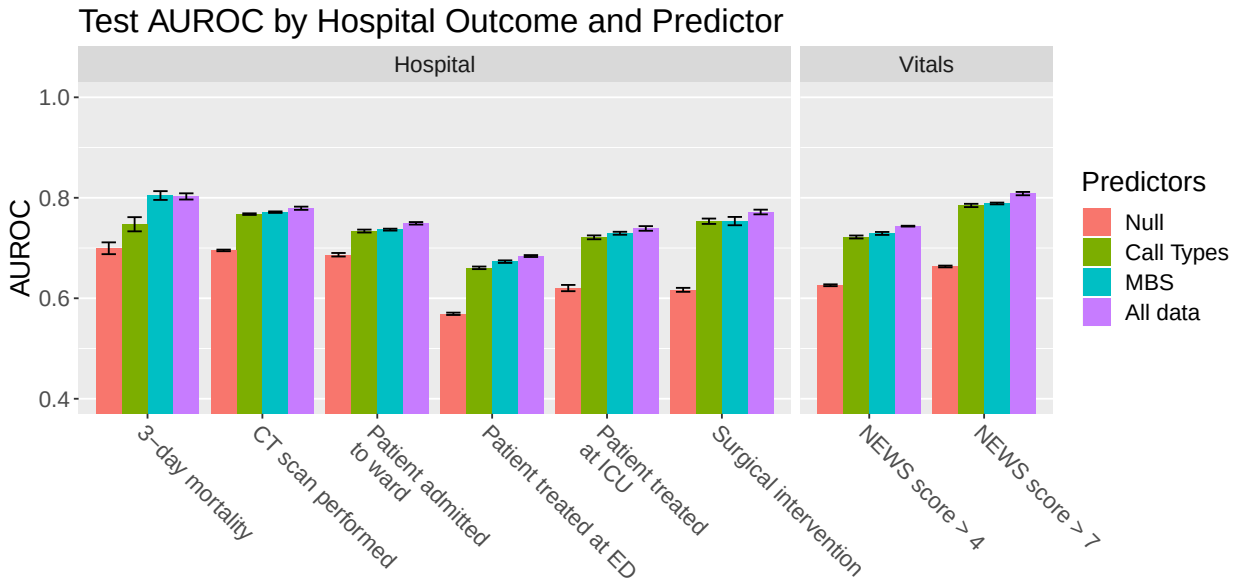
We see that each additional predictor adds value to the model. We see that for the simpler models, gradient boosting quickly maximizes it's ability to predict the outcome, with additional iterations only increasing the training AUROC. For others, additional training iterations could lead to small gains in cross-validated performance. The model containing all predictors achieved a maximum test AUROC of 0.717 - Frankly, not a fantastic showing.

Model summaries

With an understanding of how these models are generated, let's have a look at how these models perform for each of our outcomes of interest. To summarize our results so far, we can extract the cross-validated AUROC for each of the nested sets of predictors for each outcome. As noted previously, There are some issues with AUROC as a measure of predictive value, but we will use it nonetheless given it's prevalence in the literature. We can begin with ambulance interventions (error bars again indicate one standard error from the mean):



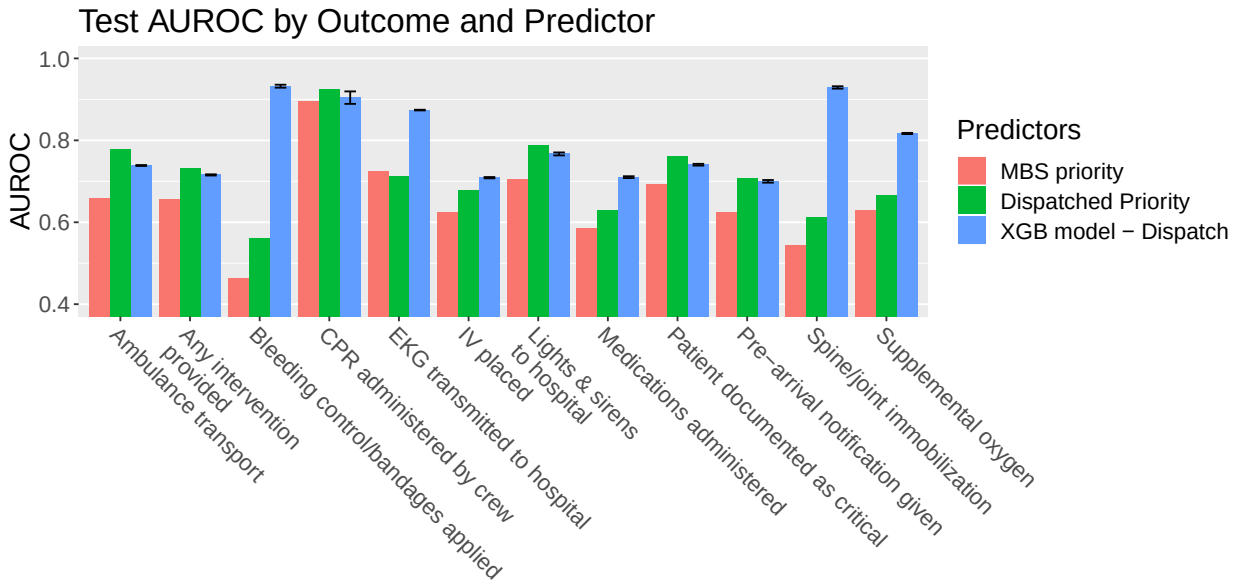
Across the board, we see modest improvements upon adding additional predictors, though for some interventions, adding the text data decreased test AUROC, indicative of overfitting. We see similar patterns for NEWS and hospital outcomes:



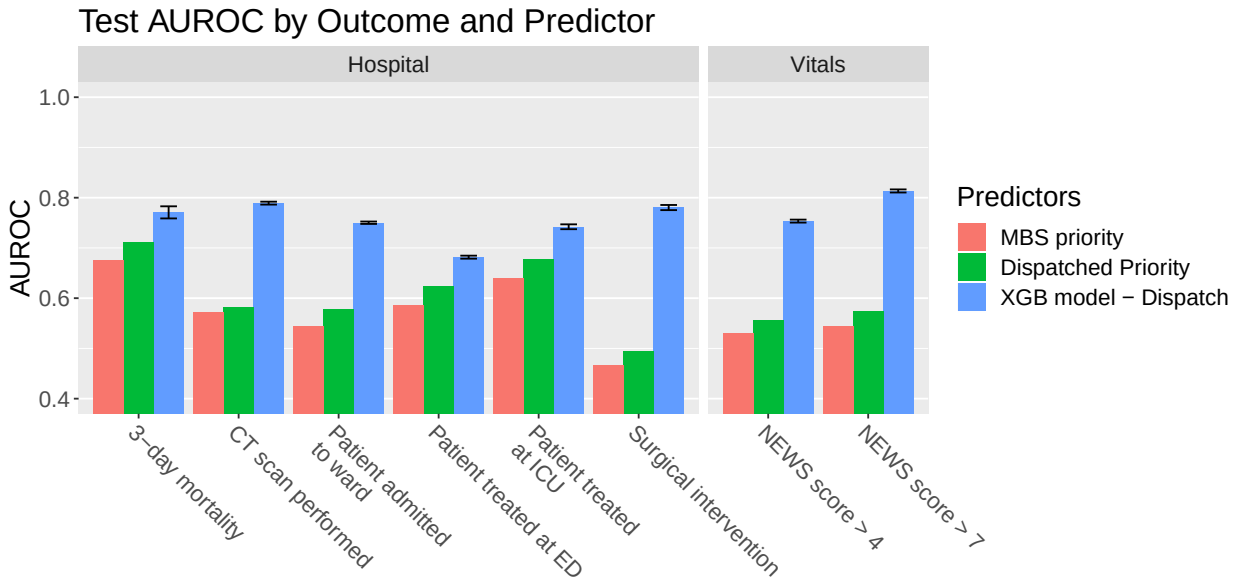
Each indicator improves steadily with the addition of additional predictors, with the exception of 3 day mortality which does not appear to improve with the addition of free-text questions.. Recall that AUROC is sensitive to unbalance in datasets (i.e., having large numbers of negative cases improves AUROC), and as such, comparisons between outcomes here are unreliable.

Comparative analysis

We can also compare the performance of these models with the MBS recommended and dispatched priorities. Here we'll include only calls where the MBS was used, thereby generating a recommended priority:

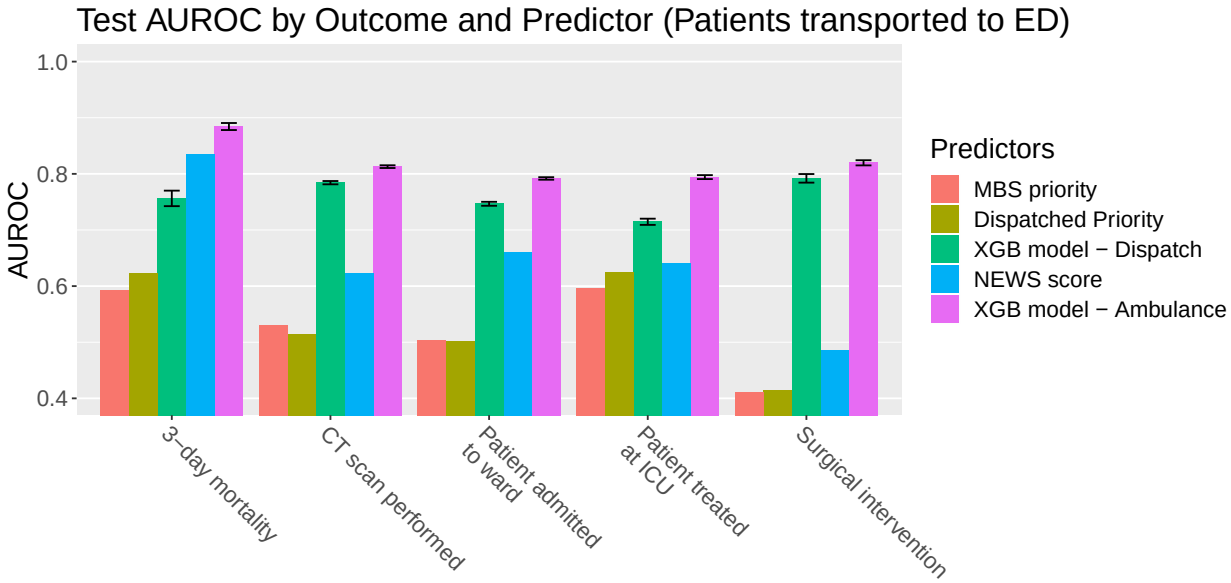


We see that for many ambulance interventions, these models perform better than the recommended priority of the MBS at predicting the likelihood of the outcome occurring, but worse than the actual dispatched priority. Among measures related to specific interventions however, the models perform better than both.



Among NEWS and hospital outcome measures, these models outperform current prioritizations consistently. Overall, we find that we have quite high AUROC values for some specific types of interventions, in general our models have AUROC value of 0.7 to 0.8.

In addition to their use in EMD, similar models could be applied to predict hospital interventions based on data gathered by the ambulance. We can get a sense of how including these data might affect the predictive value of these models by investigating the subset of our patients who are transported by ambulance to an ED. By including only patients for whom we have enough vitals recorded to have a NEWS score imputed, we can also investigate the performance of our measures in comparison with this benchmark:



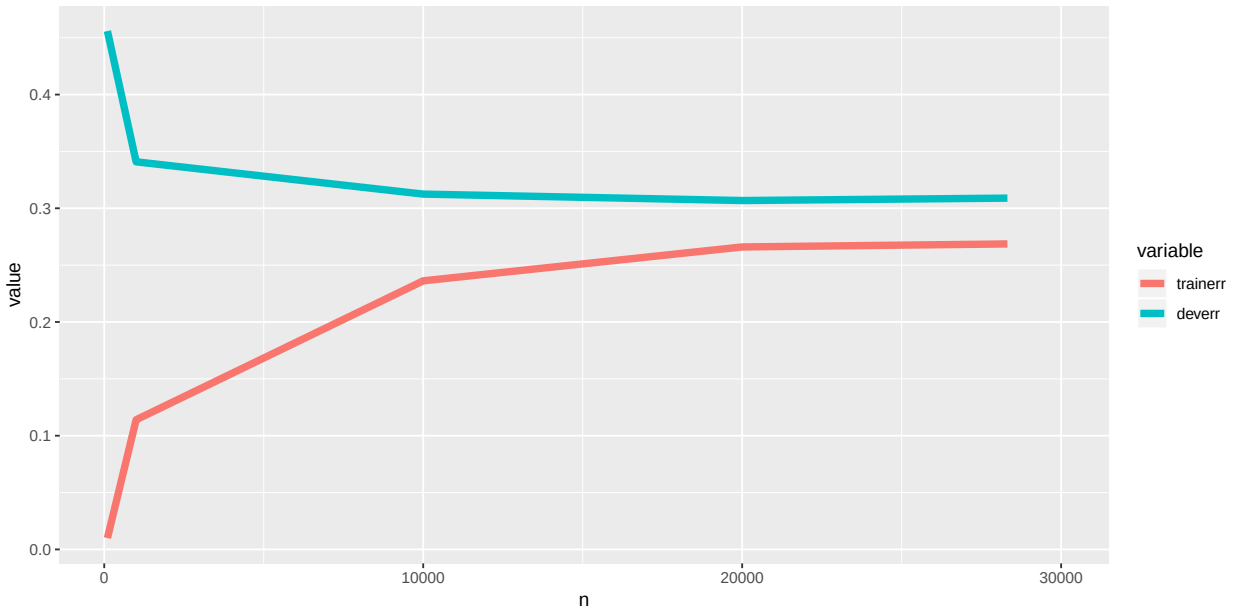
We see that with the exception of 3-day morality, the dispatch data based model outperforms NEWS, with the model containing predictors from the ambulance journal performing better still, outperforming NEWS scores in terms of 3-day mortality as well. Interestingly, while NEWS has a good predictive value for mortality, it performs less well in terms of predicting for instance admission and ICU treatment.

Model validation

So far, we've limited this investigation to the discriminatory ability of our models. That is to say, we've only looked at the overall ability of our models to predict outcomes of interest. While this is useful in guiding exploratory analysis, once an approach has been selected we need to apply some additional investigative tools to optimize our models and determine that they perform well in terms of calibration and clinical usefulness.

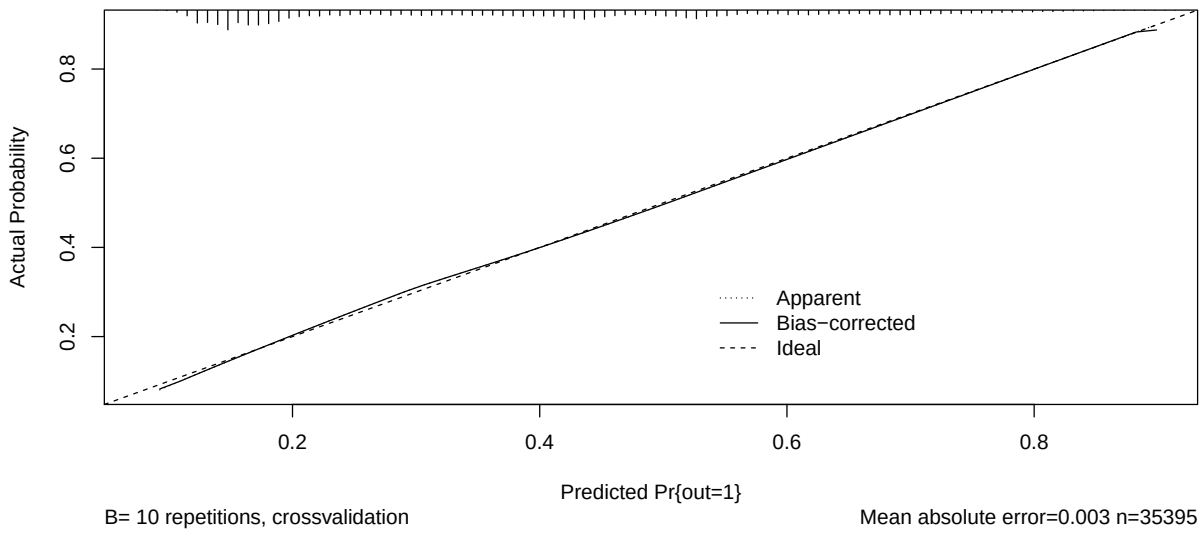
Based on the analysis presented above, we believe that an intervention to improve the ability of nurses to distinguish between patients requiring hospital care and those that may be definitively treated in primary care. This distinction is of particular interest among patients determined not to require an ambulance, as the current decision support system makes no distinction between these outcomes. We believe that hospital admission is the best marker available to us to determine this, and as such, we will here focus our validation efforts on this measure.

Let's begin by using a learning curve to understand how our preliminary model for admission lays in terms of the bias/variance trade off. We saw that the models tended to overfit the data based on the cross-validation results shown above, but a learning curve will assist us in diagnosing whether further data collection and a more or less complex might produce better results. For this purpose, we'll split our data into a training dataset and development dataset consisting of the last 20% of collected observations (in this case, the last 3 months of the data we collected in 2017). We then plot how the predictive value of our dataset changes as we add more data to the training set. Note that we'll here move from reporting the mode AUROC, to investigating the error rate of the model (lower is better):



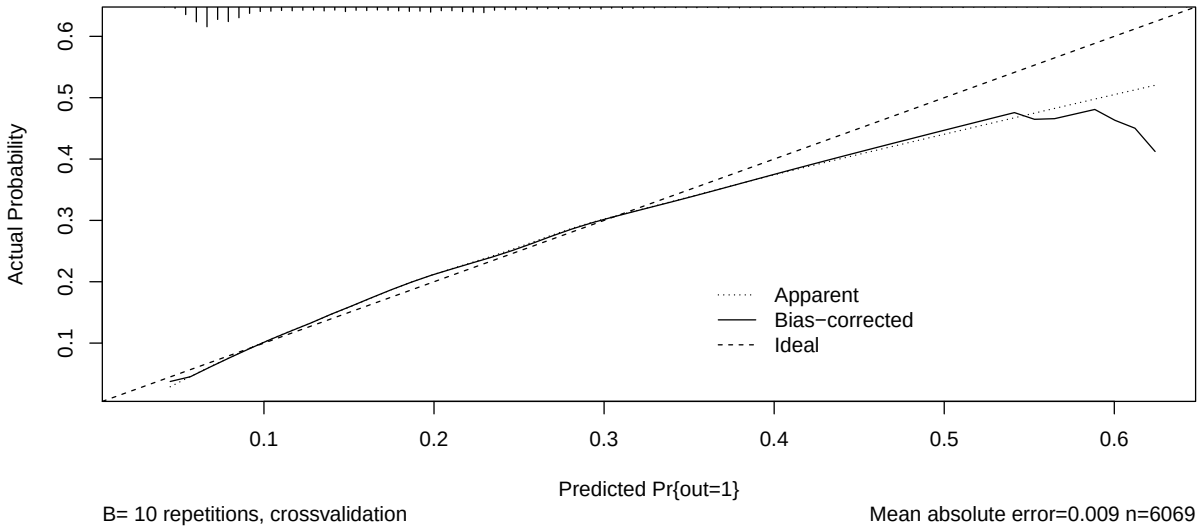
We see that the error in our models in out-of-samle data generally plateaus at an error rate of around 0.3 after ~10 000 observations, and that the error is due to bias in the model, rather than high variance. As such, adding additional data to our model is unlikely to improve performance, and that the data could likely support a more complex model.

We also need to check that our model is well calibrated - That is to say, that the admission risks predicted by the model lack systematic bias across the range of predicted values. We can do this using the cross validation approach we used earlier, and we'll use the the RMS package calibrate function to do the prior.




```
##
## n=35395   Mean absolute error=0.003   Mean squared error=0.00002
## 0.9 Quantile of absolute error=0.008
```

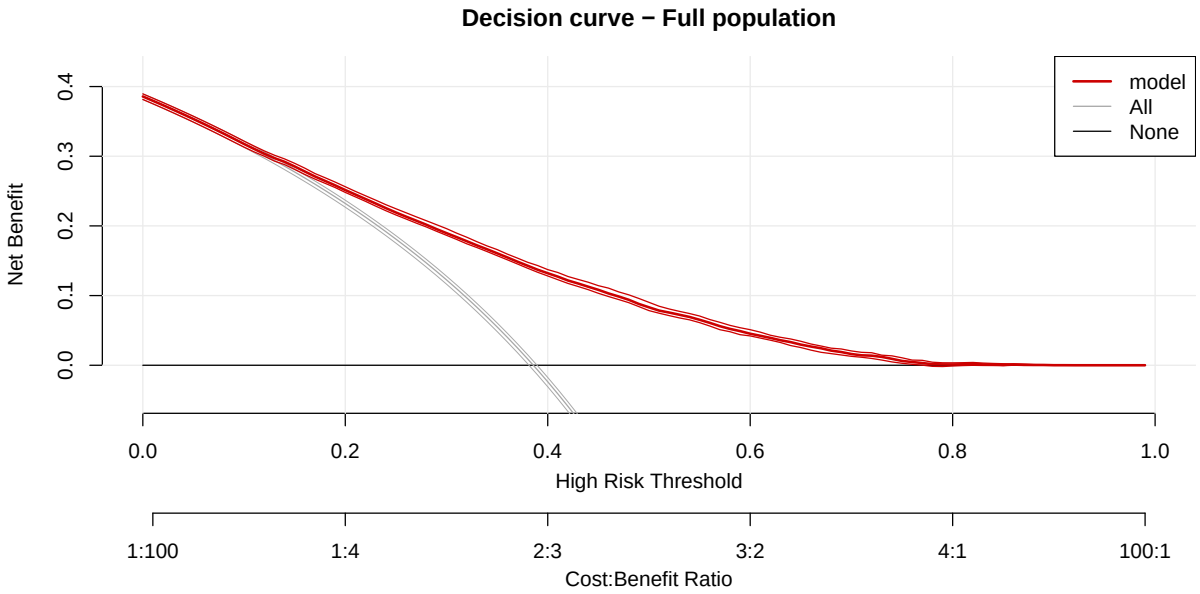
We find that based on cross-validation, the model is well calibrated across the full population. However, we're interested in predicting outcomes among a certain sub-group of patients, namely those who are determined not to need an ambulance. Let's have a look at what the calibration looks like when we apply the model trained on the full dataset to this sub-population:

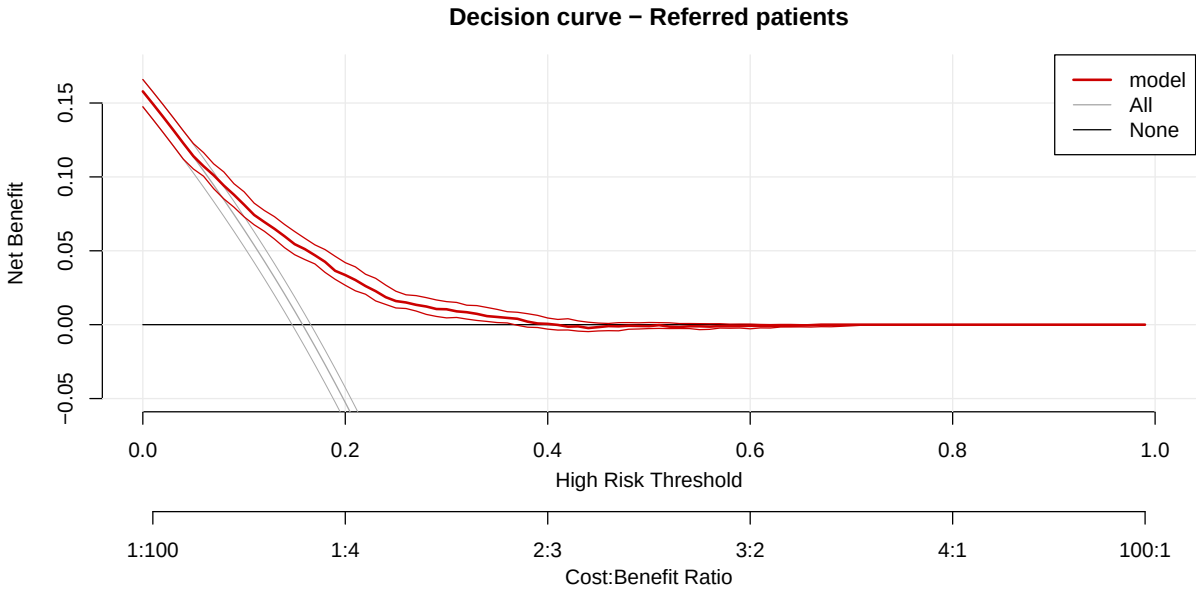



```
##
## n=6069   Mean absolute error=0.009   Mean squared error=0.00018
## 0.9 Quantile of absolute error=0.012
```

We see that the model retains a reasonable degree of calibration within this sub-population, though in the region of high-risk patients where there are very few observations, the model tends to overestimate risks.

Finally, we're interested in assessing whether these models have the ability to provide a clinical benefit to patients who are triaged based on model findings. To do this, we can perform a decision curve analysis to evaluate the potential benefit of a predictive model across various risk thresholds.





We see that in the plots above, the net benefit of applying the model to determine whether a patient will be admitted to the hopsital beings to diverge in terms of net benefit from assuming that all patients will be admitted to the hospital at threshold value corresponding to 1:7 among the full population, and 1:16 among referred patients only.

Thus the questions becomes a clinical and operational one: How much worse is it to refer a patient who should be admitted to primary care, than to send a patient who doesn't need in-patient care to the hospital? If the answer to this question is "Less than 16 times as bad among referred patients", then this model may be able to deliver a clinical benefit.

4) Discussion

In this analysis, we have presented the distribution of a number of measures of patient acuity in our data as they are dispatched today. We continued on to describe our modelling approach, and the performance of a set of preliminary models with regards to these markers. This is by no means an exhaustive validation of these markers as measures of patient acuity, nor of the validity of the models. There is much still to be done - While these results perhaps demonstrate the feasibility of the approach, only validation in an unexamined dataset can address the "researcher degrees of freedom" problem inherent in this type of analysis. Furthermore, the predictions made for one outcome should be validated against other outcomes to establish their generalizability (e.g., do the model predictions for lights and sirens to the hospital also predict ICU treatment?). The residuals of the models must also be examined in detail with regards to various characteristics - Do our models over- or under-estimate risks for older patients or patients with multiple contacts? Do the models work better for patients calling from different areas with different socioeconomic conditions? Do the models work better for certain call types? Upon answering these questions to our satisfaction, we plan to validate our models in data from 2018 and publish the results in a peer-reviewed article.

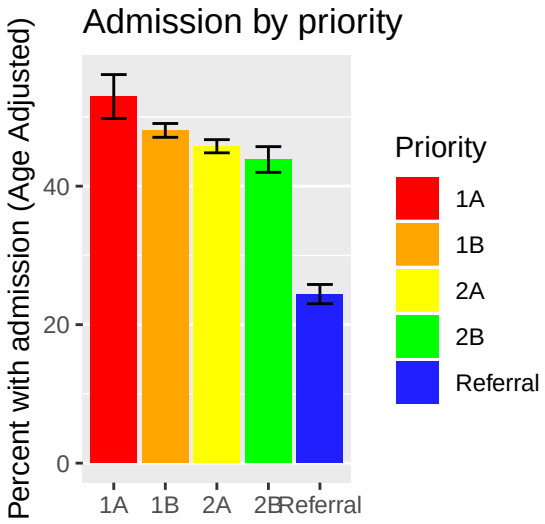
A number of interesting effects have been identified which could yield interesting follow-up investigations. Our investigation of data quality reveals a somewhat troubling trend, whereby missingness is associated with patient acuity. In most pre-hospital care research, observations with missing data are excluded. If the pattern of higher levels of missingness for high-acuity patients also occurs at other agencies, this practice could introduce bias to model estimates. In terms of outcomes, we found that 65% ambulance calls had no documented interventions beyond transportation. While we are certainly missing some concrete medical interventions with these measures, and we cannot hope to capture "softer" services that ambulance crews provide (e.g. patient navigation, psychological comfort, etc.), this suggests that there is room to improve the specificity of the dispatching process. In our manual review of ambulance journals, we've found that about 1/3 to 1/2 of calls dispatched priority 1 but receiving no interventions nonetheless had some form of ambulance need. Unfortunately, these forms of soft care are difficult to document in a structured manner, consequently making them difficult to capture using predictive models.

In our descriptive analysis, we found that the prioritizations made by ambulance nurses demonstrated a comparatively high degree of differentiation among outcomes relating to high acuity conditions, but had a lower degree of predictive value with regards to outcomes beyond the patients immediate care needs. Ambulance priority was for instance perhaps even negatively associated with hospital admission (though patients referred to alternate care did have lower admission rates). It is of course appropriate that nurses focus on the acute condition of the patient when determining response priority, though the results suggest that there is substantial room for improvement with regards to improving differentiation among these measures.

This lack of differentiation could also be explained by the large number of elderly patients transported with a low priority, whose more comprehensive care needs lead to higher rates of admission:

Priority	Mean age
1A	47.9
1B	57.0
2A	62.7
2B	65.6
Referral	52.8

Indeed, in a simple logistic model controlling for patient age, we see that the effect is reversed, though differentiation is still quite poor:



With regards to our predictive models, we found a number of interesting effects. It can be seen that the bulk of the predictive value of the models are drawn from the patient characteristics and call type documented by the nurse. While adding MBS question data and text data does increase the predictive value somewhat in the majority of cases, the value added is marginal. This could be due to a number of reasons - Documentation completeness of the MBS leaves something to be desired, while our text data is rife with spelling errors and idiosyncratic terminology and abbreviations. We plan on addressing these data quality issues as we move forwards, and additional text pre-processing could improve the yield of this unstructured data. Interestingly, it seems that cases where text data tends towards overfitting are outcomes where there is a high degree of discretion on the part of ambulance crews - e.g., whether to drive priority 1 to the hospital, the documentation of a "critical" patient in the ambulance record, or notification of the receiving facility. It is also almost certainly the case that we have not yet identified the optimal methods for modelling the data, and a few more percentage points of AUROC can likely be garnered by tuning model hyperparameters.

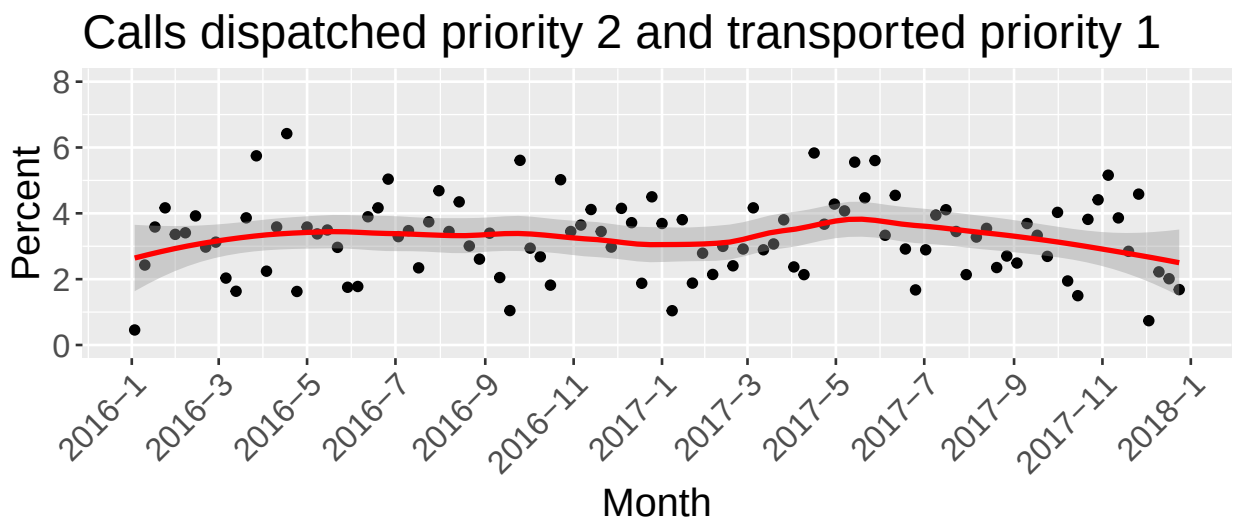
The overall predictive value of our models based on dispatch data is lower than that found in models based on data gathered at the ED. This is unsurprising, given that information based on visually observing or physically interacting with the patient are typically unavailable at the time of ambulance dispatch. While Hong et al. (2018) achieved an impressive AUROC of 0.92 for hospital admission, while we achieved an AUROC of .75 using only data from the EMD center, and 0.77 upon including data from the ambulance. Levin et al. (2017) on the other hand achieved AUROCs of 0.82-0.84 for hospital admission, much closer to the results found here. Given this range of predictive values, it seems uncertain as to what an "acceptable" level of differentiation in this context is. While we had hoped for AUROCs on the order of 0.8, lower levels of predictive value could be compensated for by sacrificing specificity (AKA over-triaging) in order to ensure the safety of decisions made based on model predictions.

While the models were able to out-perform triage determinations with regards to hospital outcomes and vital signs, they struggled to do so for many ambulance-based indicators of patient acuity. This suggests that these models may be able to deliver the most value to nurses in terms of assessing longer-term care needs, with the rule-based system sufficing to identify conditions necessitating an emergency ambulance response. It is also interesting to note that in every case, actual dispatched priorities outperformed MBS recommended priorities in terms of predictive value. However, it may also be noted that in many cases where patients have been inappropriately referred to non-emergency care, the MBS rules would have recommended a higher priority. This illustrates the point that improved accuracy could be undesirable if it comes at the cost of patient safety!

Further development

A major task which lies before us is to aggregate these measures and present risk estimates in a manner which makes them intelligible to nurses without statistical expertise. Our goal is to present a maximum of 3 indicators to the nurse in the user interface of the MBS. A number of approaches could be taken to achieve this. The predicted value of each indicator could for instance be weighted based on some measure of the acuity the outcome represents - A patient with high risk of ICU treatment is likely more acute than a patient with a high risk of admission to any in-patient ward. The weighted sum of all predictors could then be presented to the nurse in the form of a final risk score. Another approach could involve establishing appropriate levels of sensitivity for each outcome based on clinical expertise, with an alert appearing if any measure exceeds this level.

It also lays before us to define a set of quality measures to track during the course of our evaluation of these tools. Several of these measures have face validity as measures of triage error; for instance patients dispatched priority 2 but transported to the hospital priority 1, and referred patients treated at an ICU. Others have value as measures of system specificity, for instance the percentage of priority 1 calls with no documented interventions. While it is the goal of this project to develop a tool to prospectively assist nurses in appropriately characterizing patient care needs, another important use of such measures are in retrospective quality assurance and follow-up. These measures could for instance be tracked over time allowing us to characterize the performance of our system:



In this plot, we see points representing weekly average, plus a smooth locally fitted function (loess) and it's 95% confidence interval. Under-triage rates by this measure lay around 3-4%, while there appears to be (perhaps) a slightly higher rate during the summer months. Perhaps this has something to do with the utilization of replacement staff while many of the regular staff nurses are on summer vacation?

Finally, the protocol for a clinical trial to evaluate the effectiveness of the tool must be finalized. We plan to focus this investigation on low acuity calls, excluding patients determined to be in need of a priority 1 ambulance. In this study we will investigate two dimensions of decisions made by nurses at the EMD center: Whether the patient condition requires an ambulance or can be transported via alternate means to the hospital, and whether the patient condition requires specialist care at a hospital or can be treated within the primary care system. Pending ethics board approval, we plan to perform a fully randomized trial among this cohort of low-acuity patients to investigate whether providing nurses with risk estimates based on these models can lead to improvements in measures of both sensitivity and specificity in the triage process.

Conclusion / Summary

In this analysis, we raise important questions of data quality and missingness which must be carefully considered if valid predictive models are to be developed in the domain of Emergency Medical Dispatching. We presented a set of outcome measures which we believe can provide a useful model of the care provided by ambulance crews, the patient' condition, and their hospital outcomes. We then demonstrated in our data how these measures could be used to characterize the ability of an EMD system to direct patients to an appropriate level of care. We went on to present a brief exploratory analysis using our data. Finally, we present a set of predictive models, which we found to be particularly useful in predicting patient vital signs and hospital outcomes. We hope that the methods and tools provided will be useful to other ambulance agencies wishing to analyze data in a similar manner.